Techniques of Data Mining In Healthcare: A Review

Parvez Ahmad
Dept. Of Computer Science
Aligarh Muslim University
Aligarh

Saqib Qamar
Dept. Of Computer Science
Aligarh Muslim University
Aligarh

Syed Qasim Afser Rizvi
Dept. Of Computer Science
Aligarh Muslim University
Aligarh

ABSTRACT

Data mining is gaining popularity in disparate research fields due to its boundless applications and approaches to mine the data in an appropriate manner. Owing to the changes, the current world acquiring, it is one of the optimal approach for approximating the nearby future consequences. Along with advanced researches in healthcare monstrous of data are available, but the main difficulty is how to cultivate the existing information into a useful practices. To unfold this hurdle the concept of data mining is the best suited. Data mining have a great potential to enable healthcare systems to use data more efficiently and effectively. Hence, it improves care and reduces costs. This paper reviews various Data Mining techniques such as classification, clustering, association, regression in health domain. It also highlights applications, challenges and future work of Data Mining in healthcare.

Keywords

Data Mining, Classification, Clustering, Association, Healthcare

1. INTRODUCTION

In the early 1970's, it was very costly to store the data or information. But due to the advancement in the field of information gathering tools and WWW in the last twenty-five years, we have seen huge amount of information or data are available in electronic format. To store such a large amount of data or information the sizes of databases are increased very rapidly. Such type of databases consist very useful information. This information may be very useful for decision making process in any field. It becomes possible with the help of data mining or Knowledge Discovery in Databases (KDD). Data mining is the process of extracting the useful information from a large collection of data which was previously unknown [1]. A number of relationships are hidden among such a large collection of data for example a relationship between patient data and their number of days of stay [2].

With the help of figure 1 five stages are identified in knowledge discovery process [3, 4, and 5].

With the help of raw data the first stage starts and ends with extracted knowledge which was captured as a result of following stages as shown in figure 1:

- Selection

The data is selected according to some criteria in this stage. For example, a bicycle owns by all those people, we can determine subsets of data in this way.

- Preprocessing

This stage removes that information which is not necessary for example while doing pregnancy test it is not necessary to note the sex of a patient. It is also known as data cleansing stage.

- Transformation

This stage transformed only those data which are useful in a particular research for example only data related to a particular demography is useful in market research.

- Data mining

Data mining is a stage knowledge discovery process. This stage is useful for extracting the meaningful patterns from data.

- Interpretation and evaluation

The meaningful patterns which the system identified are interpreted into knowledge in this stage. This knowledge may be then useful for making useful decisions.

1.1 Significance of Data Mining in Healthcare

Generally all the healthcare organizations across the world stored the healthcare data in electronic format. Healthcare data mainly contains all the information regarding patients as well as the parties involved in healthcare industries. The storage of such type of data is increased at a very rapidly rate. Due to continuous increasing the size of electronic healthcare data a type of complexity is exist in it. In other words, we can say that healthcare data becomes very complex. By using the traditional methods it becomes very difficult in order to extract the meaningful information from it. But due to advancement in field of statistics, mathematics and very other disciplines it is now possible to extract the meaningful patterns from it. Data mining is beneficial in such a situation where large collections of healthcare data are available.

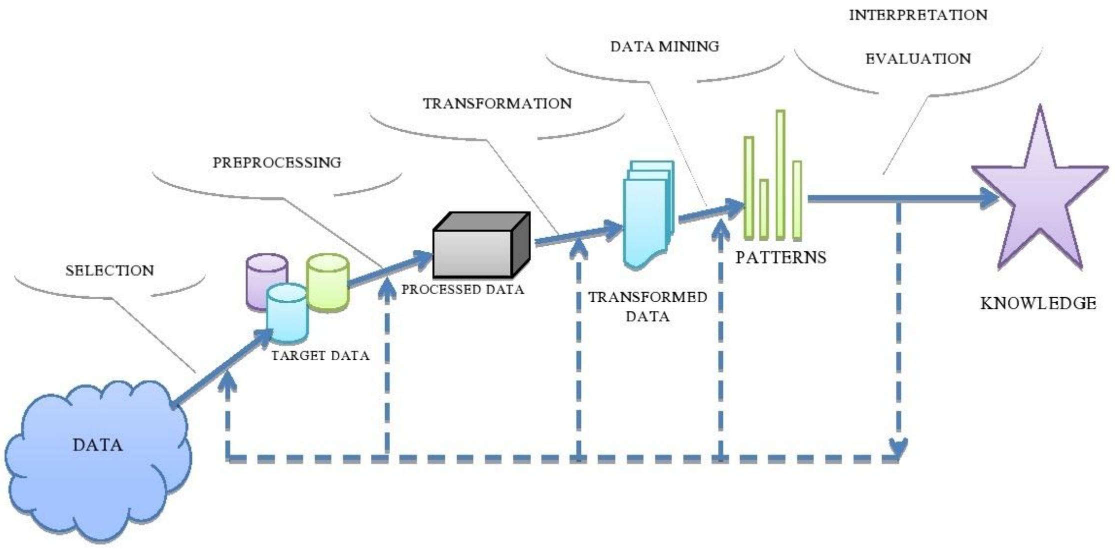

Figure 1: Stages of Knowledge Discovery Process

Data Mining mainly extracts the meaningful patterns which were previously not known. These patterns can be then integrated into the knowledge and with the help of this knowledge essential decisions can becomes possible. A number of benefits are provided by the data mining. Some of them are as follows: it plays a very important role in the detection of fraud and abuse, provides better medical treatments at reasonable price, detection of diseases at early stages, intelligent healthcare decision support systems etc. Data mining techniques are very useful in healthcare domain. They provide better medical services to the patients and helps to the healthcare organizations in various medical management decisions. Some of the services provided by the data mining techniques in healthcare are: number of days of stay in a hospital, ranking of hospitals, better effective treatments, fraud insurance claims by patients as well as by providers, readmission of patients, identifies better treatments methods for a particular group of patients, construction of effective drug recommendation systems, etc [2]. Due to all these reasons researchers are greatly influenced by the capabilities of data mining. In the healthcare field researchers widely used the data mining techniques. There are various techniques of data mining. Some of them are classification, clustering, regression, etc. Each and every medical information related to patient as well as to healthcare organizations is useful. With the help of such a powerful tool known as data mining plays a very important role in healthcare industry. Recently researchers uses data mining tools in distributed medical environment in order to provide better medical services to a large proportion of population at a very low cost, better customer relationship management, better management of healthcare resources, etc. It provides meaningful information in the field of healthcare which may be then useful for management to take decisions such as estimation of medical staff, decision regarding health insurance policy, selection of treatments, disease prediction *etc.*, [6-9]. Dealing with the issues and challenges of data mining in healthcare [10, 11]. In order to predict the various diseases effective analysis of data mining is used [12-21]. Proposed a data mining methodology in order to improve the result [22-24] and proposed new data mining methodology [25, 26] and proposed framework in order to improved the healthcare system [27-31].

2. Data Mining Techniques

2.1 Classification

Classification is one of the most popularly used methods of Data Mining in Healthcare sector. It divides data samples into target classes. The classification technique predicts the target class for each data points. With the help of classification approach a risk factor can be associated to patients by analyzing their patterns of diseases. It is a supervised learning approach having known class categories. Binary and multilevel are the two methods of classification. In binary classification, only two possible classes such as, “high” or “low” risk patient may be considered while the multiclass approach has more than two targets for example, “high”, “medium” and “low” risk patient. Data set is partitioned as training and testing dataset. It consists of predicting a certain outcome based on a given input. Training set is the algorithm which consists of a set of attributes in order to predict the outcome. In order to predict the outcome it attempts to discover the relationship between attributes. Goal or prediction is its outcome. There is another algorithm known as prediction set. It consists of same set of attributes as that of training set. But in prediction set, prediction attribute is yet to be known. In order to process the prediction it mainly analyses the input. The term which defines how “good” the algorithm is its accuracy. Consider a medical database of Pawti Medical Center, training set consists all the information regarding patient which were recorded previously. Whether a patient had a heart problem or not is the prediction attribute there. With the help of table 1 given below we demonstrates the training sets of such database.

