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

A survey on Data Mining approaches for Healthcare

Divya Tomar and Sonali Agarwal

Indian Institute of Information Technology, Allahabad, India
divyatomar26@gmail.com,sonali@iiita.ac.in

Abstract

Data Mining is one of the most motivating area of research that is become increasingly popular in health organization. Data Mining plays an important role for uncovering new trends in healthcare organization which in turn helpful for all the parties associated with this field. This survey explores the utility of various Data Mining techniques such as classification, clustering, association, regression in health domain. In this paper, we present a brief introduction of these techniques and their advantages and disadvantages. This survey also highlights applications, challenges and future issues of Data Mining in healthcare. Recommendation regarding the suitable choice of available Data Mining technique is also discussed in this paper.

Keywords: *Data Mining, Classification, Clustering, Association, Healthcare*

1. Introduction

Data Mining is one of the most vital and motivating area of research with the objective of finding meaningful information from huge data sets. In present era, Data Mining is becoming popular in healthcare field because there is a need of efficient analytical methodology for detecting unknown and valuable information in health data. In health industry, Data Mining provides several benefits such as detection of the fraud in health insurance, availability of medical solution to the patients at lower cost, detection of causes of diseases and identification of medical treatment methods. It also helps the healthcare researchers for making efficient healthcare policies, constructing drug recommendation systems, developing health profiles of individuals *etc*. [1]. The data generated by the health organizations is very vast and complex due to which it is difficult to analyze the data in order to make important decision regarding patient health. This data contains details regarding hospitals, patients, medical claims, treatment cost *etc*. So, there is a need to generate a powerful tool for analyzing and extracting important information from this complex data. The analysis of health data improves the healthcare by enhancing the performance of patient management tasks. The outcome of Data Mining technologies are to provide benefits to healthcare organization for grouping the patients having similar type of diseases or health issues so that healthcare organization provides them effective treatments. It can also useful for predicting the length of stay of patients in hospital, for medical diagnosis and making plan for effective information system management. Recent technologies are used in medical field to enhance the medical services in cost effective manner. Data Mining techniques are also used to analyze the various factors that are responsible for diseases for example type of food, different working environment, education level, living conditions, availability of pure water, health care services, cultural ,environmental and agricultural factors as shown in Figure 1.

ISSN: 2233-7849 IJBSBT

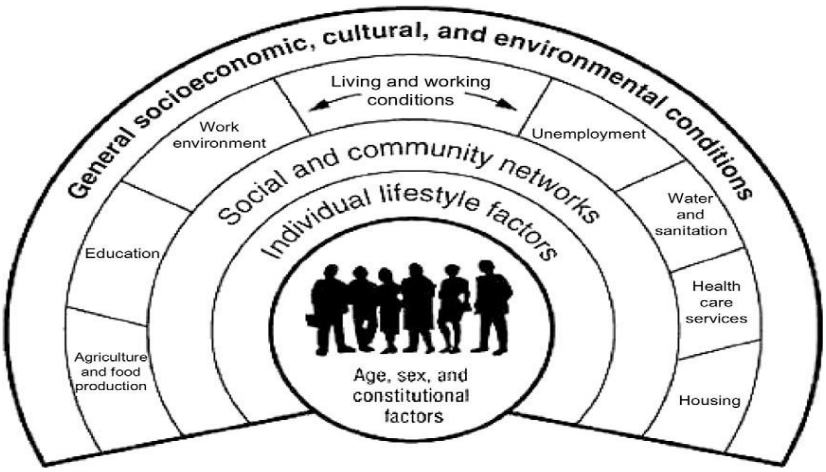

Figure 1. Factors Responsible for Diseases [2]

2. Data Mining

Data Mining came into existence in the middle of 1990's and appeared as a powerful tool that is suitable for fetching previously unknown pattern and useful information from huge dataset. Various studies highlighted that Data Mining techniques help the data holder to analyze and discover unsuspected relationship among their data which in turn helpful for making decision [3]. In general, Data Mining and Knowledge Discovery in Databases (KDD) are related terms and are used interchangeably but many researchers assume that both terms are different as Data Mining is one of the most important stages of the KDD process [4, 5]. According to Fayyad *et al.*, the knowledge discovery process are structured in various stages whereas the first stage is data selection where data is collected from various sources, the second stage is pre-processing of the selected data , the third stage is the transformation of the data into appropriate format for further processing, the fourth stage is Data Mining where suitable Data Mining technique is applied on the data for extracting valuable information and evaluation is the last stage as shown in Figure 2 [4, 6].

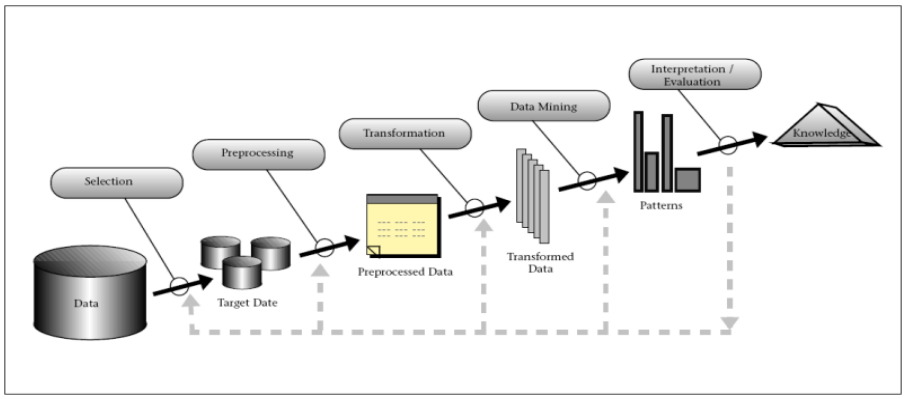

Figure 2. Stages of Knowledge Discovery Process

Skills and knowledge are essential requirement for performing the Data Mining task because the success and failure of Data Mining projects is greatly dependent on the person who are managing the process due to unavailability of standard framework. The

CRISP-DM (CRoss Industry Standard Process for Data Mining) provides a framework for carrying out Data Mining activities. CRISP-DM divides the data mining task into 6 phases. The first phase is the understanding of the business activities while the data for carrying out business activities are collected and analyzed in the second phase. Data pre-processing and modelling is done in the third and fourth phase respectively. Fifth phase evaluates the model and last phase is responsible for deployment of the construed model. McGregor *et al.*, proposed an extended CRISP-DM framework for improving clinical care through integrating the temporal and multidimensional aspects. This model supports the process mining in critical care [7]. Figure 3 represents the CRISP_TDM model for patient care in clinical environment.

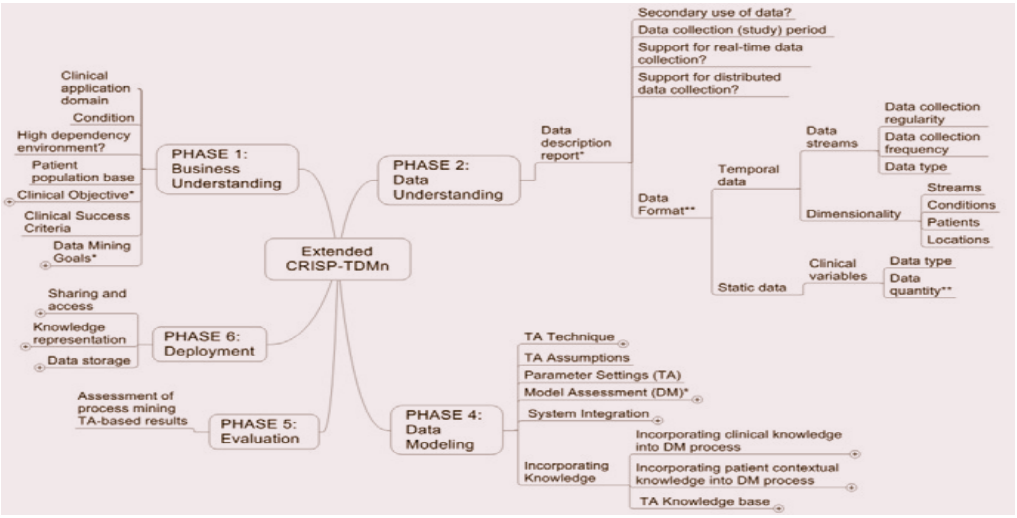

Figure 3. CRISP-TDM Model for Patient Care [7]

In present era various public and private healthcare institutes are producing enormous amounts of data which are difficult to handle. So, there is a need of powerful automated Data Mining tools for analysis and interpreting the useful information from this data. This information is very valuable for healthcare specialist to understand the cause of diseases and for providing better and cost effective treatment to patients. Data Mining offers novel information regarding healthcare which in turn helpful for making administrative as well as medical decision such as estimation of medical staff, decision regarding health insurance policy, selection of treatments, disease prediction *etc.*, [8-11]. Several studies identified with primary focus on various challenges and issues of data mining in healthcare [12, 13]. Data Mining are also used for both analysis and prediction of various diseases [14-23]. Some research work proposed an enhancement in available Data Mining methodology in order to improve the result [24-26] and some studies develop new methodology [27, 28] and framework for healthcare system [29-33]. It is also found that various Data Mining techniques such as classification, clustering and association are used by healthcare organization to increase their capability for making decision regarding patient health. There are ample of research resources available regarding Data Mining tasks which are presented in subsequent sections with their advantages and disadvantages.

