Utilization of Data Mining Techniques for Prediction and Diagnosis of Major Life Threatening Diseases Survivability-Review

K. Rama Lakshmi and S.Prem Kumar

Abstract— Data mining is defined as shifting through very large amounts of data for useful information. Some of the most important and popular data mining techniques are association rules, classification, clustering, prediction and sequential patterns. Data mining techniques are used for variety of applications. In health care industry, data mining plays an important role for predicting diseases. For detecting a disease number of tests should be required from the patient. But using data mining technique the number of test should be reduced. This reduced test plays an important role in time and performance. This technique has an advantages and disadvantages. This research paper analyzes how data mining techniques are used for predicting different types of major life threatening diseases. It reviewed the research papers which mainly concentrated on predicting heart disease, Diabetes, Breast cancer, HIV/AIDS and Tuberculosis.

Index Terms— Breast Cancer, Data mining techniques, Diabetes, Heart disease, HIV/AIDS and Tuberculosis.

1 INTRODUCTION

Data mining is a broad area that integrates techniques from several fields including machine learning, statistics, pattern recognition, artificial intelligence, and database systems, for the analysis of large volumes of data. There have been a large number of data mining algorithms rooted in these fields to perform different data analysis tasks. Data Mining is the process of extracting hidden knowledge from large volumes of raw data. The knowledge must be new, not obvious, and one must be able to use it. Data mining has been defined as "the nontrivial extraction of previously unknown, implicit and potentially useful information from data. It is "the science of extracting useful information from large databases". It is one of the tasks in the process of knowledge discovery from the database. Data Mining is used to discover knowledge out of data and presenting it in a form that is easily understood to humans. It is a process to examine large amounts of data routinely collected. Data mining is most useful in an exploratory analysis because of nontrivial information in large volumes of data. It is a cooperative effort of humans and computers. Best results are achieved by balancing the knowledge of human experts in describing problems and goals with the search capabilities of computers. There are two primary goals of data mining tend to be prediction and description. Prediction involves some variables or fields in the data set to predict unknown or future values of other variables of interest. On the other hand Description focuses on finding patterns describing the data that can be interpreted by humans. The Disease Prediction plays an important role in data mining. There are different types of diseases predicted in data mining namely Hepatitis, Lung Cancer, Liver disorder, Breast cancer, Thyroid disease, Diabetes, HIV/AIDS and Tuberculosis etc…

- *Director, IERDS, Maddur Nagar, Kurnool, Andhra Pradesh, India, Phone: +918374529162. e-mail: krlakshmi_cse@yahoo.com*
- *Professor&Head, Department of CSE&IT, G.Pullaiah college of Engineering & Technology, Nandikotkur Road, Kurnool, Andhra Pradesh, India, Ph-+919866504950, e-mail: mcahod@gpcet.ac.in*

2 REVIEW OF THE RELATED LITERATURE

2.1 Heart Disease Prediction

Medical data mining has high potential for exploring the hidden patterns in the data sets of the medical domain. These patterns can be utilized for clinical diagnosis for widely distributed in raw medical data which is heterogeneous in nature and voluminous. These data should be collected in an organized form. This collected data can be integrated to form a hospital information system. Data mining technology provides a user oriented approach to novel and hidden patterns in the data. From the analysis of World Health Organization, they estimated 12 million deaths occur worldwide, every year due to the Heart diseases. Half the deaths occur in United States and other developed countries due to cardio vascular diseases. On the above discussion, it is regarded as the primary reason behind deaths in adults. Heart disease kills one person every 34 seconds in the United States. The following paper reviewed about predicting of heart disease using data mining technique. Jyoti Soni et. al [3] proposed three different supervised machine learning algorithms. They are Naïve Bayes, *k*-NN, and Decision List algorithm. These algorithms have been used for analyzing the heart disease dataset [14]. Tanagra data mining tool is used for classifying these data. These classified data is evaluated using 10 fold cross validation and the results are compared. Decision tree is one of the popular and important classifier which is easy and simple to implement. It doesn"t have domain knowledge or parameter setting. It handle huge amount of dimensional data. It is more suitable for exploratory knowledge discovery. The results attained from Decision Tree are easier to interpret and read [1]. Naive Bayes is a statistical classifier which assigns no dependency between attributes. To

determine the class the posterior probability should be maximized. The advantages are one can work with the naïve bayes model without using any Bayesian methods. Here Naïve Bayes Classifiers performs well. [1] *k-nearest neighbor's* algorithm (*k*-NN) is the one of the important method for classifying objects based on closest training data in the feature space. It is simplest among all machines learning algorithm but, the accuracy of k-NN algorithm can be degraded by presence of noisy features. This observation is performed using training to consist 3000 instances with 14 different attributes. The dataset is divided into two testing and training i.e. 70% of data are used for training and 30 % is used for testing. The authors concluded that Naïve Bayes algorithm performs well when compared to other algorithms.

Jyoti Soni et.al [3] proposed for predicting the heart diseases using the association rule data mining technique. In their work, unfortunately they have produced a large number of rules when association rules are applied to medical dataset. Most of the rules are medically irrelevant to the data. In [15], the authors proposed four constraints to reduce the number of rules i.e., item filtering, attribute grouping, maximum item set size and antecedent/consequent rule filtering. The important issue is without validation, the association rules are mined on the entire dataset. To solve these limitations, the author introduced an algorithm that uses search constraints to decrease the number of rules. The training set searches association rules and test set to check the validation. Here, a new parameter „lift" is used instead of support and confidence. Lift has been used as the metrics to evaluate the reliability and medical significance of association rules. To validate the results the two basic statistics sensitivity and specificity are used by medical doctors. The chance of correctly identifying sick patients are defined by sensitivity and chance of correctly identifying healthy individuals is defined by specificity. To find predictive association rules in medical dataset the algorithm has three steps: [15]

1. In medical dataset both the categorical and numeric attribute are transformed into transaction dataset.
2. To find the predictive association rules with medically relevant attributes the search process should be incorporate with the above mentioned four constraints.
3. To validate the association rules the train and test approach should be used.

Genetic algorithm have been used in [5], to reduce the actual data size to get the optimal subset of attributed sufficient for heart disease prediction. Classification is one of the supervised learning methods to extract models describing important classes of data. Three classifiers e.g. Decision Tree, Naïve Bayes and Classification via clustering have been used to diagnose the Presence of heart disease in patients. Classification via clustering: Clustering is the process of grouping same elements. This technique may be used as a preprocessing step before feeding the data to the classifying model. The attribute values need to be normalized before clustering to avoid high value attributes dominating the low value attributes. Further, classification is performed based on clustering. Experiments were conducted with Weka 3.6.0 tool [13]. Data set of 909 records with 13 attributes. All attributes are made categorical and inconsistencies are resolved for simplicity. To enhance the prediction of classifiers, genetic search is incorporated. Observations exhibit that the Decision Tree data mining technique outperforms other two data mining techniques after incorporating feature subset selection but with high model construction time. Naïve Bayes performs consistently before and after reduction of attributes with the same model construction time. Classification via clustering is not performing well when compared to other two methods.