Table 1 – TRAINING AND PREDICTION SETS FOR PAWTI MEDICAL DATABASE
Training Set

Age	Heart rate	Blood pressure	Heart problem
45	75	140/64	Yes
28	85	101/60	No
38	62	105/55	No

Prediction Set

Age	**Heart rate**	**Blood pressure**	**Heart problem**
33	**89**	**142/82**	**?**
45	**52**	**102/56**	**?**
87	**83**	**138/61**	**?**

In order to disembosom the knowledge, classification predicts rules. Prediction rules are divulged in the form of IF-THEN rules. With the help of above example, a rule predicting the first row in the training set may be represented as follows:

IF (Age=45 AND Heart rate>75) OR (Age>44 AND Blood pressure>139/60) THEN Heart problem=yes

Following are the various classification algorithms used in healthcare:

2.1.1 K-Nearest Neighbor (K-NN)

K-Nearest Neighbor (K-NN) classifier is one of the simplest classifier that discovers the unidentified data point using the previously known data points (nearest neighbor) and classified data points according to the voting system [6]. Consider there are various objects. It would be beneficial for us if we know the characteristics features of one of the objects in order to predict it for its nearest neighbors because nearest neighbor objects have similar characteristics. The majority votes of K-NN can play a very important role in order to classify any new instance, where k is any positive integer (small number). It is one of the most simple data mining techniques. It is mainly known as Memory-based classification because at run time training examples must always be in memory [32]. Euclidean distance is calculated when we take the difference between the attributes in case of continuous attributes. But it suffers from a very serious problem when large values bear down the smaller ones. Continuous attributes must be normalized in order to take over this major problem so that they have same influence on the distance measure between distances [33].

K-NN has a number of applications in different areas such as health datasets, image field, cluster analysis, pattern recognition, online marketing *etc*. There are various advantages of KNN classifiers. These are: ease, efficacy, intuitiveness and competitive classification performance in many domains. If the training data is large then it is effective and it is robust to noisy training data. A main disadvantage of KNN classifiers is the large memory requirement needed to store the whole sample. If there is a big sample then its response time on a sequential computer will also large.

2.1.2 Decision Tree (DT)

DT is considered to be one of the most popular approaches for representing classifier. We can construct a decision tree by using available data which can deal with the problems related to various research areas. It is equivalent to the flowchart in which every non-leaf nodes denotes a test on a particular attribute and every branch denotes an outcome of that test and every leaf node have a class label. Root node is the top most node of a decision tree. For example, with the help of medical readmission decision tree we can decide whether a particular patient requires readmission or not. Knowledge of domain is not required for building decision regarding any problem. The most common use of Decision Tree is in operations research analysis for calculating conditional probabilities [34]. Using Decision Tree, decision makers can choose best alternative and traversal from root to leaf indicates unique class separation based on maximum information gain [35]. Decision Tree is widely used by many researchers in healthcare field. Several advantages of decision tree as follows: Decision trees are self–explanatory and when compacted they are also easy to follow. Even set of rules can also be constructed with the help of decision trees. Hence, representation of decision tree plays a very important role in order to represent any discrete-value classifier because it can be capable to handle both type of attributes, nominal as well as numeric input attributes. If any datasets have missing or erroneous values, such type of datasets can be easily handled by decision trees. Due to this reason decision tree can be considered to be nonparametric method. The meaning of above sentence is that there is no need to make assumptions regarding distribution of space and structure of classifier. Decision trees have several disadvantages. These are as follows: Most of the algorithms (like ID3 and C4.5) require that the target attributes have only discrete values because decision trees use the divide and conquer method. If there are more complex interactions among attributes exist then performance of decision trees is low. Their performance is better only when there exist a few highly relevant attributes. One of the reasons for this is that other classifiers can compactly describe a classifier that would be very challenging to represent using a decision tree. A simple illustration of this phenomenon is replication problem of decision trees [36], and the greedy characteristic of decision trees leads to another disadvantage. This is its over-sensitivity to the training set, irreverent attributes and to noise [37].

2.1.3 Support Vector Machine (SVM)

Vapnik *et al.* [38- 39] was the first one who gave the notion of SVM. Among all the available algorithms it provides very accurate results. It is gaining popularity because it can be easily extended to problems related to multiclass though it was mainly developed for problems related to binary classification [40-41]. In order to be useful for various effective and efficient tasks it is capable of creating single as well as multiple hyper planes in high dimensional space. The main aim of creating hyper plane by SVM in order to separate the data points. There are two ways of implementing SVM. The first technique employs mathematical programming and second technique involves kernel functions. With the help of training datasets non linear functions can be easily mapped to high dimensional space. Such situation can only be possible when we use kernel functions. Gaussian, polynomial, sigmoid etc. are some examples of kernel functions. For the classification of data points hyper plane is used. The primary task of hyper plane is to maximize the separation between data points. Support vectors are used in order to construct the hyper plane. There are various advantages of SVM. Some of them as follow: First one is that, it is effective in high dimensional spaces. Another one is that, it is effective in cases where number of dimensions is greater than the number of samples. Third one is that, it is memory efficient because it uses a subset of training points in the decision function (called support vectors), and next is that, it is versatile because different kernel functions can be specified for the decision function. Some of the disadvantages of SVM as follow: First one is that, if the number of features is much greater than number of samples then SVM is more likely to give poor performances, and the another one is that, it does not directly provide probability estimates. These are calculated using an expansive five –fold cross-validation.

2.1.4 Neural Network (NN)

In the early 20th century it was developed [42]. Before the introduction of decision trees and the Support Vector Machine (SVM) it was regarded as the best classification algorithm [43]. This was one of the reasons which encouraged NN as the most widely used classification algorithm in various biomedicine and healthcare fields [44, 45, 46]. For example, NN has been widely used as the algorithm supporting the diagnosis of diseases including cancers [47-51] and predict outcomes [52-54]. In NN, basic elements are neurons or nodes. These neurons are interconnected and within the network they worked together in parallel in order to produce the output functions. From existing observations they are capable to produce new observations even in those situations where some neurons or nodes within the network fails or go down due to their capability of working in parallel. An activation number is associated to each neuron and a weight is assigned to each edge within a neural network. In order to perform the tasks of classification and pattern recognition neural network is mainly used [55]. The basic property of an NN is that it can minimize the error by adjusting its weight and by making changes in its structure. It minimizes the error only due to its adaptive nature. NN are capable to produce predictions of greater accuracy. One of the major advantages of NN is that it can properly handle noisy data for training and can reasonably classify new types of data that is different from training data. There are also various disadvantages of NN. First is that it requires many parameters, including the optimum number of hidden layer nodes that are empirically determined, and its classification performance is very sensitive to the parameters selected [56]. Second is that its training or learning process is very slow and computationally very expensive. Another is that they do not provide any internal details regarding to that phenomena which is currently under investigation. Hence, this approach is like a “black-box” approach for us.

2.1.5 Bayesian Methods

For probabilistic learning method Bayesian classification is used. With the help of classification algorithm we can easily obtained it [6]. Bayes theorem of statistics plays a very important role in it. While in medical domain attributes such as patient symptoms and their health state are correlated with each other but Naïve Bayes Classifier assumes that all attributes are independent with each other. This is the major disadvantage with Naïve Bayes Classifier. If attributes are independent with each other then Naïve Bayesian classifier has shown great performance in terms of accuracy. In healthcare field they play very important roles. Hence, researchers across the world used them there are various advantages of BBN. One of them is that it helps to makes computation process very easy. Another one is that for huge datasets it has better speed and accuracy.

Classification Techniques Examples in Healthcare

Hu et al., used different types of classification methods such as decision trees, SVMs, Bagging, Boosting and Random Forest for analyzing microarray data [57]. In this research, experimental comparisons of LibSVMs, C4.5, BaggingC4.5, AdaBoostingC4.5, and Random Forest on seven micro-array cancer data sets were conducted using 10-fold cross validation approach on the data sets obtained from Kent Ridge Bio Medical Dataset repository. On the basis of the experimental results, it has been found that Random Forest classification method performs better than all the other used classification methods [57].

Breast cancer is one of the fatal and dangerous diseases in women. Potter *et al*., had performed experiment on the breast cancer data set using WEKA tool and then analyzed the performance of different classifier using 10-fold cross validation method [58].

Huang *et al.*, constructed a hybrid SVM-based diagnosis model in order to find out the important risk factor for breast cancer because in Taiwan, women especially young women suffered from breast cancer. In order to construct the diagnosis model, several types of DNA viruses in this research are studied. These DNA viruses are HSV-1 (herpes simplex virus type-1), EBV (Epstein-Barr virus), CMV (cytomegalovirus), HPV (human papillomavirus), and HHV-8 (human hepesvirus-8). On the basis of experimental results, either {HSV-1,HHV-8} or {HSV-1,HHV-8,CMV}can achieved the identical high accuracy. The main aim of the study was to obtain the bioinformatics about the breast cancer and DNA viruses. Apart from SVM-based model, another type of diagnosis model called LDA (Linear discriminate analysis) was also constructed in this research. After comparing the accuracies of both SVM and LDA, the accuracy of SVM was far better than that of LDA [59].

