Data Mining Prediction in Healthcare System

1Anushree Ashok Wasu, 2Prof.J.S.Karnewar
Department of Computer Science and Engineering, J.C.O.E.T. Yavatmal
anushreewasu@gmail.com

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 The main purpose of data mining application in healthcare systems is to develop an automated tool for identifying and disseminating relevant healthcare information. The medical industries come across with new treatments and medicine every day. The healthcare industries should provide better diagnosis and therapy to the patients to attaining good quality of service. This paper explores different data mining techniques which are used in medicine field for good decision making. Neural Networks are one of many data mining analytical tools that can be utilized to make predictions for medical data. From the study it is observed that Hybrid Intelligent Algorithm improves the accuracy of the heart disease prediction system. In this paper, we have focused to compare a variety of techniques, application, advantage and disadvantage and different tools and its impact on the healthcare sector.**

Keywords— **Data Mining, Data Mining Task, Application, Data Mining Healthcare Technique, Advantage And Disadvantage Of Data Mining.**

I. INTRODUCTION

Data mining is the methodology for finding hidden values from enormous amount of data. As the patients population increases the medical databases also growing every day. The transactions and analysis of these medical data is complex without the computer based analysis system. Data mining is the massive areas for the doctors to handling the huge amount of patient's data sets in many ways such as make sense of complex diagnostic tests, interpreting previous results, and combining the different data together. Heart Diseases remain the biggest cause of deaths for the last two decades. Recently computer technology and machine learning techniques to develop software to assist doctors in making decision of heart disease in the early stage. In biomedical field data mining plays an essential role for prediction of diseases In biomedical diagnosis, the information provided by the patients may include redundant and interrelated symptoms and signs especially when the patients suffer from more than one type of disease of the same category. Data mining with intelligent algorithms can be used to tackle the said problem of prediction in medical dataset involving multiple inputs. Now a day's Artificial neural network has been used for complex and difficult tasks. The healthcare industry collects huge amounts of healthcare data and that need to be mined to discover hidden information for effective decision making. The risk factors for heart disease can be divided into modifiable and non modifiable. Modifiable risk factors include obesity, smoking, lack of physical activity and so on. The non modifiable risk factors for heart disease are like age, gender, and family history. Data mining algorithms applied in healthcare industry play a significant role in prediction and diagnosis of the diseases. In health care institutions leak the appropriate information systems to produce reliable reports with respect to other information in purely financial and volume related statements. Data mining tools to answer the question that traditionally was a time consuming and too complex to resolve

II. DATA MINING

Data mining is the process of combining the different data source and derives the new pattern from that data collection. Data Mining came into existence in the middle of 1990's and appeared as a powerful tool that is suitable for fetching previously unknown pat tern 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. 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. 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. The following diagram represents different stages of data mining process

III. DATA MINING TASK

Data mining tasks are mainly classified into two broad categories:

- Predictive model
- Descriptive model

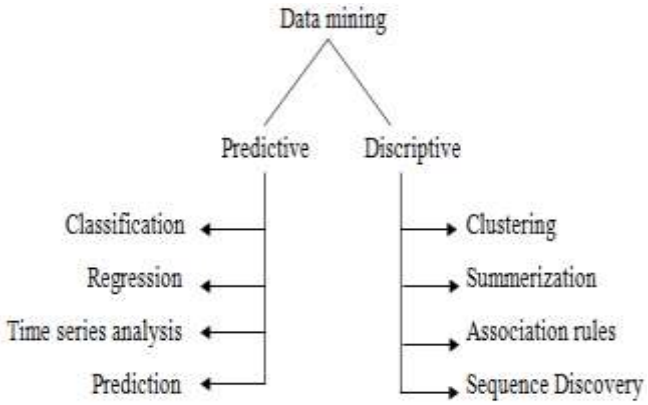

Fig. 1 Data mining models and tasks

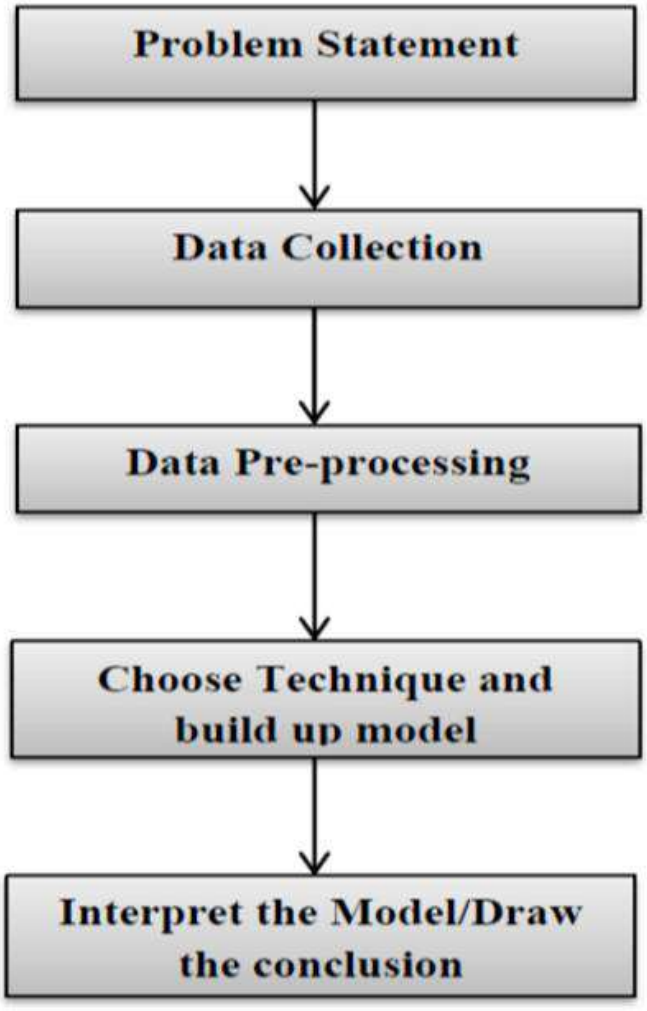

Fig. 2 Data mining process

IV. APPLICATION OF DATA MINING IN HEALTH

Data mining provides several benefits to healthcare industry. Data Mining helps the healthcare researchers to make valuable decision. Healthcare industry today generates large amounts of complex data about patients, hospital resources, disease diagnosis, electronic patient records, medical devices etc. Larger amounts of data are a key resource to be processed and analyzed for knowledge extraction that enables support for cost-savings and decision making. Data mining applications in healthcare can be grouped as the evaluation into broad categories. Following are the several applications of Data Mining in healthcare:

A. Treatment effectiveness

Data mining applications can develop to evaluate the effectiveness of medical treatments. Data mining can deliver an analysis of which course of action proves effective by comparing and contrasting causes, symptoms, and courses of treatments.

B Healthcare management

Data mining applications can be developed to better identify and track chronic disease states and high-risk patients, design appropriate interventions, and reduce the number of hospital admissions and claims to aid healthcare management. Data mining used to analyze massive volumes of data and statistics to search for patterns that might indicate an attack by bio-terrorists.

C Customer relationship management

Customer relationship management is a core approach to managing interactions between commercial organizations-typically banks and retailers-and their customers, it is no less important in a healthcare context. Customer interactions may occur through call centers, physicians' offices, billing departments, inpatient settings, and ambulatory care settings.

D Fraud and abuse

Detect fraud and abuses establish norms and then identify unusual or abnormal patterns of claims by physicians, clinics, or others attempt in data mining applications. Data mining applications fraud and abuse applications can highlight inappropriate prescriptions or referrals and fraudulent insurance and medical claims

E Medical Device Industry

Healthcare system's one important point is medical device. For best communication work this one is mostly used. Mobile communications and low-cost of wireless bio-sensors have paved the way for development of mobile healthcare applications that supply a convenient, safe and constant way of monitoring of vital signs of patients. Ubiquitous Data Stream Mining (UDM) techniques such as light weight, one-pass data stream mining algorithms can perform real-time analysis on-board small/mobile devices while considering available resources such as battery charge and available memory.