In the survey of [6] Naïve bayes have been used to predict attributes such as age, sex, blood pressure and blood sugar and the chances of a diabetic patient getting a heart disease. The clinical dataset is having been collected from one of the leading diabetic research institute in Chennai. The records of 500 patients are taken. The data is analyzed and implemented in WEKA ("Waikato Environment for Knowledge Analysis") tool. Data mining finds out the valuable information hidden in huge volumes of data. Weka tool is a collection of machine learning algorithms for data mining techniques, written in Java. It consists of data pre-processing, classification, regression, association rules, clustering and visualization tools. We have used Naïve bayes method to perform the mining and classification process. We have used 10 folds cross validation to minimize any bias in the process and improve the efficiency of the process. From the experiment the result of bayes model was able to classify 74% of the input instances correctly. It exhibited a precision of 71% in average, recall of 74% in average, and F-measure of 71.2% in average. The results show clearly that the proposed method performs well compared to other similar methods in the literature, taking into the fact that the attributes taken for analysis are not direct indicators of heart disease.

Applying data mining in the medical field is a very challenging task in medical profession. In medical research the data mining begins with a hypothesis and results are adjusted to fit the hypothesis. This differs from standard data mining practice, which simply starts with datasets without an apparent hypothesis. [11]Patterns and trends in dataset are mainly concerned with traditional data mining, but in medical data mining they are not conformed. According to the doctor intuition the clinical decision are often made. The quality of service provided to patients is affected due to unwanted bias, errors and excessive medical cost. Data mining have the capacity to generate a knowledge-rich environment. It can help to improve the significant quality of clinical decision. [5] In the survey of [3] the three supervised machine learning algorithms are used. These algorithms have been used for analyzing the heart disease dataset. The Classification Accuracy should be compared for this algorithm. This work should be extended to predict the heart disease with reduced number of attributes. In

the survey of [3] the heart disease is predicted by using association rule data mining technique. The author introduced an algorithm that uses search constraint to decrease the number of rules. In future this work should be extended by using fuzzy learning models to find the accuracy of time to decrease the number of rules. In the survey of [4] the author proposed a new concept that uses weighted association rule for classification. In future this work can be extended by using association rule hiding technique in data mining. In the survey of [5] the author proposed the minimal subset of attributes for predicting heart disease. In future this work can be expanded and enhanced for the automation of heart disease prediction. Real data should be collected from health care organizations and agencies are taken to compare the optimum accuracy with all data mining technique. In the survey of [6] the author predicts attributes of a diabetic patient getting a heart disease. Weka tool is performed as a result bayes model was able to classify 74% of the input instances correctly. In future this work is extended by using other data mining techniques.

The Heart Disease Data Prediction is designed to support clinicians in their diagnosis for heart disease prediction. They typically work through an analysis of medical data and a knowledge base of clinical expertise. The quality of medical diagnostic decisions for heart disease can be increased by improvements to these Predicting systems [66]. Data mining provides a way to get the information buried in the data. Numerous experiments were conducted on linear and nonlinear characteristics of HRV (Heart Rate Variability) indices to assess several classifiers, e.g., Bayesian classifiers [67], CMAR (Classification based on Multiple Association Rules) [72], C4.5 (Decision Tree) [68] and SVM (Support Vector Machine) [64]. SVM surmounted the other classifiers. The problem of identifying constrained association rules for heart disease prediction was studied by Carlos Ordonez [63]. The assessed data set encompassed medical records of people having heart disease with attributes for risk factors, heart perfusion measurements and artery narrowing. Three constraints were introduced to decrease the number of patterns. First one necessitates the attributes to appear on only one side of the rule. The second one segregates attributes into uninteresting groups. The ultimate constraint restricts the number of attributes in a rule. Experiments illustrated that the constraints reduced the number of discovered rules remarkably besides decreasing the running time. Two groups of rules envisaged the presence or absence of heart disease in four specific heart arteries. Data mining methods may aid the clinicians in the prediction of the survival of patients and in the adaptation of the practices consequently. The work of Franck Le Duff et al. [65] might be executed for each medical procedure or medical problem and it would be feasible to build a decision tree rapidly with the data of a service or a physician. Comparison of traditional analysis and data mining analysis illustrated the contribution of the data mining method in the sorting of variables and concluded the significance or the effect of the data and variables on the condition of the study.The main drawback of the process was knowledge acquisition and the need to collect adequate data to create an appropriate model. In [70] Latha Parthiban et al. projected an approach on basis of coactive neuro-fuzzy inference system (CANFIS) for prediction of heart disease. L. Goodwin et.al [71] discussed different Data mining issues and opportunities for building nursing knowledge. In Lei Yu and Huan Liu [72] introduced a novel concept, predominant correlation, and proposed a fastfilter method which can identify relevant features as well as redundancy among relevant features without pair wise correlation analysis. The efficiency and effectiveness of their method is demonstrated through extensive comparisons with other methods using real world data of high dimensionality. A methodology for comparing classification methods through the assessment of model stability and validity in variable selection was proposed by J. Shreve et.al [69]. This study provides a systematic design for comparing the performance of six classification methods using Monte Carlo simulations and illustrates that the variable selection process is integral in comparing methodologies to ensure minimal bias, enhanced stability, and optimize performance. They quantify the variable selection bias and show that, for sufficiently large samples, this bias is minimized so that methods can be compared. Later John peter and Somasundaram [73] discussed the hybrid attribute selection method combining CFS and Filter Subset Evaluation gives better accuracy for classification. Mohammad Taha Khan [74] discussed a prototype model for the breast cancer as well as heart disease prediction using data mining techniques is presented. The data used is the Public-Use Data available on web, consisting of 909 records for heart disease and 699 for breast cancer. Two decision tree algorithms C4.5 and the C5.0 have been used on these datasets for prediction and performance of both algorithms is compared. Adithya Sundar et.al [75] discussed a prototype using data mining techniques, namely Naïve Bayes and WAC (weighted associative classifier). Srinivas et.al [76] briefly examine the potential use of classification based data mining techniques such as Rule based, Decision tree, Naïve Bayes and Artificial Neural Network to massive volume of healthcare data. Jaya Rama Krishnaiah [77] discussed how the data classification is based on supervised machine learning algorithms which result in accuracy, time taken to build the algorithm. Tanagra tool is used to classify the data and the data is evaluated using entropy based cross validations and partitioned techniques and the results are compared.

