# Improving Indian Healthcare using Data Mining

Vivek Sundar Magesh
TVS Logistics Services Limited
Chennai, India
viveksundarm@gmail.com

Harinivas Bhaskaran
State Bank of India
Chennai, India

Beri Pradeep
Murugappa Group
Chennai, India

Sadhanand Lakshmi Narayana Moorthy
BPCL
Mangalore, India

Thomas Greefin Franco
HCL
Chennai, India

Nishant Kumar
Dun & Bradstreet
India

***Abstract*— Data mining provides automatic pattern recognition and uncovers patterns in data that are difficult to detect with traditional statistical methods. For developing countries with limited resources, data mining can be an answer towards improving the overall health of the population at an affordable cost.**

**In this paper, we initially analyze how data mining can transform various aspects of healthcare in general such as in evaluating treatment effectiveness, in healthcare management by identifying high risk patients, in pharmaceutical industry etc. We then focus on two specific diseases viz. Tuberculosis (an infectious disease) and Cardiovascular disease (lifestyle disease) and analyze how data mining can equip public systems better to tackle them.**

**In the case of Tuberculosis, given a set of variables such as Gender, Loss of Appetite, Erythrocyte, Age group, Weight, Sweating at Night, Fever Cavity, Exhaustion etc. one can estimate the probability of the presence of the disease in a patient. Various techniques such as ANFIS, Multilayer Perception, Rough-Neural-Networks etc. are available for this purpose. Additionally we also analyze the data from India's National Family and Health Survey and provide our findings and policy recommendations for tackling Tuberculosis.**

**Similarly, data mining can aid in early detection of Cardiovascular diseases. The information collected from patients is already digitized in most hospitals. Tools such as cluster analysis, decision trees, neural networks etc. can help in identifying patterns among sufferers of cardiovascular diseases.**

**Keywords— Data Mining, Tuberculosis, Cardiovascular, Healthcare**

## I. Introduction

Data mining is the process of analysing data from various perspectives and summarizing it into information - information typically used to increase and enhance the revenue, reduce costs, or to provide a new understanding of and solution to a problem. Data mining software are tools for which help in performing the data mining function. It allows users to analyse data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. [1]

Data mining has been successfully used in industries such as e-commerce, retail, and social media. Usage of data mining in the healthcare domain is still at a nascent stage. The scope of medical data mining is huge. With rising healthcare expenditures around the world, "medical data mining has great potential for exploring the hidden patterns in the data sets of the medical domain." [8]

In this paper we look at how data mining has influenced the health care industry as well as the future prospects of data mining in the health care industry. For a start, there has been a large growth of medical databases in advanced countries. Data mining techniques are used on these datasets to help in answering many questions such as:

- "Given historical records of dialysis patients, what ought to be done to improve the condition and treatment of people?
- Given records on cancer, should the treatment include chemotherapy alone, radiation alone, or both chemotherapy and radiation?
- Can human DNA be characterized as a model called genetic coding model?"

Data mining provides automatic pattern recognition and attempts to uncover patterns in data that are difficult to detect with traditional statistical methods. Most statistical methods require a hypothesis which is proved or disproved by the data. In data mining however, patterns are automatically recognised and the strength of the associations is evaluated too. This is advantageous as it is not practical for a human being to develop and evaluate hypothesis as the number of such hypothesis could theoretically be very large. Again it is quite possible that there is more knowledge hidden in the data than what is visible outside. Data mining techniques have a group of various tools and techniques and are used for different purposes. One can however start a data mining process with some simple statistical techniques such as averages. One could then use descriptive statistics techniques like plotting a histogram for the parameter under consideration, or plotting pie charts to get a sense of the data. This could provide a general 'feel' for the data under consideration and this is important as it decides the data mining technique to be used as well as other aspects such as sampling techniques etc.

In health care for example, one could do the following:

- Find out average age of people suffering from diabetes
- Find out the average haemoglobin of women of a particular community
- Plot the histogram of heart patients with respect to gender

Statistics provide a strong background for quantifying and evaluating results. Data mining goes beyond statistics and uses various algorithms to understand underlying relations between data. However, algorithms ought to be modified and scaled before they are applied to data mining. Data mining involves using different algorithms to do different tasks. Basically, algorithms are used to try to fit to a model closest to the characteristics of data under consideration.

