Indian Journal of Science and Technology, *Vol 11(15), DOI: 10.17485/ijst/2018/v11i15/121111,* April 2018

ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Survey of Classification Based Prediction Techniques in Healthcare

Sujata Joshi[1] and Mydhili K. Nair[2]

[1]Department of Computer Science and Engineering, Nitte Meenakshi Institute of Technology, Bengaluru - 560064, Karnataka, India; sujata_msrp@yahoo.com
[2]Department of Information Science & Engineering, M. S. Ramaiah Institute of Technology, Bengaluru - 560054, Karnataka, India; mydhili.nair@gmail.com

Abstract

Data mining is used extensively and is applied successfully in various fields like market-basket analysis, e-business, fraud detection, quality control, cross-selling of products etc. More recently, data mining has been successfully applied to healthcare sector and healthcare applications. **Objectives**: The objective of this research is to study the classification based prediction techniques as applied to healthcare. It also aims at finding the different applications and tools used in classification based prediction in the healthcare sector. **Methods**: Prevalently the prediction techniques used are Decision Trees, Naive Bayes classifier, Bayesian networks, k-Nearest neighbour and artificial neural networks. A few researchers also have used support vector machines, genetic algorithm and decision rules for prediction. Feature selection techniques have been applied to extract relevant features required for the purpose of prediction. **Findings**: It is found that there is no single algorithm or technique that is the best of all the other algorithms/technique on any given medical dataset and application. Always there is a need to explore the right technique for the given dataset. A detailed review of the research on classification based prediction techniques reveal that the algorithms and techniques are applied on different data sets, which also has heterogeneous data types. It is observed that work is done on improving the predictive accuracy by applying attribute selection measures and feature selection techniques. Techniques have been developed to diagnose diseases, predict the occurrence of diseases, assess the gravity of the diseases such as cancer, heart, skin, liver, SARS, diabetes to name a few. The various applications explored are SMARTDIAB, H-Cloud, Medical Decision Support System, Evidence based medicine, adverse drug events, Passive In-home Health and Wellness monitoring, Healthcare management are a few applications developed in support of Medical data mining. **Application:** SMARTDIAB is an automated system for monitoring and management of type 1 Diabetic patients which supports monitoring, management and treatment of patients with type 1 diabetes. Passive In-home health and wellness monitoring is an application for monitoring older adults passively in their own living settings through placing sensors in their living environment.

Keywords: Bayesian, Challenges, Classification, Data Mining, Decision Tree, Healthcare, Survey

1. Introduction

Data is being generated, collected and accumulated at a very fast pace across a wide variety of fields. Since the data is massive, it is potentially not feasible for humans to analyse it manually. Hence there is a need for new computational theories and tools to assist humans in extracting useful information from large volumes of data. KDD-Knowledge Discovery from Data - is the nontrivial process of identifying valid, novel, potentially useful, understandable patterns in data. Data mining is a step in the KDD process that consists of applying data analysis and discovery algorithms over the data[1]. It is the computational process of discovering interesting patterns or extracting useful information from large data sets. It is required to mine data because massive amounts of

**Author for correspondence*

data is being collected and warehoused in the form of Web data, e-commerce data, purchases at departmental stores, Bank transactions, Hospital data etc. and is available. At the same time computers are much cheaper and more powerful as compared to the scenario a few years ago. Many times, information is “hidden” in the data that is not readily evident. Human analysts may take several weeks to discover useful information. With data mining, massive datasets can be automatically analysed thus extracting useful and novel patterns and information. The information thus extracted can be further used in decision making, predictions etc. In today’s fast world, hospitals generate enormous amounts of data which is in the form of patient information, electronic patient records, hospital resources, disease diagnosis, medicine, medical devices, treatment plans etc. This data can be processed and analysed, and useful information can thus be extracted which can support decision making. These hidden patterns provide healthcare professionals an additional source of knowledge for making decisions.

1.1 KDD Process

The KDD process is shown in Figure 1 and involves the following steps[1].

Data Selection: The relevant data required for data mining is selected from the database.

Data Pre-processing: Noise and inconsistency in data is removed.

Data Transformation: The selected data is transformed into appropriate form required.

Data Mining: The data mining methods are applied and patterns are extracted.

Pattern Evaluation: The patterns are evaluated and truly representative and interesting patterns are retained.

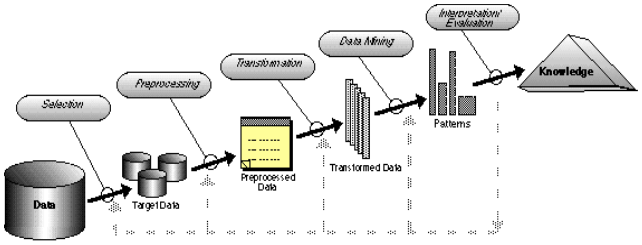


Figure 1. An overview of the steps in KDD process.

Knowledge Presentation: The extracted knowledge is presented using visualization techniques.

2. Data Mining Tasks

The prevalent data mining tasks are classification, association analysis, clustering andoutlier detection.

2.1 Classification

Classification is a process grouping data objects into one of the predefined classes. It is a supervised method. Here a classification model is derived for the data objects for which the class labels are known. The derived model is then used to predict the class label of data objects for which the class label is unknown. For example, a patient can be classified as “Diabetic” or “Non diabetic” based on the disease pattern. This is a typical case of binary classification where only two possible classes are considered. It is also possible to classify based on multiple classes as “Established diabetes”, “Marginally diabetic” or “Non diabetic”. A classification model for heart disease prediction is developed[2] to predict the occurrence of heart disease.

2.2 Association Analysis

Association Analysis is one of the vital data mining tasks which involve finding interesting patterns and relationships among attributes in a dataset. The importance of Association analysis lies in market basket analysis where the customer buying habits are analyzed based on the items the customers place in their shopping carts. An Efficient Incremental associative classification (EIAC) is proposed[3] which can keep previous mining results and learn from new data set to avoid repetitive relearning of data[3]. An ICU clinical decision support system - icu-ARM - based on Association Rule mining is designed and developed to perform real time data mining in ICU setting. The system provides valuable insights for physicians to prepare treatment plans based on the patients’ requirements[4].

2.3 Clustering

Clustering is an important data mining task which analyses data objects without referring to predefined class labels. It is an unsupervised method. Here the data objects are grouped into clusters based on some distance measure. The clusters of data objects are formed in such way that the objects within a cluster have high similarity as compared to objects in other clusters. Research work on tracing of contacts controlling infectious diseases and quarantine management is described in[5]. It finds clusters of cases and transmission routes of infectious diseases by applying clustering algorithms to activities of patients and their social interaction along with characteristics of SARS[5]. An analysis on impact of fluoride on human health is described in[6]. Here K-Means clustering algorithm is used to analyse the effect of fluoride on people who use underground water with high levels of fluoride[6].

2.4 Outlier Detection

In a dataset, there may be data that do not comply with the general behaviour of the data. Such data objects are called outliers. In some applications like fraud detection, such rare events are interesting and important. Observing the general behaviour of data objects and finding whether the object is an outlier is known as outlier detection/analysis. It has extensive use in a wide variety of applications such as intrusion detection in cyber security, military surveillance for enemy activities, insurance or health care, fraud detection for credit cards, and fault detection in safety critical systems[7]. Figure 2 illustrates outliers in a 2-dimensional data set. It shows that the data has two normal regions, N1 and N2, since most observations lie in these two regions. Those points that are far away from these regions, e.g., point's o1 and o2, and points in region O3, are outliers.

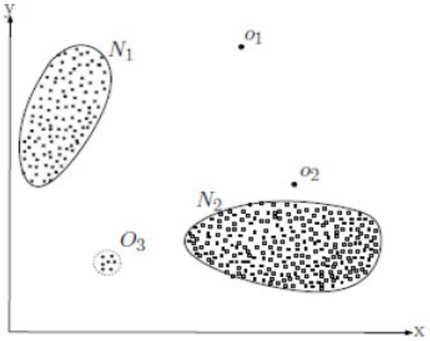


Figure 2. A simple example of outliers in a 2-dimensional data set[7].