F Pharmaceutical Industry

The technology is being used to help the pharmaceutical firms manage their inventories and to develop new product and services. A deep understanding of the knowledge hidden in

the Pharmacy data is vital to a firm's competitive position and organizational decision-making

G Hospital Management

Organizations including modern hospitals are capable of generating and collecting a huge amount of data. Application of data mining to data stored in a hospital information system in which temporal behavior of global hospital activities is visualized. Three layers of hospital management:

- Services for hospital management
- Services for medical staff
- Services for patients

H System Biology

Consequently multi-relational data mining techniques are frequently applied to biological data Biological databases containing a wide variety of data types, often with rich relational. Systems biology is at least as demanding as, and perhaps more demanding than, the genomic challenge that has fired international science and gained public attention.

S.No	Type of disease	Data mining tool	Technique	Algorithm	Traditional Method	Accuracy level(%) from DM application
1	Heart Disease	ODND, SCC2	Classification	Naive	Probability	60
2	Cancer	WEKA	Classification	Rules, Decision Table		97.77
3	HIV/AIDS	WEKA 3.6	Classification, Association Rule Mining	J48	Statistics	81.8
4	Blood Bank Sector	WEKA	Classification	J48		89.9
5	Brain Cancer	K-means Clustering	Clustering	MAFIA		85
6	Tuberculosis	WEKA	Naive Bayes Classifier	KNN	Probability, Statistics	78
7	Diabetes Mellitus	ANN	Classification	C4.5 algorithm	Neural Network	82.6
8	Kidney dialysis	RST	Classification	Decision Making	Statistics	75.97
9	Dengue	SPSS Modeler		C5.0	Statistics	80
10	IVF	ANN, RST	Classification			91
11	Hepatitis C	SNP	Information Gain	Decision rule		73.20

TABLE 1. Data mining application in healthcare

V. DATA MINING TECHNIQUES

Healthcare data mining is the growing research area in data mining technology. Data mining holds great promising for healthcare management to allow health system to systematically use data and analysis to improve the care and reduce the cost concurrently could apply to as much as 30% of overall healthcare spending. In the healthcare management data mining prediction are playing active role. Some of the prediction based data mining techniques are as follows:

A. Decision tree

B. Bayesian Classifiers

C. Neural network

D. Support Vector Machine

A Decision Tree:

The decision tree is the model that consists of root node, branch and leaf node. It 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. 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 The root node is the top most nodes in the tree structure, each internal node specifies the test on attributes, the class label is hold by the leaf node, and the branch node is used to hold the test results. Decision tree is easy and fast method since it does not require any domain knowledge. . They further improved the existing decision tree model to classify different activities of patients in more accurate manner. 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 . Chang *et al.*, also used an integrated decision tree model for characterize the skin diseases in adults and children. In the decision tree inputs are divided into two or more groups repeat the steps till complete the tree as shown on Some of the decision tree algorithms as follows:

a) ID3 (Iterative Dichotomiser 3)
b) C4.5 (Successor of ID3)
c) CART (Classification & Regression Tree)
d) CHAID (CHI-squared Automatic Interaction Detector)

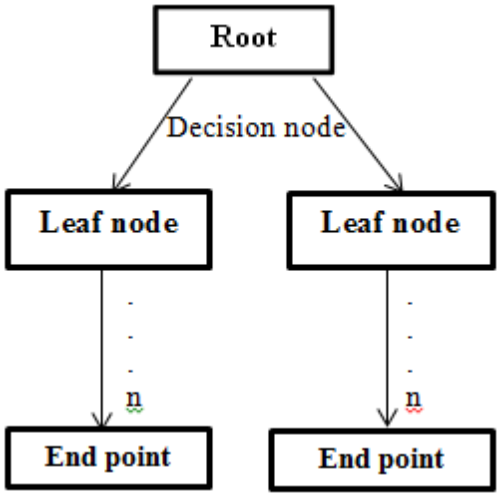

Fig. 3 Decision tree Structure

B. Bayesian Classifier:

The classification based on bayes theory is known as Bayesian classification. 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.. Bayes

theorem concentrates on prior, posterior and discrete probability distributions of data items. Figure 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. 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.

Theorem: P (B given A) = P (A and B)/P (A) to calculate probability of A given B, the algorithm counts the number of cases where A and B occurs together and divides it by the number of cases where A occurs alone. Let X be a data tuple, In Bayesian terms, X is considered "Evidence". Let H be some hypothesis, such that the data tuple X belongs to class C. P (H|X) is posterior probability, of H conditioned on X. In contract, P (H) is the prior probability of H.