Models can be predictive or descriptive. Predictive models are used to make predictions, for example, to make a quick diagnosis of certain element. A patient may be subjected to particular treatment not because of his own history but because of results of treatment of other patients with similar symptoms. Classification, regression, neural networks and time series analysis are some of the tasks of predictive modelling. Descriptive models are used to identify patterns in data. Clustering rules, association rules, and visualization rules are some are the tasks of descriptive modelling.

Data mining typically involves developing rules for classifications based on features in the training set, which is used to classify future data and develop a better understanding of each class (which have been classified) in the database. Regression is one method used to map target data using some known function. It deals with estimating output value based on input values. Time series analysis is examining the value of an attribute over a time period usually at evenly spaced time intervals. Similarly, predictive data modelling is an important data mining task to determine future values of data based on past values. Predictions are made on the basis of regression, time series analysis, neural networks or other methods. Visualization techniques are methods to discover patterns typically in a medical data set. Scatter diagrams in a Cartesian plane of two interesting medical attributes can be used to identify interesting subsets of medical data sets. Once interesting subsets are obtained, other data mining techniques may be used on these subsets to discover further knowledge. Clustering is the identification of classes or clusters for a set of unclassified objects based on their attributes.

Thus one can see that data mining is a term given to any method which helps to classify, explain and predict situations using data. Unlike statistics, data mining doesn't require us to hypothesize about data. It can produce patterns out of large amounts of data mostly by learning with new data. In many cases, some of the underlying logics may be difficult to understand for humans. Take the case of neural networks. It involves using functions called neurons on linear combinations of all possible causal parameters. Many such neurons exist in one particular layer, and many such layers exist in a model. Suffice to say that human beings may not be able to comprehend about what is happening beyond a point. Technological innovations leading to increased processing power of Computers, larger data storage facilities and improved statistical techniques have brought the field of data mining into the limelight. Today all of this is available and one could thus use them to get new insights on data which were hitherto unavailable and unthought-of!

We now briefly describe the application of data mining in major industries.

## II. INDUSTRIAL APPLICATIONS OF DATA MINING

Data Mining is widely used in many areas. We describe four of them here.

A) Financial Data Analysis

B) Retail Industry

C) Telecommunications Industry

D) Biological Data Analysis

### *A. Financial Data Analysis*

The financial data in banking and financial industry is generally reliable and of high quality which facilitates the systematic data analysis and data mining.

- Construction of warehouses for data analysis and data mining.
- Predicting loan payment and estimating customer credit policy.
- Classifying and to cluster of customers for targeting.
- Detecting of financial issues like financial crimes.
- Predicting stock prices. [19]

### *B. Retail Industry*

Data Mining is applied in retail industry due to its nature of collecting large amount data from on sales, customer purchasing history, goods transportation etc. The Data Mining in retail industry helps in identifying customer buying patterns and trends. That leads to an improved quality of customer service and good customer retention and satisfaction.

- Design of warehouses and even construction.
- Multidimensional analysis of sales, customers, etc.
- To analyse how effective are sales campaigns.
- Customer Retention. [19]

### *C. Telecommunications Industry*

Telecommunication industry is one of the fastest growing industries which provide various services such as pager, mobile phone, Internet messenger etc. Due to the advancement of new computer and communication, the telecom industry is growing fast. This is the reason why data mining has been important to help and understand the business. Data mining in telecommunication industry helps in identifying the patterns in communication, catching fraudulent activities and making quality use of the resources under consideration.

- Multidimensional analysing of data.
- Fraudulent pattern analysis.
- Identification of unusual patterns.
- Association and sequential patterns to analyse.(Multidimensional) [19]

### *D. Biological Data Analysis*

Now there is a vast growth in field of biology such as genomics, protein science, and biomedical research. Biological data mining is a key part of bioinformatics.