3. Healthcare Data Mining

Data mining usage has witnessed unprecedented growth in the last few years. Recently the usefulness of data mining techniques has been realized in Healthcare domain. Healthcare industry today generates huge amounts of complex data about patients, hospitals resources, disease diagnosis, electronic patient records, medical devices etc. This large amount of data can be processed and analyzed to extract knowledge that can support cost-savings and decision making. In this scenario, data mining provides algorithms, tools and techniques that can be applied to this data to discover hidden and useful patterns that provide healthcare professionals an additional source of knowledge for making decisions. Medical data mining can discover the hidden patterns present in massive medical data which otherwise would be left undiscovered. Data mining techniques which are applied to medical data include association rule mining for finding frequent patterns, prediction of diseases, classification and clustering. This has led to the development of intelligent systems and decision support systems in Healthcare domain for accurate diagnosis of skin diseases, cancer and predicting the severity of respiratory illness, diabetes, heart diseases and remote health monitoring.

In traditional data mining patterns and trends in datasets are discovered, whereas in medical data mining more emphasis is on the minority that do not conform to the trends and patterns[1]. Sequential mining technique can be used to equip patients for managing their health for specific disease, better care and understanding through e-health[8]. Physicians can plan for differential diagnosis and treatment planning based on similar patients' diagnosis, treatments and outcomes[9].

A few applications of data mining in healthcare are:

- Diagnosis and Treatment of diseases
- Prediction of spread of epidemics
- Evidence based medicine
- Decision Support Systems for Physicians
- Detection of Adverse Drug Events
- Fraud and Anomaly Detection in Health Insurance Claims

4. Major Challenges of Data Mining in Healthcare

Data mining in healthcare has very specific features when compared to other fields. The major uniqueness features of healthcare data with respect to data mining are:

- **Heterogeneity of Healthcare Data[10,11]:** Healthcare data may be collected from electronic medical records, various images, patient interviews, reports, laboratory tests, physician observations and interpretations. This requires efficient mining in image databases, high capacity storage devices, and new tools for analysis of such data and visualization techniques, computer translation for processing physician's interpretation. The data-thus collected may be structured, unstructured or semi-structured. Therefore mining knowledge from them adds challenges to data mining.
- **Voluminous Data**[1]: Since healthcare data is voluminous it is required to extract samples from a database such that the results are representative for the entire database. Dimensionality reduction also needs to be done.
- **Change Capture**[1,10]: Healthcare data are constantly updated. This requires methods that are able to incrementally update the knowledge learnt so far without having to learn from scratch.
- **Noisy, Redundant, Inconsistent, Incomplete Data**[11]: Healthcare data may contain noisy, redundant, incomplete or inconsistent data objects and attributes. This requires suitable techniques to handle redundant, insignificant or inconsistent data objects and attributes. Without effective data scrubbing methods, the accuracy of the discovered patterns will be poor.
- **Incorporation of Constraints, Expert Knowledge, and Background Knowledge in Data Mining**[1,11]: The patterns discovered may not always be interpretable. Hence to guide the discovery process and to express the discovered patterns, the background knowledge, constraints, expert knowledge can be used.
- **Efficiency and Scalability of Data Mining Algorithms**[11]: In order to effectively extract the information from huge amount of data in databases, data mining algorithm must be efficient and scalable.
- **Parallel, Distributed, and Incremental Mining Algorithms**[1]: Since the size of the data is huge and the data is widely distributed across many computers, it requires parallel and distributed data mining algorithms to extract patterns quickly. Incremental algorithms are used to mine data from data updates.
- **Protection of Security, Integrity, and Privacy in Healthcare Data[11]:** The data mining techniques must incorporate proper security and privacy mechanisms to ensure patient confidentiality.
- **Imbalanced Data**[1]: A dataset is imbalanced if the classification categories are not approximately equally represented. The data mining techniques to learn classifiers for classes with very imbalanced distributions are required to handle imbalance data.
- **Usefulness of Interesting Patterns[11]:** Generally data mining techniques generate large number of patterns. Hence the major challenge here is to extract useful and interesting patterns from this large pattern set.
- **Interpretation and Analysis of Patterns**[1]: The patterns extracted require right interpretation for their proper analysis. Hence there is need for a domain expert to interpret and analyze the patterns.
- **Privacy and Ethical Use of Patient Information**[10]: As medical data is collected on humans, there are privacy and legal issues related to misuse of data. These issues are regarding data ownership, privacy and security of human data.

5. Prediction

Prediction is a data mining task which tends to predict an outcome of interest. Statistical techniques are used in predictive modelling predict future behaviour. Predictive modelling involves the following steps:

- Data collection
- Data analysis
- Formulation of a statistical model
- Prediction
- Model validation

With respect to predictive data mining in clinical medicine, the objective is to derive models using patient specific information to predict the outcome of interest and to thus support clinical decision-making. Predictive data mining techniques may be applied to construct decision models for prognosis, diagnosis and treatment planning for patients. These models can be incorporated within clinical information systems after evaluation and verification[12]. Prediction may be done using classification as well as clustering data mining techniques. In classification based prediction, a model is learnt with the training data using one of the many algorithms. The model is and then evaluated for its usefulness for test data based on its classification accuracy and other measures. Once evaluated, the model is ready for usage and can perform predictions for unknown data. In clustering based prediction, the data objects are grouped based on some distance measure, which form clusters of objects. A data object belongs to a particular cluster if it is in close proximity with that cluster as compared to other clusters. Following is the review of different classification based prediction techniques particularly in the healthcare domain and medical fields.

6. Classification based Prediction Techniques

6.1 Decision Tree

Decision Trees are powerful classification algorithms that are predominantly used in data mining. A Decision tree is a tree structure which comprises of the root, non-leaf nodes and leaf nodes. Each non leaf node denotes a test on an attribute, each branch represents an outcome of the test and each leaf node holds a class label[13]. The basic idea of decision tree is to split data recursively into subsets such that each subset contains almost homogenous states of target variable[14]. To select the splitting criterion, an attribute selection measure is chosen such that it "best" separates a given dataset. Some of the popular decision tree algorithms include ID3, C4.5 and CART which use Information gain, Gain Ratio, and Gini Index as their attribute selection measure respectively. C5 is also a decision tree based algorithm which is an improved version of C4.5[15]. Once the decision tree is built; it can be used to classify a new instance by traversing from root to the leaf, applying the test criterion at every non leaf node. The class for the instance is the class of the leaf node. Figure 3 shows decision tree that can be used to classify a patient into high risk and low risk class.

6.2 Rule based Classification

In rule based classification, the learned model is represented as a set of IF-THEN rules. The rules are composed of two parts namely rule antecedent - which is the If part, and rule consequent - which is the else part. An IF-THEN rule is of the form

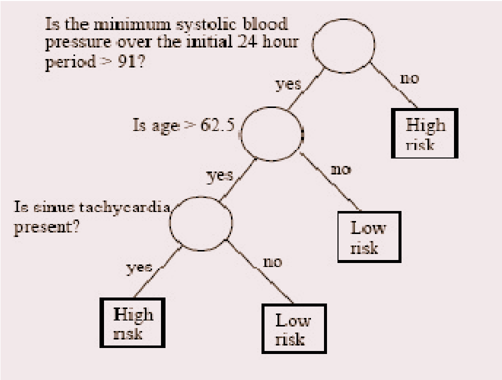

Is the minimum systolic blood pressure over the initial 24 hour period > 91?
yes
no
High risk
Is age > 62.5
yes
no
Low risk
Is sinus tachycardia present?
yes
no
High risk
Low risk

Figure 3. Decision tree for classification of high risk and low risk heart patients.

IF condition THEN conclusion

The condition in the rule antecedent has attribute tests and the consequent contains class prediction. As an example consider the rule R1

R1: High CHD Risk* ← *gender = female ^ age > 63 ^ body mass index >25 kg/m^2

This rule describes a group of patients who are female, overweight, and older than 63 years as having high risk of Coronary Heart Disease (CHD)[16]. The rules can be derived from induced decision trees or they can be derived from the data using sequential covering algorithm as in AQ, CN2 and RIPPER. To extract rules from a decision tree, one rule is created for each path from the root to a leaf node. Each splitting criterion is logically ANDed to form the rule antecedent and the leaf node has class prediction which is the rule consequent[13]. Rules can also be extracted directly from data. Here one rule is learned at a time. When a rule is learned, the data tuples covered by the rule are removed and the process repeats on the remaining data tuples. The rules thus generated can then be used to classify a new data tuple based on the rule triggered by the data tuple.

6.3 Logistic Regression

Logistic regression is a powerful statistical method for analysing a dataset which has one or more independent variables determining an outcome. It measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function. The outcome is two-valued which is categorical in nature. It can be used for predicting the probability of occurrence or non-occurrence of an event. Logistic Regression and Naive Bayes is used to identify risk factors associated with Type 2 Diabetes Risk Factors. This research work used Anthropometry and Triglycerides to assess the association between the HW(Hypertriglyceridemic waist) phenotype and type 2 diabetes[17]. Logistic Regression and Naive Bayes to predict fasting plasma glucose using anthropometric measures for type 2 diabetic patients[18] this research work used 37 anthropometric measures like weight, BMI, NeckC, ChestC, waist to name a few. The results obtained for type 2 diabetes in Men is shown in Table 1[18].