Classification techniques were used for predicting the treatment cost of healthcare services which was increased with rapid growth every year and was becoming a main concern for everyone [60].

Khan *et al.*, used decision tree for predicting the survivability of breast cancer patient [61].

Chang *et al.*, used an integrated decision tree model for characterize the skin diseases in adults and children. The main focus of this research was to analyze the results of five experiments on the six major skin diseases. The main aim of this research was to construct the best predictive model in dermatology by using the decision tree and combine this decision tree with the neural network classification methods. On the basis of experimental result, it has been found that neural network has 92.62% accuracy in prediction of skin diseases [62].

Das *et al.*, proposed a intelligent medical decision support system based on SAS software for the diagnosis of heat diseases. In order to construct the proposed system, neural networks method was mainly used. In this research, experiments were performed on the data taken from Cleveland heart disease database. On this basis of experiments, it has been found that neural networks have 89.01% accuracy [63].

Curiac *et al.*, analyzed the psychiatric patient data using BBN in order to identify the most significant factors of psychiatric diseases and their correlations by performing experiment on real data obtained from Lugoj Municipal Hospital. In this research, it has been found that BBN plays a very important role in medical decision making process in order to predicate the probability of a psychiatric patient on the basis detected symptoms [64].

Liu *et al.*, develop a decision support system using BBN for better analyzing risks that were associated with health. With the help of using BBN in order to construct dose-response relationship and in order to predict the human disease and cancer risks due to specific toxic substance are the major objectives of this research [65].

E.Avci *et al.,* proposed an intelligent system on the basis of genetic support vector machines (GVSM) for better analyzing the heart valve disease. This system extracts the important feature and classifies the signal obtained from the ultrasound of heart valve. In this research, the proposed system was mainly used for the diagnosis of the heart valve diseases. In this research, the proposed system was evaluated in 215 samples. On the basis of samples result, it has been found that the proposed system was very effective to detect Doppler heart sounds [66].

Gunasundari *et al.,* used ANN for discovering the lung diseases. This research work analyze the chest Computed Tomography (CT) and extract significant lung tissue feature to reduce the data size from the Chest CT and then extracted textual attributes were given to neural network as input to discover the various diseases regarding lung [67].

Soni *et al.,* proposed the associative classification approach for better analyzing the healthcare data. The proposed approach was the combined approach that integrated the association rules as well as classification rules. This integrated approach was useful for discovering rules in the database and then using these rules to construct an efficient classifier. In this research, experiments on the data of heart patients were performed in order to find out that accuracy of associative classifiers was better than accuracy of traditional classifiers. Apart from this, the research also generated the rules using weighted associative classifier [68].

Fei *et al.,* proposed Particle Swarm Optimization – Support Vector Machines (PSO-SVM) model for better analyzing the arrhythmia cordis to ensure the health of humans and save humans life. In this research, PSO was used to determined the parameters of SVM. This research demonstrates the performance of proposed model by using MIT-BIH ECG database on which experiments were performed. On the basis of experimental results, it has been found that the accuracy of proposed model was better than the accuracy of artificial neural network in diagnosis of arrhythmia cordis [69].

Er *et al.*, constructed a model using Artificial Neural Network (ANN) for analyzing chest diseases and a comparative analysis of chest diseases was performed using multilayer, generalized regression, probabilistic neural networks [70].

DNA repair genes were considered by Chuang *et al.,* fin order to better prediction of oral cancer by choosing a single nucleotide polymorphisms (SNPs) dataset. The chosen dataset had certain samples of oral cancer's patients. In this research, by using the support vector machines all prediction experiments were performed. On the basis of experimental result, it has been found that the performance of holdout cross validation was better than the performance of 10-fold cross validation. Apart from this, it has been also found that the accuracy of classification was 64.2% [71].

Bakar *et al.*, proposed predictive models by using multiple rule based classifiers for the better early detection of dengue disease. In this research, the multiple rule based classifiers which were used in the proposed models were decision tree, rough set classifier, naïve bayes, and associative classifier. In order to predict the early detection of dengue disease several classifiers were investigated in this research. The classifiers were investigated individually and also in combination in order to study their performance. On the basis of experimental results, it has been found that the accuracy of multiple classifiers was better than the accuracy of single classifier [72].

Moon *et al.*, used decision tree algorithm in order to characterize the smoking behaviors among smokers by assessing their psychological distress, psychological health status, consumption of alcohol, and demographic variables. The classification analysis was conducted on the basis of decision tree algorithm in order to find the relationship between the average numbers of cigarette consumption per day [73].

Jena *et al.*, used K-NN and Linear Discriminate Analysis (LDA) for classification of chronic disease in order to generate early warning system. This research work used K-NN to analyze the relationship between cardiovascular disease and hypertension and the risk factors of various chronic diseases in order to construct an early warning system to reduce the complication occurrence of these diseases [74].

Chien *et al.*, proposed a universal hybrid decision tree classifier for classifying the activity of patient having chronic disease [75]. They further improved the existing decision tree model to classify different activities of patients in more accurate way.

Shouman *et al.*, used K-NN classifier for analyzing the patients suffering from heart disease [76]. The data was collected from UCI and experiment was performed using without voting or with voting K-NN classifier and it was found that K-NN achieved better accuracy without voting in diagnosis of heart diseases as compared to with voting K-NN.

Liu *et al.*, proposed an improved Fuzzy K-NN classifier for diagnosing thyroid disease. Particle Swarm Optimization (PSO) was also used for specifying fuzzy strength constraint and neighborhood size [77].

Hattice *et al.*, discussed the classifier in medical field to diagnosis the skin disease using weighted KNN classifier [78].

Abdi *et al.*, was constructed a PSO based SVM model for identifying erythemato-squamous diseases which consists two stages. In the first stage optimal feature were extracted using association rule and in second phase the PSO was used to discovered best kernel parameters for SVM in order to improve the accuracy of classifier model [79].

Zuoa *et al.*, introduced an adaptive Fuzzy K-NN approach for Parkinson disease [80].

Rusdah *et al.*, reviewed the various latest data mining methods in order to better diagnosis of tuberculosis. After reviewing, it has been found that the support vector machines (SVM) outperformed the other methods in the diagnosis of tuberculosis [81].

Johnson *et al.*, proposed a multistage methodology in order to better detection of fraud committed by patients as well as by providers for healthcare insurance companies. The proposed methodology also helps in reducing significant costs for insurance companies. The proposed methodology was made up of various stages including risk determination stage. The risk threshold was determined by using a decision tree based method. The proposed methodology was compared with unsupervised and supervised neural network techniques. After comparison, it has been found that the proposed methodology outperforms unsupervised and supervised neural network techniques in order to detect the fraud. Apart from this, proposed methodology plays a significant role in validated the legitimated claims by obtaining the information from insurance claim forms [82].

Govaert *et al.,* reviwed the relationship between surgical auditing and healthcare costs in order to evaluate that the surgical auditing has the potential to reduce the overall costs of hospitals only if when it focused on high-risk procedures like colorectal cancer surgery [83].

Peng Z *et al.,* explored the embryonic stem cell (ESC) gene signatures importance in order to estimated the survival of prostate cancer (PCa) patients at the time of their diagnosis. In the research, a total of 641 ESC gene predictors (ESCGPs) were identified by using microarray data sets. For estimating the survival a k-nearest neighbor (K-NN) algorithm was used to estimate the overall survival [84].

2.2 Regression

Regression is very important technique of data mining. With the help of it, we can easily identify those functions that are useful in order to demonstrates the correlation among different various variables. It is mainly a mathematical tool. With the help of training dataset we can easily construct it. Consider two variables ‘P’ and ‘Q’. These two types of variables are mainly used in the field of statistics. One of them is known as dependent and another one is independent variables. The maximum number of dependent variables cannot be more than one while independent can be exceeds one. Regression is mostly used in order to inspect the certain relationship between variables. With the help of regression technique we can easily entrenched the addiction of one variable upon others [85]. Regression can be classified into linear and non-linear on the basis of certain count of independent variables. In order to appraisal associations between two types of variables in which one is dependent variable and another one is independent variables (one or more), linear regression used. In order to construct the linear model, linear function is utilized by linear regression. But there is limitation while we use linear approach because both types of variables are known already and hence, its main purpose is to trace a line that correlates between both these variables [86]. We cannot use linear regression for categorized data. It is restricted only to numerical data. With the help of logistic regression the categorical data can be used. Such type of data is used by non-linear regression and logistic regression is basically a type of non-linear regression. Logistic regression with the help of logit function can predict the probability of occurrence. However, between variables logistic regression cannot consider linear relationship [87]. Due to all these reasons regression is widely used in medical field for predicting the diseases or survivability of a patient.