2.1. Classification

Classification divides data samples into target classes. The classification technique predicts the target class for each data points. For example, patient can be classified as "high risk" or "low risk" patient on the basis of their disease pattern using data classification approach. 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. Using training dataset we trained the classifier. Correctness of the classifier could be tested using test dataset. Classification is one of the most widely used methods of Data Mining in Healthcare organization. Hu et al. used different classification method such as decision tree, SVM and ensemble approach for analyzing microarray data [34]. This research work performed comparative analysis of above mentioned classification method using 10-fold cross validation approach on the data set obtained from Kent Ridge Bio Medical Dataset repository. The experiment results indicate that among all classification method ensemble achieved good accuracy [34]. Further use of classifier in medical field is discussed by Hatice *et al.*, to diagnosis the skin diseases using weighted KNN classifier [35]. Breast cancer is one of the fatal and dangerous diseases in women. Potter et al. has performed experiment on the breast cancer data set using Weka tool and then analyze the performance of different classifier using 10-fold cross validation method [36]. The research work revealed that there is no single best algorithm which yields better result for every dataset. Classification techniques are also used for predicting the treatment cost of healthcare services which is increases with rapid growth every year and is becoming a main concern for everyone [37]. Bestsimas et al. used classification tree approach to predict the cost of healthcare [38] by using the dataset of 3 years collected from the insurance companies to perform the experiment. The first two year data was used to train the classifier and last one year data was used for comparing the predicted results of classifier. Following are the various classification algorithms used in healthcare:

K-Nearest Neighbour (K-NN)

K-Nearest Neighbour (K-NN) classifier is one of the simplest classifier that discovers the unidentified data point using the previously known data points (nearest neighbour) and classified data points according to the voting system [8]. K-NN classifies the data points using more than one nearest neighbour. K-NN has a number of applications in different areas such as health datasets, image field, cluster analysis, pattern recognition, online marketing *etc*. Jen *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 as shown in figure 4 [39]. Shouman *et al.*, used K-NN classifier for analyzing the patients suffering from heart disease [40]. The data was collected from UCI and experiment was performed using without voting or with voting K-NN classifier and it is found that K-NN achieve better accuracy without voting in diagnosis of heart diseases as compare 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 neighbourhood size [41]. Zuo *et al.*, also introduced an adaptive Fuzzy K-NN approach for Parkinson disease [42].

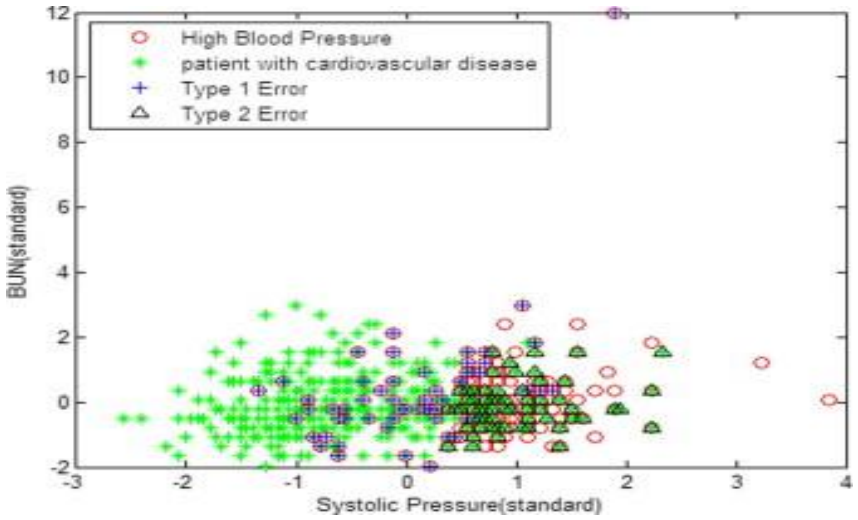

Figure 4. K-NN Classifier for Chronic Disease [39]

Decision Tree (DT)

DT is similar 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. The node at the top most labels in the tree is called root node. For example we have a financial institution decision tree which is used to decide that a person must grant the loan or not. Building a decision for any problem doesn't need any type of domain knowledge. Decision Trees is a classifier that use tree-like graph. The most common use of Decision Tree is in operations research analysis for calculating conditional probabilities [43]. Using Decision Tree, decision makers can choose best alternative and traversal from root to leaf indicates unique class separation based on maximum information gain [44]. Decision Tree is widely used by many researchers in healthcare field. Figure 5 shows classification of a patient into high risk and low risk category using decision tree.

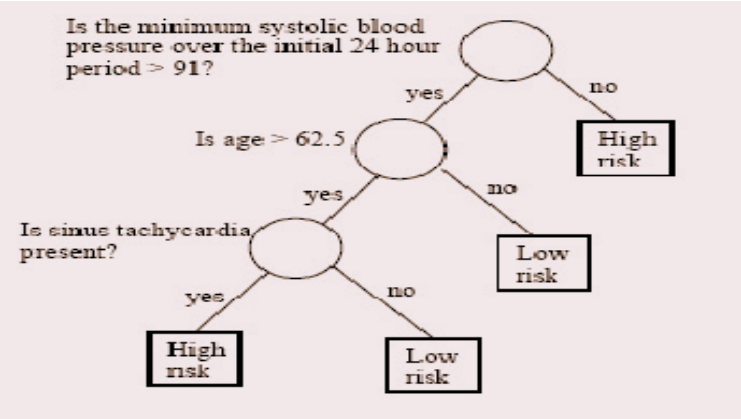

Figure 5. Classification by Decision Tree Induction

Khan *et al.*, used decision tree for predicting the survivability of breast cancer patient [45] and Chien *et al.*, proposed a universal hybrid decision tree classifier for classifying the activity of patient having chronic disease. They further improved the existing decision tree model to classify different activities of patients in more accurate manner [46]. In the similar domain, Moon et al. exemplify the patterns of smoking in adults using decision tree for better understanding the health condition, distress, demographic and alcohol [47]. Chang *et al.*, also used an integrated decision tree model for characterize the skin diseases in adults and children [48].

Support Vector Machine (SVM)

The concept of SVM is given by Vapnik *et al.*, which is based on statistical learning theory [49- 50]. SVMs were initially developed for binary classification but it could be efficiently extended for multiclass problems [51-52]. The support vector machine classifier creates a hyper plane or multiple hyper planes in high dimensional space that is useful for classification, regression and other efficient tasks. SVM have many attractive features due to this it is gaining popularity and have promising empirical performance. SVM constructs a hyper plane in original input space to separate the data points. Some time it is difficult to perform separation of data points in original input space, so to make separation easier the original finite dimensional space mapped into new higher dimensional space. Kernel functions are used for non-linear mapping of training samples to high dimensional space. Various kernel function such as polynomial, Gaussian, sigmoid *etc.*, are used for this purpose. SVM works on the principal that data points are classified using a hyper plane which maximizes the separation between data points and the hyper plane is constructed with the help of support vectors. Figure 6 shows the working of SVM classification algorithm.

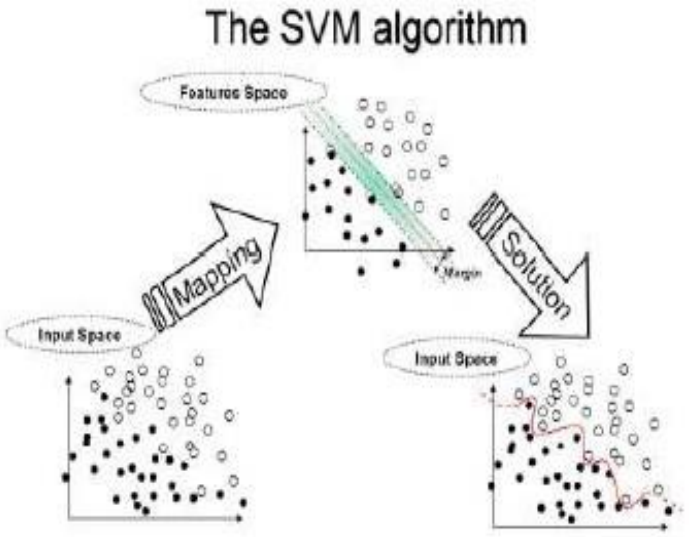

Figure 6. Support Vector Machine Classification

Figure 7 shows the classification of diabetic patients using SVM. In order to classify the patient into high risk and low risk of having diabetes, SVM constructs a hyper-plane that maximizes the distance between two classes.