6.4 k-Nearest Neighbor (k-NN)

k-Nearest Neighbour classifier represents each tuple as a data point in a d-dimensional space where d is the number of attributes. In this way all tuples are stored and thus remembered during learning stage. When a new tuple whose class is unknown is given, the k-NN classifier compares its proximity with the k-nearest training tuples and assigns the class of k-nearest neighbours with majority vote or distance weighted vote to the new tuple. Some of the widely used proximity measures for finding nearest neighbours include Euclidean distance, Manhattan distance, Simple Matching coefficient, Jaccard similarity coefficient, Cosine similarity, and correlation coefficient. K-NN has a wide number of applications which include cluster analysis, image analysis, pattern recognition, prediction and economic forecasting. A diagnostic software tool to obtain correct diagnosis of Skin diseases is developed[19]. A prediction model for inference of missing ICD 9 (International Classification of Diseases) coded based on attributes like medical diagnosis, medical remarks, and patient statements aredeveloped by[20]. An early warning system for chronic illnesses is developed using K-NN. It determines the critical value of the important risk factors of each chronic illness[21]. Figure 4 shows the classification result for hypertension and diabetes mellitus using k-NN.

6.5 Bayesian Classifiers

Bayesian classifiers are statistical classifiers and are based on Bayes theorem. These classifiers can predict class membership probabilities. The probability that the data tuple X belongs to class C[2,13].

Let X be a data tuple which is called evidence, H is some hypothesis

According to Bayes theorem

Table 1. Results showing association between type 2 diabetes and anthropometric measures

Index	Unadjusted		Adjusted	
	p	OR	*p**	OR*
HW phenotype	<0.001	2.01(1.68-2.41)	<0.0001	2.07(1.72-2.49)
Weight	<0.001	1.31(1.22-1.40)	<0.0001	1.56(1.45-1.68)
BMI	<0.001	1.39(1.29-1.49)	<0.0001	1.53(1.42-1.65)
NeckC	<0.001	1.49(1.39-1.60)	<0.0001	1.61(1.49-1.73)
ChestC	<0.001	1.51(1.41-1.62)	<0.0001	1.60(1.49-1.73)
RibC	<0.001	1.61(1.50-1.73)	<0.0001	1.64(1.52-1.77)
WaistC	<0.001	1.59(1.48-1.71)	<0.0001	1.60(1.48-1.72)
HipC	<0.001	1.19(1.11-1.28)	<0.0001	1.30(1.21-1.40)
Neck_Hip	<0.001	1.35(1.26-1.44)	<0.0001	1.33(1.24-1.43)
Rib_Hip	<0.001	1.63(1.52-1.76)	<0.0001	1.60(1.48-1.73)
Waist_Hip	<0.001	1.73(1.60-1.86)	<0.0001	1.66(1.54-1.80)
Forehead_Waist	<0.001	0.60(0.56-0.65)	<0.0001	0.62(0.58-0.68)
Forehead_Rib	<0.001	0.59(0.55-0.64)	<0.0001	0.61(0.56-0.66)
Forehead_Neck	<0.001	0.65(0.60-0.69)	<0.0001	0.64(0.59-0.69)
WHtR	<0.001	1.36(1.27-1.46)	<0.0001	1.57(1.46-1.69)
TG	<0.001	1.34(1.26-1.44)	<0.0001	1.38(1.28-1.47)

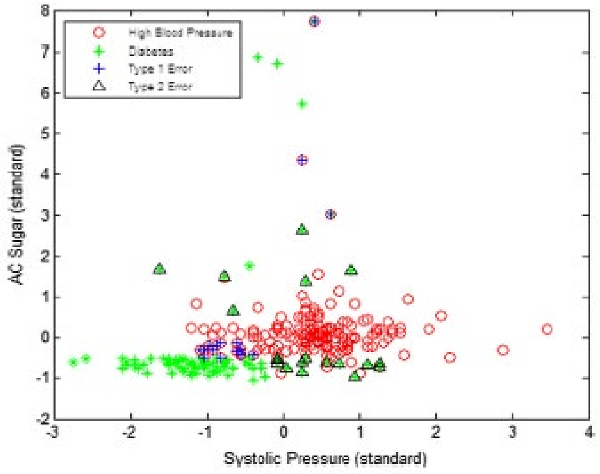


Figure 4. The classification result for hypertension and diabetes mellitus using k-NN[21].

$$P(H\,|\,X) = \frac{P(X\,|\,H)P(H)}{P(X)}$$

Where P(H | X) - is the probability that the hypothesis H holds for the given evidence X, and is called Posterior probability of H conditioned on X,

P(H) - is the Prior probability of H,

P(H | H) - is the posterior probability of X conditioned on H,

P(X) is the prior probability of X

Bayes theorem provides basis for two classifiers viz. Naive Bayes classifier and Bayesian Networks.

6.5.1 Naive Bayesian Classifier

Naive Bayesian Classifier presumes that the attributes values are conditionally independent of one another and there exists no dependence relationships among the attributes. This is the most accurate classifier when the assumption holds true[13,22].

Suppose there are m classes C_1, C_2, … C_m and X is data tuple to be classified, $X = (x_1, x_2, \ldots x_n)$ where x_i are attribute values then by Bayes theorem,

$$P(C_i\,|\,X) = \frac{P(X\,|\,C_i)P(C_i)}{P(X)}$$

$$\text{Where } P(X\,|\,C_i) = \prod_{k=1}^{n} P(x_k\,|\,C_i)$$

$$=P(x_1 \mid C_i)\ xP(x_2 \mid C_i)\ x\ldots x\ P(xn \mid C_i)$$

The Naive Bayesian classifier predicts tuple X belongs to class $C_i$iff

$$P(C_i \mid X) > P(C_i \mid X) \text{ for} 1 \le j \le m, j \ne i$$

Naive Bayes classifier and logistic regression is used to evaluate prediction power of anthropometric measures from body to diagnose Fasting Plasma Glucose status in Korean adults. Results indicate that combination of anthropometric measures was superior to individual measures[18]. ODAB (One Dependency Augmented Naïve Bayes classifier) and NCC2 (Naive Creedel Classifier) is applied to diagnose lung cancer disease by[23]. A new algorithm cNk is proposed which is a combination of Naive Bayes and k-NN and is applied to different datasets from UCI learning repository[24,25]. Table 2 shows the results of computation of probabilities for the attributes chills, runny nose, and swine flu using Naive Bayes classifier[39].

6.5.2 Bayesian Belief Networks (BBN)

Naive Bayes classifier assumes that the attributes are independent of each other but in real world scenario, the attributes may be correlated as in the medical domain where a patient's symptoms and health state are correlated[26]. In practice dependencies can exist and hence Bayesian belief networks are used to model the dependencies between attributes using joint conditional probability distributions. A Bayesian Belief network consists of a) directed acyclic graph which shows dependencies among attributes and b) Conditional Probability Table (CPT) associated with each attribute. Here the nodes represent attributes and arcs represent dependencies.

Table 2. Computed results in the prediction of swine flu using Naive Bayes classifier[39]

P(swine flu)=Y	0.625	P(swine flu)=N	0.375
P(chills=Y | swine flu=Y)	0.6	P(chills=Y | swine flu=N)	0.333
P(chills=N | swine flu=Y)	0.4	P(chills=N| swine flu=N)	0.666
P(runny nose=Y | swine flu=Y)	0.8	P(runny nose=Y | swine flu=N)	0.333

Suppose $X=(x_1,\ldots x_n)$ *be a data tuple with attributes* $Y_1\ldots\ Y_n$.

A node in a Bayesian network is conditionally independent of its non-descendants if its parents are known. The following equation represents existing joint probability distribution.

$$P(x_1,\ldots \mathrm{X}_n) = \prod_{k=1}^{n} P(x_i \mid Parens(Y_i))$$

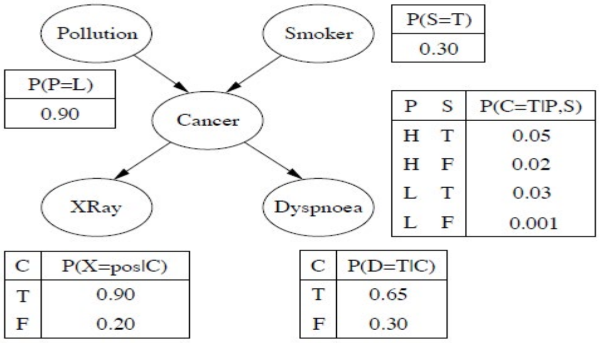
Where $P(x_1,\ldots.\ X_n)$ *is the probability of a particular combination of values of X and*

$P(x_i \mid Parents(Y_i))$ *are the entries in conditional probability table for* Y_i.

