Utilization of Data mining Approaches for Prediction of Life Threatening Diseases Survivability

A.Sudha
Student, M.Tech (CSE)
Vellore Institute of Technology
Vellore, India

P.Gayathri
Asst.Professor
Vellore Institute of Technology
Vellore, India

N.Jaisankar
Professor
Vellore Institute of Technology
Vellore, India

ABSTRACT
Data mining now-a-days plays an important role in prediction of diseases in health care industry. The Health care industry utilizes data mining Techniques and finds out the information which is hidden in the data set. Many diagnoses have been done for predicting diseases. Without knowing the knowledge of profound medicine and clinical experience the treatment goes wrong. The time taken to recover from diseases depends on patients' severity. For finding out the disease, number of test needs to be taken by patient. In most cases not all test become more effective. And at last it leads to the death of the patient. Many experiments have been conducted by comparing the performance of predictive data mining for reducing the number of test taken by the patient indirectly. This research paper is to present a survey on predicting the presence of life threatening diseases which causes to death and list out the various classification algorithms that has been used with number of attributes for prediction.

General Terms
Data mining applications, Review on Lifethreatning diseases, Prediction algorithm.

Keywords
Data mining, Classification algorithm, life threatening diseases.

1. INTRODUCTION
The rise of health care cost is one of the world's most important problems. Due to increase in world's population, the health care industries are facing many challenges and issues based on patient's severity is to be reduced and detect it earlier in a more effective way. The hospitals are trying to improve and increase the efficiency of medical resource utilization in order to reduce the health care cost. For example, according to the report [1] mentions from the Institute of Medicine (IOM), Due to medical errors Americans die each year in hospitals from 44,000-98,000. The report says medical error costs in the U.S. may range from 37.6 dollars billion each year and remaining 45% of those are said to be preventable errors. In order to save the life of patients and reduce the healthcare cost the medical errors should be prevented. This survey is the extension of previous works. A predictive model for 1. Heart diseases, 2.Lung cancer, 3.Cerebrovascular disease, 4.Stroke diseases, 5.predicting multiple diseases of Hypertension and Hyperlipidemia, 6.Diabetes, 7.Tuberculosis using classification algorithms in data mining. In order to predict these diseases the efficient classification algorithms is discussed and used with various techniques. The Classification algorithms adopts and makes use of decision trees, Bayesian classifier, back propagation neural network, multivariate adaptive regression splines,Adaptive-network-based fuzzy inference system ,genetic algorithm, Fuzzy rulebase,Association rule and k means clustering for prediction of the diseases mentioned above. It consists of attributes containing patient's medical history and symptoms. The records used for this study of prediction of diseases are cleaned and filtered with the data's which are irrelevant and aims to analyze the predictive/descriptive data mining techniques developed for the diagnosis of life threatening diseases.

2. METHODOLOGY
The Methodology described in this paper is survey of lifethreatning diseases focuses on recent trends from journals and publications in field of data mining and healthcare.

2.1 Life threatening diseases an overview
The clinical research has mentioned the very common medical errors occur due to expiry of medicines, incorrect drugs, incorrect dosages, and medicines given to the wrong patient and Diagnosis test report errors. To avoid this type of medical errors care should be taken and prediction should be done earlier. Reports mentions that patients suffer from most common diseases are listed in Table 2.1.

Table 2.1. Major life threatening Diseases

Rating	Cause	Deaths
1	Heart Disease	739,000
2	Cancer	538,000
3	Stroke	158,000
4	Lung disorder	105,000
5	Diabetes	59,000

1. Every 34 seconds in the United States kills the person suffering from heart diseases and prediction has been presented in [5, 6, 13, and 14]. Coronary and cardiovascular diseases are most important categories of heart diseases. The term "cardiovascular disease" means affecting the heart and the blood vessels which circulates throughout the body. Cardiovascular disease (CVD)

results in severe illness, disability and death. Half the deaths in the United States and other in the developed countries occur due to Heart disease.