- Semantic integrating of heterogeneous, distributed varied genomic databases.
- Aligning, indexing, similarity searching and comparative analysis of various multiple nucleotides.
- Find out structural patterns and analysing genetic networks and protein paths.
- Association and path analysis. [19]

## III. GENERAL APPLICATIONS OF DATA MINING IN HEALTHCARE

Data mining can be a great tool to be used in the health care industry. For this to happen, the minimum requirement is that the healthcare industry has to develop data about patients, resources in hospital, disease diagnosis, and records in electronic form. Data are a key resource to be processed and analysed for knowledge extraction that enables support for cost-savings and decision making. Data mining applications in healthcare can be grouped as follows.

### *A. Treatment Effectiveness*

Data mining applications can be developed to evaluate the effectiveness of medical treatments. Data Mining can deliver an analysis about which course of action proves effective by comparing causes, symptoms etc. [3]

### *B. Healthcare Management*

Data mining applications can be developed to identify chronic states of disease and high-risk patients, to design appropriate Intravenous (IVs) therapies, and reduce the number of hospital admissions and claims to aid healthcare management. Data mining can be used to analyse massive volumes of data and statistics to search for patterns which could prove a feed for the bio terrorists. [3]

### *C. Customer relationship management*

Customer relationship management is an approach to managing interactions between commercial organizations typically banks and retailers-and their customers, and also very important in a healthcare context. Customer interactions may occur through call centres, physicians' offices, billing departments, and ambulatory care settings. [3]

### *D. Fraud and Abuse*

Detect fraud and abuses establish norms and then identify unusual patterns of claims by physicians, clinics etc. Data mining applications fraud and abuse applications can highlight inappropriate prescriptions or referrals and fraudulent insurance and medical claims. [3]

### *E. Medical device industry*

Mobile communications and wireless biosensors have paved path for development of mobile healthcare applications to 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 will perform real-time analysis on-board small/mobile devices while considering available resources such as battery charge etc. [3]

### *F. Pharmaceutical Industry*

The technology is being used to help the pharmaceutical firms manage their inventories and to develop new product and services. An understanding of the knowledge hidden in the Pharma data is vital to a firm's competitive position and organizational decision-making. [3]

### *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 behaviour of global hospital activities is visualized. Three layers of hospital management are services for hospital management, services for medical staff and services for patients [3]

### *H. System Biology*

Biological databases contain a variety of data types, with rich relational structure. Consequently multi relational data mining techniques are frequently applied to biological data. Systems biology is has gained a lot of attention in recent years. [3]

### *I. Future of data mining in healthcare*

Applications of data mining in healthcare have great potential in the near future. The success of healthcare data mining solely depends on the availability of good and clean data for inspection. In this respect, it is critical that the healthcare industry consider how data can be better captured, stored, prepared, and mined. Possibly include the standardization of clinical vocabulary and the sharing of data across organizations to enhance the benefits of healthcare data mining applications. Further, as given that healthcare data are not only reduced to quantitative data, it is mandatory to also explore the use of text mining to expand the scope of the data. In particular, it is useful to data and text mining.

It also makes sense to look at how digital diagnostic images can be brought into healthcare data mining applications. Good progress has been made in these areas. [2]

From an Indian perspective, the most important usage of data mining is in diagnosis of chronic and acute diseases quickly. Data mining can definitely aid this process. In our paper we discuss the use of data mining towards detection of two diseases. One is Tuberculosis which is a disease primarily affecting developing countries. The other is Heart related cardiovascular diseases which is a life style disease.

## IV. TACKLING TUBERCULOSIS WITH THE AID OF DATA MINING

Tuberculosis (TB) is a bacterial infection that causes more deaths in the world than any other infectious disease. The bacterium is called Mycobacterium tuberculosis and it usually affects the lungs. Pulmonary tuberculosis refers to disease involving the lung parenchyma. Extra-pulmonary tuberculosis refers to tuberculosis of organs other than the lung, e.g. pleura,

lymph nodes, abdomen, genitourinary tract, skin, joints and bones, meninges. Only people who have pulmonary tuberculosis are infectious. One-third of the world's population is currently infected and new infections are occurring at a rate of one per second [10]. India leads in the number of TB cases in the world and thus it is important to use all the available resources in early detection and cure of the disease. This is especially true because TB spreads quite easily from person to person and early diagnosis will enable quarantine of patients if necessary. TB as of now is a public health emergency in India!