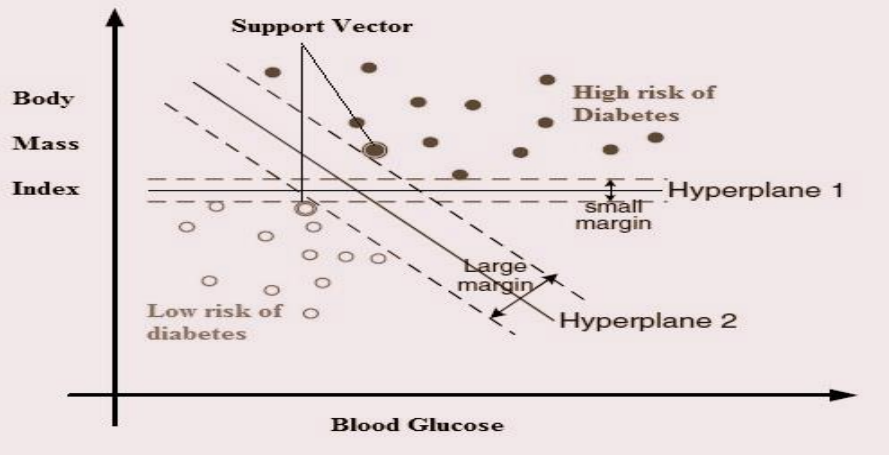

Figure 7. Classification of Diabetic Patients using Support Vector Machine

Soliman *et al.,* used SVM classification approach for classification of various diseases and SVM together with k-means clustering was applied on microarray data for identifying the diseases [53]. SVM is one of the most popular approaches that are used by researcher in healthcare field for classification. Fei proposed Particle Swarm Optimization SVM (PSO-SVM) approach for analyzing arrhythmia cordis [54] and Huang *et al.*, build a predictive model for breast cancer diagnosis using hybrid SVM based strategy [55]. E.Avci proposed a system using genetic SVM classifier for analyzing the heart valve disease. This system extracts the important feature and classifies the signal obatined from the ultrasound of heart valve [56]. A PSO based SVM model is constructed by Abdi et al. for identifying erythemato-squamous diseases which consists two stages. In the first stage optimal feature are extracted using association rule and in second phase the PSO is used to discover best kernel parameters for SVM in order to improve the accuracy of classifier model [57].

Neural Network (NN)

It is an algorithm for classification that uses gradient descent method and based on biological nervous system having multiple interrelated processing elements known as neurons, functioning in unity to solve specific problem. Rules are extracted from the trained Neural Network (NN) help to improve interoperability of the learned network [8]. To solve a particular problem NN used neurons which are organized processing elements. Neural Network is used for classification and pattern recognition [58]. An NN is adaptive in nature because it changes its structure and adjusts its weight in order to minimize the error. Adjustment of weight is based on the information that flows internally and externally through network during learning phase. In NN multiclass, problem may be addressed by using multilayer feed forward technique, in which Neurons have been employed in the output layer rather using one neuron. Er *et al.*, construct 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 [59]. Figure 8 shows diagnosis of various chest diseases such as Lung Cancer, Asthma, Pneumonia *etc.*, using Multilayer Neural Network [59].

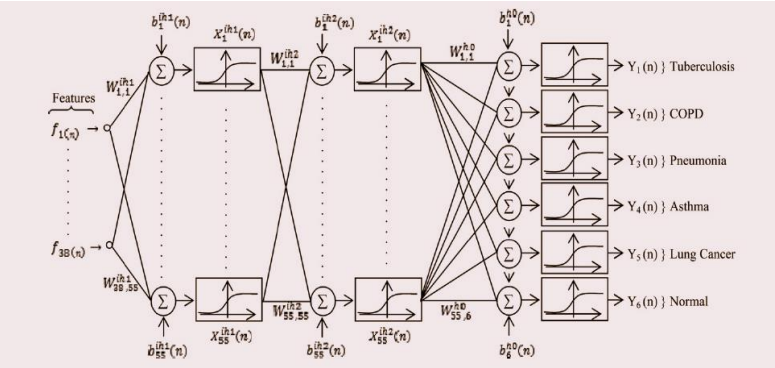

Figure 8. Classification of Chest Diseases using Multilayer Neural Network [59]

An ensemble neural network methodology is proposed by Das *et al.*, for diagnosis of heart disease in order to develop effective decision support system [60]. 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 [61].

Bayesian Methods

The classification based on bayes theory is known as Bayesian classification. It is a simple classifier which is achieved by using classification algorithm [8]. Bayes theorem provides basis for Naive Bayesian Classification and Bayesian Belief Networks (BBN). The main problem with Naïve Bayes Classifier is that it assumes that all attributes are independent with each other while in medical domain attributes such as patient symptoms and their health state are correlated with each other. In spite of assumption of attribute independence, Naïve Bayesian classifier has shown great performance in terms of accuracy so if attributes are independent with each other then we can use it in medical field. Bayes theorem concentrates on prior, posterior and discrete probability distributions of data items. Figure 9 shows the Bayesian Belief Network for patients suffering from lung cancer. Bayesian Belief Network is widely used by many researchers in healthcare field. Liu et al. develop a decision support system using BBN for analyzing risks that are associated with health [62]. Curiac *et al.*, analyze the psychiatric patient data using BBN in making significant decision regarding patient health suffering from psychiatric disease and performed experiment on real data obtain from Lugoj Municipal Hospital [63].

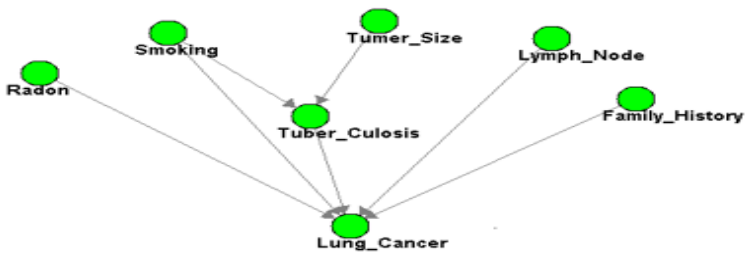

Figure 9. Bayesian Belief Network for Lung Cancer Patients

Advantage and disadvantage of different classification techniques are indicated in Table 1.

Table 1. Advantage and Disadvantage of Various Classification Techniques

Methods	Advantage	Disadvantage
K-NN	1. It is easy to implement. 2. Training is done in faster manner.	1. It requires large storage space. 2. Sensitive to noise. 3. Testing is slow.
Decision Tree	1. There are no requirements of domain knowledge in the construction of decision tree. 2. It minimizes the ambiguity of complicated decisions and assigns exact values to outcomes of various actions. 3. It can easily process the data with high dimension. 4. It is easy to interpret. 5. Decision tree also handles both numerical and categorical data.	1. It is restricted to one output attribute. 2. It generates categorical output. 3. It is an unstable classifier i.e. performance of classifier is depend upon the type of dataset. 4. If the type of dataset is numeric than it generates a complex decision tree
Support Vector Machine	1. Better Accuracy as compare to other classifier. 2. Easily handle complex nonlinear data points. 3. Over fitting problem is not as much as other methods.	1. Computationally expensive. 2. The main problem is the selection of right kernel function. For every dataset different kernel function shows different results. 3. As compare to other methods training process take more time. 4. SVM was designed to solve the problem of binary class. It solves the problem of multi class by breaking it into pair of two classes such as one-against-one and one-against-all.
Neural Network	1. Easily identify complex relationships between dependent and independent variables. 2. Able to handle noisy data.	1. Local minima. 2. Over-fitting. 3. The processing of ANN network is difficult to interpret and require high processing time if there are large neural networks.
Bayesian Belief Network	1. It makes computations process easier. 2. Have better speed and accuracy for huge datasets.	1. It does not give accurate results in some cases where there exists dependency among variables.