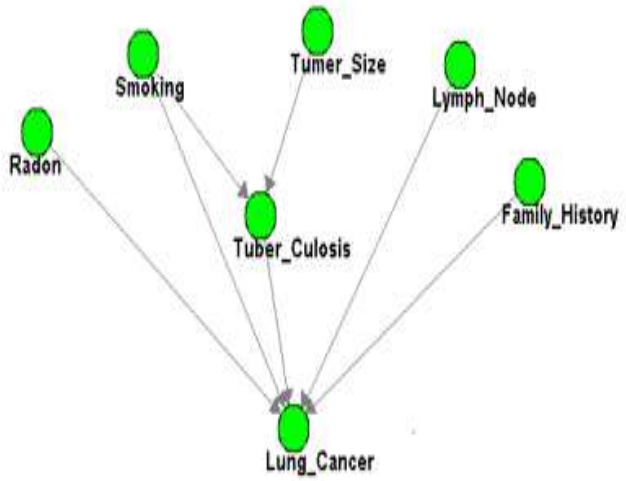

Fig 4 Bayesian Belief Network for Lung Cancer Patients

C Artificial Neural Network

Neural network is a widely used decision making technique. Since 1959 neural network are proposed for healthcare decision making. In neural network the neurons are started with random weights. Neuron doesn't know anything they have to train. 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 . To solve a particular problem NN used neurons which are organized processing elements. Neural Network is used for classification and pattern recognition. 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. 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. 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 . in the output layer rather using one neuron

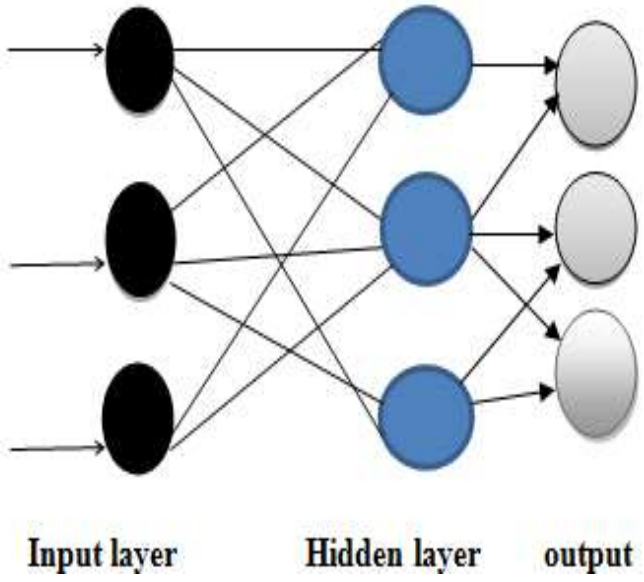

Fig.5 Neural Network

D Support Vector Machine (SVM)

Normally SVM is the classification technique. Initially it developed for binary type classification later extended to multiple classifications. This SVM creates the hyper plane on the original inputs for effective separation of data points.