### *A. Tuberculosis and Data Mining*

Data mining techniques are very popular for solving various problems. As a brief description, data mining is a mechanism for obtaining patterns from an existing data set. Those extracted patterns are used to interpret the new or existing data into useful information. In most of the areas, large scaled data is collected. To convert these data into information, many different algorithms and approaches are used. Tuberculosis is a disease where correct diagnosis of patients is very important. To make a correct diagnosis of tuberculosis, a medical test must be applied to patient's phlegm. The result of this test is obtained about after a time period of 45 days. Data mining solutions can be developed which makes diagnosis of tuberculosis as accurate as possible and helps deciding if it is reasonable to start tuberculosis treatment on suspected patients without waiting the exact medical test results or not. It is imperative that, there must be a very accurate classification of patients. Because false positive classified patients will use strong antibiotics for 45 days for nothing and they have to deal with its side effects. And the false negative classified patients' treatment plan will be suspended for 45 days and within this untreated period their disease will get even worse than it is as well as spread to other people.

The international standard for tuberculosis control is the World Health Organization's DOT (Direct Observation of Therapy) strategy that aims to reduce the transmission of the infection through prompt diagnosis and effective treatment of symptomatic tuberculosis patients. The treatment is based on the strict supervision of medicines intake. The supervision is possible thanks to the availability of an information system that records the individual patient data. These data can be used at the facility level to monitor treatment outcomes, at the district level to identify local problems as they arise, at provincial or national level to ensure consistently high-quality tuberculosis control across geographical areas. [13] Additionally, the use of chest radiography for diagnosis of pulmonary tuberculosis can be compromised by poor film quality, low specificity, and difficulties with interpretation. [12]

Physicians are concerned about the poor specificity of current methods. In particular, they want to analyse diagnosis of childhood tuberculosis. In addition, the notification rate of relapsed cases is slightly increasing so they are interested in finding patterns that can explain this trend.

Data analysis is vital for answering these questions. The availability of DOTs records gives an opportunity to use Data Mining techniques such as demographic clustering, or decision trees. In particular, decision trees have an immense prediction power and provide an explanation of the results. If one can predict the prevalence of the disease in an individual based on his/her symptoms and living conditions for example, it would go a huge way to solve this public health crisis.ata mining applications can be developed to evaluate the effectiveness of medical treatments. Data Mining can deliver an analysis about which course of action proves effective by comparing causes, symptoms etc. [10]

### *B. Tuberculosis Data Variables*

There are plenty of variables which can be considered for obtaining data for tuberculosis. Some commonly used variables are:

Gender, Loss of appetite, Erythrocyte, Age group, Loss in weight, Haematocrit, Weight, Sweating at nights, Haemoglobin, Smoke addiction, Chest pain, Leucocyte, Alcohol addiction, Back pain, Number of leucocyte types, BCG vaccine, Coughing, Active specific lung lesion, Malaise, Haemoptysis, Calcific tissue, Arthralgia, Fever Cavity, Exhaustion, Sedimentation, Pneumonic infiltration, Unwillingness for work, PPD, Pleural effusion etc.

The variables can contain direct values or cluster values like Female=0, Male=1, No=0, Yes=1, None=0, Little (<5 items)=1, Moderate(6-10 items)=2, Very Much(11+ items)=3, Normal=0, Moderate=1, High=2 etc. Ranking of the variables is done in an appropriate way and a rank percentage is assigned. After that some variables whose rank percentage is less than a critical value is eliminated and then data mining techniques are applied to get insights about the data. [11]

### *C. Methods of analysing data*

There are many methods which can be applied for analysis of tuberculosis data like ANFIS, Bayesian Network, Multilayer Perception, Ripper Algorithm (JRIP), Partial Decision Trees, Rough Neural Networks, Statistical Accuracy Metrics, Receiver Operating Characteristic (ROC) etc.