Figure 5 shows the Bayesian Belief Network and the associated CPT for all the attributes for lung cancer problem[27]. From the CPT for cancer attribute, we see that

$$P(cancer = T \mid Pollution = H,\ Smoker = T) = 0.05$$

Bayesian belief networks are now increasingly used in prediction data mining by researchers in the healthcare sector. They are widely applied to classification problems, decision support systems, genetic linkage analysis and text analysis.



P(S=T)
0.30

P(P=L)
0.90

P	S	P(C=T|P,S)
H	T	0.05
H	F	0.02
L	T	0.03
L	F	0.001

C	P(X=pos|C)
T	0.90
F	0.20

C	P(D=T|C)
T	0.65
F	0.30

Figure 5. Bayesian belief network and CPT for lung cancer problem[27].

6.6 Artificial Neural Network (ANN)

An artificial neural network is a computational model based on biological neural systems. It consists of interconnected processing elements called nodes or neurons that work together to produce an output function. It is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase[26].

An ANN can be a simple perceptron model or a more complex multilayer perceptron model. The perceptron model consists of input nodes and output nodes where each input node is connected to output node through weights. The model is trained by adjusting the weights until they fit the input output relationship of the data[22]. The multilayer ANN consists of several intermediate hidden layers between input and output layers and can be used to model complex relationships between input and output variables. ANN has been successfully applied in clinical medicine in classification and pattern recognition. Neural Networks with Back propogation and association rule mining is used for tumor classification in mammograms[28]. ANN is used in lung abnormality diagnosis to find whether it is cancerous or benign. An ANN model is developed which represents 1-year mortality in elderly patients with intertrochanteric fracture. The model has 8 input nodes, 6 nodes in hidden layer, and 1 output node, which represents 1-year mortality in elderly patients with intertrochanteric fracture[21] as shown in Figure 6.

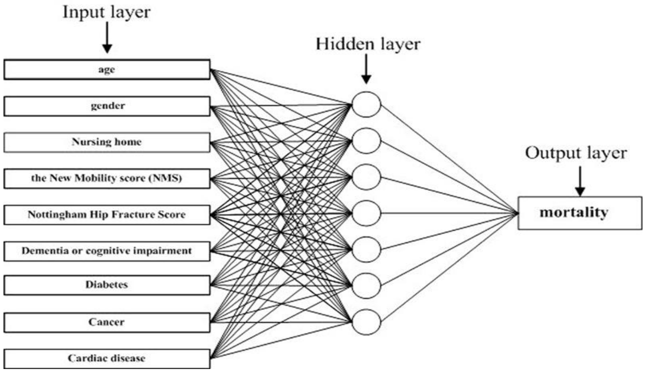


Figure 6. A multilayer ANN representing 1-year mortality in elderly patients with intertrochanteric fracture[21].

6.7 Support Vector Machine

Support Vector Machines (SVM) is most powerful classification algorithms in terms of predictive accuracy. They are based on strong mathematical foundations and statistical learning theory[29]. They can classify both linear and nonlinear data. SVMs were initially designed for two-class problems but later used for multi-class problem also. The basic principle of SVM is to find an optimal hyper plane with a maximum distance to the closest point of the two classes. A set of tuples that is closest to the optimal hyperplane is called a support vector. SVM uses these support vectors to find the optimal hyperplane. Finding the optimal hyperplane provides a linear classifier, whereas to classify nonlinear data, the original training data is transformed into higher dimension using nonlinear kernel functions such as polynomial, radial, Gaussian, sigmoid etc. 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[12,29]. Figure 7 shows the working of SVM classification algorithm. SVMs can be applied for numeric prediction as well as classification. They have a wide area of application which includes pattern recognition, medicine, bioinformatics, object recognition and prediction. SVM is used to cluster microarray data and extract associated genes for classifying cancer related documents[30]. A combination of Kernelized fuzzy rough set and SVM is proposed to identify cancer biomarkers from microarray data[31]. Biomarkers are discovered from one miRNA and three gene expression data sets.

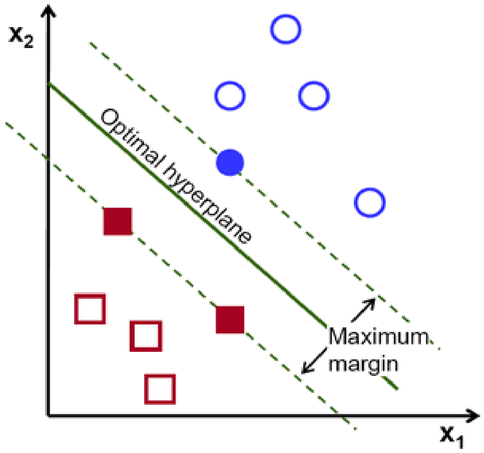


Figure 7. Optimal hyperplane with maximum margin between support vectors in SVM.

6.8 Genetic Algorithm

Genetic Algorithms (GA) are search algorithms based on natural selection and natural genetics. They provide robust search capabilities in complex spaces. GA is an iterativeprocess that operates on a population. An initial population consisting of randomly generated rules is created, where each rule is represented as string of bits. Every individual in the population is assigned a fitness value by means of a fitness function. Based on the theory of survival of the fittest, a new population (offsprings) is formed consisting of the fittest rules in the current population. To create the offsprings, the genetic operators - crossover and mutation are used. In cross over, the substrings from pairs of rules from the parent are swapped to form new rules. Cross over can be done in many ways which include single point crossover, two point cross over, uniform cross over and arithmetic cross over. In mutation, one or more randomly selected bits are altered[13,32,33] thus giving rise to a new rule. Genetic algorithms have been used for classification and optimization problems. GA is used for Cancer Detection using serum proteomic profiling. In this research work GA is used for feature selection from 3 serum datasets in the detection of cancer[12].

6.9 Fuzzy Logic

Fuzzy logic is a multivalued logic and is an effective tool to handle problems of uncertainty. In crisp logic, the truth values of the predicates take only to 2 values, either a 1 or a 0, whereas in fuzzy logic, the truth values are multivalued and hence take values in the range 0 - 1. Fuzzy logic is being used to handle the concept of partial truth, where the truth value may range between completely true and completely false. The fuzzy logic employs membership function to represent the degree of truth. The membership function defines how each point in the input space is mapped to a membership value (or degree of membership) between 0 and 1. Initially all input values are fuzzified into fuzzy membership functions. Then fuzzy output functions are computed with all the applicable rules in the rule base. Finally the fuzzy output functions are de-fuzzified to get "crisp" output values[34]. A heart disease prediction system is developed using fuzzy c means clustering[35]. The system uses data from Cleveland Clinic Foundation to predict heart disease symptoms.

6.10 Rough Sets

Rough set theory is a mathematical approach to imperfect knowledge or vagueness. Rough sets are imprecise sets as opposed to crisp sets which are precise. Each rough set has boundary line cases where objects cannot be classified with certainty. Here the vagueness or imprecision is expressed by a boundary region of a set. To establish the boundary region of the set, lower approximation and upper approximation are defined. The lower approximation consists of all objects which certainly belong to the set and the upper approximation contains all objects which possibly belong to the set. The difference between the upper and the lower approximation constitutes the boundary region of the rough set. It finds its applications in artificial intelligence, machine learning, knowledge acquisition, expert reasoning, and knowledge discovery from databases and pattern recognition[36]. A rough set based classification model to identify hospice candidates within a group of terminally ill patients is presented in[37]. Here subgroups of patients possessing common characteristics are identified and a collection of decision rules is derived for classifying new patients based on the subgroup. Among many feature selection techniques, a group incremental approach of feature selection using rough sets is presented in[38]. This method uses rough set technique to find new subset of features when a group of objects are added to a decision table.

7. Advantages and Disadvantages of Classification Techniques

The advantages and disadvantages of different classification techniques are listed in Table 3.