2.2 Breast Cancer Prediction

Breast cancer has become a common cancer in women. For instance, it affects one in every seven women in the United

State [16]. The mammography is the traditional method for breast cancer diagnosis. However, the radiologists show considerable variability in how they interpret a mammogram. Moreover, Elmore indicated that 90% of radiologists recognized fewer than 3% of cancers and 10% recognized about 25% of the cases. The fine needle aspiration cytology is another approach adopted for the diagnosis of breast cancer with more precise prediction accuracy. However, the average correct identification rate is around 90% [17]. Generally, the purpose of all the related research is identical to distinguish between patients with breast cancer in the malignant group and patients without breast cancer in the benign group. There are three predictive focus of cancer prognosis: 1) prediction of cancer susceptibility (risk assessment), 2) prediction of cancer recurrence and 3) prediction of cancer survivability. The accepted prognostic factor for breast cancer is the American Joint Commission on Cancer (AJCC). It is staging system based on the TNM system (T, tumor; N, node; M, metastasis) [7] and survival is considered as any incidence of breast cancer where the person is still living from the date of diagnosis. The objective is to handle cases for which cancer has not recurred (censored data) as well as case for which cancer has recurred at a specific time. This section describes various technical and review articles on data mining techniques applied in breast cancer prognosis. C4.5 is a well known classification technique in decision tree induction which has been used by Abdelghani Bellaachia and Erhan Gauven [8] along with two other techniques i.e. Naïve Bayes and Back-Propagated Neural Network. They conduct an analysis of the prediction of survivability rate of breast cancer patients using above data mining techniques and it is used in the new version of the SEER Breast Cancer Data. The preprocessing data set consists of 151,886 records, which are available in 16 fields from the SEER database. They have adopted a different category in the pre-classification process by including three fields: STR(Survival Time Recode), VSR(Vital Status Recode), and COD(Cause Of Death) and it is used by Weka toolkit to experiment these three data mining algorithms. Some of the experiments were conducted using these algorithms. The attained prediction performances are compared to existing techniques. However, the author found the model generated by C4.5 algorithm for the given data has a much better performance than the other two techniques. In [9] M. Lundin et al has applied ANN on 951 instances dataset of Turku University Central Hospital and City Hospital of Turku. To evaluate the accuracy of neural networks in predicting 5, 10 and 15 years breast cancer specific survival. From the experiment the values of ROC curve for 5 years was evaluated as 0.909, for 10 years 0.086 and for 15 years 0.883, these values were used as a measure of accuracy of the prediction model. They author compared 82/300 false prediction of logistic regression with 49/300 of ANN for survival estimation and found ANN predicted survival with higher accuracy. In [10] Delen et al compared ANN, decision tree and logistic regression techniques for breast cancer prediction analysis. They used the SEER data of twenty variables in the prediction models. From the experiment the author found that the decision tree with 93.6% accuracy and ANN with 91.2% are more superior to logistic regression with 89.2% accuracy.

In [8] the author discussed and resolved the algorithms, issues and techniques for the problem of breast cancer prediction. This analysis does not include records with missing data so in future this work is enhanced by including the missing data. In [9] by analyzing the artificial neural network, trained on a number of clinic pathological variables of patients with breast cancer, predicted survival with high accuracy. The author concluded that the consistent accuracy over time and the good predictive performance of a network trained without information on nodal status. It shows that neural networks are valuable tools in cancer survival prediction. In future the study should concentrate on collecting data from a more recent time period and find new potential prognostic factors to be included in a neural network model. In the survey [10], the study is based on multiple prediction models for breast cancer survivability using large datasets along with 10 fold cross validation method. It provides a relative prediction ability of different data mining methods. In future this work is extended by collecting real dataset in the clinical laboratory.

Recently Dursun Delen [78] et.al discussed the comparative study of multiple prediction models for breast cancer survivability using a large dataset along with a 10-fold cross-validation provided us with an insight into the relative prediction ability of different data mining methods. Shweta Kharya [79] discussed various data mining approaches that have been utilized for breast cancer diagnosis and prognosis. Breast Cancer Diagnosis is distinguishing of benign from malignant breast lumps and Breast Cancer Prognosis predicts when Breast Cancer is to recur in patients that have had their cancers excised. Ramachandran [80] has evolved from an experimental analysis to a predicting methodology due to highly precise algorithms and high performance data mining tools. Knowledge discovery in database has been used to predict survivability and diagnosis of diseases in the field of medicine. This can prove helpful for prevention of epidemics.

2.3 Diabetes Prediction

Insulin is one of the most important hormones in the body. It aids the body in converting sugar, starches and other food items into the energy needed for daily life. However, if the body does not produce or properly use insulin, the redundant amount of sugar will be driven out by urination. This disease is referred to diabetes. The cause of diabetes is a mystery, although obesity and lack of exercise appear to possibly play significant roles. Based on the American Diabetes Association [4] in November 2007, 20.8million children and adults in the United States (i.e., approximately 7% of the population) were diagnosed with diabetes. In early the ability to diagnose diabetes plays an important role for the patient"s treatment process.

In [18] the author predicts whether a new patient would test positive for diabetes. This paper studied a new approach, called the Homogeneity- Based Algorithm (or HBA) to determine optimally control the over fitting and overgeneralization behaviors of classification on this dataset (Pima Indian diabetes data set). The HBA is used in conjunction with classification approaches (such as Support Vector Machines (SVMs), Artificial Neural Networks (ANNs), or Decision Trees (DTs)) to enhance their classification accuracy. Some expermental results seem to indicate that the proposed approach significantly outperforms current approaches. From the experiment the author concluded that it is very important both for accurately predicting diabetes and also for the data mining community, in general. In [19] Data mining algorithm is used for testing the accuracy in predicting diabetic status. Fuzzy Systems are been used for solving a wide range of problems in different application domain Genetic Algorithm for designing. Fuzzy systems allows in introducing the learning and adaptation capabilities. Neural Networks are efficiently used for learning membership functions. Diabetes occurs throughout the world, but Type 2 is more common in the most developed countries. The author implemented in Genetic Algorithm. The steps involved in this algorithm namely selection, crossover, mutation, fitness and population statistics. As a result the author concluded that the optimization of chromosome using GA is obtained and it is based on the rate of old population diabetes can be restricted in new population to get chromosomal accuracy. Recently Karthikeyini et.al [57 & 58] discussed comparison a performance of data mining algorithms for diabetes disease based on computing time, precision value , the data evaluated using 10 fold Cross Validation error rate, error rate focuses True Positive, True Negative, False Positive and False Negative, bootstrap validation and accuracy. In [18] the author proposed a new algorithm Homogeneity-Based Algorithm to determine over fitting and over generalization behavior of classification. The algorithms used in this paper are Support Vector Machine, Decision Tree and Artificial Neural Networks. In future this work is enhanced by using any optimization techniques. In [19] for predicting diabetic status the author uses data mining algorithm for testing the accuracy. The author implemented using genetic algorithm. In future this work is extended by using other optimization technique.

