# A Survey on Role of Data Mining Techniques in Healthcare Domain

**Prateek Pandey[1]* Dr. P. K. Rai[2]**

[1] Research Scholar, APS University, Rewa (MP)

[2] Head, Department of Computer Centre, APS University, Rewa (MP)

***Abstract – In today's world application of data mining in healthcare is growing faster as the health sector is rich in data and data mining has become a necessity. Healthcare organizations generate and collect large volumes of information to a daily basis. The healthcare environment is generally perceived as being 'information rich' however 'knowledge poor'. Automated data mining and knowledge discovery techniques aid in fetching some interesting patterns.***

***Data mining can enable healthcare organizations to anticipate trends in the patient's medical condition and behaviour proved by analysis of prospects different and by making connections between seemingly unrelated information. The raw data from healthcare organizations are voluminous and heterogeneous. It needs to be collected and stored in organized form and their integration allows the formation unite medical information system.***

***Data mining in healthcare domain offers boundless opportunities for analyzing databases to discover patterns that are hidden in traditional data analysis techniques. These patterns can be used by healthcare practitioners to make predictions, perform diagnoses and set treatments for patients.***

***In this survey paper we intend to study some of the data mining techniques such as classification, clustering, association, regression in healthcare field.***

***Keywords: Healthcare data mining, Weka, Predictive Analytics, KDD***

----------X----------

## 1. INTRODUCTION

The purpose of data mining application in a particular field is to find the useful and hidden knowledge from the available dataset. This paper mainly deliberates the Data Mining applications in the healthcare sector. In this work, a detailed survey is carried out on data mining applications in the healthcare domain, types of data used and facts of the information mined. Data mining algorithms applied in healthcare industry has shown a substantial role in prediction and diagnosis of the diseases. There are a large number of data mining applications in healthcare areas such as Medical device industry, Pharmaceutical Industry and Hospital Management. Popularly data mining is used as a step in knowledge discovery from the data.

The knowledge discovery is an interactive process, consisting by developing an understanding of the application domain, selecting and creating a data set, preprocessing, data transformation. Data Mining has been used in a variety of applications such as marketing, customer relationship management, engineering, and medicine analysis, expert prediction, web mining and mobile and mobile computing.

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. They prepare databases for finding predictive information.

Data mining tasks are Association Rule, Patterns, Classification and Prediction, Clustering. Most common modeling objectives are classification and prediction. The reason that attracted a great deal of attention in information technology for the discovery of useful information from large collections is due to the perception that we are data rich but information poor. Some of the data mining applications are:

- Developing models to detect fraudulent phone or credit-card activity
- Predicting good and poor sales prospectus.
- Predicting whether a heart attack is likely to recur among those with cardiac disease.
- Identifying factors that lead to defects in a manufacturing process.

Expanding the health coverage to as many people as possible, and providing financial assistance to help those with lower incomes purchase coverage. Eliminating current health disparities would decrease the costs associated with the increased disease burden borne by certain population groups. Health administration or healthcare administration is the field relating to leadership, management, and administration of hospitals, hospital networks, and health care systems. In the Healthcare sector Government spends more money.

Data mining is an algorithmic techniques for extracting new useful patterns from collected raw data .Today, huge amount of data is produced by healthcare industries. The data produced includes data about hospitals, resources, electronic patient records, disease diagnosis, etc. Data mining pro-cases includes the hypothesis framing, data gathering, pre-processing, model estimation, model understanding and then finding the conclusions. Healthcare industries have data that clutches complex information about the patients and their medical conditions. Data mining is becoming popular in various research arenas due to its wide variety of applications and system of methods used to mine information that are useful in correct manner.

Data mining techniques have the ability to detect hidden patterns or relationships among the objects in medical records. Data mining techniques have been applied on medical data during the last few decades. This is for determining useful trends or patterns that are used in analysis and making proper decisions. The infinite potential of data mining is utilized for predicting the different kind of diseases more efficiently and effectually in health care data. Some data mining methods used in medical field includes Association, Clustering, and Classification.