Regression Examples in Healthcare

Divya *et al.,* proposed Weighted Support Vector Regression (WSVR) which used weight factor on the basis of sensor reading for providing continuous monitoring to patients in order to provide them better healthcare services. In this research, on the basis of experimental result, it has been found that the proposed approach had better accuracy than simple vector regression [88].

Logistic regression for the estimation of relative risk for various medical conditions such as Diabetes, Angina, stroke etc [89].

Xie *et al.,* proposed a regression decision tree algorithm in order to predict the number of hospitalization days in a population. Proposed algorithm was developed using extensive health insurance claims data sets. Experimental results displayed that proposed algorithm was performed significantly in general population as well as in sub-populations in order to predict future hospitalization [90].

Alapont *et al.,* used WEKA for different learning methods like LinearRegression, LeastMedSq, SMOreg, Multilayer Percepton, KStart, Tree M5P etc. In the research, experiments were carried out using 10-fold cross validation. After experiments were carried out, it has been found that LinearRegression and Tree M5P gave best results for the effective utilization of hospital resources, improved hospital ranking and better customer relationship services [91].

2.3 Clustering

Clustering is defined as unsupervised learning that occurs by observing only independent variables while supervised learning analyzing both independent and dependent variables. It is different from classification which is a supervised learning method. It has no predefined classes. Because of this reason, clustering may be best used for studies of an exploratory nature, mainly if those studies encompass large amount of data, but not very much known about data (such as mass of data generated by microarray analysis). The goal of clustering is descriptive while goal of classification is predictive (Veyssieres and Plant, 1998). The main task of unsupervised learning method means clustering method is to form the clusters from large database on the basis of similarity measure [6]. The goal of clustering is to discover a new set of categories, the new groups are of interest in themselves, and their assessment is intrinsic. In classification tasks, an important part of the assessment is extrinsic. Clustering partitioned the data points based on the similarity measure [6]. Clustering groups data instances into subsets in such a manner that similar instances are grouped together, while different instances belongs to different groups. Clustering approach is used to identify similarities between data points. Each data points within the same cluster are having greater similarity as compare to the data points belongs to other cluster. Clustering of objects is as ancient as the human need for describing the salient characteristics of men and objects and identifying them with a type. Therefore, it grasp various scientific disciplines: from mathematics and statistics to biology and genetics, each of which uses different terms to describe the topologies formed using this analysis. From biological “taxonomies”, to medical “syndromes” and genetic “genotypes” to manufacturing “group technology”— the problem is identical: forming categories of entities and assigning individuals to the proper groups within it. Following are the various clustering algorithms used in healthcare:

2.3.1 Partitional Clustering

The maximum number of data points in the datasets is ‘n’. With the help of ‘n’ data points the maximum possible number of ‘k’ clusters is obtained. In order to obtained the ‘k’ clusters from ‘n’ data points partitional clustering method is used. In this method, each ‘n’ data points is relates to one and only ‘k’ clusters while each ‘k’ clusters can relates to more than ‘n’ data points. Partitional clustering algorithms require a user to input k, (which is the number of clusters). Generally, partitional algorithms directly relocate objects to k clusters. Partitional algorithms are categorized according to how they relocate objects, how they select a cluster centroid (or representative) among objects within a (incomplete) cluster, and how they measure similarities between objects and cluster centroids. Before we obtained the clusters this method requires to define the required number of cluster which we may have to obtained from datasets. On the basis of similarities between objects and cluster centroids this method is partitioned into two categories. These are K-means and K-Mediods. One of the most popular algorithms of this approach is K-means [3]. First of all it randomly selects k objects and then decomposes these objects into k disjoint groups by iteratively relocating objects based on the similarity between the centroids and objects [92, 93]. In k-means, a cluster centriod is mean value of objects in the cluster. The next algorithm is K-mediods. In order to group the cluster it used mediods. Mediod is very important because in the database it is that data point which is most centrally located. In order to improve the healthcare services related to public healthcare domain Lenert *et al.*, utilize the application of k-means clustering [94] and by using the clustering technique Belciug *et al*. detect the recurrence of breast cancer [95]. Escudero *et al.*, used the concept of Bioprofile and K-means clustering for early detection of Alzheimer's disease [96]. The major advantage of partitional clustering algorithms is their superior clustering accuracy as compared with hierarchal clustering algorithms that is the result of their global optimization strategy (i.e., the recursive relocations of objects) [97,98]. Another advantage is, partitional algorithms can handle large data sets which hierarchal algorithms cannot (i.e., better scalability) and can more quickly cluster data [97, 98]. In other words we can say that, partitional algorithms are more effective and efficient than hierarchical algorithms. One major drawback to the use of partitional algorithms is that their clustering results depend on the initial cluster centroids to some degree because the centroids are randomly selected. Each time when partitional algorithms run different clustering results are obtained.

2.3.2 Hierarchical Clustering

In order to partition the data points this method can be used two approaches. Data points can be partitioned in a tree way known as hierarchical way by using either top down or bottom up approaches. On the basis of partitioned process this method can be classified as Agglomerative and Divisive. The maximum number of data points can be n. A number of data points among n data points may have similarity with each other. The main aim of agglomerative approach is to merge such data points into a single group [3]. Divisive approach initially takes this single group and iteratively partitioned it into smaller group until and unless each data point relates to one and only one cluster [3]. There are three types of hierarchical clustering algorithms: First one is single-link, another one is complete-link, and the last one is average-link. The single-link clustering algorithm select the closet pair of objects from two groups and measure the similarity between objects as group similarity. The complete-link algorithm calculates the similarity between the most distant pair of objects from two groups. The average-link algorithm selects all pairs of objects from two groups and averages all possible distances between objects. The most similar two groups, those having the shortest distance are merged together after calculating similarities or distances between groups. Among the various hierarchical algorithms the average-link algorithm provides the best accuracy in most cases [97]. A main advantage of the use of hierarchical clustering algorithms is the visualization capability that shows how much objects in the data set are similar one another. In addition, with the utilization of a dendrogram, researchers can reasonably guess the number of clusters. This is a distinguishing feature of hierarchical algorithms because other clustering algorithms cannot provide this very useful feature, especially when there is no additional information available about the data itself.

2.3.3 Density Based Clustering

Density based clustering methods play a very important role in biomedical research because they are capable of handle any cluster of arbitrary shape. Recent researches prove that this method can be efficiently and effectively beneficial in order to extract the meaningful patterns from a very large database which mainly consist biomedical images. Besides density based clustering method, partitional clustering and hierarchical clustering methods do not extract the meaningful patterns from biomedical images database because these two methods are capable to handle only the clusters of spherical shape not the clusters of arbitrary shape. To remove the problem of patitional clustering and hierarchical clustering methods density based clustering method evolved. On the basis of density distribution function following are the main approaches of density based clustering: DBSCAN, OPTICS, and DENCLUE [3]. Celebi *et al.*, used density based clustering approach in order to obtained the useful patterns from a very large biomedical images database. These patterns play a very important role in order to determine homogeneous colour [99]. There are various advantages of the density based clustering. First one is that in density based clustering approach in advance the number of clusters does not required. Another one is that, it can easily handle the clusters of any arbitrary shape. And the last one is that it can be used very effectively and efficiently even in noisy situations. In other words it is performed equally well in noisy situations. One major disadvantage of density based clustering is that, if there is a lot of variation in densities along with data points then it cannot be able to handle such variations in data points. Another one disadvantage is that, distance measure is the primary factor which calculates its result.

Clustering Examples in Healthcare

Chen *et al.*, proposed hierarchical K-means regulating divisive or agglomerative approach for better analyzing large micro-array data. It was reported that divisive hierarchical K-means was superior to hierarchical and K-means clustering on cluster quality as well as on computational speed. Apart from this, it was also mentioned that divisive hierarchical K-means establishes a better clustering algorithm satisfying researcher's demand [100].

Chipman *et al.*, proposed the hybrid hierarchical clustering approach for analyzing microarray data [101]. In this research, the proposed hybrid clustering approach combines bottom-up as well as top-down hierarchical clustering concepts in order to effectively and efficiently utilizes the strength of both concepts for analyzing micro-array data. The proposed approach was built on a mutual cluster. A mutual cluster is a group of points closer to each other than to any other points. The research demonstrates the proposed technique on simulated as well as on real micro-array data.

Bertsimas *et al*., used the approaches of classifications trees and clustering algorithms in order to predict the cost of healthcare [102] by using the dataset of three years collected from the insurance companies to perform the experiment. On the basis of analysis, following results were obtained in this research. First result shows that in order to provide accurate prediction of medical costs and to represent a powerful tool for prediction of healthcare costs data mining methods provide better accuracy. Another result shows that in order to predict the future costs pattern of past data was useful.