2.4 HIV/AIDS Prediction

AIDS, the Acquired Immune Deficiency syndrome is a set of symptoms and infections resulting from the damage done to the human immune system caused by the Human Immunodeficiency Virus (HIV) [20]. HIV is transmitted through direct contact of mucous membrane or the blood stream with a bodily fluid containing HIV [21 & 22].There is currently no vaccine or cure to both HIV and AIDS. However, treatment such as the antiretroviral therapy slows down the course of the disease thus reduces mortality and morbidity of HIV infection. The analysis of survival data has long been the domain of statistics as can be observed through the number of medical statistics textbooks and journals dedicated to the field [20-23]. Statistical methods such as the life-table, the Kaplan-Meier method and regression models such as the Cox Proportional Hazard are usually used to model and predict survival data with the ability to handle censored data [24 & 25]. However, these methods are usually used to explain the data and to model the progression of the disease rather than to make survival predictions for populations or individual patients. A major part of the work done in the area of survival analysis is that carried by Ohno-Machado as part of her PhD thesis. Ohno-Machado made comparisons of standard and sequential neural network models in the survival of coronary heart disease, the comparisons of neural network model with that of Cox Proportional Hazards and the use of modular neural networks to predict HIV survival [25-29]. In writing this paper, the literature review that was carried out on computer-based prognostic systems, namely, on the use of techniques of database research, AI and statistics, in prognostic systems has resulted in at least fifty different articles/papers published by international researchers. In the Malaysian scenario, the number of papers published does not even come close to this figure. A literature search in the National Library of Medicine database (PUBMED), for "Malaysia and AIDS" yielded 169 citations. A PUBMED search on "Malaysia and AIDS and prognosis" resulted in a mere 4 articles, while a search for "Malaysia and AIDS and prognosis and computer", "Malaysia and AIDS and prognosis and artificial intelligence" and "Malaysia and AIDS and prognosis and neural network" yielded "0" (zero) articles [30-32]. Thus, there is a great potential for research in this area in the local scenario. In this paper we describe our research on using CART in the prediction of survival of AIDS in the Malaysian scenario.

HIV/AIDS has emerged as one of the leading challenges for global public health. The number of reported HIV cases in Malaysia has increased from a mere three (3) cases in 1986 to 6427 cases in 2004. During the period of 1986 to 2004 the total number of HIV infections in Malaysia is a staggering 64, 439 (59,962 Male, 4477 Female) of which 9442 had developed into AIDS (8596 Male, 846 Female) with a total of 7195 cases reported to be AIDS related deaths (6661, 534 Female). The ages of these HIV/AIDS related deaths vary from as young as less than 2 years of age to that of above 50 years of age [33]. Since HIV/AIDS poses a challenge to the general development of human resource in Malaysia, any research in HIV/AIDS should thus be duly encouraged. The prediction of survival is not the only research that can be carried out in the area as researchers could also investigate the association of certain risk factors to the disease, matching patients to treatment protocols and the application of AI techniques, database research and

statistics to the follow up and management of AIDS patients [33].

One of the most challenging tasks in carrying out research in the clinical scenario is the availability of clinical data sets. This is especially so, here in Malaysia, where medical data banks do not currently exist. Medical data are usually only accessible to hospital authorities; even then, inter-hospital data accesses may not be possible. Most researchers will have to depend on the goodwill of clinicians, who may do their own medical record keeping, in order to access medical data. Furthermore, even if a researcher is able to access these data they may not be in an electronic form. For chronic illnesses in particular, clinical data sets may not contain an appreciable number of cases, some data collection may be incomplete or some attributes may not be well represented. Those that are complete may not be error-free. All these add to the difficulties of conducting research using real data sets as opposed to artificial data sets. However, recently local computer scientists are beginning to make attempts in carrying out researches in medical informatics through collaborations with their medical counterparts.

The various mathematical and statistical approaches have been proposed for the prediction of survival in HIV/AIDS. M. Bonarek used Kaplan-Meier method to determine prognostic factors associated with in-hospital survival in HIV-infected patients admitted to MICUs (medical intensive care units) [34]. A logistic regression model has been used in prognosis of HIV patient's survival [35], Weibull, loglogistic, lognormal distributions has been used in formulating the prognostic model for HIV survivals [36]. Cox models are applied and STATA 7 for statistical analyses is used as well in a study to determine the survival in HIV [37]. The purpose of this study is to examine the use of CART as a tool for data mining, predictive modeling and data processing in the prognosis of AIDS. The goal of any modeling exercise is to extract as much information as possible from available data and provide an accurate representation of both the knowledge and uncertainty about the epidemic. Recently Sameem A.K. et.al [59] studied certain experiments on the application of Classification And Regression Tree (CART) to predict the survival of AIDS. CART builds classification and regression trees for predicting continuous dependent variables and categorical or predictor variables, and by predicting the most likely value of the dependent variable. A total of 998 patients who had been diagnosed with AIDS were grouped according to prognosis by CART. We found that CART were able to predict the survival of AIDS with an accuracy of 60-93% based on selected dependent variables, validated using Receiver Operating Characteristics (ROC). Rosma mohd dom et.al [60] was to describe the feasibility of applying a predictive data mining technique to predict the survival of AIDS. An adaptive fuzzy regression technique, FuReA, was used to predict the length of survival of AIDS patients based on their CD4, CD8 and viral load counts. Predictive ability of FuReA was measured and compared with fuzzy neural network prediction models.We found that both FuReA and fuzzy neural network models were able to predict the survival of AIDS with an accuracy of 60% to 100% based on selected dependent variables. Saravanan. A.M. et.al [61] was used Dempster Shafer (DS) theory and focused to identify the hidden information from the data set using the concept theory of evidence. Later Santhosh Kumar and Ramraj [62] discussed several classifier approaches are used to classify the HIV1 and HIV2. For implementation of the work a real time patient database is taken and the patient records are experimented and the final best classifier is identified with quick response time and least error rate.