2. Lung cancer is the second most leading cause of death in World. Lung cancer occurs without uncontrollable growth of lung tissues and form tumors and prediction for lung cancer is presented in paper [7]. It spreads between the age of 45 and 70.We may think most of the people who get lung cancer are due to cigarette smoking. But report behind is some people who smoke may not get lung cancer and the people who have never smoked get lung cancer.
3. The Stroke is a third leading cause of death in the world and makes an attack in brain which does not circulate blood and enough oxygen to the brain cells. It can be prevented through the manipulation of therapy and risk factors behind stroke are very crucial .The time taken to recover from stroke disease depends on patients' severity. For stroke prediction the algorithms are described in [2, 8].Cerebrovascular diseases are major disease which affects the people easily and 15 to 58 percent of the patients die within 30 days. Only 15 percent of the people return to work without disability, 40 percent suffer with minor disability, 40 percent are majorly disabled and 10 percent may not be cured at any time. If the patient suffers from heart-related disease, coronary sclerosis, ventricular fibrillation arrhythmia, then the condition is easily complicated by cerebrovascular disease or known as Stroke. Hypertension and Hyperlipidemia are the symptoms that lead to CVD [3]. Hyperlipidemia occurs when blood levels (i.e. fatty substances) are too high which includes cholesterol and triglycerides. Hypertension and Hyperlipidemia are the risk factors of stroke.
4. Tuberculosis is serious worldwide problem caused by bacterium who suffer from this disease may die unless they given right treatment[10] .Tuberculosis (TB) is the most common cause of infectious disease and caused by various strains of mycobacterium. It mostly affects the lungs and other parts of the body. The World Health Organization (WHO) reports that 2 billion people are suffering from TB, in 2009, and kills 1.7 million people.
5. Diabetes is the fifth leading cause of death in developed countries. Diabetes mellitus is known as diabetes occurs due to high blood sugar and less insulin production. It is a most common disease in the United States. The World Health Organization reports over 16 million Americans are suffering with diabetes and 5.4 million peoples are not aware of the diseases [24]. It is one of the most important criteria for death in United States every year. Eating of sugar may not cause diabetes but when there is high level of sugar in the blood may leads to Diabetes.

3. PREDICTION MODELS

3.1 Data Mining in the Lifethreatning Disease Prediction

The prediction model uses Data mining classification algorithm. It consists of two methods showed in [fig 3.1],1. Prediction model and 2.Evaluation method. In the first method, it uses the training dataset for screening the attributes and build classification predictive model. In the second method, it uses testing dataset for finding classification efficiency. The classification algorithm can be divided by many types based on the survey that has been reviewed then input variables, measured and indicates whether the patient disease can be predicted with high or low level and finds out whether patient suffering from diseases or not..

Decision tree looks like a tree structure. It is very simple and easy to implement. It is presented in [2, 5, 18].Many works has been carried out to predict the life threatening diseases using decision tree and shown efficiency.

A naïve Bayesian classifier depends on Bayes' theorem which works by using probabilistic statistical classifier. The major advantage of using this naïve Bayesian classifier includes rapidity of use and simplicity for handling the dataset containing many attributes with easy and simple way.

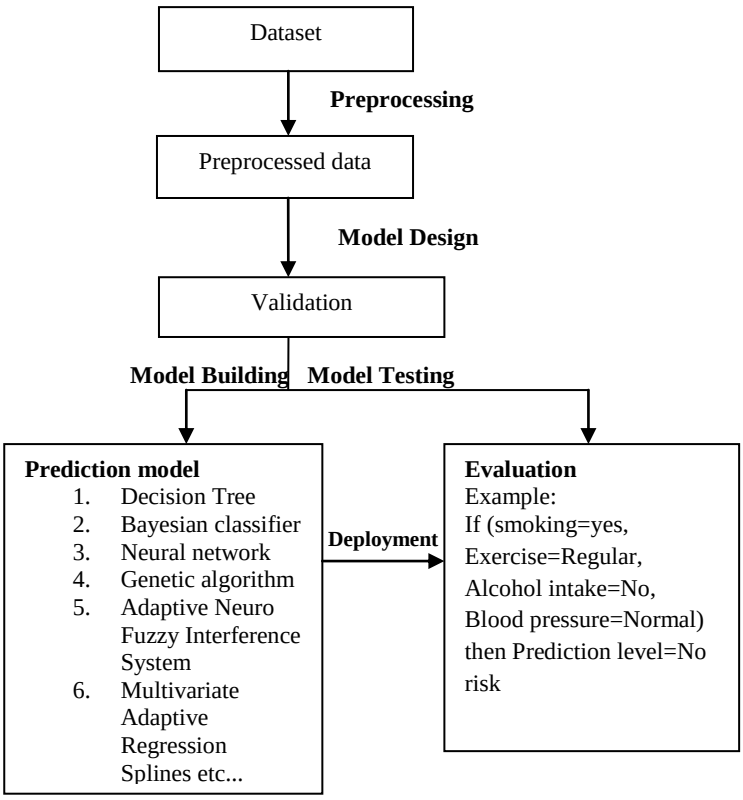
Clustering refers to grouping of similar records [5, 6]. This is used as a preprocessing stage before the data is added into classifying model. The values should be normalized before clustering by avoiding the domination of high value attribute to low value attributes.