ANFIS is a neural-fuzzy system which contains both neural networks and fuzzy systems. This mapping is done by converting the inputs from numerical domain to fuzzy domain. Bayesian Networks produce probability estimates as network output like logistic regression models. The main purpose is to estimate the probability of an instance for each class value if that value suits for a class or not. In JRIP classes are considered by their sizes. Receiver Operating Characteristic ROC is a

visualization technique for experimental results. ROC plot is drawn with the help of sensitivity and specificity values which are calculated by true positive, true negative, false negative and false positive values. Confusion matrix is created with this. [11]

### *D. Conclusion and Findings*

First the benchmarking of different methods is done with the help of some statistical matrix like Root Mean Squared Error and then relevant insights are drawn from the results and some policy recommendations are made. With the help of different data mining techniques the population is divided into different clusters or classes which have varied probabilities of existence of tuberculosis disease. Among all the methods it has been observed that ANFIS is an accurate and reliable method compared to Bayesian Network, Multilayer Perceptron, Part, JRIP and RSES methods for classification of tuberculosis patients [11]

## V. UNDERSTANDING TUBERCULOSIS IN INDIA: INSIGHTS FROM INDIA'S NATIONAL FAMILY AND HEALTH SURVEY

We studied the factors which contribute and correlate with Tuberculosis in India. We took the help of the National Family Health survey conducted in 2006 to do our analysis. Our conclusions may help in effective targeting of populations who have a higher likelihood of the disease. Preventive health care systems can be developed keeping the vulnerable populations in mind. We present the key findings here.

We used three different approaches to understand Tuberculosis in India. These were the decision tree approach, the logistic regression approach and the neural networks approach. The foundation for the study for the preliminary descriptive statistical anaylsis carried out.

We found that the following were the most important variables which determine the presence or absence of Tuberculosis in a household.

1. HV 270 - Wealth index
2. HV 205 - Type of toilet facility
3. SH 46 - Type of caste or tribe of the household head
4. SH 44 - Household head's religion
5. HV 025 - Type of place of residence
6. SHSTRUC - Household structure
7. SH56A - House has any windows

A few fascinating insights about Tuberculosis obtained are:

A) If the cooking fuel is either LPG or Kerosene, 3239 person with 97% confidence fall in no-Tuberculosis category even if they belong to the poorest category and family status is non-nuclear (This is inspite of the fact that being poor and in a non-nuclear family environment enhances the probability of presence of Tuberculosis!)

B) If the family head's religion is Hindu and family structure is nuclear, 7500 records with 98% confidence fall in no-TB category even if they are poor. (One however has to understand that in certain parts of India, religion still correlates with standards of living. Thus the conclusion from here cannot be that a particular religious way of life reduces the possiblity of Tuberculosis. It still requires a deeper analysis)

C) It was found that Crowded cities and small villages both have high incidence of Tuberculosis. The rationale behind this finding is that Tuberculosis being an air-borne disease thrives in crowded cities. In small villages, absence of health care centres can contribute to high incidence of Tuberculosis.

D) The key finding from Neural Networks and Logistic regression was the importance of the variable corresponding to the number of members in a family. If the number of members increases, then probability of there being Tuberculosis also increases. This is a very important finding. It proves that Tuberculosis is quite wide spread and its prevalence increases by increase in number of people..

Any public policy initiative to get the most returns should therefore target crowded cities and small villages as they seem to show more incidence of Tuberculosis. The Indian Government should consider health aspects of the population before taking decisions about removal of LPG subsidies, since the cooking medium used has a direct impact on prevalence of Tuberculosis.

Admittedly, this paper focuses purely on the scope of data mining in improving healthcare in general. Therefore we restrict ourselves to the few fascinating insights stated above, so that this may serve as a motivation for further studies on using Big

data for solving specific problems relating to healthcare. A detailed analysis of our study on Tuberculosis is a fit publication by itself.

## VI. DATA MINING AND CARDIVASCULAR DISEASES

The term cardiovascular disease refers to diseases that involve the heart, the blood vessels, or both. The widely prevalent types of heart diseases are Coronary heart disease, Heart failure, Coronary artery disease, Ischemic heart disease, Hypo plastic left heart syndrome, Atherosclerosis, Chronic obstructive pulmonary disease, Congenital heart disease, Valvular heart disease etc. According to the World Health Statistics Report published by the WHO in 2012, the largest proportion of non-communicable diseases is caused by cardiovascular disease (48%). According to the report, it is projected that the annual number of deaths due to cardiovascular disease will increase from 17 million in 2008 to 25 million in 2030. Half the deaths in the United States and other developed countries occur due to cardiovascular diseases [9]. In countries such as India, which has a large population of low-income groups, quality health care is still a luxury for millions. In such a scenario, it is necessary that cheap methods of diagnosing fatal cardiovascular diseases are available for the general population. The reality is that a large proportion of the population cannot afford to undertake expensive diagnostic tests for detection of cardiovascular diseases. Misdiagnoses only complicate the situation. If one can determine the prevalence of this disease using a patient's past history, symptoms, life style etc. it will be a great step towards both preventing and treating heart related ailments.