Belciug *et al.,* used the agglomerative hierarchical clustering approach for grouping the patients according to their length of stay in the hospital in order to provide better utilization of hospital resources and provide better services to patients [103].

Tapia *et al.*, analyzed the gene expression data with the help of a new hierarchical clustering approach using genetic algorithm. In this research, the main focus was on regeneration of protein-protein functional interactions from genomic data. In this research, the proposed algorithm can predicate the functional associations accurately by considering genomic data [104].

Soliman *et al.,* proposed a hybrid approach for better analyzing the cancer diseases on the basis of informative genes. The proposed approach used the K-means clustering with statistical analysis (ANOVA) for gene selection and SVM to classify the cancer diseases. On the basis of experiments that were performed on micro-array data, it has been found that the accuracy of K-means clustering with the combination of statistical analysis was better [105].

Schulam *et al.,* proposed a Probabilistic Subtyping Model (PSM) which was mainly designed in order to discovered subtypes of complex, systematic diseases using longitudinal clinical markers collected in electronic health record (EHR) databases and patient registries. Proposed model was a model for clustering time series of clinical markers obtained from routine visits in order to identify homogeneous patient subgroups [106].

Belciug *et al.,* concluded that among hierarchical, partitional, and density based clustering, the hierarchical clustering was provided effective utilization of hospital resources and provided improved patient care services in healthcare [103].

Lu *et al.,* proposed an Adaptive Benford Algorithm in the application area of healthcare insurance claims. The proposed algorithm was a digital analysis technique that utilizes an unsupervised learning approach in order to handle incomplete or missing data. This technique was applied to the detection of fraud and abuse in the health insurance claims using dataset, real health insurance data. The dataset was analyzed. After the experimental analysis of dataset, it has been found that the proposed algorithm has significantly enhanced precision than the traditional Benford Approach in the detection of fraud and abuse in health insurance claims. Apart from this, it has been also found that the proposed algorithm was not restricted to already known instances of fraud [107].

Peng Y *et al.,* used clustering technique in order to detect the suspicious healthcare fraudsfrom large databases. In the research, two clustering methods, SAS EM and CLUTO were used to a large real-life health insurance dataset. After comparison, it has been found that SAS EM outperforms the CLUTO [108].

2.4 Association

In 1991, Piatetsky-Shapiro introduced the first association rule mining algorithm (called KID3) [109]. But this algorithm did not receive much attention because of its inefficiency and its serious scalability issues. But after R.Agarwal and his colleagues at IBM Almaden Research Center introduced a novel association rule algorithm called Apriori [110,111], association mining has received significant attention. This attention occurred because Apriori resolved the issues identified in KID3 using the "Apriori property" so that association mining can be applied to real databases to extract association rules. In order to find out the frequent patterns, interesting relationships among a set of data items in the data repository association is one of the most essential approaches of data mining is used. It has great impact in the healthcare field to detect the relationships among diseases, health state and symptoms. Researchers currently used this approach in order to determine the relationships between various diseases and their prescribed drugs. This approach is widely used by the healthcare insurance companies in order to determine the fraud and abuse. Accuracy is not an evaluation factor in association mining because every association algorithm mines all association rules. Efficiency is the only evaluation factor and the main goal of association mining algorithms.

2.4.1 Apriori Algorithm

It was coined by R.Agarwal *et al* [110,111], in 1994. The Apriori algorithm requires two user inputs: first one is support and second is confidence (as percentages). This is because users are interested in association rules (sets of transactions) that frequently occur in a database (i.e., support) and which are accurate (i.e., confidence). Thus, support and confidence are used to filter out many uninteresting association rules. The core property of the algorithm is its use of the Apriori property. Thus, if an item is not frequent (i.e., not satisfying support in terms of transaction), its descendants are not frequent (e.g., if male breast cancer cases are not frequent, no association rules related to the disease are generated.). This property significantly limits the search for frequent item sets and considerably improves the efficiency of the algorithm. Hash table and other methods are currently analyzed in order to improve the efficiency of this approach. [112] [113].

Association Examples in Healthcare

Medical practices, insurance companies and various other types of health organizations are involved in order to collect very huge amounts of data. Due to this a number of researchers are attracted towards it in order to explorer it and tries to find out something beneficent from it. In order to find out some beneficent from such a huge amounts of data apriori algorithm is very useful. Abdullah *et al.*, used this algorithm in order to find out the associations between diagnosis and treatments [114]. In this research, on the basis of results it has been found that the apriori algorithm was equally beneficent for finding large item sets as well as for generating associations rules in medical billing data. In this research, the main motivation of using apriori algorithm was the similarity between medical bill and purchase bill [114].

Patil *et al.*, used apriori algorithm in order to generate association rules. In this research, by using association rules the patients suffering from type-2 diabetes were classified. In this research, an approach for discretizing the attributes having continuous value using equal width bining interval which was selected on the basis of medical expert's opinion has been proposed [115].

Ying *et al.*, proposed a data mining association approach on the basis of fuzzy recognition-prime decision (RPD) model for developing the relationship between drugs and their associated adverse drug reactions (ADRs). The proposed approach was tested on the real patient data which is obtained from Veterans Affair Medical Center in Detroit, Michigan [116].

Ilayaraja *et al.*, used Apriori algorithm to discover frequent diseases in medical data. This study proposed a method for detecting the occurrence of diseases using Apriori algorithm in particular geographical locations at particular period of time [117].

Nahar *et al.,* used predictive apriori approach for generating the rules for heart diseasepatients. In this research work rules were produced for healthy and sick people. Based on these rules, this research discovered the factors which caused heart problem in men and women. After analyzing the rules authors conclude that women have less possibility of having coronary heart disease as compare to men [118].

Kai *et al.,* proposed a clinical decision support system in order to helps the healthcare workers to identify the noncommunicable diseases in non-reachable communities. Proposed system was a remote healthcare consultancy system. Associate rule technique was applied in order to build proposed system [119].

3. DATA MINING CHALLENGES IN HEALTHCARE

As we know that a lot of healthcare data is generated and stored by various healthcare organizations. But there are various challenges related to healthcare data which may play serious hurdles in the making proper decisions. The first challenge with healthcare data is the format of data being stored is different in different healthcare organizations. Till date there is no standard format is laid down for data being stored. In epidemic situations this lack of standard format can make the epidemic situations even more worse. Suppose that an epidemic disease is spread within a country at its different geographical regions. The country health ministry requires that all the healthcare organizations must share their healthcare data with its centralized data warehouse for analysis in order to take all the essential steps so that epidemic situation may get resolve. But since the formats of data is different. Hence, the analysis of data may take longer time than usual. Due to this it may be possible that the situation may become out of control. The healthcare data is very useful in order to extract the meaningful information from it for improving the healthcare services for the patients. To do this quality of data is very important because we cannot extract the meaningful information from that data which have no quality. Hence, the quality of data is another very important challenge. The quality of data depends on various factors such as removal of noisy data, free from missing of data etc. All the necessary steps must be taken in order to maintain the quality in healthcare data. Data sharing is another major challenge. Neither patients nor healthcare organizations are interested in sharing of their private data. Due to this the epidemic situations may get worse, planning to provide better treatments for a large population may not be possible, and difficulty in the detection of fraud and abuse in healthcare insurance companies etc. Another challenge is that in order to build the data warehouse where all the healthcare organizations within a country share their data is very costly and time consuming process.

4. CONCLUSION AND FUTURE WORK

For any algorithm its accuracy and performance is of greater importance. But due to presence of some factors any algorithm can greatly lost the above mentioned property of accuracy and performance. Classification is also belongs to such an algorithm. Classification algorithm is very sensitive to noisy data. If any noisy data is present then it causes very serious problems regarding to the processing power of classification. It not only slows down the task of classification algorithm but also degrades its performance. Hence, before applying classification algorithm it must be necessary to remove all those attributes from datasets who later on acts as noisy attributes. Feature selection methods play a very important role in order to select those attributes who improves the performance of classification algorithm.

Clustering techniques are very useful especially in pattern recognitions. But they suffer from a problem on choosing the appropriate algorithm because regarding datasets they do not have information. We can choose partitional algorithm only when we know the number of clusters. Hierarchical clustering is used even when we do not know about the number of clusters. Hierarchical clustering provides better performance when there is less datasets but as soon as volume of datasets increases its performance degrades. To overcome this problem random sampling is very beneficial.