Neural Network (NN) is a collection of neurons interconnected between two or more network layers implemented in various diseases prediction used in paper [2, 8, 14, 15, and 18]. It is made up of three layers input layer, hidden layer and output layer. It uses linear transfer function as input layer and non linear transfer function as output layer. In the first stage it sets transfer function and network parameters and calculates output of every neuron in hidden layer and estimates output in hidden layer.

AdaptiveNeuroFuzzyInterferencesystem combines both neural networks and fuzzy systems. This is done by converting the inputs from numerical domain to fuzzy domain. Fuzzy and genetic algorithm use fuzzy logic framework and significantly improves the performance to diagnose diseases of patients and minimizes cost and maximize accuracy.

The Multivariate Adaptive Regression Spline Technique approach uses number of Basis Functions (BFs)[3] to construct MARS predictive models. It is very simple and easy to use in order to select variables and widely used in the fields of healthcare and engineering.

Figure 3.1: Data Mining Classification Techniques for Life threatening Disease Diagnosis



4. DATAMINING TECHNIQUES IN HEALTHCARE

4.1 Survival Analysis

Prediction model are the core data mining methods mainly used in healthcare and engineering field and the techniques used are shown in figure 4.1. In the first step data collection for predicting lifethreatning disease are collected from medical institute. During data preprocessing step, some inappropriate and inconsistent data are deleted. Then prediction algorithm are used .The performance evaluation is done by comparing with various models used and accuracy is measured. Then it is compared with existing model and validated how the proposed model is better than existing models. The Intelligent and Effective Heart Attack Prediction System uses data mining methods like artificial neural network techniques. In first stage it extracts the data from data warehouse of prediction of heart diseases and then it was preprocessed and clustered by means of K-means clustering algorithm [6]. The neural network is trained with the selected significant patterns and Multi-layer Perceptron Neural Network with Back-propagation used as training algorithm. The results show that the designed prediction system is capable of predicting the heart diseases more effectively. The prediction of heart diseases significantly uses 15 attributes, with basic data mining technique like ANN, Clustering and Association Rules, soft computing approaches etc. The outcome shows that Decision Tree performance is more and some time Bayesian classification is having similar accuracy as of decision tree but other predictive methods like KNN, Neural Networks, Classification based on clustering will not perform well [5], By using the Weighted Associative Classifier (WAC), a slight changes has been made, instead of considering 5 class label, only 2 class labels used 1 for "Heart Disease" and 2 for "No Heart Disease". The maximum accuracy (81.51%) has been achieved .After applying genetic algorithm the accuracy of the Decision Tree and Bayesian Classification further improves and reduces the actual data size. The dataset of 909 patient records has been collected and 13 attributes has been used for consistency. The patient records have been splitted equally as 455 records for training dataset and 454 records for testing dataset. After applying genetic search the attributes has been reduced to 6 and decision tree performs more efficiency with 99.2% accuracy when compared with other algorithms. It uses two phases: 1.The generation of weighted fuzzy rules 2.To develop a fuzzy rule-based decision support system [17]. The attribute selection and attribute weight age methods are used to obtain the weighted fuzzy rules. Here, the prediction results are analyzed with the k-fold cross validation. Then, it is constructed in accordance with the weighted fuzzy rules and chosen attributes .It is compared with neural networks by utilizing accuracy and sensitivity.