2.2. Regression

Regression is used to find out functions that explain the correlation among different variables. A mathematical model is constructed using training dataset. In statistical modeling two kinds of variables are used where one is called dependent variable and another one is called independent variable and usually represented using 'Y' and 'X'. There is always one dependent variable while independent variable may be one or more than one. Regression is a statistical method which investigates relationships between variables. By using Regression dependences of one variable upon others may be established [64]. Based on number of independent variables regression is of two types, one is Linear and another one is Non-linear. Linear regression identifies relation of a dependent variable and one or more independent variables. It is based on a model which utilizes linear function for its construction. Linear regression finds out a line and calculates vertical distances of points from the line and minimize sum of square of vertical distance. In this approach dependent and independent variables are already known and purpose is to spot a line that correlates between these variables [64]. But, linear regression is limited to numerical data only and cannot be use for categorical data. Logistic regression, a type of non-linear regression can accept categorical data and predicts the probability of occurrence using logit function. Logistic regression is of two types, one is Binomial and other is multinomial. Binomial regression predicts the result for a dependent variable when there occurs only two possible outcomes such as either a person is dead or alive while the multinomial handles the situation when dependent variable has three or more outcome. For example either a patient is at 'low risk', 'medium risk' and 'high risk'. Logistic regression does not consider linear relationship between variables [65]. Regression is widely used in medical field for predicting the diseases or survivability of a patient. Figure 10 represents an application of logistic regression for the estimation of relative risk for various medical conditions such as Diabetes, Angina, stroke etc [66]. In another research work, Weighted Support Vector Regression (WSVR) is used for monitoring the daily activities of patient [67]. This paper presents a model based on WSVR to overcome the over-fitting problem occurred due to noise and outliers.

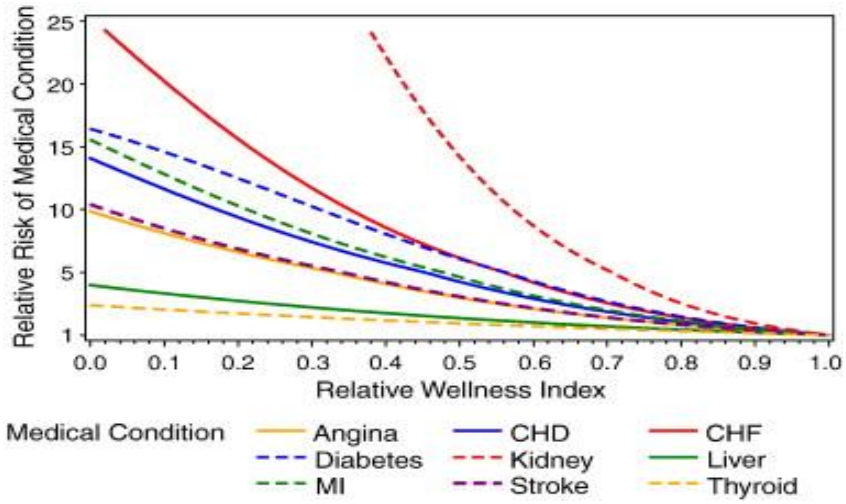

Figure 10. Example of Logistic Regression [66]

In Figure 11, we represent the functioning of classification and regression techniques. Both, classification and regression are used for predicting the class or outcome of a function. The only difference between them is the nature of attributes. If the attributes are categorical then one can use classification algorithms such as Naïve Bayes, SVM *etc*., and if the attributes are continuous then regression model using SVM or linear regression achieves great performance.

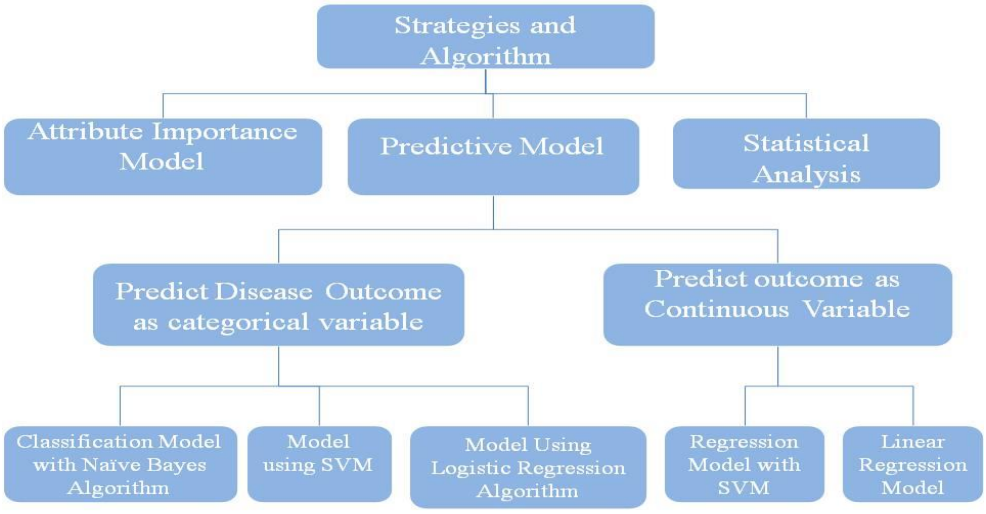

Figure 11. Functioning of Classification and Regression Techniques

2.3. Clustering

Clustering is an unsupervised learning method that is different from classification. Clustering is unlike to classification since it has no predefined classes. In clustering large database are separated into the form of small different subgroups or clusters. Clustering partitioned the data points based on the similarity measure [8]. 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. Various clustering techniques are established and used over the last few decades. As pointed out earlier clustering need less or no information for analyzing the data. So it is mainly used for analyzing microarray data because very little details are available for genes. Tapia et al. analyzed the gene expression data with the help of a new hierarchical clustering approach using genetic algorithm [68].

Partitioned Clustering

In this clustering method the datasets having ‘n’ data points partitioned into ‘k’ groups or clusters. Each cluster has at least one data point and each data point must belong to only one cluster. In this clustering approach there is a need to define the number of cluster before partitioning the datasets into groups. Based on the choice of cluster centroid and similarity measure, partition clustering method is divided into two categories-K-means and K-Mediods. K-Means clustering approach is one of the most widely used approach that partition the given ‘n’ data points into ‘k’ cluster based on similarity measure in such a way that data points belong to the same cluster have high similarity as compare to the data points of other cluster [5].

It first selects the k-centroid randomly and then assign the data points to these 'k' centroid based on some similarity measure. For every iteration, a data point is handed over to the cluster based on similarity of cluster mean (the distance between the data points) [69, 70]. The latest mean is calculated and this step is recurred to accommodate every newly arrived data points. The approach is intended to form compact clusters of similar data points with fare dissimilarity with other clusters. Cluster similarity could be characterized in the form of cluster mean which is also considered as centroid of the cluster. It is a self organized approach and easily initiates clustering process, so many complex clustering approach uses K means as beginning process. Unlike K-Means, K-mediods used medoids instead of mean for grouping the cluster. Medoid is one of the most centrally located data point in the database. Initially arbitrarily select the medoids for each cluster and after that data point is grouped with that medoid to which it is most similar. Figure 12 represents the grouping of person on the basis of high blood pressure and cholesterol level into high risk and low risk of having heart disease using K-means clustering. Lenert *et al.*, utilize the application of k-means clustering in the health services of public domain [71] and Belciug et al. detect the recurrence of breast cancer with the help of clustering technique [72]. Another research work explores the application of Data Mining techniques in healthcare. Balasubramanian *et al.*, analyze the impact of ground water on human health using clustering technique. They discovered the causes of risk related with the fluoride content in water with the help of k-means clustering. Using this, author identified the valuable information in order to make decision regarding human health [73]. Escudero *et al.*, used k-means clustering to classify the Alzheimer's disease (AD) data feature into pathologic and non-pathologic groups. This research work used the concept of Bioprofile and K-means clustering for early detection of AD [74].

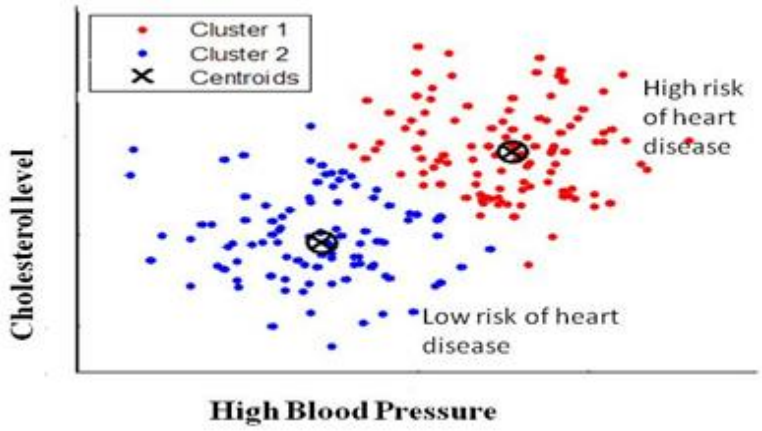

Figure 12. K-means Clustering for Heart Disease Patients

Hierarchical Clustering

Unlike partitioned clustering there is no need to define the number of cluster in advance. Hierarchical Clustering algorithm decomposes the data points in hierarchical way. It decompose the data points either using bottom up approach or top down approach. Hierarchical clustering is classified into two categories –Agglomerative and Divisive that depends on the decomposition process. Agglomerative approach initially

consider each data point as a separate group and further it merges the data points that have some similarity with each other and repeat this process until all the data points are merged into one group or class or until it gets some termination condition [5]. On the other hand divisive approach assume all the data points as one group initially and further it splits the data points into small group until it satisfy some termination condition or each data point belongs to single cluster. Chipman *et al.*, proposed the hybrid hierarchical clustering approach for analyzing microarray data [75]. The research work combines both top-down and bottom-up hierarchical clustering concepts in order to effectively utilize the strength of this clustering approach. Chen *et al.*, proposed an integrated approach for analyzing micro- array data. This study combined both k-means and hierarchical clustering in order to improve the performance of analyzing large micro array data [76]. Belciug use the hierarchical clustering approach for grouping the patients according to their length of stay in the hospital that enhance the capability of hospital resource management [77]. Figure 13 shows the grouping of the patients into two cluster using 192-gene expression profile. Liu *et al.*, predict the severity of disease in patients using gene expression profile having Rheumatoid Arthritis [78].

