International Journal of Advance Research in Science and Engineering
Volume No.06, Special Issue No.(01), December 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354

# A Study on Detection and Prevention on Infectious disease using Data Mining Techniques

**Zubair Ahmad[1], Sheeraz Ahmad Peerzada[2], Dr.Jitendra Seethalni[3], Gh.Hassan[4]**

[1,2,3,4]*Department of Computer Science,*
*Sri Satya Sai University of Technology & Medical Science,*
*Sehore, MP, (INDIA)*

**ABSTRACT**

*Data mining techniques plays an imperative role in the field of biomedical sector and they have turned into the subfield of medical research. This data mining techniques is utilized for the detection and anticipation of both kind of communicable and non-communicable disease. In communicable disease that is transmitted through direct contact with a tainted individual or by implication through a vector. Likewise called contagious disease example ordinary cold, gastroenteritis, hepatitis and so forth while non-communicable disease are of long span and by and large moderate movement. The four principle sorts of non-communicable infections are cardiovascular disease (like heart attack and stroke), tumor, constant respiratory ailments, (for example, perpetual discouraged pulmonary disease and asthma) and diabetes. The reason for this development is that these days the vast majority of the well being related information are put away in little datasets scattered through different Hospitals, Clinics and Research centers. The health care sector is by and by confronting both the financial consistency and the methodological prospect of an information based approach for quality organization. Thus, data mining techniques has been proposed to process medical data stream. Applying data mining procedures to the incorporated database will offer to specialists investigative and anticipating devices from the surface of the information. In spite of the fact that data mining strategies and instruments have been connected in different areas Therefore, this paper presents a survey on the significance of Data Mining methods in detection and mitigation of communicable disease.*

## I.INTRODUCTION

Infectious disease ranks among the gravest threats to human health, alongside global warming and terrorism. New disease strains are naturally continuously emerging and resulting outbreaks can spread rapidly, with the potential to cause enormous losses to health and livelihood. Worldwide, however, many infections remain undetected. Due to poor diagnostic tools, infection is often undiagnosed and therefore untreated, or diagnosed at a late stage when treatment is less effective. This results in on-going transmission of serious infections (for example common cold and hepatitis) and delays in the identification of emerging threats (for example pandemic influenza), which may lead to major human and economic consequences for millions of people.

The best strategic approach to control any outbreak is to identify the source of infection at an early stage and halt its spread, or prevent the outbreak altogether. To do this, we rely on very early detection. Our research into early warning sensing systems for infectious diseases focuses on communicable disease - though data mining techniques could potentially be applied to other disease areas.

A range of factors is responsible for the (re-)emergence of infectious disease threats, including antimicrobial resistance, altering the epidemiology and spread of disease in a changing global environment. These include drivers such as climate change and associated environmental impacts, population growth, unplanned urbanisation and high mobility, as well as animal husbandry or intensive farming practices.

It is expected that proposals develop: the technology to allow the pooling, access, analysis and sharing of relevant data, including next generation sequencing; the innovative bio-informatics and modelling methodologies that enable risk modelling and mapping; and the analytical tools for early warning, risk assessment and monitoring of (re-)emerging infectious disease threats.

Data mining has become important to the healthcare world. On the one hand, EHR offers the data that gets data miners excited. However, on the other hand, it is accompanied by challenges such as 1) the unavailability of large sources of data to academic researchers, and 2) limited access to data-mining experts [1]. But, still the quality of health care service at a cheap cost continues to be a difficult issue in developing countries. There are various data mining techniques which is used for extracting an essential information among the large set of patient dataset such as clustering, classification, decision tree etc. In this paper we present the study for the detection and diagnosis of infectious disease using several data mining technique and applications of data mining in the various field of health sector. Here fig. 1 shows the steps involved in the data mining process.