Cancer gene search with data-mining and genetic algorithms [19] describes only through the analysis of genetic data the cancer can be easily identified. The quality of genes has been analyzed and uses as a training and test data set for each cancer types. The integrated gene-search algorithm involves with genetic algorithm for data preprocessing and data mining classification methods for making predictions. It has high classification accuracy with the ability to identify the most significant genes. It was compared with other reports and shows that mapping to genotype information to the phenotype parameters will easily reduces the cost and detects cancer in earlier stage. An intelligent model for liver disease diagnosis implemented in paper [20] describes Liver disease may not be easily found at early stage. Classification and regression tree (CART) and case-based reasoning (CBR) techniques are used as a classification tool in many medical applications. In the first stage, CART classifies the presence and absence of liver disease and builds a set of rules, which consists of predictable and target variables for designing proper diagnosis. In the second Stage, CBR is used to identify the type of liver and shows accuracy of CART is 94% and CBR diagnostic accuracy rate is 90%.

An expert system for detection of breast cancer based says women's are nowadays suffering from breast cancer and its has become the most important leading cause of death in developed countries. In order reduce the breast cancer deaths, it should be detected earlier and awareness should be created to the women's .The accurate and reliable diagnosis Procedure makes physicians to differentiate breast tumors from malignant ones without going for surgical operations [21]. The data's are collected from 257 women diagnosed with primary breast cancer. The Model Selection Tool is used as a more reliable predictive outcome model. Whereas the database was relatively small, the highest classification accuracy obtained was 80%.The automatic diagnosis system for predicting breast cancer [22] based on association rules and neural network. Association Rule is used for reducing the dimension of breast cancer and neural network is used for intelligent classification. AR and NN system performance is compared with NN model. The input attributes taken for diagnosing the breast cancer has been reduced from nine to four by using Association rule and detection are done earlier.

A predictive model for cerebrovascular disease using data mining [2] shows the cerebrovascular disease major disease affects the human health seriously which uses 493 valid samples and contains 29 attributes which includes blood test, physical exam results and diagnosis results. It adopts three classification algorithms, decision tree, and Bayesian classifier and back propagation neural network, to construct classification models. After applying the classification rules the decision tree shows 99.59% accuracy and could construct classification model with stable classification efficiency. The result indicates that the decision tree is the best classification algorithm when compared with other algorithms. The stroke is a major leading cause of serious long term disability [8]. The time taken to recover from stroke depends on patients' severity. This model makes use of Artificial Neural Networks (ANN) in predicting the Thrombo-embolic stroke disease. The dataset consists of 50 patient records. This research work demonstrates about ANN based prediction of stroke disease by improving the accuracy obtained 89%with higher consistent rate. The ANN exhibits good performance level for prediction of stroke disease.

Multivariate Adaptive Regression Splines (MARS) method is used, to construct a predictive model for multidisease prediction. In their examination of studies Patient with Hypertension or Hyperlipidemia is more likely to suffer from cardiovascular disease. This study uses six data mining approaches to construct a predictive model for hypertension and hyperlipidemia based on the common risk factors of these two diseases. The proposed multi-diseases predictor method has a classification accuracy rate of 93.07% [3]. The results of this paper provide an effective method for predicting Hypertension and Hyperlipidemia.

Tuberculosis is serious worldwide problem caused by bacterium who suffer from this disease may die unless they

given right treatment [10]. This study uses the adaptive-network-based fuzzy inference system (ANFIS) to predict the existence of mycobacterium tuberculosis. They have collected 667 different patient records which are obtained from a clinic and used for the entire process of this research the patient records consist of 30 input parameters and reduced to 20 input variables. ANFIS models is generated by using 500 of those records and also implemented by using Multi-layer Perceptron. ANFIS is accurate and reliable when compared with Multilayer Perceptron for classification of tuberculosis patients.