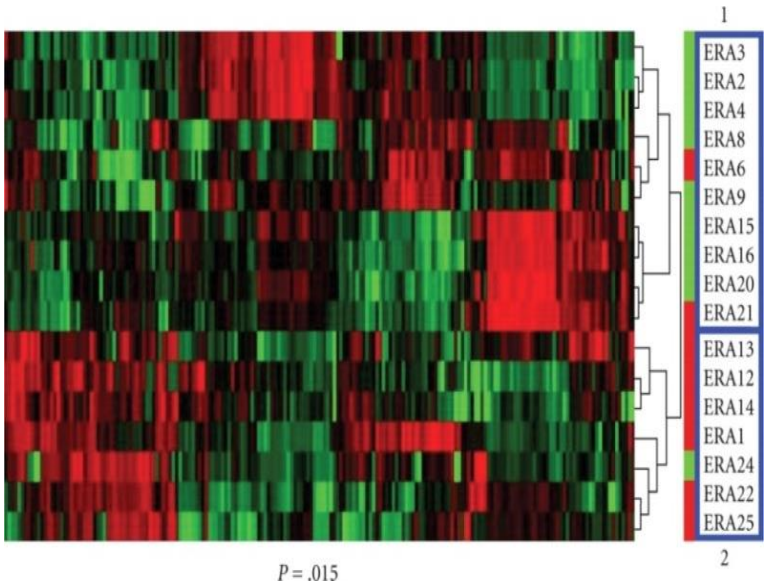

Figure 13. Hierarchical Clustering for Grouping the Patients into Two Cluster using 192-gene Expression Profile [78]

Density Based Clustering

The problem with partition and hierarchical clustering method is that they can handle only spherical shaped cluster and are not suitable for discovering cluster of arbitrary shapes. Density clustering methods remove this drawback and efficiently handle outliers and arbitrary shaped cluster. DBSCAN and OPTICS are two approach of Density based clustering which discover cluster on the basis of density connectivity analysis. DENCLUE is another approach of density based clustering methods that form the grouping of data points on the basis of distribution value analysis of density function [5]. The research work [79] extracts the useful and interesting patterns from biomedical images using density based clustering. This research discovers the area of homogeneous colour in biomedical images. This method separates the unhealthy skin or

wound from healthy skin and discovers the sub regions of varied colour or spotted part inside the unhealthy skin which is again useful for classification and association task [79]. Figure 14 represents the clustering of wounded skin images using DBSCAN algorithm.

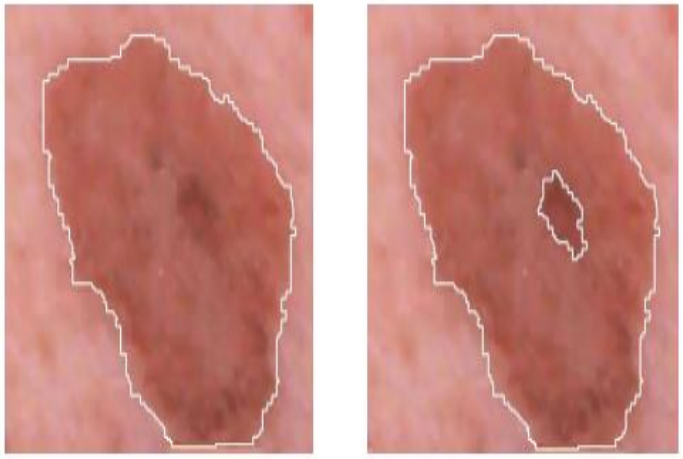

Figure 14. Clustering of Skin Wound Image using DBSCAN [79]

Advantage and disadvantage of different clustering techniques are indicated in Table 2.

Table 2. Advantage and Disadvantage of various Clustering Techniques

Methods	Advantage	Disadvantage
K-means Clustering	1. Simple clustering approach. 2. Efficient. 3. Less complex method.	1. Requires number of cluster in advance. 2. Problem with handling categorical attributes. 3. Not discover the cluster with non-convex shape. 4. Result varies in the presence of outlier.
Hierarchical Clustering	1. Easy to implement. 2. Having good visualization capability. 3. There is no need to specify the number of clusters in advance.	1. Have cubic time complexity in many cases so it is slower. 2. Decision regarding selection of merge or split point. Once a decision is made it cannot be undone. 3. Not work well in the presence of noise and outlier. 4. Not scalable.
Density Based Clustering	1. No need to specify number of cluster in advance. 2. Easily handle cluster with arbitrary shape. 3. Worked well in the presence of noise.	1. Not handle the data points with varying densities. 2. Results depend on the distance measure.

2.4. Association

Association is one of the most vital approach of data mining that is used to find out the frequent patterns, interesting relationships among a set of data items in the data repository. It is also known as market basket analysis due to its capability of discovering the association among purchased item or unknown patterns of sales of customers in a transaction database. For example if a customer is buying a computer then the chance of buying antivirus software is high. This information helps the storekeeper to further enhance their sales [80-81]. Association also has great impact in the healthcare field to detect the relationships among diseases, health state and symptoms. Ji *et al.*, used association in order to discover infrequent casual relationships in Electronic health databases [82]. Healthcare organization widely used Association approach for discovering relationships between various diseases and drugs. It is also used for detecting fraud and abuse in health insurance. Association is also used with classification techniques to enhance the analysis capability of Data Mining. Soni *et al.*, used an integrated approach of association and classification for analyzing health care data. This integrated approach is useful for discovering rules in the database and then using these rules an efficient classifier is constructed. This study performed experiment on the data of heart patients and also generate rules using weighted associative classifier [83]. Bakar *et al.*, also construct a predictive model using various rule based classifier for dengue occurrence. In this research work authors combine rough set, naïve bayes, decision tree and associative classifier to build a predictive model for enhancing the early detection of dengue occurrence [84]. Doctor's prescriptions and treatment materials are produced large amount of data. Utah Bureau of Medicaid Fraud used this data to discover hidden and useful information in order to detect fraud. This approach is also helpful for identifying the improper prescriptions, irregular or fake patterns in medical claims made by physicians, patients, hospitals *etc*.

Apriori Algorithm

Apriori algorithm for association is proposed by R.Agarwal *et al.*, in 1994. It finds out the relationships among item sets using two inputs-support and confidence. These two inputs help to discriminate the frequent and infrequent item sets. The research work filtered out those item from transaction database that are not satisfy some given criteria such as frequent item set satisfy the minimum support and confidence constraint. This algorithm is based on the principle that if an item does not fulfils minimum support constraint or not frequent then its descendants are also not frequent so remove this item from the transaction database because this item does not contribute in the construction of association rules. Unlike classification and clustering, efficiency is the evaluation factor of association mining. Various methods are used to improve the efficiency of Apriori algorithm such as Hash table, transaction reduction, partitioning *etc.*, [81] [82]. Patil et al. used apriori algorithm for generating association rule. Using these rules they classify the patients suffering from type-2 diabetes. In this research, authors proposed an approach for discretizing the attributes having continuous value using equal width bining interval which is selected on the basis of medical expert's opinion [85]. Figure 15 indicates the association rules for patients having diabetes. Another research work analyzes the medical bill using apriori algorithm [86]. Abdullah *et al.*, proposed some modification in existing Apriori algorithm and then utilize its effectiveness in constructed useful information in medical bill. Ilayaraja *et al.*, also 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 [87].