**Fig.1 steps involved in data mining process**

## II. DATA MINING TECHNIQUES

Data mining technique is defined as the process of discovering interesting patterns and knowledge from the large amounts of data. Technique refers to analyzing data from different viewpoints and abstracting it to get the necessary information. DM technique provides an important means for extracting valuable medical rules hidden in medical data and acts as an important role in clinical diagnosis.

### 2.1 Classification

Classification is the process of predicting output based on some given input data. The goal of classification is to accurately predict the target class for each case in the data. In order to predict the data, it processes the training set and predictive set. It first develop relationships between the attributes of training data set .Then it is provided with the predictive data set, which contains similar attributes but with different data values, Then it analyze the given data and produce prediction by placing the different data sets in different classes based on the relationship of attributes [2][3]. Classification uses predictive rules expressed in the form of IF-THEN rules where the first part (IF part) consist of conjunction of conditions and the second part(THEN part) predict a certain prediction attribute value that satisfy the first part

#### 2.1.1 Decision Tree

Decision tree is similar to flow chart in which every non-leaf node denote a test on a particular attribute and every branch represent a outcome of the test. Root node is the topmost node in the decision tree. For example, with the help of readmission tree, we can decide whether a patient needs to be readmitted or not. Using Decision Tree, a decision maker can choose best alternative and traversal from root to leaf indicates unique class separation based on maximum information gain [4]. Decision tree are self explanatory and easy to follow. Set of rules can also be constructed with the help of decision tree. Decision Tree can be considered as nonparametric method because there is no need to make assumptions regarding distribution of space and structure of classifier. Decision tree have several disadvantages. These are: Most of the algorithm like ID# and C4.5 require target attributes to have discrete values as decision tree use divide and conquer strategy. More the complex relationship among attributes lesser is the performance.

#### 2.1.2 Support Vector Machines

Vladimir Vapnik first introduced idea of Support Vector Machine [5]. Its accuracy is better than all other available techniques. It was first introduced for binary classification problems; but it can be further extended to multi class problems. It creates hyper-planes to separate data points.

It can be implemented in 2 ways:

1. Mathematical programming
2. Using kernel functions

With the help of training data sets, non linear functions can be easily mapped to high dimensional space. This can only be possible using kernel functions like Gaussian, sigmoid etc.

**2.1.3 Artificial Neural Network**

Artificial neural networks models have been studied for many years in the hope of achieving human like performance in several fields. In Neural Networks, basic elements are neurons or nodes. These neurons are interconnected and within the network they worked together in parallel in order to produce the output functions. From existing observations they are capable to produce new observations even in those situations where some neurons or nodes within the network fails or go down due to their capability of working in parallel. An activation number is associated to each neuron and a weight is assigned to each edge within a neural network. In order to perform the tasks of classification and pattern recognition neural network is mainly used [6]. ANN is based on the biological neural networks in the human brain and described as a connectionist model.