Table 3. Advantages and disadvantages of classification techniques

Method	Advantages	Disadvantages
Decision Trees	Easy to understand. Fast learning and prediction Lower memory requirements Does not need domain knowledge in the construction of a decision tree Can handle data with high dimension. Easy to interpret. Can handle both numerical and categorical data	Parts of the trees are replicated. Limited rule representation Numeric attributes can lead to large branching factors Restricted to one output attribute Unstable classifier For numeric dataset, it generates complex decision tree.
Decision Rules	Highly expressive Can produce descriptive models that are easy to interpret Easy to generate Can classify new instances very fast Performance is comparable to decision trees Can handle large data sets with imbalanced class distributions	May produce rules with very small coverage Rules may not be exhaustive
Logistic Regression	Can avoid overfitting Feature selection can be done The output can be interpreted as a probability Robust to noise	It requires more data to achieve stable, meaningful results Identifying right independent variables Overfitting
Nearest Neighbour	Fast training Complex target functions No loss of information Easy to implement	Slow at query time (have to evaluate all instances) Sensitive to correct choice of similarity metric Easily fooled by irrelevant attributes Needs large storage space to remember all instances Noise sensitive Slow testing
Bayesian Classifier	Yields optimal prediction given the assumptions Has high speed Can handle discrete or numeric attribute values Naive Bayes classifier easy to compute	Optimal Bayes classifier computationally intractable Naive Bayes assumption usually violated Does not give accurate results in some cases where dependency exists among variables
ANN	Adaptive learning Self organization Fast prediction Easily identify complex relationships Can handle noisy data	All inputs have to be translated into numeric inputs Training is slow Learning might result in a local optimum Overfitting Difficult to interpret Processing time for large neural networks is high Cannot be initialized with prior knowledge

SVM	Deterministic algorithm Uses maximum marginal hyperplane for classifying linearly separable data Hard to train Can learn very complex functions using kernels	Computationally expensive Training process takes more time Can solve binary class problem
GA	Concepts are easy to understand Easily parallelizable They find the best solution	They are very slow Fitness function must be accurate Crossover rate must be high Mutation rate must be low
Fuzzy Logic	Intelligent approach and simple Easy to understand and implement Provides user friendly approach of presentation	To develop a model for a fuzzy system is hard Requires fine tuning and simulation before its operational use
Rough sets	Programs implementing methods of RST may easily run on parallel computers Useful rule induction from incomplete datasets Can identify partial or total dependencies in data Does not need any prior information about data Identifies such relationships that would not be found using statistical methods Permits both quantitative and qualitative data Easy to understand and good interpretation	Cannot handle noisy data Inefficient computation & time consuming Not suitable for large datasets in real world applications

8. Developments and Work Done

The research work and developments in data mining, particularly in the healthcare sector are summarized in Table 4.

9. Research Applications in Healthcare

Some of the research applications developed in healthcare are listed in Table 5.

Table 4. Research work in healthcare domain

Objective	Methods	Dataset	Health Issue	Results
Evaluate factors for preterm birth[40]	Exploratory factor analysis	TMR- Duke University Medical Centre – 20000 records	Pregnancy	3 Factors responsible for preterm birth were identified
Develop predictive models for breast cancer survivability[15]	Artificial Neural networks, Decision Tree, Logistic Regression	SEER – Surveillance, Epidemiology and End Results- National Cancer Institute	Breast Cancer	C5 – 93.6% accuracy ANN- 91.2 % Logistic Regression – 89.2%
i) Review of data mining technique in cancer detection /diagnosis ii) Explore new analytic method iii) Compare results on different datasets[41]	SVM Genetic Algorithm Used for feature selection	3 serum SELDI MS datasets	Cancer	GA performed better
To cluster microarray data and extract associated genes for classify cancer related documents[30]	SVM	Gene Microarray	Cancer	Can extract potential patterns for cancer
Predict medical costs based on previous years' data[42]	Clustering, Classification	800,000	Claims data	Clustering has better predictions

Find clusters of cases and transmission routes. Tracing of contact by applying algorithms to model patients activities, their social interaction and characteristics of infectious disease[5]	Clustering	Hong Kong SARS outbreak	SARS – Severe Acute Respiratory Syndrome	Contact tracing done successfully
i) Evaluate Decision tree and Association Rule mining ii) Predict occurrence of route of transmission[14]	Microsoft Decision Tree, Microsoft Association Rules	HIV Database - AIIMS Size - 672	HIV	DT,AR effectively used
Develop a system to assess heart related risk factors with the aim to reduce Chronic heart diseases[43]	C4,5 –Decision tree	Paphos in Cyprus, 528	Heart Disease	MI- 66% PCI-75% CABG-75%
To predict whether the patient is diabetic[44]	Apriori	UCI	Pima Indian Diabetes	Association rules generated
An Active learning model to discover new classes A classifier to switch between generative and discriminative classifiers during learning[45]	New algorithm for active learning	UCI	9 data sets	Model superior over other methods
To obtain correct diagnosis of skin diseases[19]	Basic KNN, Weighted KNN	UCI	Skin diseases	Weighted KNN with Manhattan distance gives better accuracy
Develop predictive model for ICD 9 codes[20]	Nearest Neighbour	Occupational injury data	All	70% accuracy
Evaluate the increased tendency of ectopic pregnancy and liver disease[46]	Regression Analysis	WEKA 162	Pregnancy &Liver Disease	Modern adaptive lifestyle& Cultural changes
Determine fitness of a person based on historical and real time data[47]	K Means, D Stream	Switzerland Dataset, 107 records	General	K-Means – 83% D-Stream – 87%
To predict and detect early onset of diabetes and cardiac diseases[48]	C4.5,	WEKA UCI	Heart &Diabetes	Prediction accuracy 93%
Discover interesting rules and relations among attributes in a database[49]	Association Rule Mining	2102 records	Malaysian Ministry of Health	90% accuracy
Presents a wireless sensor network based Mobile Real time Health Care Monitoring(WMRHM) framework which can give predictions based on real time vital body signals[50]	K Means	MIMIC II 64 records	ICU	Proposed framework is validated
Predict the diagnosis of heart disease with reduced number of attributes[51]	Decision Tree, Naïve Bayes, Clustering	WEKA UCI	Heart	DT-99.2% NB-96.5% C-88.3%
Demonstrate the accuracy level of different classifiers[52]	Decision Tree, Naïve Bayes, Multilayer Perceptron(MLP)	WEKA UCI	Heart	MLP is better than CDP

To develop predictive models for heart disease survivability[53]	CART, ID3, Decision Tree	Cleveland Clinic Foundation, 303 records	Heart	CART – 83% ID3 -72% DT- 82.5
Build a new classifier that combines statistical based and distance based classifier[25]	K-NN, Naïve Bayes	UCI	Hepatitis Heart	86%, 85%
Identify frequency of diseases in particular geographical areas[54]	Association rule based Apriori Algorithm WEKA tool used	1246 records	All	Frequent Diseases identified
Automatically label groups of segmentations of different structures from a radiation therapy plan[55]	Conditional Random forests	Princess Margaret Hospital Toronta, 6844 records	Lung Cancer	Accuracy of classification – 91.58%
A model for early detection and correct diagnosis of lung cancer is proposed[23]	One Dependency Augmented Naïve Bayes Classifier (ODANB)and Naïve Creedal Classifier(NCC2)	UCI	Lung Cancer	ODANB- 80.46% NB – 84.14%
Discover Adverse drug effect[56]	Likelihood ratio model, Bayesian network model	Simulated Dataset	Drug	Usefulness of proposed method demonstrated Results improved by 23.83%
Develop heart disease prediction system[22]	Decision Tree, Naive Bayes, Neural Networks	Cleveland Clinic Foundation	Heart	Model proposed - Study
To find effect of diabetes on Kidney[57]	Decision Tree C4.5	Jyoti diagnostic & Research Centre 148 records Tanagra Tool	Diabetes	Total Protein is the attribute which has greatest effect on kidney due to diabetes
Design and develop diagnosis and prediction system for heart diseases[58]	Naïve Bayes , Decision Tree	UCI	Heart	Naïve Mayes – 85% DT - 84%
A Predictive adoption model for people with dementia is proposed which is based on video streaming using mobile phone[59]	Compares C4.5, CART, KNN, Naïve Bayes, NN, SVM	Data collected through questionnaires	Dementia	KNN perform better 84% accuracy
Predict survival of patients undergoing bone marrow transplant[60]	Logistic Regression, Random Forest, SVM	Sharaiati Hospital	Cancer	Accuracy 92-97%
Predict biomarkers for cancer disease from one miRNA and three gene expression datasets[31]	Kernelized Fuzzy Rough Set , Semisupervised SVM	Gene and miRNA cancer datasets	Cancer	Effectiveness of KFRS and SVM demonstrated
Predict the status of fasting plasma glucose(FPG) using 37 anthropometric measures for diagnosis of Type 2 Diabetes[18]	Logistic Regression, Bayes,	Korean Health and Genome Epidemiology study database (KHGES) 4870 SPSS	General	Bayes – 0.739 Logistic Regression 0.741