Knowledge discovery in databases is well-defined process consisting of several distinct steps. Data mining is the core step, which results in the discovery of hidden but useful knowledge from massive databases. A formal definition of Knowledge discovery in databases is given as follows: “Data mining is the non-trivial extraction of implicit previously unknown and potentially useful information about data”. Data mining technology provides a user-oriented approach to novel and hidden patterns in the data. The discovered knowledge can be used by the healthcare administrators to improve the quality of service [1]. The discovered knowledge can also be used by the medical practitioners to reduce the number of adverse drug effect, to suggest less expensive therapeutically equivalent alternatives. Anticipating patient’s future behavior on the given history is one of the important applications of data mining techniques that can be used in health care management.

The availability of integrated information via the huge patient repositories, there is a shift in the perception of clinicians, patients and payers from qualitative visualization of clinical data by demanding a more quantitative assessment of information with the supporting of all clinical and imaging data. For instance it might now be possible for the physicians to compare diagnostic information of various patients with identical conditions. Likewise, physicians can also confirm their findings with the conformity of other physicians dealing with an identical case from all over the world. Medical diagnosis is considered as a significant yet intricate task that needs to be carried out precisely and efficiently. The automation of the same would be highly beneficial [2].

## 2. LITERATURE REVIEW

A literature review is a text written by critical points of current knowledge including substantive find theoretical and methodological contributions to a particular topic. Literature reviews are secondary sources and do not report any new or original experimental work.

**Hian Chye Koh and Gerald Tan** mainly discusses data mining and its applications with major areas like Treatment effectiveness, Management of healthcare, Detection of fraud and abuse, Customer relationship management[3].

**Jayanthi Ranjan** presents how data mining discovers and extracts useful patterns of this large data to find observable patterns. This paper demonstrates the ability of Data mining in improving the quality of the decision making process in pharma industry. Issues in the pharma industry are adverse reactions to the drugs [4].

**M. Durairaj, K. Meena** illustrates a hybrid prediction system consists of Rough Set Theory (RST) and Artificial Neural Network (ANN) for dispensation medical data. The proposed hybrid tool incorporates RST and ANN to make proficient data analysis and indicative predictions. The projected hybrid prediction system is applied for pre-processing of medical database and to train the ANN for production prediction. The prediction accuracy is observed by comparing observed and predicted cleavage rate [5].

**K. Srinivas, B. Kavitha Rani and Dr. A. Goverdhan** mainly examine the potential use of classification based data mining techniques such as Rule Based, Decision tree, Naïve Bayes and Artificial Neural Network to the massive volume of healthcare data. Using an age, sex, blood pressure and blood sugar medical profiles it can predict the likelihood of patients getting a heart disease [6].

**Shweta Kharya** discussed various data mining approaches that have been utilized for breast cancer diagnosis and prognosis Decision tree is found to be the best predictor with 93.62% Accuracy on benchmark dataset and also on SEER data set [7].

**Elias Lemuye** discussed abot that AIDS is a disease that is caused by HIV, which weakens the body's immune system until it can no longer fight off the simple infections that most healthy people's immune system can resist. Apriori algorithm is used to discover association rules. WEKA 3.6 is used as the data mining tool to implement the Algorithms. The J48 classifier performs classification with 81.8% accuracy in predicting the HIV status [8].

**Arvind Sharma and P.C. Gupta** argued that data mining can contribute with important benefits to the blood bank sector. J48 algorithm and WEKA tool have been used for the complete research work. Classification rules performed well in the classification of blood donors, whose accuracy rate reached 89.9% [9].

## 3. DATAMINING TECHNIQUES IN HEALTH CARE SECTOR

Data mining techniques are very useful in healthcare; these techniques are helpful in diagnosis and treatment of diseases, resource management in healthcare, fraud detection, customer relationship management etc. Two strategies are used in data mining i.e. supervised learning and unsupervised learning. In supervised learning a training set is there with the help of which model parameters are learned [10].

On other hand there is absence of training set in unsupervised learning, no training set is present therefore learning is modeled with unknown target parameter. The models are in descriptive form which describes the interesting and valuable information present in data [11].

Descriptive and predictive are the two categories in which data mining tasks are classified. The goal of descriptive tasks is to review the data and construction of entire model and find out human interpreted forms and associations. While in predictive task the goal is focused to find out the interesting outcomes. Also it find out there is any relationship present between dependent and independent variables [12].