2.5 Tuberculosis Prediction

Data classification process using knowledge obtained from known historical data has been one of the most intensively studied subjects in statistics, decision science and computer science. Data mining techniques have been applied to medical services in several areas, including prediction of effectiveness of surgical procedures, medical tests, medication, and the discovery of relationships among clinical and diagnosis data. In order to help the clinicians in diagnosing the type of disease computerized data mining and decision support tools are used which are able to help clinicians to process a huge amount of data available from solving previous cases and suggest the probable diagnosis based on the values of several important attributes Ensemble of classifiers has been proved to be very effective way to improve classification accuracy because uncorrelated errors made by a single classifier can be removed by voting. A classifier which utilizes a single minimal set of classification rules to classify future examples may lead to mistakes. An ensemble of classifiers is a set of classifiers whose individual decisions are combined in some way to classify new example. Many research results illustrated that such multiple classifiers, if appropriately combined during classification, can improve the classification accuracy. India has the world's highest burden of tuberculosis (TB) with million estimated incident cases per year. It also ranks among the world's highest HIV burden with an estimated 2.3 million persons living with HIV/AIDS. Tuberculosis is much more likely to be a fatal disease among HIV-infected persons than persons without HIV infection**.** It is a disease caused by mycobacterium which can affect virtually all organs, not sparing even the relatively inaccessible sites. The microorganisms usually enter the body by inhalation through the lungs. They spread from the initial location in the lungs to other parts of the body via the blood stream. They present a diagnostic dilemma even for physicians with a great deal of experience in this disease.

Minou Rabiei et.al.[39] use tree based ensemble classifiers for the diagnosis of excess water production. Their results

demonstrate the applicability of this technique in successful diagnosis of water production problems. Hongqi Li, Haifeng Guo and team present[40] a comprehensive comparative study on petroleum exploration and production using five feature selection methods including expert judgment, CFS, LVF, Relief-F, and SVMRFE, and fourteen algorithms from five distinct kinds of classification methods including decision tree, artificial neural network, support vector machines(SVM), Bayesian network and ensemble learning. Zhenzheng Ouyang, Min Zhou, Tao Wang, Quanyuan Wu[41] propose a method, called WEAP-I, which trains a weighted ensemble classifier on the most n data chunks and trains an averaging ensemble classifier on the most recent data chunk. All the base classifiers are combined to form the WEAP-I ensemble classifier. Orhan Er. And temuritus [42] present a study on tuberculosis diagnosis, carried out with the help of multilayer neural networks (MLNNs). For this purpose, an MLNN with two hidden layers and a genetic algorithm for training algorithm has been used. Data mining approach was adopted to classify genotype of mycobacterium tuberculosis using c4.5 algorithm [43]. Evaluation of the performance of two decision tree procedures and four Bayesian network classifiers as potential decision support systems in the cytodiagnosis of breast cancer was carried out [44]. Paper on "Mining Several Data Bases with an Ensemble of Classifiers"[45] analyze the two types of conflicts, one created by data inconsistency within the area of the intersection of the data bases and the second is created when the meta method selects different data mining methods with inconsistent competence maps for the objects of the intersected part and their combinations and suggest ways to handle them. Referenced paper [46] studies medical data classification methods, comparing decision tree and system reconstruction analysis as applied to heart disease medical data mining. Under most circumstances, single classifiers, such as neural networks, support vector machines and decision trees, exhibit worst performance. In order to further enhance performance combination of these methods in a multi-level combination scheme was proposed that improves efficiency [47]. Paper [48] demonstrates the use of adductive network classifier committees trained on different features for improving classification accuracy in medical diagnosis. Paper on "MReC4.5: C4.5 ensemble classification with MapReduce" [49] takes the advantages of C4.5, ensemble learning and the MapReduce computing model, and proposes a new method MReC4.5 for parallel and distributed ensemble classification. Seppo Puuronen and team [50] propose a similarity evaluation technique that uses a training set consisting predicates that define relationships within the three sets: the set of instances, the set of classes, and the set of classifiers. Lei Chen and Mohamed S. Kamel [51] propose the scheme of Multiple Input Representation-Adaptive Ensemble Generation and Aggregation (MIR-AEGA) for the classification of time series data. Kai Jiang et.al. [52] propose a neural network ensemble model for classification of incomplete data. In the method, the incomplete dataset is divided into a group of complete sub datasets, which is then used as the training sets for the neural networks. Recently Asha et.al [81] discussed the comparison of classification techniques for TB based on two categories namely pulmonary tuberculosis (PTB) and retroviral PTB using ensemble classifiers such as Bagging, AdaBoost and Random forest trees.

3 Conclusions

In this survey paper the problem of summarizing the different algorithms of data mining for the major life threatening diseases are used in the field of medical prediction are discussed. The main focus is on using different algorithm and combination of several targets attributes for different types of disease prediction using data mining. First we discuss about the heart disease prediction, in that machine learning algorithms namely naïve bayes, *k*-NN, Decision List. Of these the classification accuracy of the naïve bayes algorithm is better when compared to other algorithm. In Weighted Associative Rule Classifier, the GUI has been designed to enter the patient record and the presence of Heart Disease for a patient is predicted by using the rules stored in the rule base. Next we discuss the feature subset selection using genetic algorithm. In this attributes are reduced using genetic search. Here the accuracy is compared to the three classifiers namely Decision Tree, Naïve bayes and classification via clustering. Association rule discovery is mainly based on four constraints namely item filtering, attribute grouping, maximum item set size and antecedent/consequent rule filtering. To find predictive rules in medical data set the three important steps are generated in this algorithm. The heart disease is diagnosed for diabetic patients using Naïve Bayes technique. Of these the author concluded that naïve bayes classify 74% of input instances correctly. Next we discuss about the breast cancer prediction. It is performed by using various data mining techniques namely C4.5, ANN and fuzzy decision trees. By using C4.5 the author discussed and resolved the issues and algorithms of the problem. Using ANN the author concluded that the consistent accuracy over time and good performance of the network is trained. The fuzzy decision tree survives by using 10 fold cross validation method. Finally we discuss about diabetes prediction, by using homogeneity based algorithm the author find over fitting and overgeneralization behavior of classification. By using genetic algorithm the author predicts accuracy of the class. In future the work can be expanded and enhanced for the automation of various types of disease prediction. It also extended to find other types of diseases with the uses of these attributes. The various mathematical and statistical approaches have been proposed for the prediction and diagnosis of survival in case of HIV/AIDS. An ensemble of classifiers is a set of classifiers whose individual decisions are combined in some way to classify new example. Many research results illustrated that such multiple classifiers, if appropriately com-

bined during classification, can improve the classification accuracy. India has the world's highest burden of tuberculosis (TB) with million estimated incident cases per year.