Predict the occurrence of heart disease using relevant attributes[61]	Decision Trees, Naïve Bayes, K-NN	UCI 303 records WEKA	Heart	Decision Tree- 92.2% Naïve Bayes – 84.2
Predict the occurrence of heart disease[62]	Decision Table, Naïve Bayes, Hybrid	1080 WEKA	Heart	DT-93.5% NB-95.8% Hybrid – 97.5%
Implementation of different classification methods on diabetes dataset. Comparative study and analysis of the techniques[63]	CART, Naïve Bayes, Bayesian, J48, Random Forest, Random Tree, K-NN	UCI 768 records WEKA	Diabetes	Comparative analysis done.
Assess the association between the HW(Hypertriglyceridemic waist) phenotype and type 2 diabetes[17]	Naïve Bayes, Logistic Regression	Korean Health and Genome Epidemiology study database (KHGES) 2099 SPSS WEKA	General	Waist-to hip ratio and TG – best predictors
Developed a tool to predict leukaemia. Most significant genes are identified and a classifier is built for prediction[64]	Decision Tree	Golub dataset	Cancer	Decision Tree – 94.11%

Table 5. Research applications in healthcare

Sl No	Area	Details
1	Evidence Based Medicine[65]	It is an approach taken by medical practitioners to diagnose and medicate diseases using medical evidence. It involves collection, interpretation, and summarization of evidences followed by systematic retrieval of the best evidence available and then applying them in practice.
2	Adverse Drug Events[65,66]	It is a damage incurred by a patient resulting from the use of a drug. It may occur due to medication errors like overdose, wrong medication, adverse drug reactions or wrong patient. Identification of factors relating to adverse drug events is one of the major areas of research in medicine.
3	Healthcare Management[67,68,69]	It is health management using monitoring equipment and devices whose objective is to help patients manage their medical conditions at home. Systems are designed which help to identify chronic disease state and keep track of high risk patients, and design appropriate interventions and reduce number of hospital admissions.
4	Predictive Data Mining in clinical medicine[12,70,71,72]	Uses patient specific information to derive models that can predict the outcome of interest and thus support clinical decision making.
5	Medical Decision Support System[73,74]	It is a framework which enables Medical decision making in the presence of partial information. A medical decision support system is a health information technology system designed to provide doctors and health professionals with assistance during decision making.
6	Passive In-Home Health and Wellness Monitoring[75]	These systems are used for monitoring older adults passively in their own living settings through placing sensors in their living environment. Their routine activities are analysed and mined to detect changes in their health conditions or indicators of early onset of disease.
7	SMARTDIAB :A Communication and Information Technology Approach for the Intelligent monitoring, Management and Follow-Up of Type -1 Diabetes Patients[76,77]	An automated system for monitoring and management of type 1 Diabetic patients is developed which supports monitoring, management and treatment of patients with type 1 diabetes.
8	HCloud : A Novel Application Oriented Cloud Platform for Preventive Healthcare[78,79]	HCloud – A healthcare system is developed for preventive healthcare service. It is used by many healthcare workers to access and consolidate all patient electronic medical records which gives fast access to patient

10. Data Mining Tools

Some of the most common Data Mining tools are listed in Table 6.

11. Data Repositories

Availability of the right data is a key factor in any data mining task. A list of publicly available datasets is presented in Table 7.

Table 6. Data mining tools

Sl No	Data Mining Tool	Description
1	SPSS / SPSS Clementine[80,81]	Statistical Package for the Social Sciences Now Developed and owned by IBM Includes Descriptive statistics, Bivariate statistics, Prediction
2	Salford systems[82] SPM/CART/MARS/ TreeNet/RF	Data mining and predictive analytics software Developed by Salford systems It offers advanced data mining software and consulting services and has powerful new automation and modeling capabilities
3	Rapid Miner[80,83]	Open source predictive analytics platform that provides an integrated environment for data mining, predictive analytics, machine learning, text mining and business analytics
4	SAS /SAS Enterprise Miner[84]	SAS (Statistical Analysis System) is a software suite developed by SAS Institute for advanced analytics, multivariate analyses, business intelligence, data management, and predictive analytics.
7	Weka[85]	*Waikato Environment for Knowledge Analysis* *Data mining with open source machine learning software* *It* is a collection of machine learning algorithms for data mining tasks. It contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning algorithms.
8	R[86]	R is a free software environment for statistical computing, data mining and graphics. Users can statistically explore data sets and can make many graphical displays of data.
9	Microsoft SQL Server[87]	Microsoft SQL Server Analysis Services, SSAS is an online analytical processing and data mining tool in Microsoft SQL Server. It includes Integration Services, Reporting Services and Analysis Services. Analysis Services includes a group of OLAP and data mining capabilities for Multidimensional and Tabular data.
10	MATLAB[88]	**MATLAB offers a** full set of statistics and machine learning functionality in addition to advanced methods and prebuilt algorithms for image and video processing, financial modeling, control system design.

Table 7. Publicly available datasets

Sl No	Name
1	UCI Machine Learning Repository[24]
3	KDD Cup 2008 -Siemens(Requires registration)[89]
4	MIT-BIH Arrhythmia Database[24]
5	ECML/PKDD discovery challenge dataset[89]
6	Healthcare Cost and Utilization Project (H-CUP)[89]
7	HIV Prevention Trials Network -Vaccine Preparedness Study/Uninfected Protocol Cohort[90]
8	National Trauma Data Bank (NTDB)[91]
9	Behavioural Risk Factor Surveillance System (BRFSS)[92]

12. Conclusion

The growth of data mining particularly in the healthcare field has reached great horizons in the recent years. It is evident that data mining helps in extracting useful and interesting patterns from massive data. The four important data mining tasks viz. Classification, Association analysis, Clustering and Outlier Detection are discussed keeping in mind the healthcare domain. Further the challenges of data mining in healthcare are explored. The heterogeneity of medical data is a major challenge because data is collected in the form of images, interviews with patient, laboratory tests, reports, physician observations and interpretations, which require appropriate techniques to handle them. The patterns thus discovered have to be interpretable which may require incorporation of expert knowledge, background knowledge and constraints to guide in the discovery process or to express the discovered patterns for the purpose of human interpretation.

Further a detailed review of the different classification techniques used for prediction and also their merits and demerits are explored. It is found thatthere is no single algorithm or technique that is the best of all theother algorithms/technique on any given dataset and application. Always there is a need to explore the right technique for the given dataset. A detailed review of the research on classification based prediction techniques and applications by various researchers are done extensively. It is found that the algorithms and techniques are applied on different data sets, sometimes from publicly available datasets and sometimes on data collected personally. It is observed that work is done on improving the predictive accuracy by applying attribute selection measures and feature selection techniques. Techniques have been developed to diagnose diseases, predict the occurrence of diseases, assess the gravity of the diseases such as cancer, heart, skin, liver, SARS, diabetes to name a few.

Applications have been developed for health monitoring from remote places, decision making, self-healthcare, preventive care and many more. These applications can be used by a wide variety of users which includes doctors, patients, healthcare insurers and healthcare organizations. With these applications, the doctors can identify effective cure and treatment plans for diseases, patients can obtain cost effective and non-invasive treatments, healthcare insurers can discover fraudulent medical claims. To conclude, always there is a need for effective and efficient data mining technique in order to uncover hidden information which could be useful for the medical fraternity as a whole.