The fuzzy and genetic algorithms are used for the prediction of diabetes. This work is applied to minimize cost and maximize accuracy [9].With the help of genetic algorithm, the computation cost decreases and the classification performance increases. It shows that by applying this principle within a fuzzy logic framework can significantly improve the mechanism's performance to diagnose patients having diabetes. The artificial neural network (ANN) and fuzzy neural network (FNN) was developed. The classification accuracies of these datasets were obtained by k-fold Cross-validation. It achieves accuracy values of 84.24%. A wide variety of classification methods have been applied previously in order to find the best performing algorithm on this dataset. The accuracies achieved have been in the range of 59.4–84.05%. Efficient predictive modeling is required for predicting diabetes, many models has been designed. [24] Uses Hybrid Prediction Model (HPM) which uses Simple K-means clustering algorithm for validation and C4.5 algorithm is used to build the final classifier model and uses the k-fold cross-validation method. This HPM shows the Classification accuracy of 92.38%.

The construction and training of an artificial neural network for diagnosis of Skin diseases, constructs by a feed-forward architectural design shows successful result on diagnosing skin diseases in the important areas such as Nigeria with 90 percent accuracy [25]. The medical diagnosis for skin diseases uses testing evaluation, treatment evaluation, and treatment effectiveness. The result shows that it reduces cost and maximizes accuracy with 90%.

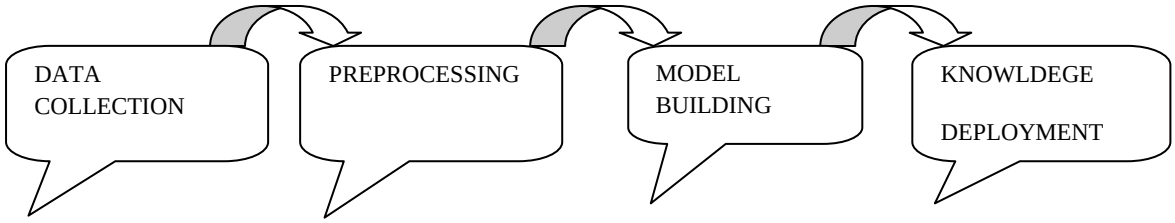


Figure 4.1: Steps involved for predicting life threatening diseases

5. DATAMINING APPLICATIONS

The Data mining techniques are nowadays mostly used in healthcare industry for predicting diseases, Therapy manipulation, Biomedical/Biological Analysis, Research Studies, Hospital Management, .These application makes use of prediction model and reduces cost and increases efficiency Health care industry has been involved in several data mining Efforts to give better healthcare at lower costs. Healthcare Industry uses Data mining Techniques for creating awareness for people for detecting diseases earlier. The purpose of using this data mining application is to identify risk levels of patients and reduce the risk .It consists of various datasets for example which includes variables like gender, alcohol intake, smoking status, the number of times a patient exercises per week, and diabetes level, are the variable, measured and indicates whether the patient disease can be predicted with high or low level and finds out whether patient suffering from diseases or not.

6. CONCLUSION

In healthcare, data mining is becoming increasingly more essential. Data mining has come into existence in the mid of 1990 and widely used in the fields of biomedical, healthcare and engineering. Using data mining technologies, we can predict the diseases earlier. This paper provides an idea about major life-threatening diseases and their diagnosis using data mining with minimum number of attributes and creates awareness about diseases which leads to death. This paper has addressed the widely used classification algorithms for diagnosis of various life threatening diseases. With the help of data mining algorithms, the computation cost decreases and also the classification performance increases. This can be further enhanced and expanded with more prediction algorithm for major lifethreatning diseases.

7. ACKNOWLEDGMENTS

I thank Dr.N.Jaisankar and Prof: P.Gayathri VIT University for giving me full support for involving me into this work and making it in successful way.

8. REFERENCES

[1] Kohn, L. T., Corrigan, J. M., and Donaldson, M. S., to err is human: building a safer health system. Institute of Medicine (IOM). National Academies Press, Washington, 1999

[2] Duen-Yian Yeh a, Ching-Hsue Cheng b, Yen-Wen Chen b 2011 A predictive model for cerebrovascular disease using data mining'Science, Vol. 8970-8977.

[3] Cheng-Ding Chang a, Chien-Chih Wang b, Bernard C. Jiang 2011 Using data mining techniques for multi-diseases prediction modeling of hypertension and hyperlipidemia by common risk factors Vol 38 ,5507–5513.