ACKNOWLEDGMENTS

The authors are thankful to Prof. C.Uma Shankar, Dept. of OR&SQC and Dr. M. Veera Krishna, Department of Mathematics, Rayalaseema University, Kurnool, Andhra pradesh, India, for their valuable guidance and suggestions with thought provoking discussions throughout the period of my research and in the preparation of this paper, and IJSER Journal for the support to develop this document.

REFERENCES

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

[2]. Margaret H. Dunham, "Data mining: Introductory and Advanced Topics".

[3]. Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita Soni "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", *International Journal of Computer science and Engineering*, Vol. 3, No. 6, June 2011.

[4]. Carloz Ordonez, "Association Rule Discovery with Train and Test approach for heart disease prediction", *IEEE Transactions on Information Technology in Biomedicine*, Vol. 10, No. 2, April 2006, pp. 334-343.

[5]. M. Anbarasi, E. Anupriya, N.CH.S.N.Iyengar, "Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm", *International Journal of Engineering Science and Technology*, Vol. 2(10), 2010, pp. 5370- 5376

[6]. G. Parthiban, A. Rajesh, S.K.Srivatsa, "Diagnosis of Heart Disease for Diabetic Patients using Naive Bayes Method".

[7]. Choi J.P., Han T.H. and Park R.W., "A Hybrid Bayesian Network Model for Predicting Breast Cancer Prognosis", *J Korean Soc Med Inform*, 2009, pp. 49-57.

[8]. Bellaachia Abdelghani and Erhan Guven, "Predicting Breast Cancer Survivability using Data Mining Techniques", *Ninth Workshop on Mining Scientific and Engineering Datasets in conjunction with the Sixth SIAM International Conference on Data Mining*, 2006.

[9]. Lundin M., Lundin J., BurkeB.H.,Toikkanen S., Pylkkänen L. and Joensuu H. , "Artificial Neural Networks Applied to Survival Prediction in Breast Cancer- Oncology", *International Journal for Cancer Resaerch and Treatment*, Vol. 57, 1999.

[10]. Delen Dursun , Walker Glenn and Kadam Amit, "Predicting breast cancer survivability: a comparison of three data mining methods," *Artificial Intelligence in Medicine*, Vol. 34, June 2005, pp. 113-127.

[11]. Ruben.D. Canlas. Jr., "Data mining in healthcare: current applications and issues", August 2009.

[12]. Michael Feld, Dr. Michael Kipp, Dr. Alassane Ndiaye and Dr. Dominik Heckmann "Weka: Practical machine learning tools and techniques with Java implementations".

[13]. K.P Soman, Shyam Diwakar, V.Vijay "Insight into Data mining theory and practice".

[14]. Asha Rajkumar, G.Sophia Reena, "Diagnosis Of Heart Disease Using Datamining Algorithm", *Global Journal of Computer Science and Technology*, Vol. 10, Issue. 10, September 2010.

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

[16]. Wingo PA, Tong T, Bolden S, "Cancer statistics", *CA Cancer J Clin*, Vol. 45, No. 1 (1995), pp. 8-30.

[17]. Fentiman IS, "Detection and treatment of breast cancer", *London: Martin Duntiz* (1998).

[18]. Huy Nguyen Anh Pham and Evangelos Triantaphyllou "*Prediction of Diabetes by Employing a New Data Mining Approach Which Balances Fitting and Generalization*" Department of Computer Science, 298 Coates Hall, Louisiana State University, Baton Rouge, LA 70803.

[19]. Ms.S.Sapna, Dr.A.Tamilarasi "Data mining-Fuzzy Neural Genetic Algorithm in predicting diabetes" Department Of Computer Applications (MCA), K.S.R College of Engineering "BOOM 2K8" *Research Journal on Computer Engineering*, March 2008.

[20]. Breiman, L., J. H. Fried man, R. A. Olshen, and C. J. Stone. 1984. Classification and regression trees. Monterey, Calif., U.S.A.: Wadsworth, Inc.

[21]. Collet, D., "Modeling Survival Data in Medical Research", Ed. C. Chatfield, J.V. Zidek. *London: Chapman & Hall*, 1994.

[22]. Elandt-Johnson, R.C., Johnson, N.L., "Survival Models and Data Analysis". *John Wiley & Sons*, 1976, pp. 50-83.

[23]. Lee, E.T., "Statistical Methods for Survival Analysis", *Lifetime Learning Publications*, California, 1980, pp. 9-18.

[24]. CART FAQ http://clearinghouse.wayne.edu/oldsite/downloads/CARTFAQ.pdf Accessed, May 2009.

[25]. Swinscow, T.D.V., "Survival analysis", Statistics at Square One, *British Medical Journal* (electronic version. Available: http://www.bmj.com/collections/statsbk/12.html. 1999.

[26]. Ohno-Machado, L.,"Neural network techniques: utilization in medical prognosis", *Electronic Copy of Review*, [Emailed by Ohno-Machado], 1999.

[27]. Ohno-Machado, L., Musen, M.A., "Sequential versus standard neural networks for temporal pattern recognition: an example using the domain of coronary heart disease", Stanford University. Technical Report. *School of Medicine. Medical Computer Science. Knowledge Systems Laboratory*, 1996.

[28]. Ohno-Machado, L., "Identification of low frequency patterns in backpropagation neural network", Technical Report, *Section in Medical Informatics, Stanford University*, 1994.

[29]. Ohno-Machado, L., Walker, M.G., Musen, M.A., "Hierarchical neural networks for survival analysis", Technical Report, *Section in Medical Informatics, Stanford University*, 1994.

[30]. Tan DB, Yong YK, Tan HY, Kamarulzaman A, Tan LH, Lim A, James I, French M, Price P. "Immunological profiles of immune restoration disease presenting as mycobacterial lymphadenitis and cryptococcal meningitis", *HIV Med*. Vol. 9(5) May 2008, pp. 307-16.

[31]. Fadzelly AB, Asmah R, Fauziah O. "Effects of Strobilanthes crispus tea aqueous extracts on glucose and lipid profile in normal and streptozotocin-induced hyperglycemic rats". *Plant Foods Hum Nutr*. Vol. 61(1), March 2006, pp. 7-12.

[32]. Mohammad Z, Naing NN. "Characteristics of HIV-infected tuberculosis patients in Kota Bharu Hospital, Kelantan from 1998 to 2001". *Southeast Asian J Trop Med Public Health*, Vol. 35(1), March 2004, pp. 140-3.