13. References

1. Fayyad U, Piatetsky-Shapiro G, Smyth P. From Data Mining to Knowledge Discovery in Databases. AI Magazine.1996; 17(3):37-54.
2. Jenzi IS, Priyanka P, Alli P. A Reliable Classifier Model Using Data Mining Approach for Heart Disease Prediction. International Journal of Advanced Research in Computer Science and Software Engineering. 2013 Mar; 3(3):20-4.
3. Shaorong F, Zhixue H. An Incremental Associative Classification Algorithm used for Malware Detection. Proceedings of International Conference on Future Computer and Communication. 2010; 1:757-60. crossref.
4. Cheng CW, Chanani N, Venugoplan J, Maher K. icu-ARM-An ICU Clinical Decision Support System Using Association Rule Mining. IEEE Journal of Translational Engineering in Health and Medicine. 2013; 1:1-10. crossref.
5. Leong KI, Si YW, Biuk-Aghai RP, Fong S. Contact tracing in healthcare digital ecosystems for infectious disease control and quarantine management. Proceedings of International Conference on Digital Ecosystems and Technologies. 2009; p. 306-11.
6. Balasubramanian T, Umarani R. An analysis on the impact of fluoride in human health (dental) using clustering data mining technique. Proceedings of Pattern Recognition, Informatics and Medical Engineering (PRIME). 2012; p. 370-5.
7. Singh K, Upadhyaya S. Outlier detection: applications and techniques. International Journal of Computer Science Issues. 2012 Jan; 9(1):307-23.
8. Rahaman B, Shashi M. Sequential mining equips e-Health with Knowledge for mining diabetes. New Trends in Information Science and Service Science. 2010; p. 65-71.
9. Chignell M, Rouzbahman M, Kealey R, Samavi R, Yu E, Sieminowski T. Nonconfidential patient types in emergency clinical decision support. IEEE Security & Privacy. 2013; 11(6):12-8. crossref.
10. Cios KJ, William Moore G. Uniqueness of medical data mining. Journal of Artificial Intelligence in Medicine. 2002 Sep-Oct; 26(1-2):1-24. crossref.
11. Yang Q, Wu X. 10 Challenging Problems in Data Mining Research. International Journal of Information Technology & Decision Making.2006; 5(4):597-604. crossref.
12. Bellazzi R, Zupan B. Predictive data mining in clinical medicine: current issues and guidelines. International Journal of Medical Informatics. 2008 Feb; 77(2):81-97. crossref PMid:17188928.

13. Han J, Kamber M, Pei J. Data Mining: Concepts and Techniques, 3rd Edition. Morgan Kaufmann. 2011; p. 1-703.
14. Gosain A, Kumar A. Analysis of health care data using different data mining techniques. International Conference on Intelligent Agent & Multi-Agent Systems. 2009; p. 1-6. crossref.
15. Dursun D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. Artificial Intelligence in Medicine. 2005 Jun; 34(2):113-27. crossref PMid:15894176.
16. Lavrac N, Novak P. Relational and Semantic Data Mining for Biomedical Research. Informatica. 2013; 37(1):35-9.
17. Lee BJ, Kim JY. Identification of Type 2 Diabetes Risk Factors Using Phenotypes Consisting of Anthropometry and Triglycerides based on Machine Learning. IEEE Journal of Biomedical and Health Informatics. 2016; 20(1):39-46. crossref. PMid:25675467.
18. Lee BJ, Ku B, Nam J, Pham DD, Kim JY. Prediction of Fasting Plasma Glucose Status Using Anthropometric Measures for Diagnosing Type 2 Diabetes. IEEE Journal of Biomedical and Health Informatics. 2014 Mar; 18(2):555-61. crossref. PMid:24608055.
19. Cataloluk H, Kesler M. A diagnostic software tool for skin diseases with basic and weighted K-NN. Proceedings of Innovations in Intelligent Systems and Applications. 2012; p. 1-4.
20. Erraguntla M,Gopal B, Ramachandran S, Mayer R. Inference of Missing ICD 9 Codes Using Text Mining, and Nearest Neighbor Techniques. Proceedings of Hawaii International Conference on System Sciences. 2012; p. 1060-9. crossref.
21. Jen CH, Wang CC, Jiang BC, Chu YH, Chen MS. Application of classification techniques on development an early-warning system for chronic illnesses. Expert Systems with Applications. 2012 Aug; 39(10):8852-8. crossref.
22. Tan PN, Steinbach M, Kumar V. Introduction to Data Mining, 1st Edition. Pearson Addison Wesley. 2005; p. 1-769.
23. Krishnaiah V, Narsimha G, Subhash Chandra N. Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques. International Journal of Computer Science and Information Technologies. 2013; 4(1):39-45.
24. UCI machine learning repository. Available from: crossref. Date accessed: 30/10/2017.
25. Ferdousy EZ, Islam MM, Matin A. Combination of Naive Bayes Classifier and K-Nearest Neighbour(cNK) in the Classification Based Predictive Models. Journal of Computer and Information Science. 2013; 6(3):48-56.
26. Tomar D, Agarwal S. A survey of Data Mining approaches for Healthcare. International Journal of Bio-Science and Bio-Technology. 2013; 5(5):241-66. crossref.
27. Graphical models. Available from: crossref. Date accessed: 8/01/2018.
28. Shi L, Wang XC, Wang YS. Artificial neural network models for predicting 1-year mortality in elderly patients with intertrochanteric fractures in China. Brazilian Journal of Medical and Biological Research. 2013 Nov; 46(11):993-9. crossref. PMid:24270906 PMCid:PMC3854329.
29. Vapnik V. The support vector method of function estimation. Nonlinear Modeling. 1998; p. 55-85. crossref.
30. Lee CH, Wu CH, Chung HH. A Unified Multilingual and Multimedia Data Mining Approach for Cancer Knowledge Discovery. Intelligent Information Hiding and Multimedia Signal Processing. 2007; 2(2):241-4. crossref.
31. Chakraborty D, Maulik U. Identifying Cancer Biomarkers from Microarray Data Using Feature Selection and Semisupervised Learning. IEEE Journal of Translational Engineering in Health and Medicine. 2014; 2:1-11. crossref. PMid:27170887 PMCid:PMC4848046.
32. Holland JH. Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence. University of Michigan Press. 1992.
33. Goldberg DE. Genetic algorithms in search, optimization and machine learning. Pearson. 1989; p. 1-372.
34. Ross TJ. Fuzzy logic with engineering applications. John Wiley & Sons. 2009; p. 1-606.
35. Kaur L. Predicting Heart Disease Symptoms using Fuzzy C-Means Clustering. International Journal of Advanced Research in Computer Science Engineering & Technology. 2014 Dec; 3(12):4232-5.
36. Pawlak Z. Rough Sets. International Journal of Computer and Information Sciences. 1982 Oct; 11(5):341-56. crossref.
37. Herrera G, Tsalatsanis Y. Towards a classification model to identify hospice candidates in terminally ill patients. Engineering in Medicine and Biology Society. 2012; p. 1278-81.
38. Liang J, Wang F, Dang C. A Group Incremental Approach to Feature Selection Applying Rough Set Technique. IEEE Transactions on Knowledge and Data Engineering. 2014; 26(2):294-308. crossref.
39. Borkar AR, Deshmukh PR. Naive Bayes Classifier for Prediction of Swine Flu Disease. International Journal of Advanced Research in Computer Science and Software Engineering. 2015 Apr; 5(4):120-3.
40. Prather JC, Lobach DF, Goodwin LK, Hales JW, Hage ML, Hammond WE. Medical data mining: knowledge discovery in a clinical data warehouse. Proceedings of the AMIA annual fall symposium American Medical Informatics Association. 1997; p. 101-5. PMid:9357597 PMCid:PMC2233405.