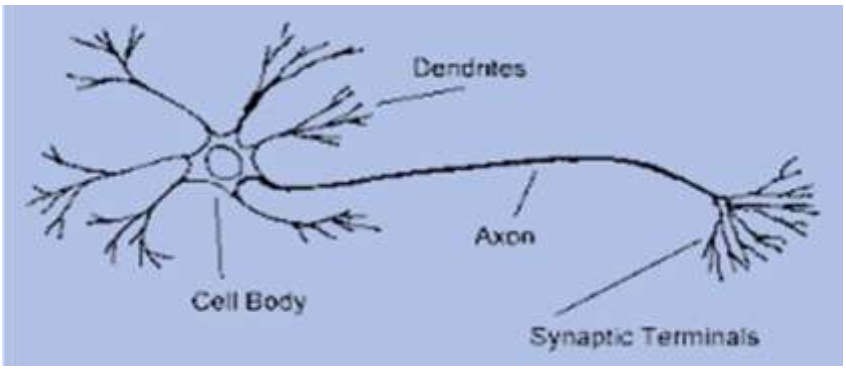



**Fig.2 Biological Neuron**

It is based on the neuron, a cell that processes information in the human brain [7]. The neuron cell body contains the nucleus, and has two types of branches, the axon and the dendrites. The axon transmits signals or impulses to other neurons while the dendrites receive incoming signals or impulses from other neurons. Every neuron is connected and communicates through the short trains of pulses [7]. The nodes are the artificial neuron and the directed edges represented the connection between output neurons and the input neurons. In training phase, the internal weights of the neural network are adjusted according to the transactions used in the learning process. For each training transaction the neural network receives in addition the expected output. This allows modification of weight.

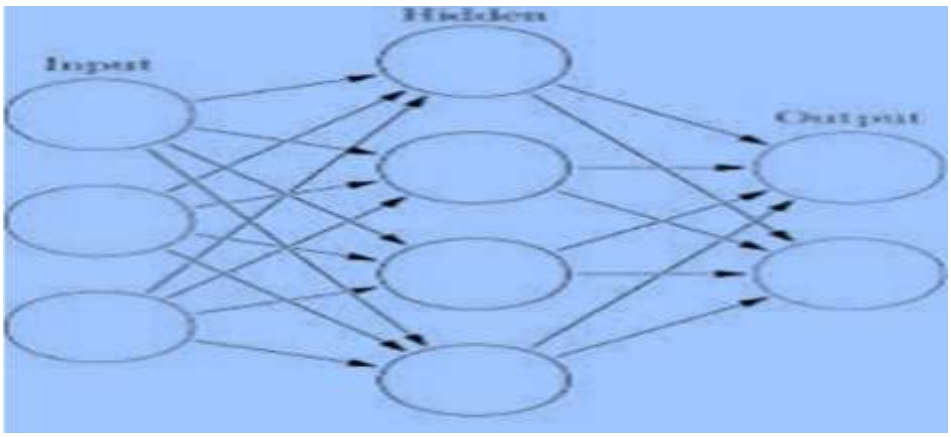


**Fig.3 Artificial Neural Networks**

### 2.2 ARM (Association Rule Mining)

The aim of ARM is to identify the useful rules from the large amounts of data. Association rule mining has following logical process and appeal.

- Logical process: Interesting rules are determined in terms of support and confidence.

1) Support: It reflects the usefulness of discovered rules.

2) Confidence: It reflects certainty of discovered rules.

- Appeal: Association rules are considered as interesting if they satisfy both minimum support threshold and minimum confidence threshold.

These thresholds set by users or domain experts. Additional analysis can be performed to discover interesting statistical correlations between associated items.

### 2.3 Probabilistic Learning Method (Bayesian Classifier)

For probabilistic learning method Bayesian classification is used. Bayes theorem of statistics plays a very important role in it. While in medical domain attributes such as patient symptoms and their health state are correlated with each other but Naïve Bayes Classifier assumes that all attributes are independent with each other. This is the major disadvantage with Naïve Bayes Classifier. If attributes are independent with each other then Naïve Bayesian classifier has shown great performance in terms of accuracy. There are two types of probabilities

- Posterior Probability [P(H/X)]
- Prior Probability [P(H)]