Sr.no	Rules	Coverage	Confidence
1	Plasma–Glucose = high and Age = 40–59 ⇒ yes	60	84 %
2	lasma–Glucose = high and BMI = severely obese ⇒ yes	66	82%
3	Plasma–Glucose = high and BMI = obes ⇒ yes	56	78%
4	Pregnant = high and Plasma–Glucose =high ⇒ yes	75	77%

Figure 15. Association Rules for Diabetic Patients [85]

Nahar *et al.*, used Apriori, predictive apriori for generating the rules for heart disease patients. In this research work rules are produced for healthy and sick people. Based on these rules, this research discovered the factors which cause 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 [88]. Figure 16 indicates the rules generation for healthy and sick people with the help of Apriori algorithm.

Algorithms	Rules
Apriori	Healthy rules: Healthy rule: If {Sex = female ∩ exercise_induced_angina = fal ∩ number_of_vessels_colored=0 ∩ thal = nom} => class healthy (conf., 0.98). Healthy rule: If {Sex = female ∩ fasting_blood_sugar = fal ∩ exercise_induced_angina = fal ∩ number_of_vessels_colored = 0} => class healthy (conf., 0.98). Healthy rule: If {Sex=female ∩ exercise_induced_angina = fal ∩ number_of_vessels_colored = 0} => class healthy (conf., 0.98). Healthy rule: If {Sex = female ∩ fasting_blood_sugar = fal ∩ exercise_induced_angina = fal ∩ thal = norm} => class healthy (conf., 0.95). Healthy rule: If {Resting_blood_pres less or = '(115.2,136.4]' ∩ exercise_induced_angina = fal ∩ number_of_vessels_colored = 0 ∩ thal = norm} => class healthy (conf., 0.94). Sick Rules: Sick rule: If {Chest_pain_type = asympt ∩ slope = flat ∩ thal = rev} => class sick (conf., 0.96). Sick rule: If {Chest_pain_type=asympt ∩ exercise_induced_angina=TRUE ∩ thal=rev} => class sick (conf., 0.94).

Figure 16. Rules Generation using Apriori for Healthy and Sick People [88]

Frequent Pattern Tree Algorithm

FP-tree algorithm identifying the frequent item sets without generating candidate item-set. This algorithm has two steps-in the first step FP tree data structure is

constructed and in the second step frequent item set is fetched from this data structure. Association analysis is helpful in finding out the hidden or previously unseen relationship among attributes. Due to this nature it is widely used in medical field to discover the correlation among different diseases and drugs. Noma *et al.*, used FP-tree algorithm for identifying interesting patterns in medical audiology data. This research work proposed a knowledge discovery model containing five steps which is further implemented using FP-tree technique in order to discover valuable information from audiometric datasets [89].

The following table describes the summary of data mining approaches that are used in health domain:

Table 3. Summary of Data Mining Approaches Used in Healthcare

Author	Publication Year	Approaches	Accuracy
Yan et al.[100]	2003	Multilayer Perceptron	63.6%
Andreeva, P [101]	2006	Naïve Bayes	78.56%
		Decision Tree	75.73%
		Neural Network	82.77%
		Kernel Density	84.44%
Hara et al.[102]	2008	Automatically Defined Groups	67.8%
		Immune Multi-agent Neural Network	82.3%
Sitar-Taut et al. [103]	2009	Naïve Bayes	62.03%
		Decision Tree	60.40%
Chang et al.[48]	2009	Decision Tree	90.89%
		Artificial Neural Network	92.62%
		Decision Tree combined with ANN	86.89%
		Decision Tree with sensitivity Analysis	80.33%
		ANN with sensitivity Analysis	83.61%
Rajkumar et al.[104]	2010	Naïve Bayes	52.33%
		Decision tree	52%
		KNN	45.67%
Srinivas et al. [105]	2010	Naïve Bayes	84.14%
		One Dependency Augmented Naïve Bayes classifier	80.46%
Kangwanariyakul, et al.[106]	2010	Back-Propagation Neural Network	78.43%
		Bayesian Neural Network	78.43%
		Probabilistic Neural Network	70.59%
		Linear Support Vector Machine	74.51%
		Polynomial Support Vector Machine	70.59%
		RBF- kernel Support Vector Machine	60.78%
Anbarasi, et al.[107]	2010	Genetic with Decision tree	99.2%
		Genetic with Naïve Bayes	96.5%
		Genetic with Classification via Clustering	88.3%
Fan et al. [108]	2010	CHAID	69.75%

		C & RT	69.73%
		QUEST	67.25%
		C 5.0	71.17%
Sonali et al.[109]	2010	One-against-many with POLY kernel	85.14%
		One-against-many with Gaussian kernel	95.98%
		M-SVM with polynomial kernel	83.25%
		M-SVM with Gaussian kernel	97.19%
Osareh et al. [110]	2010	PNN	92.86%
		KNN	94.06%
		SVM-RBF	95.45%
		SVM-POLY	95.19%
Fei [54]		RBF-NN	89.13%
		PSO-SVM	95.65%
		BP-NN	83.7%
		Selective Base Classifier on Bagging	96.98%
Abdi et al.[57]	2013	SVM	94.56%
		AR_MLP	97.28%
		AR_PSO-SVM	98.91%

3. Application of Data Mining in Health

Data mining provides several benefits to healthcare industry. Data Mining helps the healthcare researchers to make valuable decision. Following are the several applications of Data Mining in healthcare:

Effective management of Hospital resource: Data mining provides support for constructing a model for managing the hospital resources which is an important task in healthcare. Using data mining, it is possible to detect the chronic disease and based on the complication of the patient disease prioritize the patients so that they will get effective treatment in timely and accurate manner. Fitness report and demographic details of patients is also useful for utilizing the available hospital resources effectively. An automated tool using data mining is proposed by J.Alapont *et al.*, for managing hospital resources such as physical and human resources [90]. Group Health Cooperative provides various healthcare services at lower cost using data mining techniques [1]. It is a non-profit organization of healthcare that offers patients to online access their medical information, online fill the prescription form and allow safe exchanging of e-mail with the healthcare provider. Seton Medical centre also used data mining to enhance the healthcare quality, provide various details regarding patient's health and reduce admitted duration of the patients in the hospitals [91]. With the help of data mining Blue Cross provide a system for managing the diseases efficiently and improve the results and lower the cost of expenditure. Sierra Health Centre provides guidelines for treatment, managing the cost of treatment and detects the areas for improving the health quality using data mining [92].

Hospital Ranking: Different data mining approaches are used to analyze the various hospital details in order to determine their ranks [93]. Ranking of the hospitals are done on the basis of their capability to handle the high risk patients. The hospital with higher rank handles the high risk patient on its top priority while the hospital with lower rank does not consider the risk factor.

Better Customer Relation: Data Mining helps the healthcare institute to understand the needs, preferences, behavior, patterns and quality of their customer in order to make better relation with them. Using Data Mining, Customer Potential Management Corp. develops an index represent the utilization of Consumer healthcare. This index helps to detect the influence of customer towards particular healthcare service.

Hospital Infection Control: A system for inspection is constructed using data mining techniques to discover unknown or irregular patterns in the infection control data [93]. Association rules are used to produce unexpected and interesting information from the public surveillance and hospital control data. To control the infection in the hospitals, this information is reviewed further by an Expert.

Smarter Treatment Techniques: Using Data Mining, physicians and patients can easily compare among different treatments technique. They can analyze the effectiveness of available treatments and find out which technique is better and cost effective. Data Mining also helps them to identify the side effects of particular treatment, to make appropriate decision to reduce the hazard and to develop smart methodologies for treatment.

Improved Patient care: Large amount of data is collected with the advancement in electronic health record. Patient data which is available in digitized form improve the healthcare system quality. In order to analyze this massive data, a predictive model is constructed using data mining that discover interesting information from this huge data and make decision regarding the improvement of healthcare quality. Data mining helps the healthcare providers to identify the present and future requirements of patients and their preferences to enhance their satisfaction levels. Milley has also recommended that data mining are useful to determine the requirement of particular patients for enhancing the services provided by healthcare organization [94]. Hallick has suggested that Data mining techniques are helpful to provide the information to patient regarding various diseases and their prevention [95]. Kolar has identified that healthcare organization used data mining techniques for patient grouping [96].

Decrease Insurance Fraud: Healthcare insurer develops a model to detect the fraud and abuse in the medical claims using data mining techniques. This model is helpful for identifying the improper prescriptions, irregular or fake patterns in medical claims made by physicians, patients, hospitals etc. US taxpayers also reported to lost hundred dollars in 1997 due to fraudulent in the hospitals bill. ReliaStar financial corp. has improved the annual savings by 20% by detected the fraud and abuse. Doctor's prescriptions and treatment materials are produced large amount of data. Utah Bureau of Medicaid Fraud used this data to discover hidden and useful information in order to detect fraud [94]. Australian Health Insurance Commission has also mined the huge data and reported millions of dollars of annual saving [97]. Texas Medicaid Fraud and Abuse Detection System have also used data mining techniques to discover the fraud and abuse and saved million dollars in 1998 [98].