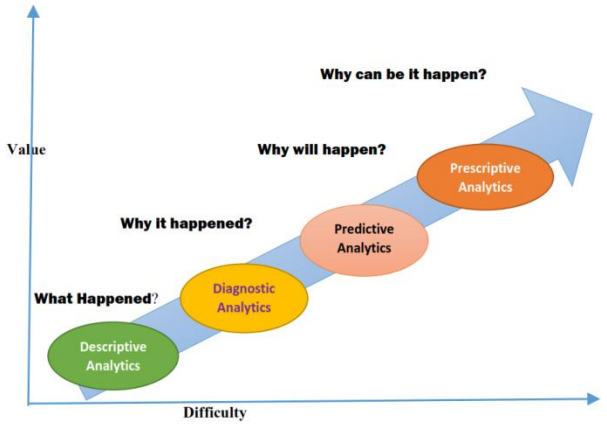


**Figure 1: Steps in the advanced analytics**

There are four main categories when it comes to data analytics: descriptive, diagnostic, predictive and prescriptive as shown in figure 1. Every category is distinct in the value it offers and in how it could be used in business to advance productivity and revenue. It is crucial to understand each category and know the right timing for using a category.

**Predictive analytics:** In predictive analytics, we need to make sense of why certain things happened and then build a model to project what could happen in the future, hence the name.

**Descriptive Analytics:** As the name suggests, descriptive analytics are more about summarizing and reporting data. This type of data analytics is geared towards what is currently happening or what has already happened.

**Diagnostic Analytics:** In contrast to descriptive analytics, diagnostic analytics is less focused on what has occurred but rather focused on why something happened. In general, these analytics are looking on the processes and causes, instead of the result.

**Prescriptive data analytics:** It is a final step of an advanced data analysis that consists of the application of the predictive model to determine the best solution or outcome among various choices, given the known parameters. In this phase, not only is predicted what will happen in the future using our predictive model, but also is shown to the decision maker the implications of each option.

All these advanced data analytics approaches are based on different data science algorithms that can be classified in the following categories: classification, clustering, regression, and forecasting algorithms. It can be noticed that there are several programming languages, which implement these methods as packages or libraries.

Some Data mining techniques and tools that are being widely used in health care sectors are discussed below:

- **Classification** –It is the task of generalizing a well-known structure to apply to new data. It is the process of identification of a set of categories on the basis of a training set of data containing observations whose category of membership is known. For example, some of e-mails are classified as “legitimate" and other are classified as "spam" on the basis of content present or on the basis of some other characteristic.

- **Clustering** –Cluster analysis or clustering is the task of alliance more similar objects in same group (known as cluster). It is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.

- **Association rule learning** – It is an admired and well researched technique for discovering attention grabbing relations between variables in large databases. It searches for relationships between variables.

- **Regression** – It is the process to find a function which models the data with the least error. It includes various techniques for modeling and analyzing several variables. The task mainly spotlight on the correlation between a dependent variable and one or more independent variables.

- **Anomaly detection** – Also known as Outliner detection or Deviation detection. The task involves the discovery of unusual data records, measures or annotations that might be exciting or data errors that require further exploration.

- **Summarization –**Automatic summarization is the process of reducing a text document using a computer program in order to generate a summary that retains the most significant points of the original document. It provides a more dense demonstration of the data set which includes visualization and report generation. The interest in automatic summarization has increased nowadays due to the increase in information overload and the quality of data has increased.

- **Time Series Analysis**- Time series analyses consist of methods for analyzing time series data consecutively to dig out meaningful statistics and other characteristics of the data. Time series forecasting is to utilize a model to forecast upcoming values based on formerly observed values. Data of Time series includes natural temporal ordering. Additionally, time series models will frequently take advantage of the usual one-way ordering of time so that values for a known period will be expressed as deriving in several ways from past values, rather than from future values. Time series analysis can be applied to continuous data, real-valued, discrete numeric data, or discrete symbolic data.

- **The prediction task-** It is a supervised learning task which works on direct data and there is no explicit model for new instance of class value prediction. Some of the approaches which are used for prediction task are:

  ► Instance-based (nearest neighbor)

  ► Statistical (naive bayes)

  ► Bayesian networks

  ► Regression (a kind of concept learning for continuous class)

- **Sequence Discovery**- Also known as Sequential Pattern mining is a topic of data mining concerned with discovery of statistically significant patterns among data examples where the values are conveyed in a sequence. It is usually assumed that the values are discrete. Sequential pattern mining is a special case of structured data mining. Each and every technique has its own importance. All of these tasks can be efficiently used in healthcare field. Many of the researchers are currently working on these techniques for various purposes [11].