[33]. United Nations General Assembly Special Session on HIV/AIDS, December 2005, "Monitoring the Declaration of Commitment on HIV/AIDS" Country Report, MALAYSIA.

[34]. M. Bonarek, "Prognostic score of short-term survival in HIV-infected patients admitted to medical intensive care units (MICUs)". International Conference AID; abstract no: MoPeB2183, July 2000. Classification And Regression Tree In Prediction Of Sur-

vival Of AIDS Patients, *Malaysian Journal of Computer Science*, Vol. 23(3), 2010, p. 163.

[35]. Hanson DL; Horsburgh CR Jr; Fann SA; Havlik JA; Thompson SE 3d; "Survival prognosis of HIV-infected patients", Division of HIV/AIDS, *Centers for Disease Control, Atlanta, Georgia* 30333. AIDSLINE MED/93267399. Jun 1993.

[36]. Matthias Egger, Margaret May, Geneviève Chêne, Andrew N Phillips, Bruno Ledergerber, François Dabis, Dominique Costagliola, Antonella D"Arminio Monforte, Frank de Wolf, Peter Reiss, Jens D Lundgren, Amy C Justice, Schlomo Staszewski, Catherine Leport, Robert S Hogg, Caroline A Sabin, M John Gill, Bernd Salzberger, Jonathan A C Sterne, and the ART Cohort Collaboration," Prognosis of HIV-1-infected patients starting highly active antiretroviral therapy: a collaborative analysis of prospective studies". *THE LANCET*, Vol. 360, July 13, 2002.

[37]. K. Porter, A G Babiker, J H Darbyshire, P Pezzotti, K Bhaskaran, A S Walker, "Determinants of survival following HIV-1 seroconversion after the introduction of HAART". *THE LANCET*, Vol. 362 , October 2003.

[38]. Ohno-Machado, L.,"Medical applications of neural networks: connectionist models of survival". *PhD Thesis, Section in Medical Informatics, Stanford University*, Palo Alto, California, 1996.

[39]. Minou Rabiei, Ritu Gupta "Excess Water Production Diagnosis in Oil Fields using Ensemble Classifiers" *IEEE*, 2009.

[40]. Hongqi Li, Haifeng Guo, Haimin Guo and Zhaoxu Meng, "Data Mining Techniques for Complex Formation Evaluation in Petroleum Exploration and Production: A Comparison of Feature Selection and Classification Methods" in *proceedings of 2008 IEEE Pacific-Asia, Workshop on Computational Intelligence and Industrial Application,* Vol. 1 , pp. 37-43.

[41]. Zhenzheng Ouyang, Min Zhou, Tao Wang and Quanyuan Wu "Mining Concept-Drifting and Noisy Data Streams using Ensemble Classifiers", *International Conference on Artificial Intelligence and Computational Intelligence*, Nov. 2009, pp. 360-364.

[42]. Orhan Er, Feyzullah Temurtas and A.C. Tantrikulu, "Tuberculosis disease diagnosis using Artificial Neural networks ", *Journal of Medical Systems*, Springer, 2008, DOI 10.1007/s10916-008-9241-x online.

[43]. M. Sebban, I. Mokrousov, N. Rastogi and C. Sola " A data-mining approach to spacer oligo nucleotide typing of Mycobacterium tuberculosis" *Bioinformatics, oxford university press*, Vol. 18, Issue 2, 2002, pp. 235-243.

[44]. Nicandro Cruz-Ram_rez , Hector-Gabriel Acosta-Mesa, Humberto Carrillo-Calvet , Roc_o-Erandi Barrientos-Mart_nez "Discovering interobserver variability in the cytodiagnosis of breast cancer using decision trees and Bayesian networks" *Applied Soft Computing*, Elsevier, Vol. 9, Issue 4, September 2009, pp. 1331-1342.

[45]. Seppo Puuronen, Vagan Terziyan and Alexander Logvinovsky "Mining Several Data Bases With an Ensemble of Classifiers" in *Proc. 10th International Conference on Database and Expert Systems Applications*, Vol.1677, 1999, pp. 882-891.

[46]. Tzung-I Tang, Gang Zheng ,Yalou Huang ,Guangfu Shu "A Comparative Study of Medical Data Classification Methods Based on Decision Tree and System Reconstruction Analysis", *IEMS*, Vol. 4, Issue 1, June 2005, pp. 102-108.

[47]. Tsirogiannis, G.L. Frossyniotis, D. Stoitsis, J. Golemati, S. Stafylopatis, A. Nikita, K.S., " Classification of medical data with a robust multi-level combination scheme" in *Proceeding of 2004 IEEE International Joint Conference on Neural Networks*, Vol. 3, 25-29, July 2004, pp 2483- 2487.

[48]. R.E. Abdel-Aal "Improved classification of medical data using abductive network committees trained on different feature subsets", *Computer Methods and Programs in Biomedicine*, Volume 80, Issue 2, 2005, pp 141-153.

[49]. Gongqing Wu, Haiguang Li, Xuegang Hu, Yuanjun Bi, Jing Zhang and Xindong Wu "MReC4.5: C4.5 ensemble classification with MapReduce" in *Proceeding of 2009 Fourth ChinaGrid Annual Conference*, 2009, pp. 249-255.

[50]. Seppo Puuronen and Vagan Terziyan "A Similarity Evaluation Technique for Data Mining with an Ensemble of classifiers", *Cooperative Information Agents III, Third International Workshop, CIA*, 1999, pp. 163-174.

[51]. Lei Chen and Mohamed S. Kamel "New Design of Multiple Classifier System and its Application to the time series data" *IEEE International Conference on Systems, Man and Cybernetics*, 2007, pp. 385-391.

[52]. Kai Jiang, Haixia Chen, Senmiao Yuan "Classification for Incomplete Data Using Classifier Ensembles", *Neural Networks and Brain*, 2005.

[53]. X. Wu, V. Kumar, Ross, J. Ghosh, Q. Yang, H. Motoda, G.Mclachlan, A. Ng, B. Liu, P. Yu, Z.-H. Zhou, M. Steinbach, D. Hand and D. Steinberg, "Top 10 algorithms in data mining," *Knowledge and Information Systems*, Vol. 14, No. 1, January 2008, pp. 1-37.

[54]. R. J. Quinlan, "Bagging, boosting, and c4.5," in AAAI/IAAI: *Proceedings of the 13th National Conference on Artificial Intelligence and 8th Innovative Applications of Artificial Intelligence Conference. Portland, Oregon, AAAI Press / The MIT Press*, Vol. 1, 1996, pp. 725-730.

[55]. Han and M. Kamber. Data mining: concepts and techniques: *Morgan Kaufmann Publications*, 2006.