Data mining can help identify a set of factors and their relative importance in predicting the risk of getting cardiovascular diseases. The automation of medical diagnosis through data mining can help a long way in reducing costs of tests. The information collected from patients is already digitized in most large hospitals. However, such valuable information has rarely been used to support decision-making. Application of data mining tools such as cluster analysis, decision trees, neural network etc. can help in identifying patterns among sufferers of cardiovascular diseases. This knowledge can help in early detection of heart diseases, identification of vulnerable groups and identification of factors chiefly responsible for causing heart diseases.

Some hospitals do use decision support systems that answer simple queries. However, they cannot answer complex queries such as "Identify the important preoperative predictors that increase the length of hospital stay", "Given patient records on cancer, should treatment include chemotherapy alone, radiation alone, or both chemotherapy and radiation?", and "Given patient records, predict the probability of patients getting a heart disease." [7]. Integrating patient information with clinical decision support could help reduce diagnostic errors and enhance patient safety.

There are several data mining techniques available in the health care domain as per the specific requirements. IF-THEN prediction rules represent discovered knowledge at a high level of abstraction. In the health care system, it can be applied as follows:

Example: If_then_rule induced in the diagnosis of level of alcohol in blood
IF
Sex = MALE AND Unit = 8.9 AND Meal = FULL
THEN
Diagnosis=Blood_alcohol_content_HIGH. [5]

Decision Tree algorithms uses a set of predictor attributes such as age, sex, weight etc. to indicate whether a patient has a particular disease or not. Neural networks have been proposed as useful tools in decision making in a variety of medical applications. Neural networks will never replace human experts but they can help in screening and can be used by experts to double-check their diagnosis. [5]

Palaniappan and Awang [7] developed a prototype Intelligent Heart Disease Prediction System (IHDPS) using three data mining modelling techniques namely Decision Trees, Naïve Bayes and Neural Network. They used 909 patient records obtained from the Cleveland Heart Disease database. Each patient record consisted of 15 attributes namely the predictable attribute – diagnosis, key attribute – patient ID and 13 input attributes – sex, chest pain type, fasting blood sugar, Restecg (Resting electrocardiographic), Exang (Exercise induced angina), Slope (the slope of the peak exercise segment), CA, Thal, Trest blood pressure, serum cholesterol, thalach, old peak and age in year. IHDPS uses the CRISP-DM methodology to build the mining models. It consists of six major phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. The research defined five mining goals based on exploration of the heart disease dataset. The goals were:

Goal 1: Given patients' medical profiles, predict those who are likely to be diagnosed with heart disease.

Goal 2: Identify the significant influences and relationships in the medical inputs associated with the predictable state – heart disease.
Goal 3: Identify the impact and relationship between the medical attributes in relation to the predictable state – heart disease.
Goal 4: Identify characteristics of patients with heart disease.
Goal 5: Determine the attribute values that differentiate nodes favouring and disfavouring the predictable states: (1) patients with heart disease (2) patients with no heart disease. [7]

These goals were reached quite comprehensively in the research using the tools of Decision Trees, Naïve Bayes and Neural Network. This research was carried out on a smaller dataset and only 13 contributing attributes were tested. Data mining tools could be applied on a larger datasets and more factors could be considered to devise a more accurate model for predicting heart diseases. Incorporating more tools and bigger data could provide valuable insight into cardiovascular diseases. Such a model could be used for training nurses, paramedical staff, medical students and even physicians may use it as a guide to diagnose patients. Building effective models will eliminate the need for elaborate testing procedures to diagnose heart diseases. [7]

Often, physicians are valued for their experience in handling vast number of cases over several years of practice. Medical data mining can provide even younger medical professionals with usable information about diseases and patients. Mining already available historical data would lessen the need for practical experience to build effective treatment qualities. Preventing wrong diagnoses would also reduce the problem of wrong medications and adverse drug effects. The impact of drugs could also be measured more qualitatively using data mining by studying the effects of different drugs among different type of patients.