Where X is data tuple and H is some hypothesis. According to Bayes' Theorem,

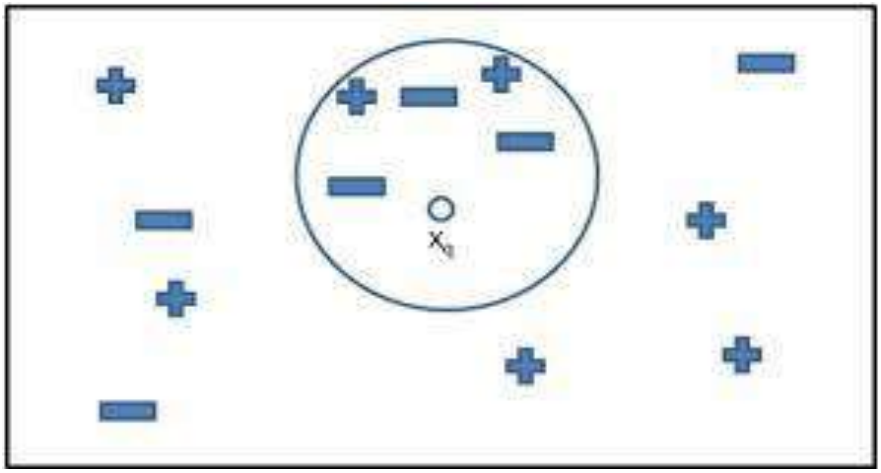
$$P(H/X)= P(X/H)P(H) / P(X)$$

Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Maximum-likelihood training can be done by evaluating a closed-form expression, which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers.

### 2.4 k-Nearest Neighbors

In pattern recognition, the k-Nearest Neighbors algorithm (or k-NN for short) is a non – parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature spaces. The output depends on whether k-NN is used for classification or regression: In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor. K-NN has a number of applications in different areas such as health datasets, image field, cluster analysis, pattern recognition, online marketing etc. There are various advantages of KNN classifiers.[8] These are: ease, efficacy, intuitiveness and

competitive classification performance in many domains. If the training data is large then it is effective and it is robust to noisy training data. A main disadvantage of KNN classifiers is the large memory requirement needed to store the whole sample. If there is a big sample then its response time on a sequential computer will also large.[9]



### 2.5 Apriori Algorithm

The objective of Apriori is to find the frequent-item sets. Apriori has following logical process and appeal.

- Logical Process: It employs an iterative approach known as level-wise search, where k-item sets are used to explore (k+1) item sets. First, the set of frequent 1-itemsets is found by scanning the database to accumulate the count for each item, and collecting those items that satisfy minimum support. The resulting set is denoted by L1. Next, L1 is used to find L2, the set of frequent 2-itemsets, which is used to find L3, and so on, until no more frequent k-item sets can be found. The finding of each Lk requires one full scan of the database.
- Appeal: The Apriori procedure performs two kinds of actions, namely, join and prune, as described before. In the join component, Lk-1 is joined withLk-1 to generate potential candidates. The prune component employs the Apriori property (all nonempty subsets of a frequent item set must also be frequent) to remove candidates that have a subset that is not frequent. Finally, all the candidates satisfying the minimum support from the set of frequent item sets is considered.

## III. APPLICATION OF DATA MINING IN HEALTH SECTOR

Business and marketing organizations may be ahead of healthcare in applying data mining to derive knowledge from data. This is quickly changing. Successful mining applications have been implemented in the healthcare arena, three of which are described below.

### 3.1 Hospital Infection Control

Nosocomial infections affect 2 million patients each year in the United States, and the number of drug-resistant infections has reached unprecedented levels.[10] Early recognition of outbreaks and emerging resistance

requires proactive surveillance. Computer-assisted surveillance research has focused on identifying high-risk patients, expert systems, and possible cases and detecting deviations in the occurrence of predefined events. A surveillance system that uses data mining techniques to identify new and interesting patterns in infection control data has been implemented at the University of Alabama.[11] The system uses association rules on culture and patient care data obtained from the laboratory information management systems and generates monthly patterns that are reviewed by an expert in infection control. Developers of the system conclude enhancing infection control with the data mining system is more sensitive than traditional infection control surveillance, and significantly more specific.

### 3.2 Ranking Hospitals

Organizations rank hospitals and healthcare plans based on information reported by healthcare providers. There is an assumption of uniform reporting, but research shows room for improvement in uniformity. Data mining techniques have been implemented to examine reporting practices. With the use of International Classification of Diseases, 9th revision, codes (risk factors) and by reconstructing patient profiles, cluster and association analyses can show how risk factors are reported.[12]

Standardized reporting is important because hospitals that underreport risk factors will have lower predications for patient mortality. Even if their success rates are equal to those of other hospitals, their ranking will be lower because they reported a greater difference between predicted and actual mortality.[12] Standardized reporting would also be important for meaningful comparisons across hospitals.