In hierarchical clustering, if the data is too large to be presented in a dendrogram, the visualization capability is very poor. One possible solution to this problem is to randomly sample the data so that users can properly understand the overall grouping/similarity of the data using the dendrogram that is generated with the sampled data. The main drawback to the use of hierarchical clustering algorithms is cubic time complexity. This complexity is such that the algorithms are very much limited for very large data sets. As the result, the hierarchical algorithms are much slower (in computational time) than partitional clustering algorithms. They also use a huge amount of system memory to calculate distances between objects.

The privacy regarding to patient's confidential information is very important. Such type of privacy may be lost during sharing of data in distributed healthcare environment. Necessary steps must be taken in order to provide proper security so that their confidential information must not be accessed by any unauthorized organizations. But in situations like epidemic, planning better healthcare services for a very large population etc. some confidential data may be provided to the researchers and government organizations or any authorized organizations.

In order to achieve better accuracy in the prediction of diseases, improving survivability rate regarding serious death related problems etc. various data mining techniques must be used in combination.

To achieve medical data of higher quality all the necessary steps must be taken in order to build the better medical information systems which provides accurate information regarding to patients medical history rather than the information regarding to their billing invoices. Because high quality healthcare data is useful for providing better medical services only to the patients but also to the healthcare organizations or any other organizations who are involved in healthcare industry.

Takes all necessary steps in order to minimize the semantic gap in data sharing between distributed healthcare databases environment so that meaningful patterns can be obtained. These patterns can be very useful in order to improve the treatment effectiveness services, to better detection of fraud and abuse, improved customer relationship management across the world.

5. REFERENCES

[1] D. Hand, H. Mannila and P. Smyth, "Principles of data mining", MIT, **(2001)**.

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

[3] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, “The KDD process of extracting useful knowledge form volumes of data.commun.”, ACM, vol. 39, no. 11, **(1996)**, pp. 27-34.

[4] J. Han and M. Kamber, “Data mining: concepts and techniques”, 2nd ed. The Morgan Kaufmann Series, **(2006)**.

[5] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, “From data mining to knowledge discovery in databases”, Commun. ACM, vol. 39, no. 11, **(1996)**, pp. 24-26.

[6] C. McGregor, C. Christina and J. Andrew, “A process mining driven framework for clinical guideline improvement in critical care”, Learning from Medical Data Streams 13th Conference on Artificial Intelligence in Medicine (LEMEDS). http://ceur-ws. org, vol. 765, **(2012)**.

[7] P. R. Harper, “A review and comparison of classification algorithms for medical decision making”, Health Policy, vol. 71, **(2005)**, pp. 315-331.

[8] V. S. Stel, S. M. Pluijm, D. J. Deeg, J. H. Smit, L. M. Bouter and P. Lips, “A classification tree for predicting recurrent falling in community-dwelling older persons”, J. Am. Geriatr. Soc., vol. 51, **(2003)**, pp. 1356-1364.

[9] R. Bellazzi and B. Zupan, “Predictive data mining in clinical medicine: current issues and guidelines”, Int. J. Med. Inform., vol. 77, **(2008)**, pp. 81-97.

[10] R. D. Canlas Jr., “Data Mining in Healthcare:Current Applications and Issues”, **(2009)**.

[11] F. Hosseinkhah, H. Ashktorab, R. Veen, M. M. Owrang O., “Challenges in Data Mining on Medical Databases”, IGI Global, **(2009)**, pp. 502-511.

[12] M. Kumari and S. Godara, “Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction”, IJCST ISSN: 2229- 4333, vol. 2, no. 2, **(2011)** June.

[13] J. Soni, U. Ansari, D. Sharma and S. Soni, “Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction”, **(2011)**.

[14] C. S. Dangare and S. S. Apte, “Improved Study of Heart Disease Prediction System Using Data Mining Classification Techniques”, **(2012)**.

[15] K. Srinivas, B. Kavihta Rani and Dr. A.Govrdhan, “Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks”, International Journal on Computer Science and Engineering, vol. 02, no. 02, **(2010)**, pp. 250-255.

[16] A. Aljumah, M. G.Ahamad and M. K. Siddiqui, “Predictive Analysis on Hypertension Treatment Using Data Mining Approach in Saudi Arabia”, Intelligent Information Management, vol. 3, **(2011)**, pp. 252-261.

[17] D. Delen, “Analysis of cancer data: a data mining approach”, **(2009)**.

[18] O. Osofisan, O. O. Adeyemo, B. A. Sawyerr and O. Eweje, “Prediction of Kidney Failure Using Artificial Neural Networks”, **(2011)**.

[19] S. Floyd, “Data Mining Techniques for Prognosis in Pancreatic Cancer”, **(2007)**.

[20] M.-J. Huang, M.-Y. Chen and S.-C. Lee, “Integrating data mining with case-based reasoning for chronic diseases prognosis and diagnosis”, Expert Systems with Applications, vol. 32, **(2007)**, pp. 856-867.

[21] S. Gupta, D. Kumar and A. Sharma, “Data Mining Classification Techniques Applied For Breast Cancer Diagnosis And Prognosis”, **(2011)**.

[22] K. S. Kavitha, K. V. Ramakrishnan and M. K. Singh, “Modeling and design of evolutionary neural network for heart disease detection”, IJCSI International Journal of Computer Science Issues, ISSN (Online): 1694-0814, vol. 7, no. 5, **(2010)** September, pp. 272-283.

[23] S. H. Ha and S. H. Joo, “A Hybrid Data Mining Method for the Medical Classification of Chest Pain”, International Journal of Computer and Information Engineering, vol. 4, no. 1, **(2010)**, pp. 33-38.

[24] R. Parvathi and S. Palaniammali, “An Improved Medical Diagnosing Technique Using Spatial Association Rules”, European Journal of Scientific Research ISSN 1450-216X, vol. 61, no. 1, **(2011)**, pp. 49-59.

[25] S. Chao and F. Wong, “An Incremental Decision Tree Learning Methodology Regarding Attributes in Medical Data Mining”, **(2009)**.

[26] Habrard, M. Bernard and F. Jacquenet, “Multi-Relational Data Mining in Medical Databases”, Springer-Verlag, **(2003)**.

[27] S. B. Patil and Y. S. Kumaraswamy, “Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network”, European Journal of Scientific Research ISSN 1450-216X, © EuroJournals Publishing, Inc., vol. 31, no. 4, **(2009)**, pp. 642-656.

[28] A.Shukla, R. Tiwari, P. Kaur, Knowledge Based Approach for Diagnosis of Breast Cancer, IEEE International Advance Computing Conference,IACC 2009.

[29] L. Duan, W. N. Street & E. Xu, Healthcare information systems: data mining methods in the creation of a clinical recommender system, Enterprise Information Systems, 5:2, pp169-181 , 2011.

[30] D. S. Kumar, G. Sathyadevi and S. Sivanesh, “Decision Support System for Medical Diagnosis Using Data Mining”, **(2011)**.

[31] S. Palaniappan and R. Awang, “Intelligent Heart Disease Prediction System Using Data Mining Techniques”, **(2008)**.

[32] Alpaydin, E. (1997), Voting over Multiple Condensed Nearest Neighbors. Artificial Intelligence Review, p. 115–132.

[33] Bramer, M., (2007) Principles of data mining: Springer.

[34] Goharian & Grossman, Data Mining Classification, Illinois Institute of Technology, http://ir.iit.edu/~nazli/cs422/CS422-Slides/DM-Classification.pdf, **(2003)**.

[35] 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/fgcsaptewe issue_with_cover.pdf, **(1997)**.

[36] Pagallo, G. and Huassler, D., Boolean feature discovery in empirical learning, Machine Learning, 5(1): 71-99, 1990.

[37] Quinlan, J. R., C4.5: Programs for Machine Learning, Morgan Kaufmann, Los Altos, 1993.

[38] V. Vapnik, “Statistical Learning Theory”, Wiley, **(1998)**.

[39] V. Vapnik, “The support vector method of function estimation”, **(1998)**.

[40] N. Chistianini and J. Shawe-Taylor, “An Introduction to Support Vector Machines, and other kernel-based learning methods”, Cambridge University Press, **(2000)**.

[41] N. Cristianini and J. Shawe-Taylor, “An Introduction to Support Vector Machines”, Cambridge University Press, **(2000)**.

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

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

[44] Bellazzi, R., and Zupan, B., Predictive data mining in clinical medicine: current issues and guidelines. Int. J. Med. Inform. 77:81–97, 2008.

[45] Übeyli, E. D., Comparison of different classification algorithms in clinical decision making. Expert syst 24(1):17–31, 2007.

[46] Kaur, H., and Wasan, S. K., Empirical study on applications of data mining techniques in healthcare. J. Comput. Sci. 2(2):194–200 2006.