41. Li L, Tang H, Wu Z, Gong J, Gruidl M, Zou J, Clark RA. Data mining techniques for cancer detection using serum proteomic profiling. Artificial Intelligence in Medicine. 2004 Oct; 32(2):71-83. crossref. PMid:15364092.
42. Bertsimas D, Margret B, Michael A, Christian Kryder J, Pandey R, Vempala S, Wang G. Algorithmic Prediction of Healthcare Costs. Operations Research. 2008 Dec; 56(6):1382-92. crossref.
43. Karaolis MA, Moutiris JA, Demetra H, Constantinos S. Assessment of Risk Factors of Coronary Heart Events on Data Mining with Decision Tree. IEEE Transactions on Information Technology in Biomedicine. 2010; 14(3):559-66. crossref. PMid:20071264.
44. Patil BM, Joshi RC. Association rule for classification of type-2 diabetic patients. Proceedings of Second international Conference on Machine Learning and Computing, India. 2012; p. 330-4. PMid:23326107 PMCid:PMC3543556.
45. Hospedales TM, Gong S, Xiang T. Finding rare classes: Active learning with generative and discriminative models. IEEE Transactions on Knowledge and Data Engineering. 2013; 25(2):374-86. crossref.
46. Aneesh Kumar AS, Venkateswaran JC. An Approach of Data Mining for Predicting the Chances of Liver Disease in Ectopic Pregnant Groups. International Journal of Computer Applications. 2013 Feb; p. 19-22.
47. Patil D, Agarwal B, Andhalkar S, Biyani R, Gund M, Wadhai VM. An Adaptive parameter free data mining approach for healthcare application. International Journal of Advanced Computer Science and Applications. 2012; 3(1):55-9. crossref.
48. Ashwinkumar UM, Anandakumar KR. Predicting Early Detection of Cardiac and Diabetes Symptoms using Data Mining Techniques. Proceedings of Computer Design and Engineering. 2012; 49:106-15.
49. Payus C, Sulaiman N, Shaahani M, Bakar AA. Association Rules of Data Mining Application for Respiratory Illness by Air Pollution Database. International Journal of Basic and Applied Sciences. 2013 Jun; 13(3):11-6.
50. Patil DD, Wadhai VM. Adaptive Real Time Data Mining Methodology for Wireless Body Area Network Based Healthcare Applications. Advanced Computing: An International Journal. 2012 Jul; 3(4):59-70.
51. Patel SB, Yadav PK, Shukla DP. Predict the Diagnosis of Heart Disease Using Classification Mining Techniques. IOSR Journal of Agriculture and Veterinary Science. 2013 Jul-Aug; 4(2):61-4.
52. Hossain J, Sani FM, Mustapha A, Lily S. Using Feature Selection as Accuracy Benchmarking in Clinical Data Mining. Journal of Computer Science. 2013; 9(7):883-8. crossref.
53. Chaurasia V, Pal S. Early Prediction of Heart Diseases Using Data Mining Techniques. Caribbean Journal of Science and Technology. 2013; 1:208-17.
54. Ilayaraja M, Meyyappan T. Mining medical data to identify frequent diseases using Apriori algorithm. Proceedings of Pattern Recognition, Informatics and Mobile Engineering (PRIME), India. 2013; p. 194-9.
55. McIntosh C, Svistoun I, Purdie TG. Groupwise conditional random forests for automatic shape classification and contour quality assessment in radiotherapy planning. IEEE Transactions on Medical Imaging. 2013 Jun; 32(6):1043-57. crossref. PMid:23475352.
56. Duan L, Khoshneshin M, Nick W, Liu M. Adverse Drug Effect Detection. IEEE Journal of Biomedical and Health Informatics. 2013 Mar; 17(2):305-11. crossref. PMid:24235108.
57. Jain D, Gautam S. Predicting the Effect of Diabetes on Kidney using Classification in Tanagra. International Journal of Computer Science and Mobile Computing. 2014 Apr; 3(4):535-42.
58. Venkatalakshmi B, Shankar MV. Heart Disease Diagnosis Using Predictive Data Mining. International Journal of Innovative Research in Science, Engineering and Technology. 2014 Mar; 3(3):1873-7.
59. Zhang S, McClean SI, Nugent CD, Donnelly MP, Galway L, Scotney BW, Cleland I. A predictive model for assistive technology adoption for people with dementia. IEEE Journal of Biomedical and Health Informatics. 2014 Jan; 18(1):375-83. crossref. PMid:24403437.
60. Taati B, Snock J, Aleman D, Ghavamzadeh A. Data mining in bone Marrow Transplant Records to Identify Patients with High odds of Survival. IEEE Journal of Biomedical and Health Informatics. 2014; 18(1):21-7. crossref. PMid:24403400.
61. Joshi S, Nair MK. Prediction of heart disease using classification based data mining techniques. Proceedings of Computational Intelligence in Data Mining, India. 2015; 2:503-11.
62. Sharma M, Kaur R. Data Mining in Healthcare using Hybrid Approach. International Journal of Computer Applications. 2015, 128(4):49-53. crossref.
63. Joshi S, Shetty SP. Performance Analysis of Different Classification Methods in Data Mining for Diabetes Dataset Using WEKA Tool. International Journal on Recent and Innovation Trends in Computing and Communication. 2015 Mar; 3(3):1168-73. crossref.
64. Joshi S, Deeptha A, Prathibha K, Hema N, Priyanka J. Classification and Prediction of Disease Classes using Gene Microarray Data. International Journal of Data Mining Techniques and Applications. 2016 Jun; 5(1):7-10. crossref.

65. Canlas RD. Data mining in healthcare: Current Applications and Issues. School of Information Systems & Management, Carnegie Mellon University, Australia. 2009 Aug; p. 1-11.
66. Trifiro G, Kors JA. Data mining on electronic health record databases for signal detection in pharmacovigilance: which events to monitor? Pharmacoepidemiology and Drug Safety. 2009; 18:1176-84. crossref. PMid:19757412.
67. Koh HC, Tan G. Data mining applications in healthcare. Journal of Healthcare Information Management. 2005; 19(2):64-72. PMid:15869215.
68. Diwani SA, Sam A. Framework for Data Mining In Healthcare Information System in Developing Countries: A Case of Tanzania. International Journal of Computational Engineering Research. 2013; 3(10):1-7.
69. Kavitha K, Sarojamma RM. Monitoring of Diabetes with Data Mining via CART Method. International Journal of Emerging Technology and Advanced Engineering. 2012 Nov; 2(11):157-62.
70. Milovic B, Milovic M. Prediction and Decision Making in Health Care using Data Mining. International Journal of Public Health Science (IJPHS). 2012 Aug; 1(2):69-78. crossref.
71. Ahmed A, Hannan SA. Data Mining Techniques to find out Heart Diseases: an Overview. International Journal of Innovative Technology and Exploring Engineering. 2012 Sep; 1(4):18-23.
72. Krishnaiah V, Narsimha G, Subhash Chandra N. A Study on Clinical Prediction Using Data Mining Techniques. International Journal of Computer Science Engineering and Information Technology Research. 2013; 3(1):239-48.
73. Khan A, Doucette JA, Cohen R. A Hybrid Design for Medical Decision Support using Data Mining to Impute Missing Data. Proceedings of HI-Knowledge Discovery in Databases, ACM. 2012.
74. Alistair J, Mohammed M, Shamim M, Katherine E, David A, Gari D. Machine Learning and Decision Support in Critical Care. Proceedings of Institute of Electrical and Electronics Engineers. 2016; 104(2):444-66. crossref. PMid:27765959 PMCid:PMC5066876.
75. Alwan M. Passive in-home health and wellness monitoring: Overview, value and examples. Proceedings of Annual International Conference of the IEEE Engineering in Medicine and Biology Society. 2009; p. 4307-10. crossref.
76. Skevofilakas M, Mougiakakou SG, Zarkogianni K, Aslanoglou E, Pavlopoulos SA, Vazeou A, Nikita KS. A communication and information technology infrastructure for real time monitoring and management of type 1 diabetes patients. Proceedings of 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. 2007; p. 3685-8. crossref.
77. Mougiakakou SG, Bartsocas CS, Bozas E, Chaniotakis N, Iliopoulou D, Kouris I, Varotsis K. SMARTDIAB: a communication and information technology approach for the intelligent monitoring, management and follow-up of type 1 diabetes patients. IEEE Transactions on Information Technology in Biomedicine. 2010; 14(3):622-33. crossref. PMid:20123578.
78. Fan X, He C, Cai Y, Li Y. HCloud : A Novel Application Oriented Cloud Platform for Preventive Healthcare. Proceedings of International Conference on Cloud Computing Technology and Science. 2012; p. 705-10. crossref. PMid:23184765.
79. Singapore's Health Cloud Edged out Hundreds of Global Submissions to Snag Prestigious DataCloud Enterprise Cloud Award in Monaco. Available from: crossref. Date accessed: 05/06/2015.
80. Top 10 open source data mining tools. Available from: crossref . Date accessed: 25/03/2017.
81. Ibm analytics. Available from: crossref. Date accessed: 03/2017.
82. SalfordSystens Data Mining and Predictive Analytics Software. Available from: crossref. Date accessed: 25/01/2018.
83. Rapid Miner: Data science platform. Available from: crossref. Date accessed: 17/10/2017.
84. Analytics Software. Available from: crossref. Date accessed: 16/10/2017.
85. Weka 3: Data mining software in Java. Available from: crossref. Date accessed: 09/10/2013.
86. The R Project for Statistical Computing. Available from: crossref. Date accessed: 30/07/2010.
87. Microsoft SQL Server Analysis Services. Available from: crossref. Date accessed: 21/02/2018.
88. Data Analytics with Mat lab. Available from: crossref. Date accessed: 06/10/2015.
89. Publicly available datasets. Available from: crossref. Date accessed: 24/05/2011.
90. HIV Prevention Preparedness Study. Available from: crossref. Date accessed: 14/05/2010.
91. NTDB Research Data Set. Available from: crossref. Date accessed: 01/2018.
92. Behavioral Risk Factor Surveillance System. Available from: crossref. Date accessed: 29/01/2018.