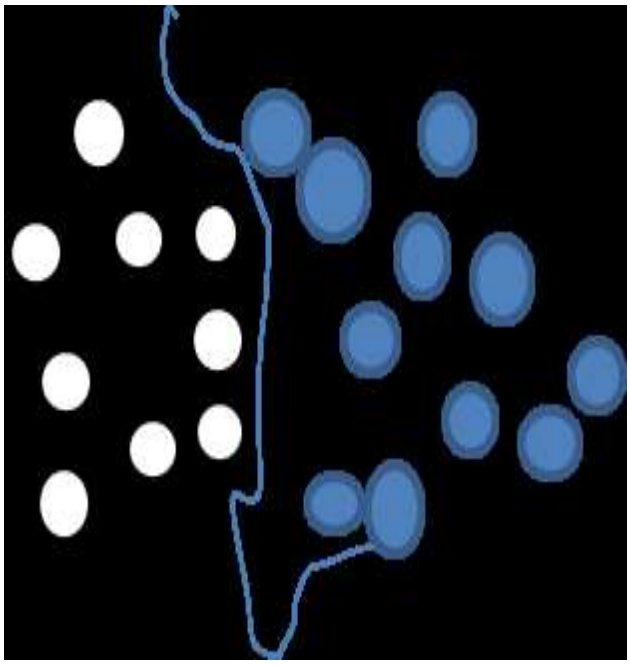

Fig. 6 Input

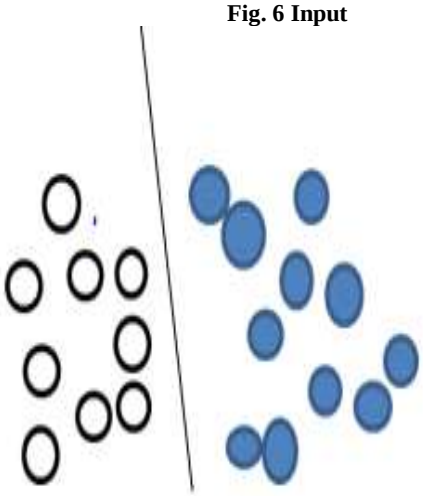

Fig. 7 Output using SVM

VI. ADVANTAGE AND DISADVANTAGE

A. Decision tree

- ***Advantage:***

1. There are no requiments of domain knowledge in the construction of decision tree.
2. It can easy process the data with high dimension.
3. It minimizes the ambiguity of complicated decision and assign exact values to outcome of varies action.
4. It is easy to interpret.
5. Decision tree also handle both numerical and categorical data.

- ***Disadvantage:***

1. It is restricted to one output attribute.
2. It generates categorical output.

It is unstable classifier i.e. performance of classifier is depend upon type of dataset.

If the type of dataset is numeric than it generates a complex decision tree.

D. Support Vector Machine

- ***Advantage:***

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.

- ***Disadvantage:***

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.

C. Neural network

- ***Advantage:***

1. Easily identify complex relationships between dependent and independent variables.
2. Able to handle noisy data

- ***Disadvantage:***

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.

D. Bayesian Classifiers

- ***Advantage:***

1. It makes computations process easier.
2. Have better speed and accuracy for huge datasets.

- ***Disadvantage:***

1. It does not give accurate results in some cases where there exists dependency among variables.

VII. CONCLUSION

This paper aimed to compare the different data mining application in the healthcare sector for extracting useful information. From the analysis it is concluded that, data mining plays a major role in heart disease classification. Neural Network with offline training is a good for disease prediction in early stage and the good performance of the system can be obtained by preprocessed and normalized dataset Exploring knowledge from the medical data is such a risk task as the data found are noisy, irrelevant and massive too. In this scenario, data mining tools come in handy in exploring of knowledge of the medical data and it is quite interesting The comparison study shows the interesting results that data mining techniques in all the health care applications give a more encouraging level of accuracy like 97.77% for cancer prediction and around 70 % for estimating the success rate of IVF treatment.

REFERENCES

[1] Neesha Jothi ,,Wahidah Husain, Nur Aini Abdul Rashid, *"Data Mining healthcare : A Review*

[2] Ravi Sanakal, Smt. T Jayakumari*," Prognosis of Diabetes Using Data mining Approach-Fuzzy C Means Clustering and Support Vector Machine*",International Journal of Computer Trends and Technology (IJCTT) – volume 11 number 2 – May 2014.

[3] V. Krishnaiah et al," *Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques*", International Journal of Computer Science and Information Technologies, Vol. 4 (1) , 2013, 39 – 45.

[4] Muhamad Hariz Muhamad Adnan,Wahidah Husain, Nur'Aini Abdul Rashid, *"Data Mining for Medical Systems: A Review*"

[5]. ShwetaKharya, ―Using Data Mining Techniques ForDiagnosis And Prognosis Of Cancer Disease‖, International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.2, April 2012. [6] N. Deepika and K. Chandra shekar, "Association rule for classification of Heart Attack Patients", *International Journal of Advanced Engineering Science and Technologies*, Vol. 11, No. 2, pp. 253 – 257, 2011.

[7]. K. Srinivas , B. Kavitha Rani and Dr. A. Govrdhan, ―Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks‖ International Journal on Computer Science and Engineering (2010). .

[8] Abdelghani Bellaachia, Erhan Guven, "Predicting Breast Cancer Survivability Using Data Mining Techniques".