Recognize High-Risk Patients: American Healthways system construct a predictive model using data mining to recognize the patients having high risk. The main concern of this system is to handle the diabetic patients, improve their health quality and also offers cost savings services to the patient. Using Predictive model, healthcare provider recognize the patient which require more concern as compare to other patients [99].

Health Policy Planning: Data mining play an important role for making effective policy of healthcare in order to improve the health quality as well as reducing the cost for health

services. COREPLUS and SAFS models were developed using data mining techniques to analyze the results of medical care services provided by hospitals and treatment cost.

4. Data Mining Challenges in Healthcare

One of the most significant challenges of the data mining in healthcare is to obtain the quality and relevant medical data. It is difficult to acquire the precise and complete healthcare data. Health data is complex and heterogeneous in nature because it is collected from various sources such as from the medical reports of laboratory, from the discussion with patient or from the review of physician. For healthcare provider, it is essential to maintain the quality of data because this data is useful to provide cost effective healthcare treatments to the patients. Health Care Financing Administration maintains the minimum data set (MDS) which is recorded by all hospitals. In MDS there are 300 questions which are answered by the patients at check-in time. But this process is complex and patients face problem to respond the entire questions. Due to this MDS face some difficulties such as missing information and incorrect entries. Without quality data there is no useful results. For successful data mining, complication in medical data is one the significant hurdle for analyzing medical data. So, it is essential to maintain the quality and accuracy data for data mining to making effective decision. Another difficulty with healthcare data is data sharing. Healthcare organizations are unwilling to share their data due to privacy concern. Most of the patients do not want to disclose their health data. So, the Health Maintenance Organization and Health insurance Organization are not distributing their data for preserving the privacy of patient. This poses hurdle in the fraud detection studies in health insurance. The startup cost of data warehouse is very high. Before applying data mining techniques in healthcare data it is essential to collect and record the data from different sources into a central data warehouse which is a costly and time consuming process. Faulty data warehouse design does not contribute to effective data mining.

5. Conclusion and Future Issues

The purpose of this section is to provide an insight towards requirements of health domain and about suitable choice of available technique. Following are the guideline for using different data mining techniques:

- Before applying classification technique there is a need to recognize the redundant and inappropriate attributes because these attributes act as a noise and outlier which in turn slow down the processing task. These attributes also had an adverse affect on the performance of classifier. Statistical methods are used for recognizing these attributes. On the other hand the most relevant and useful attributes can be recognized by feature selection methods which in turn enhance the performance and accuracy of classification model.
- We also analyzed that there is no single classifier which produce best result for every dataset. In order to check the performance of classifier, a dataset is divided into two parts- training and testing. So, a classifier is selected only when it produce better performance among all classifiers. The performance of a classifier is evaluated using testing data set. But there are also problem with testing data set. Some time it is complex and some time it becomes easy to classify the testing data set. The performance of classifier depends on testing data set. To avoid these problems we can use cross validation method so that every record of data set is used for both training and testing.

- We also analyze that clustering technique is used when there is no or less information are available regarding data set. But what type of clustering algorithm is used is still a problem. Hierarchical clustering is used when there is less information is available about data because for this algorithm there is no need to specify number of clusters in advance. Dendograms which is the output of hierarchical clustering should be analyzed to find out the suitable number of cluster. But the problem with this algorithm is that it is not scalable i.e. its performance varies as number of dataset increase. To avoid these problems random sampling should be used so that hierarchical clustering easily handles the reduced volume of data. To avoid the problem of sampling biasness there is a need to repeat the sampling process several times. Partitioned algorithm can be used after determining number of cluster.
- The main focus of classification rules is to discover the class of attributes but it does not take into account the relationships of attributes. While Association is useful for identifying the relationship or association among various attributes and generates association rules which in turn helpful for domain experts to remove insignificant association rules and consider only those rules which are useful for making vital decision.

We can also conclude that there is no single data mining techniques which give consistent results for all types of healthcare data. The performance of data mining techniques depends on the type of dataset that we have taken for doing experiment. So, we can use hybrid or integrated Data Mining technique such as fusion of different classifiers, fusion of clustering with classification or association with clustering or classification etc. for achieving better performance. Apart from this we also observe that GA with clustering or classification, PSO-SVM, Fuzzy KNN, AR-PSO_SVM, SBCB have accomplish good results as compare to single traditional approach. So hybridization is a good option for getting better results. This paper explore the application of data mining in healthcare organization, different techniques and the challenges of Data Mining in healthcare and their future issues. Data Mining provides benefit to all the people such as doctor, healthcare insurers, patients and organizations who are engaged in healthcare industry. Using Data Mining knowledge Doctor can easily recognize the effective cure, patients obtain cost effective treatments, healthcare industry manages their customer and healthcare insurers discover any cases of fraud in medical claim. Due to analytical and descriptive ability, Data Mining is widely used in medical field. Healthcare providers utilize the data mining tools to make effective decision regarding how to enhance the patient health, how to provide health care services at low cost and how to predict fraud in health insurance *etc*. Healthcare researchers also face several challenges while using Data Mining in medical field such as several Data Mining techniques required parameters from user. These techniques are sensitive to user's parameters. Its results vary according to the parameters which are given by users. Sometime users do not have sufficient information about selection and usage of parameters.

For effective utilization of data mining in health organizations there is a need of enhance and secure health data sharing among different parties. Some propriety limitations such as contractual relationships among researcher and health care organization are mandatory to overcome the security issues. There is also a need of standardized approach for constructing the data warehouse. In recent years due to enhancement of internet facility a huge datasets (text and non-text form) are also available on website. So, there is also an essential need of effective data mining techniques for analyzing this data to uncover hidden information.

References

[1] H. C. Koh and G. Tan, "Data Mining Application in Healthcare", Journal of Healthcare Information Management, vol. 19, no. 2, **(2005)**.
[2] R. Kandwal, P. K. Garg and R. D. Garg, "Health GIS and HIV/AIDS studies: Perspective and retrospective", Journal of Biomedical Informatics, vol. 42, **(2009)**, pp. 748-755.
[3] D. Hand, H. Mannila and P. Smyth, "Principles of data mining", MIT, **(2001)**.
[4] 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.
[5] J. Han and M. Kamber, "Data mining: concepts and techniques", 2nd ed. The Morgan Kaufmann Series, **(2006)**.
[6] 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.
[7] 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)**.
[8] M. Silver, T. Sakara, H. C. Su, C. Herman, S. B. Dolins and M. J. O'shea, "Case study: how to apply data mining techniques in a healthcare data warehouse", Healthc. Inf. Manage, vol. 15, no. 2, **(2001)**, pp. 155-164.
[9] P. R. Harper, "A review and comparison of classification algorithms for medical decision making", Health Policy, vol. 71, **(2005)**, pp. 315-331.
[10] 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.
[11] 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.
[12] R. D. Canlas Jr., "Data Mining in Healthcare:Current Applications and Issues", **(2009)**.
[13] F. Hosseinkhah, H. Ashktorab, R. Veen, M. M. Owrang O., "Challenges in Data Mining on Medical Databases", IGI Global, **(2009)**, pp. 502-511.
[14] 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.
[15] J. Soni, U. Ansari, D. Sharma and S. Soni, "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", **(2011)**.
[16] C. S. Dangare and S. S. Apte, "Improved Study of Heart Disease Prediction System Using Data Mining Classification Techniques", **(2012)**.
[17] 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.
[18] A. 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.
[19] D. Delen, "Analysis of cancer data: a data mining approach", **(2009)**.
[20] A. O. Osofisan, O. O. Adeyemo, B. A. Sawyerr and O. Eweje, "Prediction of Kidney Failure Using Artificial Neural Networks", **(2011)**.
[21] S. Floyd, "Data Mining Techniques for Prognosis in Pancreatic Cancer", **(2007)**.
[22] 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.
[23] S. Gupta, D. Kumar and A. Sharma, "Data Mining Classification Techniques Applied For Breast Cancer Diagnosis And Prognosis", **(2011)**.
[24] 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.
[25] 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.
[26] R. Parvathi and S. Palaniammalì, "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.
[27] S. Chao and F. Wong, "An Incremental Decision Tree Learning Methodology Regarding Attributes in Medical Data Mining", **(2009)**.
[28] A. Habrard, M. Bernard and F. Jacquenet, "Multi-Relational Data Mining in Medical Databases", Springer-Verlag, **(2003)**.
[29] 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.