[47] Romeo, M., Burden, F., Quinn, M., Wood, B., and McNaughton, D., Infrared microspectroscopy and artificial neural networks in the diagnosis of cervical cancer. Cell. Mol. Biol. (Noisy-le-Grand, France) 44(1):179, 1998.

[48] Ball, G., Mian, S., Holding, F., Allibone, R., Lowe, J., Ali, S., et al., An integrated approach utilizing artificial neural networks and SELDI mass spectrometry for the classification of human tumours and rapid identification of potential biomarkers. Bioinformatics 18(3):395–404, 2002.

[49] Aleynikov, S., and Micheli-Tzanakou, E., Classification of retinal damage by a neural network based system. J. Med. Syst. 22(3):129–136, 1998.

[50] Potter, R., Comparison of classification algorithms applied to breast cancer diagnosis and prognosis, advances in data mining,7th Industrial Conference, ICDM 2007, Leipzig, Germany, July 2007, pp.40–49.

[51] Kononenko, I., Bratko, I., and Kukar, M., Application of machine learning to medical diagnosis. Machine Learning and Data Mining: Methods and Applications 389:408, 1997.

[52] Sharma, A., and Roy, R. J., Design of a recognition system to predict movement during anesthesia. IEEE Trans. Biomed. Eng.44(6):505–511, 1997.

[53] Einstein, A. J., Wu, H. S., Sanchez, M., and Gil, J., Fractal characterization of chromatin appearance for diagnosis in breast cytology. J. Pathol. 185(4):366–381, 1998.

[54] Brickley, M., Shepherd, J. P., and Armstrong, R. A., Neural networks: a new technique for development of decision support systems in dentistry. J. Dent. 26(4):305–309, 1998.

[55] M. H. Dunham, “Data mining introductory and advanced topics”, Upper Saddle River, NJ: Pearson Education, Inc., **(2003)**.

[56] Schwarzer, G., Vach, W., and Schumacher, M., On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology. Stat. Med. 19:541–561, 2000.

[57] H. Hu, J. Li, A. Plank, H. Wang and G. Daggard, “A Comparative Study of Classification Methods For Microarray Data Analysis”, Proc. Fifth Australasian Data Mining Conference (AusDM2006), Sydney, Australia. CRPIT, ACS, vol. 61, **(2006)**, pp. 33-37.

[58] R. Potter, “Comparison of classification algorithms applied to breast cancer diagnosis and prognosis”, advances in data mining, 7th Industrial Conference, ICDM 2007, Leipzig, Germany, **(2007)** July, pp. 40-49.

[59] L. Huang, H. C. Liao and M. C. Chen, “Prediction model building and feature selection with support vector machines in breast cancer diagnosis”, Expert Systems with Applications, vol. 34, **(2008)**, pp. 578-587.

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

[61] M. U. Khan, J. P. Choi, H. Shin and M. Kim, “Predicting Breast Cancer Survivability Using Fuzzy Decision Trees for Personalized Healthcare”, 30th Annual International IEEE EMBS Conference Vancouver, British Columbia, Canada, **(2008)** August 20-24.

[62] L. Chang and C. H. Chen, “Applying decision tree and neural network to increase quality of dermatologic diagnosis”, Expert Systems with Applications, Elsevier, vol. 36, **(2009)**, pp. 4035-4041.

[63] R. Das, I. Turkoglub and A. Sengur, “Effective diagnosis of heart disease through neural networks ensembles”, Expert Systems with Applications, vol. 36, **(2009)**, pp. 7675-7680.

[64] I. Curiac, G. Vasile, O. Banias, C. Volosencu and A. Albu, “Bayesian Network Model for Diagnosis of Psychiatric Diseases”, Proceedings of the ITI 2009 31st Int. Conf. on Information Technology Interfaces, Cavtat, Croatia, **(2009)** June 22-25.

[65] K. F. R. Liu and C. F. Lu, “BBN-Based Decision Support for Health Risk Analysis”, Fifth International Joint Conference on INC, IMS and IDC, **(2009)**.

[66] Avci, “A new intelligent diagnosis system for the heart valve diseases by using genetic-SVM classifier”, Expert Systems with Applications, Elsevier, vol. 36, **(2009)**, pp. 10618-10626.

[67] S. Gunasundari and S. Baskar, “Application of Artificial Neural Network in identification of Lung Diseases”, Nature & Biologically Inspired Computing, 2009. NaBIC 2009. World Congress on. IEEE, **(2009)**.

[68] S. Soni and O. P. Vyas, “Using Associative Classifiers for Predictive Analysis in Health Care Data Mining”,

International Journal of Computer Applications (0975 – 8887), vol. 4, no. 5, **(2010)** July.

[69] S. W. Fei, "Diagnostic study on arrhythmia cordis based on particle swarm optimization-based support vector machine", Expert Systems with Applications, Elsevier, vol. 37, **(2010)**, pp. 6748-6752.

[70] O. Er, N. Yumusakc and F. Temurtas, "Chest diseases diagnosis using artificial neural networks", Expert Systems with Applications, vol. 37, **(2010)**, pp. 7648-7655.

[71] Chuang, L.Y., Wu, K.C., Chang, H.W. and Yang, C.H. (2011) "Support Vector Machine-Based Prediction for Oral Cancer Using Four SNPs in DNA Repair Genes". Proceedings of the International Multi Conference of Engineers and Computer Scientists, Hong Kong,16-18 March 2011, 16-18.

[72] A. Bakar, Z. Kefli, S. Abdullah and M. Sahani, "Predictive Models for Dengue Outbreak Using Multiple Rulebase Classifiers", 2011 International Conference on Electrical Engineering and Informatics, Bandung, Indonesia, **(2011)** July 17-19.

[73] S. S. Moon, S. Y. Kang, W. Jitpitaklert and S. B. Kim, "Decision tree models for characterizing smoking patterns of older adults", Expert Systems with Applications, Elsevier, vol. 39, **(2012)**, pp. 445-451.

[74] H. Jena, C. C. Wang, B. C. Jiangc, Y. H. Chub and M. S. Chen, "Application of classification techniques on development an early-warning systemfor chronic illnesses", Expert Systems with Applications, vol. 39, **(2012)**, pp. 8852-8858.

[75] Chien and G. J. Pottie, "A Universal Hybrid Decision Tree Classifier Design for Human Activity Classification", 34th Annual International Conference of the IEEE EMBS San Diego, California USA, **(2012)** August 28-September 1.

[76] M. Shouman, T. Turner and R. Stocker, "Applying K-Nearest Neighbour in Diagnosing Heart Disease Patients", International Conference on Knowledge Discovery (ICKD-2012), **(2012)**.

[77] Y. Liu, H. L. Chen, B. Yang, X. E. Lv, N. L. Li and J. Liu, "Design of an Enhanced Fuzzy k-nearest Neighbor Classifier Based Computer Aided Diagnostic System for Thyroid Disease", Journal of Medical System, Springer, **(2012)**.

[78] Hattice and K. Metin, "A Diagnostic Software tool for Skin Diseases with Basic and Weighted K-NN", Innovations in Intelligent Systems and Applications (INISTA), **(2012)**.

[79] M. J. Abdi and D. Giveki, "Automatic detection of erythemato-squamous diseases using PSO–SVM based on association rules", Engineering Applications of Artificial Intelligence, vol. 26, **(2013)**, pp. 603-608.

[80] W. L. Zuoa, Z. Y. Wanga, T. Liua and H. L. Chenc, "Effective detection of Parkinson's disease using an adaptive fuzzy k-nearest neighbor approach", Biomedical Signal Processing and Control, Elsevier, **(2013)**, pp. 364-373.

[81] Rusdah and Edi Winarko, "Review on Data Mining Methods for Tuberculosis Diagnosis" Information Systems International Conference (ISICO), 2 – 4 December 2013.

[82] Marina Evrim Johnson and Nagen Nagarur, "Multi-stage methodology to detect health insurance claim fraud", Health Care Management Science, DOI 10.1007/s10729-015-9317-3, Springer, 20 January 2015.

[83] Johannes Arthuur Govaert, Anne Charlotte Madeline van Bommel, Wouter Antonie van Dijk, Nicoline Johanneke van Leersum, Robertus Alexandre Eduard Mattheus Tollenaar and Michael Wilhemus Jacobus Maria Wouters, "Reducing Healthcare Costs Facilitated by Surgical Auditing: A Systematic Review", Worls J Surg, DOI 10.1007/s00268-015-3005-9, Springer, 18 February 2015.

[84] Peng Z *et al.*, "An expression signature at diagnosis to estimate prostate cancer patients' overall survival", Prostate Cancer and Prostatic Disease (2014) 17, 81-90, doi 10.1038/pcan.2013.57; January 2014, Nature.