Data mining in cardiovascular disease prediction is an area of serious research interest. Some of the important works published in the last few years are listed below.

Peter C. Austin et al [14] have described a heart disease prediction method by using data mining and machine learning approach. A substantial improvement in prediction and classification of heart failure was provided by flexible tree based method of data mining when compared with conventional classification and regression trees [14]. Jae-Hong et al [16] have introduced a system named AptaCDSS-E, a classifier ensemble-based CDSS for cardiovascular disease level prediction that can be used to identify the cardiovascular disease. Their experimental outcome showed high diagnosis accuracy that proved its usefulness in the clinical decision process of disease identification [16]. Li-Na et al [18] have suggested an investigation on cardiovascular risk prediction by exploiting genetic information. Li-Na et al [18] overview an eligible genome-wide association studies for cardiovascular disease outcomes.

Andrew Kusiak et al [17] have used a data mining approach with Combined Classification Quality (CCQ) measure to analyse data efficiently. It provides a set of rules that were easily interpretable and highly accurate. The rules predicted the health status of a new-born child, interventions, and other user-defined outcomes. The computational complexity in this method is low and it has proved to be a highly effective method to improve the care of neonates with hypo plastic left syndrome [17]. Resul Das and et al [15] developed a method to improve the ability of the physician in diagnosis of valvular heart disease. A method was introduced which uses SAS base software 9.1.3 for diagnosing valvular heart disease. The experiment was conducted on a data set containing 215 samples and it obtained 97.45% classification accuracy [15].

The advancements in data mining techniques should help solve some of the following outstanding issues,

- Data cleaning. Handling of missing values for proper diagnosis is a challenging task.
- The high dimensionality of heart disease database makes identification of significant attributes for better diagnosis of heart disease a challenging task.
- Selection of most suitable sample data for classification instead of the whole data could pose the risk of sample error.
- Selecting suitable classification techniques without much computational complexity and with high level of effectiveness.
- Improving the effectiveness of data mining is very important since the heart disease database is very sensitive and consideration and attention is needed in accuracy of diagnosis.
- Data mining techniques should have the capacity of handling multiple class labels for prediction since it can affect the performance of the medical diagnosis significantly. [6]

## VII. CONCLUSIONS

For a country like India, where public health remains a chief concern, data mining is a blessing in disguise. Most parts of India still depend on the government for health facilities. For the government, health is a priority. It needs to improve the health conditions of millions of Indians at the quickest and cheapest possible rate. One suggestion is that, the government should start using data mining for detection and diagnosis of all kinds of diseases. In most cases the problem appears to be a delay in detection of disease which results in fatalities as well as promotes spread of diseases. We saw in this paper how two different diseases can be successfully predicted using data mining viz. Tuberculosis and Cardiovascular diseases. In these two cases alone, data mining can reduce the number of fatalities by good numbers. If the government could predict that a particular group of people are more vulnerable for getting a disease, they could act and make attempts to reduce their vulnerabilities. Similarly medications can be started quickly instead of waiting for conventional reports. Apart from disease detection and treatments, data mining is a great tool for building better hospitals and smarter treatment methods based on experience. Public health systems can be revamped based on latest data. All in all, data mining is a great tool for making better public policy and it needs to be used in right earnest!

## REFERENCES

[1] Anderson Graduate School of Management. (2012). Data Mining: What is Data Mining? Retrieved from UCLA: Anderson Graduate School of Management: http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/index.htm

[2] Hian Chye Koh, G. T. (2013). Data Mining Applications in Healthcare. Journal of Healthcare Information Management, 19(2). Retrieved from http://www.himss.org/files/himssorg/content/files/jhim/19-2/datamining.pdf

[3] M. Durairaj, V. R. (2013, October). Data Mining Applications In Healthcare Sector: A Study. INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH, 2(10), 29-35.