## 4. DATABASES AND DATA ANALYSIS TOOLS

Healthcare databases are systems into which healthcare providers routinely enter clinical and laboratory data. One of the most commonly used forms of healthcare databases are:

- **Electronic health records (EHRs)**

Practitioners enter routine clinical and laboratory data into EHRs during usual practice as a record of the patient’s care.

- **Global Health Observatory (GHO)**

WHO's portal providing access to data and analyses for monitoring the global health situation. Provides critical data and analyses for over 30 health themes ranging from health systems to

disease-specific themes, as well as direct access to the full database.

- **Global Health Estimates (GHE)**

The GHE provide a comprehensive and comparable assessment of mortality and loss of health due to diseases, injuries and risk factors. Global, regional and country estimates for all-cause mortality, and deaths and disability-adjusted life years (DALYs) by age, sex and cause, are available for download.

Essentially there are many data mining and analysis tools available to use. A list with brief introduction of the most widely used ones is given here.

- **WEKA:** It is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

- **Rapid Miner:** It is a software platform developed by the company of the same name that provides an integrated environment for machine learning, data mining, text processing and data analytics.

- **R analytics suite:** It is a powerful and productive user interface for R. It's free and open source, and works great on Windows, Mac, and Linux. Actually R is the top ranked data analytics language.

- **Python scikit learn:** It is an open source Python library that implements a range of machine learning, preprocessing, cross-validation and visualization algorithms

## 5. CONCLUSION

Data mining has shown a significance role in the area of health, as it support the medical experts to find out novel lifesaving facts. Knowledge obtained due to use of data mining techniques can further be utilized to make effective decisions in order to improve success rate of healthcare organization as well as health of the patients.

Data mining requires appropriate technology and analytical techniques, as well as systems for reporting and tracking which can enable measuring of results. Healthcare institutions that use data mining applications have the possibility to predict future requests, needs, desires, and conditions of the patients and to make adequate and optimal decisions about their treatments.

As there is huge records in the healthcare industry and from this, it has become obligatory to use data mining techniques to help in decision support and prediction in the field of healthcare to identify the kind of disease. The medical data mining produces business intelligence which is useful for diagnosing of the disease. This survey paper observe some data mining techniques and tools that have been employed for medical data for several diseases which are recognized and diagnosed in patients.

## REFERENCES

[1]. Frawley and Piatetsky-Shapiro (1996). Knowledge Discovery in Databases:An Overview. The AAAI/MIT Press, Menlo Park, C.A.

[2]. Boris Milovic, Milan Milovic (2012). “Prediction and Decision Making in Health Care using Data Mining”, International Journal of Public Health Science, Vol. 1, No. 2, December 2012, pp. 69-78

[3]. Hian Chye Koh and Gerald Tan, ―Data Mining Applications in Healthcare‖, journal of Healthcare Information Management – Vol 19, No 2.

[4]. Jayanthi Ranjan (2007). Applications of data mining techniques in pharmaceutical industry‖, Journal of Theoretical and Applied Technology.

[5]. M. Durairaj, K. Meena (2011). A Hybrid Prediction System Using Rough Sets and Artificial Neural Networks‖, International Journal Of Innovative Technology & Creative Engineering (ISSN: 2045-8711) VOL.1 NO.7 JULY 2011.

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

[7]. Shweta Kharya (2012). 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.

[8]. EliasLemuye, ―Hiv Status Predictive Modeling Using Data Mining Technology‖.

[9]. Arvind Sharma and P.C. Gupta (2012). Predicting the Number of Blood Donors through their Age and Blood Group by using Data Mining Tool‖ International Journal of Communication and Computer Technologies Volume 01 – No.6, Issue: 02 September 2012.

[10]. Prasanna Desikan, Kuo-Wei Hsu, Jaideep Srivastava (2011). Data Mining For Healthcare Management‖, 2011SIAM International Conference on Data Mining, April, 2011.

[11]. Arun K Punjari (2006). Data Mining Techniques‖, Universities (India) Press Private Limited.

[12]. Margaret H. Dunham (2005). Data Mining Introductory and Advanced Topics‖, Pearson Education (Singapore) Pte. Ltd., India.

[13]. Ruban D. Canlas Jr., MSIT., MBA , — Data mining in Healthcare: Current applications and issues‖.

---

**Corresponding Author**

**Prateek Pandey***

Research Scholar, APS University, Rewa (MP)