[85] J. Fox, "Applied Regression Analysis, Linear Models, and Related Methods", **(1997)**.

[86] Gennings, R. Ellis and J. K. Ritter, "Linking empirical estimates of body burden of environmental chemicals and wellness using NHANES data", http://dx.doi.org/10.1016/j.envint.2011.09.002,2011.

[87] P. A. Gutiérrez, C. Hervás-Martínez and F. J. Martínez-Estudillo, "Logistic Regression by Means of Evolutionary Radial Basis Function Neural Networks", IEEE Transactions on Neural Networks, vol. 22, no. 2, **(2011)**, pp. 246-263.

[88] Divya and S. Agarwal, "Weighted Support Vector Regression approach for Remote Healthcare monitoring", IEEE-International Conference on Recent Trends in Information Technology, ICRTIT 2011, 978-1-4577-0590-8/11/$26.00 ©2011 IEEE MIT, Anna University, Chennai, **(2011)** June 3-5.

[89] Gennings, R. Ellis and J. K. Ritter, "Linking empirical estimates of body burden of environmental chemicals and wellness using NHANES data", http://dx.doi.org/10.1016/j.envint.2011.09.002,2011.

[90] Xie *et al., "*Predicting Days in Hospital Using Health Insurance Claims", IEEE Journal of Biomedical and Health Informatics, DOI 10.1109/JBHI.2015.2402692,ISSN 2168-291, IEEE Transactions, 2015.

[91] J. Alapont, A. Bella-Sanjuán, C. Ferri, J. Hernández-Orallo, J. D. Llopis-Llopis and M. J. Ramírez-Quintana, "Specialised Tools for Automating Data Mining for Hospital Management", http://www.dsic.upv.es/~abella/papers/HIS_DM.pdf, **(2005)**.

[92] K. Jain, M. N. Murty and P. J. Flynn, "Data Clustering: a review", ACM Compute, Surveys, vol. 31, **(1996)**.

[93] Hamerly and C. Elkan, "Learning the K in K-means", Proceedings of the 17th Annual Conference on Neural Information Processing Systems, British Columbia, Canada, **(2003)**.

[94] L. Lenert, A. Lin, R. Olshen and C. Sugar, "Clustering in the Service of the Public's Health", http://www-stat.stanford.edu/~olshen/manuscripts/helsinki.PDF.

[95] S. Belciug, F. Gorunescu, A. Salem and M. Gorunescu, “Clustering-based approach for detecting breast cancer recurrence”, 10th International Conference on Intelligent Systems Design and Applications, **(2010)**.

[96] J. Escudero, J. P. Zajicek and E. Ifeachor, “Early Detection and Characterization of Alzheimer’s Disease in Clinical Scenarios Using Bioprofile Concepts and K-Means”, 33rd Annual International Conference of the IEEE EMBS Boston, Massachusetts USA, **(2011)** August 30-September 3.

[97] Yoo, I., and Hu, X., A comprehensive comparison study of document clustering for a biomedical digital library MDELINE. ACM/IEEE Joint Conference on Digital Libraries 11–15:220–229, 2006. Chapel Hill, NC, June 11–15, 2006.

[98] Yoo, I., Hu, X., and Song, I.-Y., Biomedical ontology improves biomedical literature clustering performance: a comparison study. Int. J. Bioinform. Res. Appl. 3(3):414–428, 2007.

[99] M. E. Celebi, Y. A. Aslandogan and R. P. Bergstresser, “Mining Biomedical Images with Density-based Clustering”, Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC’05), **(2005)**.

[100] T. S. Chen, T. H. Tsai, Y. T. Chen, C. C. Lin, R. C. Chen, S. Y. Li and H. Y. Chen, “A Combined K-Means and Hierarchical Clustering Method for improving the Clustering Efficiency of Microarray”, Proceedings of 2005 International Symposium on Intelligent Signal Processing and Communication Systems, **(2005)**.

[101] Chipman and R. Tibshirani, “Hybrid hierarchical clustering with applications to microarray data”, Biostatistics, vol. 7, no. 2, **(2006)**, pp. 286-301.

[102] Bertsimas, M. V. Bjarnadóttir, M. A. Kane, J. C. Kryder, R. Pandey, S. Vempala and G. Wang, “Algorithmic prediction of health-care costs”, Oper. Res., vol. 56, no. 6, **(2008)**, pp. 1382-1392.

[103] S. Belciug, “Patients length of stay grouping using the hierarchical clustering algorithm”, Annals of University of Craiova, Math. Comp. Sci. Ser., ISSN: 1223-6934, vol. 36, no. 2, **(2009)**, pp. 79-84.

[104] J. J. Tapia, E. Morett and E. E. Vallejo, “A Clustering Genetic Algorithm for Genomic Data Mining”, Foundations of Computational Intelligence, vol. 4 Studies in Computational Intelligence, vol. 204, **(2009)**, pp. 249-275.

[105] T. H. A. Soliman, A. A. Sewissy and H. A. Latif, “A Gene Selection Approach for Classifying Diseases Based on Microarray Datasets”, 2nd International Conference on Computer Technology and Development (lCCTD 2010), **(2010)**.

[106] Schulam *et al.,* “Clustering Longitudinal Clinical Marker Trajectories from Electronic Health Data: Applications to Phenotyping and Endotype Discovery”, Associations for the Advancements of Artificial Intelligence, 2015.

[107] Fletcher Lu and J. Efrim Boritz, “Detection Fraud in Health Insurance Data: Learning to Model Incomplete Benford’s Law Distributions”, Machine Learning: EMCL 2005 16th European Conference on Machine Learning, Porto, Portugal,October 3-7, 2005 Volume 3720, pages 633-640, Springer.

[108] Peng Y, Kou G, Sabatka A, Chen Z, Khazanchil D and Shi Y,” Application Of clustering methods to health insurance fraud detection”, Int Conf Serv Syst Serv Manag 1:116-120, 2006.

[109] Piatetsky-Shapiro, G., Discovery, analysis, and presentation of strong rules. In: Piatetsky-Shapiro, G., (Ed.), Knowledge Discovery in Databases. AAAI/MIT Press, 1991, pp. 229–248.

[110] Agrawal, R., Imielinski, T., and Swami, A., Mining association rules between sets of items in large databases, Proceedings of the ACM SIGMOD International Conference on the Management of Data. ACM, Washington DC, pp. 207–216, 1993.

[111] Agrawal, R., and Srikant, R., Fast algorithms for mining association rules, Proceedings of the 20th International Conference on Very Large Data Bases (VLDB’94). Morgan Kaufmann, Santiago, pp. 487–499, 1994.

[112] J. Yanqing, H. Ying, J. Tran, P. Dews, A. Mansour and R. Michael Massanari, “Mining Infrequent Causal Associations in Electronic Health Databases”, 11th IEEE International Conference on Data Mining Workshops, **(2011)**.

[113] R. Agrawal and R. Srikant, “Fast Algorithms for Mining Association Rules”, VLDB, Chile, ISBN 1-55860-153-8, **(1994)** September 12-15, pp. 487-99.

[114] U. Abdullah, J. Ahmad and A. Ahmed, “Analysis of Effectiveness of Apriori Algorithm in Medical Billing Data Mining”, 2008 International Conference on Emerging Technologies, IEEE-ICET 2008, Rawalpindi, Pakistan, **(2008)** October 18-19.

[115] M. Patil, R. C. Joshi and D. Toshniwal, “Association rule for classification of type -2 diabetic patients”, Second International Conference on Machine Learning and Computing, **(2010)**.

[116] J. Yanqing, H. Ying, J. Tran, P. Dews, A. Mansour and R. Michael Massanari, “Mining Infrequent Causal Associations in Electronic Health Databases”, 11th IEEE International Conference on Data Mining Workshops, **(2011)**.

[117] M. Ilayaraja and T. Meyyappan, “Mining Medical Data to Identify Frequent Diseases using Apriori Algorithm”, Proceedings of the 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering, **(2013)** February 21-22.

[118] J. Nahar, T. Imam, K. S. Tickle and Y. P. Chen, “Association rule mining to detect factors which contribute to heart disease in males and females”, Expert Systems with Applications, vol. 40, pp. 1086-1093, **(2013)**.

[119] Eiko Kai *et al.,* ”Enpowering the healthcare worker using the Portable Health Clinic”, IEEE Transactions, DOI 10.1109/AINA.2014.108, 2014.

[120] Divya *et al.,*” A Survey on Data Mining Approaches for Healthcare”, International Journal Of Bio-Science and Bio-Technology Vol.5, No.5 (2013),pp.241-266 http://dx.doi.org/10.14257/ijbsbt.2013.5.5.25.