[4] Salim Diwani, ,. M. (2013, August). Overview Applications of Data Mining In Health Care. International Journal of Computational Engineering Research, 3(8), 73-77.

[5] K, S., Rani, K. B., & A, G. (2010). Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks. International Journal on Computer Science and Engineering, 251-252.

[6] P, G., & N, J. (2013, June-July). Comprehensive Study of Heart Disease Diagnosis Using Data Mining and Soft Computing Techniques. International Journal of Engineering and Technology, 5, 2956.

[7] Palaniappan, S., & Awang, R. (2008). Intelligent Heart Disease Prediction System Using Data Mining Techniques. International Journal of Computer Science and Network Security, 343-349.

[8] Soni, J., Ansari, A., Sharma, D., & Soni, S. (2011). Predictive Data Mining for Medical Diagnosis: An Overview. International Journal of Computer Applications, 43.

[9] World Health Organization. (2012). World Health Statistics. Geneva: WHO Press.

[10] Berka, P., Rauch, J., & Zighed, D. A. (2009). Data Mining and Medical Knowledge Management: Cases and Applications. New York: Medical Information Science Reference.

[11] UÇAR, T. (2009). PREDICTING THE EXISTENCE OF MYCOBACTERIUM TUBERCULOSIS ON PATIENTS BY DATA MINING APPROACH. Master Thesis. Istanbul, Turkey: Graduate School of Natural and Applied Sciences.

[12] World Health Organisation. (2004). Toman´s Tuberculosis:Case detection, treatment and monitoring-questions and answers. Geneva: WHO.

[13] World Health Organisation. (2007). Global tuberculosis control: surveillance, planning, financing. Geneva: WHO.

[14] C. Austin, P., & V. Tu, J. (2013). Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. Journal of Clinical Epidemiology, 398-407.

[15] Das, R., & Turkoglu, I. (2009). Diagnosis of valvular heart disease through neural networks ensembles. Journal of Computer Methods and Programs in Biomedicine, 185-191.

[16] Eom, J.-H., & Kim, S.-C. (2008). AptaCDSS-E: A classifier ensemble-based clinical decision support system for cardiovascular disease level prediction. Journal of Expert Systems with Applications, 2479.

[17] Kusiak, A., & A. Caldarone, C. (2006). Hypo plastic left heart syndrome knowledge discovery with a data mining approach. Journal of Computers in Biology and Medicine, 21-40.

[18] Pu, L.-N., & Zhao, Z. (2012). Investigation on Cardiovascular Risk Prediction Using Genetic Information. Journal of IEEE Transactions On Information Technology In Biomedicine, 5.

[19] Pearson,1999.Industrial Applications of Data mining. [Online] Available at: http://www.pearsonhighered.com/samplechapter/0130862711.pdf [Accessed 25 10 2014].

## BIOGRAPHY

**Vivek Sundar Magesh** is an MBA from the Indian Institute of Management at Tiruchirappalli. He is presently employed at TVS Logistics Private Limited. His interest areas are in developing better public systems and tackling global problems using techniques from

Management Science. He also has an Electrical Engineering degree from the prestigious National Institute of Technology at Tiruchirappalli. As an Engineer he has published papers on optimization using heuristics. He looks forward to researching about improving public systems to tackle global challenges by joining a doctoral program in a reputed global business school.

**Harinivas Bhaskaran** is an MBA from the Indian Institute of Management at Tiruchirappalli. He is presently employed as a credit analyst at State bank of India. He also has an Electronics and Communication Engineering degree from the prestigious National Institute of Technology at Tiruchirappalli.

**Beri Pradeep** is an MBA from the Indian Institute of Management at Tiruchirappalli. He is now a part of the Corporate Strategy and Planning Division at the Murugappa Group in India.

**Sadhanand Lakshmi Narayana Moorthy** is an MBA from the Indian Institute of Management at Tiruchirappalli. He is presently a marketing executive at Bharat Petroleum Corporation Limited.

**Thomas Greefin Franco** is an MBA from the Indian Institute of Management at Tiruchirappalli. He is presently employed at HCL in the financial services

**Nishant Kumar** is an MBA from the Indian Institute of Management at Tiruchirappalli. He works for Dun & Bradstreet.