[4] Genetics and Genomics of Stroke Novel Approaches Alison E. Baird, MBBS, PHD *Brooklyn, New York* Vol. 56, No. 4, 2010.

[5] M. Anbarasi et. al. 2010 Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm International Journal of Engineering Science and Technology Vol. 2(10), 5370-5376.

[6] ShantakumarB.Patil,Y.S.Kumaraswamy2011. 'Predictive data mining for medical diagnosis of heart disease

prediction'jyoti soni, ujma ansari, dipeshsharma IJCSE Vol .17,

[7] M.A. Saleem Durai et. al. 2010 Effective analysis and diagnosis of lung cancer using fuzzy rules International Journal of Engineering Science and Technology Vol. 2(6), 2102-2108.

[8] D.Shanthi,,Dr.G.Sahoo,,Dr.N.Saravanan,2008 'Designing an Artificial Neural Network Model for the Prediction of Thrombo-embolic Stroke (IJBB), Volume 3. pp.10-18.

[9] E.P.Ephzibah,2011 Cost effective approach on feature selection using genetic algorithm and fuzzy logic for diabetes diagnosis by International Journal on Soft Computing (IJSC), Vol.2, No.1.

[10] Tamer Uçar a, Adem Karahocaa 2011'Predicting existence of Mycobacterium tuberculosis on patients using data mining approaches' Vol .3.

[11] Han, J., Kamber, M.: "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, 2006.

[12] Kaur, H., Wasan, S. K.:2006 "Empirical Study on Applications of Data Mining Techniques in Healthcare", Journal of Computer Science 2(2), 194-200.

[13] G.Subbalakshmi et al. Apr-May 2011Decision Support in Heart Disease Prediction System using Naive Bayes / Indian Journal of Computer Science and Engineering (IJCSE) Vol. 2 No. 2.

[14] Shantakumar B.Patil 2009 Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network ISSN 1450-216X Vol.31 No.4 pp.642-656.

[15] Shantakumar B.Patil and Y.S.Kumaraswamy 2009: Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network, European Journal of Scientific Research ISSN 1450-216X Vol.31 No.4, pp. 642-656.

[16] American Heart Association. *Heart Disease and Stroke Statistics — 2004 Update*. Dallas, Tex.: American Heart Association; 2003.

[17] P.K. Anooj 2012 Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules Journal of King Saud University – Computer and Information Sciences 24, 27–40, 2011.

[18] John M. Luk a, Brian Y. Lam a,1, Nikki P.Y. Lee a, David W. Ho a, Pak C. Sham b, Lei Chen a,c, Jirun Peng c, Xisheng Leng c, Philip J. Day d, Sheung-Tat Fan 2007 Artificial neural networks and decision tree model analysis of liver cancer proteomes 68–73.

[19] Shital Shah, Andrew Kusiak 37 2007Cancer gene search with data-mining and genetic algorithms 251 – 261.

[20] An intelligent model for liver disease diagnosis 2009 Rong-Ho Lin Department of Industrial Engineering and Management, National Taipei University of Technology, Artificial Intelligence in Medicine 47, 53—62.

[21] Thora Jonsdottir a,, Ebba Thora Hvannberg b, Helgi Sigurdsson a, Sven Sigurdsson b, 2008 The feasibility of constructing a Predictive Outcome Model for breast cancer using the tools of data mining 108–118.

[22] Murat Karabatak a, M. Cevdet Ince b, 36 2009 An expert system for detection of breast cancer based on association rules and neural network 3465–3469.

[23] Humar Kahramanli , Novruz Allahverdi 2008 Design of a hybrid system for the diabetes and heart diseases 82–89.

[24] B.M. Patil, R.C. Joshi, Durga Toshniwal 2010 Hybrid prediction model for Type-2 diabetic patients 8102–8108

[25] L. G. Kabari and F. S. Bakpo,Diagnosing 2009 Skin Diseases Using an Artificial Neural Network 978-1-4244-3523-4/09/ IEEE.

[26] Data Mining Applications in Healthcare Hian Chye Koh and Gerald TanJournal of Healthcare Information Management — Vol. 19, No. 2