### 3.3 Identifying High-Risk Patients

American Heathway's provides diabetes disease management services to hospitals and health plans designed to enhance the quality and lower the cost of treatment of individuals with diabetes. To augment the company's ability to prospectively identify high-risk patients, American Heathway's uses predictive modeling technology. [13] Extensive patient information is combined and explored to predict the likelihood of short-term health problems and intervene proactively for better short-term and long-term results. A robust data mining and model-building solution identifies patients who are trending toward a high-risk condition. This information gives nurse care coordinators a head start in identifying high-risk patients so that steps can be taken to improve the patients' quality of healthcare and to prevent health problems in the future.

### 3.4 Treatment effectiveness

Data mining applications can be developed to evaluate the effectiveness of medical treatments. By comparing and contrasting causes, symptoms, and courses of treatments, data mining can deliver an analysis of which courses of action prove effective. [14] For example, the outcomes of patient groups treated with different drug regimens for the same disease or condition can be compared to determine which treatments work best and are most cost-effective.[15]

### 3.5 Healthcare management

To aid 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.

### 3.6 Customer relationship management

While customer relationship management is a core approach in 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.

## IV.CONCLUSION

Data mining is the widely used techniques for mining the essential information from large medical dataset. This technique can be used for both communicable and non-communicable disease detection and prevention. In this paper we present study of some data mining techniques and its application in health sector for early prediction and diagnosis of communicable disease. In future need to develop automated surveillance systems which offer advantages over manual ones. When analytical technologies are embedded in automated hospital infection surveillance systems, it is not clear whether data mining outperforms traditional statistical methods.

## REFERENCES

[1] Division of Health Care Statistics, "NCHS Health E-Stat Report", National Center for Health Statistics of US 2011.

[2] S. Palaniappan, and Rafiah Awang "Intelligent heart disease prediction system using data mining techniques." IEEE Conference on Computer Systems and Applications, 2008.

[3] K. B. Srinivas, Kavihta Rani, and A. Govrdhan "Applications of data mining techniques in healthcare and prediction of heart attacks." International Journal on Computer Science and Engineering (IJCSE) Vol. 2, No. 02, 2010, pp. 250-255.

[4] Parvez Ahmed, Saqib Qamar, and Syed Qasim Afser Rizvi. "Techniques of Data Mining In Healthcare: A Review." International Journal of Computer Applications Vol. 120, No.15, 2015.

[5] Vladimir Vapnik. "The support vector method of function estimation."Non-linear Modeling. Springer US, 1998, pp.55-85.

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

[7] A. K. Jain "Artificial neural network : a tutorial[Online].

[8] Bramer, M "Principles of data mining", Springer 2007

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

[10] Gaynes R, Richards C, Edwards J, et al. "Feeding back surveillance data to prevent hospital-acquired infections". Emerg Infect Dis 2001;7:295-298.

[11] Brosette SE, Spragre AP, Jones WT, Moser SA "A data mining system for infection control surveillance", Methods Inf Med 2000; 39:303-310.

[12] Cerrito P. "Using text analysis to examine ICD-9 codes to determine uniformity in the reporting of MedPAR data" Presented at the Annual Symposium of the American Medical Informatics Association; November 9-13, 2002; San Antonio, TX.

[13] Ridinger M. "American Healthways uses SAS to improve patient care. DM Review" 2002;12:139.

[14] Kincade, K. Data mining: digging for healthcare gold. Insurance & Technology,1998 23(2), IM2-IM7.

[15] Milley, A. "Healthcare and data mining" Health Management Technology, (2000). 21(8), 44-47.