[30] A.Shukla, R. Tiwari, P. Kaur, Knowledge Based Approach for Diagnosis of Breast Cancer, IEEE International Advance Computing Conference,IACC 2009.
[31] 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.
[32] D. S. Kumar, G. Sathyadevi and S. Sivanesh, "Decision Support System for Medical Diagnosis Using Data Mining", **(2011)**.
[33] S. Palaniappan and R. Awang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques", **(2008)**.
[34] 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.
[35] C. 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)**.
[36] 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.
[37] 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.
[38] D. 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.
[39] C. 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.
[40] M. Shouman, T. Turner and R. Stocker, "Applying K-Nearest Neighbour in Diagnosing Heart Disease Patients", International Conference on Knowledge Discovery (ICKD-2012), **(2012)**.
[41] D. 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)**.
[42] 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.
[43] Goharian & Grossman, Data Mining Classification, Illinois Institute of Technology, http://ir.iit.edu/~nazli/cs422/CS422-Slides/DM-Classification.pdf, **(2003)**.
[44] 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)**.
[45] 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.
[46] C. 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.
[47] 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.
[48] C. 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.
[49] V. Vapnik, "Statistical Learning Theory", Wiley, **(1998)**.
[50] V. Vapnik, "The support vector method of function estimation", **(1998)**.
[51] N. Chistianini and J. Shawe-Taylor, "An Introduction to Support Vector Machines, and other kernel-based learning methods", Cambridge University Press, **(2000)**.
[52] N. Cristianini and J. Shawe-Taylor, "An Introduction to Support Vector Machines", Cambridge University Press, **(2000)**.
[53] 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)**.
[54] 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.
[55] C. 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.
[56] E. 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.

[57] 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.
[58] M. H. Dunham, "Data mining introductory and advanced topics", Upper Saddle River, NJ: Pearson Education, Inc., **(2003)**.
[59] O. Er, N. Yumusakc and F. Temurtas, "Chest diseases diagnosis using artificial neural networks", Expert Systems with Applications, vol. 37, **(2010)**, pp. 7648-7655.
[60] 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.
[61] 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)**.
[62] 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)**.
[63] D. 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.
[64] J. Fox, "Applied Regression Analysis, Linear Models, and Related Methods", **(1997)**.
[65] 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.
[66] C. 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.
[67] 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.
[68] 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.
[69] A. K. Jain, M. N. Murty and P. J. Flynn, "Data Clustering: a review", ACM Compute, Surveys, vol. 31, **(1996)**.
[70] G. 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)**.
[71] 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.
[72] 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)**.
[73] T. Balasubramanian and R. Umarani, "An Analysis on the Impact of Fluoride in Human Health (Dental) using Clustering Data mining Technique", Proceedings of the International Conference on Pattern Recognition, Informatics and Medical Engineering, **(2012)** March 21-23.
[74] 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.
[75] H. Chipman and R. Tibshirani, "Hybrid hierarchical clustering with applications to microarray data", Biostatistics, vol. 7, no. 2, **(2009)**, pp. 286-301.
[76] 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)**.
[77] 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.
[78] Z. Liu, T. Sokka, K. Maas, N. J. Olsen and T. M. Aune, "Prediction of Disease Severity in Patients with Early Rheumatoid Arthritis by Gene Expression Profiling", Human Genomics and Proteomics, **(2009)**.
[79] 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)**.
[80] R. Agrawal, T. Imielinski and A. N. Swami, "Mining Association Rules between Sets of Items in Large Databases. SIGMOD", vol. 22, no. 2, **(1993)** June, pp. 207-16.
[81] 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.
[82] 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)**.

[83] 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.
[84] A. 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.
[85] B. 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)**.
[86] 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.
[87] 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.
[88] 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)**.
[89] N. G. Noma and M. K. A. Ghani, "Discovering Pattern in Medical Audiology Data with FP-Growth Algorithm", IEEE EMBS International Conference on Biomedical Engineering and Sciences, Langkawi, **(2012)** December 17-19.
[90] 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)**.
[91] D. R. Dakins, "Center takes data tracking to heart", Health Data Management, vol. 9, no. 1, **(2001)**, pp. 32-36.
[92] B. K. Schuerenberg, "An information excavation", Health Data Management, vol. 11, no. 6, **(2003)**, pp. 80-82.
[93] O. Mary K., Mat, "Application of Data Mining Techniques to Healthcare Data", Infection Control and Hospital Epidemiology, **(2004)** August.
[94] A. Milley, "Healthcare and data mining", Health Management Technology, vol. 21, no. 8, **(2000)**, pp. 44-47.
[95] J. N. Hallick, "Analytics and the data warehouse", Health Management Technology, vol. 22, no. 6, **(2001)**, pp. 24-25.
[96] H. R. Kolar, "Caring for healthcare", Health Management Technology, vol. 22, no. 4, **(2001)**, pp. 46-47.
[97] T. Christy, "Analytical tools help health firms fight fraud", Insurance & Technology, vol. 22, no. 3, **(1997)**, pp. 22-26.
[98] Anonymous. Texas Medicaid Fraud and Abuse Detection System recovers $2.2 million, wins national award. Health Management Technology, vol. 20, no. 10, **(1999)**.
[99] M. Ridinger, "American Healthways uses SAS to improve patient care", DM Review, vol. 12, no.139, **(2002)**.
[100] H. Yan, "Development of a decision support system for heart disease diagnosis using multilayer perceptron", Proceedings of the 2003 International Symposium, vol. 5, **(2003)**, pp. V-709- V-712.
[101] P. Andreeva, "Data Modelling and Specific Rule Generation via Data Mining Techniques", International Conference on Computer Systems and Technologies - CompSysTech, **(2006)**.
[102] A. Hara and T. Ichimura, "Data Mining by Soft Computing Methods for the Coronary Heart Disease Database", Fourth International Workshop on Computational Intelligence & Application, IEEE SMC Hiroshima Chapter, Hiroshima University, Japan, **(2008)** December 10-11.
[103] V. A. Sitar-Taut, "Using machine learning algorithms in cardiovascular disease risk evaluation", Journal of Applied Computer Science & Mathematics, **(2009)**.
[104] A. Rajkumar and G. S. Reena, "Diagnosis of Heart Disease Using Datamining Algorithm", Global Journal of Computer Science and Technology, vol. 10, no. 10, **(2010)**.
[105] K. Srinivas, B. K. Rani and A. Govrdhan, "Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks", International Journal on Computer Science and Engineering (IJCSE), vol. 02, no. 02, **(2010)**, pp. 250-255.
[106] Y. Kangwanariyakul, C. Nantasenamat, T. Tantimongcolwat and T. Naenna, "Data Mining of Magneto cardiograms For Prediction of Ischemic Heart Disease", EXCLI Journal, **(2010)**.
[107] M. Anbarasi, E. Anupriya and N. Iyengar, "Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm", International Journal of Engineering Science and Technology, vol. 2, no. 10, **(2010)**, pp. 5370-5376.
[108] Q. Fan, C. J. Zhu and L. Yin, "Predicting Breast Cancer Recurrence Using Data Mining Techniques", International Conference on Bioinformatics and Biomedical Technology, **(2010)**.
[109] S. Agarwal and G. N. Pandey Divya, "SVM based context awareness using body area sensor network for pervasive healthcare monitoring", Proceedings of the First International Conference on Intelligent Interactive Technologies and Multimedia. ACM, **(2010)**.

[110] A. Osareh and B. Shadgar, "Machine Learning Techniques to Diagnose Breast Cancer", Health Informatics and Bioinformatics (HIBIT), IEEE, **(2010)**.

Authors

Divya Tomar, she is research scholar in Information Technology Division of Indian Institute of Information Technology (IIIT), Allahabad, India under the supervision of Dr. Sonali Agarwal. She held the Bachelor of Technology (B.Tech.) degree in Computer Science and Engineering from Institute of Engineering and Technology, Lucknow, (UP) India and Masters of Technology (M.Tech.) degree in Information Technology specialized in Human Computer Interaction from Indian Institute of Information Technology (IIIT), Allahabad, India. Her primary research interests are in the areas of Data Mining, Data Warehousing, Support Vector Machine especially with the application in the area of Medical Healthcare.

Dr. Sonali Agarwal, she is working as an Assistant Professor in the Information Technology Division of Indian Institute of Information Technology (IIIT), Allahabad, India. She received the Ph. D. Degree at IIIT Allahabad and joined as faculty at IIIT Allahabad, where she was teaching since October 2009. She held the Bachelor of Engineering (B.E.) degree in Electrical Engineering from Bhilai Institute of Technology, Bhilai, (C.G.) India and Masters of Engineering (M.E.) degree in Computer Science from Motilal Nehru National Institute of Technology (MNNIT), Allahabad, India. She worked as Lecturer and Assistant Professor at B.B.S. College of Engineering and Technology, Allahabad from 2002 to 2009. Her primary research interests are in the areas of Data Mining, Data Warehousing, E Governance and Software Engineering. Her current focus in the last few years is on the research issues in Data Mining application especially in E Governance and Healthcare.