[56]. I. H. Witten and E. Frank. Data Mining: Practical Machine Learning Tools and Techniques, Second Edition: *Morgan Kaufmann Publications*, 2005.

[57]. V. Karthikeyini, Pervin begum.I., "Comparison a performance of data mining algorithms (CPDMA) in prediction of Diabetes Disease", *International journal of Computer Science and Engineering*, Vol.5, No. 03, March 2013, pp. 205-210.

[58]. V. Karthikeyini, Pervin begum.I., Tajuddin.K., Shahina Begum, "Comparative of data mining classification algorithm (CDMCA) in Diabetes Disease Prediction", *International journal of Computer Applications*, Vol.60, No. 12, Dec. 2012, pp. 26-31.

[59]. Sameem Abdul Kareem, S. Raviraja, Namir A Awadh, Adeeba Kamaruzaman, Annapurni Kajindran, "Classification And Regression Tree In Prediction Of Survival Of AIDS Patients", *Malaysian Journal of Computer Science*, Vol. 23(3), 2010, pp 153-165.

[60]. Rosma mohd dom, Sameem Abdul Kareem, Basir Abidin, Adeeba Kamaruzaman, Annapurni Kajindran, "The Prediction of AIDS Survival: A Data Mining Approach", *Proceedings of the 2nd WSEAS International Conference on Multivariate Analysis and its Application in Science and Engineering*, pp. 48-53.

[61]. A.M. Saravanan, R.Vijaya and C. Jothi Venkateswaran, "Feature Selection for Prediction of HIV/AIDS using Data Mining Technique by applying the Concept of Theory of Evidence", *International Journal of Computer Science and Network Security*, VOL.11 No.5, May 2011, pp. 285-288.

[62]. Santhosh Kumar.S and E. Ramaraj, "Analysis of Sequence Based Classifier Prediction for HIV Subtypes", *International Journal of Engineering and Technology*, Vol. 2, No. 10, October 2012, pp. 1753-1758.

[63]. Carlos Ordonez, "Improving Heart Disease Prediction Using Constrained Association Rules", Seminar Presentation at University of Tokyo, 2004.

[64]. Cristianini, N., Shawe-Taylor, J.: "An introduction to Support Vector Machines", *Cambridge University Press*, Cambridge, 2000.

[65]. Frank Lemke and Johann-Adolf Mueller, "Medical data analysis using self-organizing data mining technologies", *Systems Analysis Modeling Simulation*, Vol. 43 , No. 10 , 2003, pp: 1399 - 1408.

[66]. Frawley and Piatetsky-Shapiro, Knowledge Discovery in Databases: An Overview. *The AAAI/MIT Press*, MenloPark, C.A, 1996.

[67]. Heon Gyu Lee, Ki Yong Noh, Keun Ho Ryu, "Mining Biosignal Data: Coronary Artery Disease Diagnosis using Linear and Nonlinear Features of HRV," *LNAI 4819: Emerging Technologies in Knowledge Discovery and Data Mining*, pp. 56-66, May 2007.

[68]. Hian Chye Koh and Gerald Tan, "Data Mining Applications in Healthcare", *Journal of healthcare information management*, Vol. 19, Issue 2, Pages 64-72, 2005.

[69]. J. Shreve, H. Schneider, O. Soysal:"A methodology for comparing classification methods through the assessment of model stability and validity in variable selection", *Decision Support Systems*, Vol. 52, 2011, pp. 247–257.

[70]. Latha Parthiban and R.Subramanian, "Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm", *International Journal of Biological, Biomedical and Medical Sciences*, Vol. 3, Issue. 3, 2008.

[71]. L. Goodwin, M. VanDyne, S. Lin, S. Talbert ,"Data mining issues and opportunities for building nursing knowledge", *Journal of Biomedical Informatics*, Vol. 36, 2003, pp. 379-388.

[72]. Li, W., Han, J., Pei, J.: CMAR: Accurate and Efficient Classification Based on Multiple Association Rules. In: *Proc. of 2001 International Conference on Data Mining*, 2001.

[73]. John Peter.T and Somasundaram, "Study and Development of novel feature selection framework for Heart disease prediction", *International Journal of Scientific and Research Publications*, Vol. 2, Issue 10, October 2012, pp. 1-7.

[74]. Mohammad Taha Khan, Dr. Shamimul Qamar and Laurent F. Massin, "A Prototype of Cancer/Heart Disease Prediction Model Using Data Mining", *International Journal of Applied Engineering Research*, Vol.7 No.11 (2012), pp. 1-6.

[75]. N. Aditya Sundar, P. Pushpa Latha and M. Rama Chandra, "Performance analysis of classification data mining techniques over heart disease data base", *International journal of engineering science & advanced technology*, Vol. 2, Issue. 3, 470 – 478.

[76]. Srinivas.K, B.Kavihta Rani and A.Govrdhan, "Applications of Data mining techniques in health care and Prediction of Heart attacks", *International Journal on Computer Science and Engineering*, Vol. 02, No. 02, 2010, pp. 250-255.

[77]. Jaya Rama Krishnaiah.V.V., D.V.Chandra Sekhar and K.Ramchand H Rao, "Predicting the Heart attack symptoms using Biomedical data mining techniques", *The International Journal of Computer Science & Applications*, Volume 1, No. 3, May 2012, pp. 10-18.

[78]. Dursun Delen, Glenn Walker and Amit Kadam, "Predicting breast cancer survivability:a comparison of three data mining methods", *Artificial Intelligence in Medicine*, Vol.xxx (2004), pp. 1-15.

[79]. Shweta Kharya, "Using data mining techniques for diagnosis and prognosis of Cancer Disease", *International Journal of Computer Science, Engineering and Information Technology* (IJCSEIT), Vol.2, No.2, April 2012, pp. 55-66.

[80]. P.Ramachandran, N.Girija and T.Bhuvaneswari, "Health care Service Sector: Classifying and finding Cancer spread pattern in Southern india using data mining techniques", *International Journal on Computer Science and Engineering* (IJCSE), Vol. 4 No. 05 May 2012, pp. 682-687.

[81]. Asha.T, S. Natarajan and K.N.B. Murthy, "Diagnosis of Tuberculosis using Ensemble methods", IEEE, 2010, 978-1-4244-5539-3/10.

BIOGRAPHY OF AUTHORS

K.R.Lakshmi: She has completed Master degree in Computer Applications in 2010 from Sri Krishnadevaraya University, Anantapur, Andhra Pradesh, India. She is a Director, IERDS, Maddur Nagar, Kurnool, Andhra Pradesh, India. Her teaching and research areas Data mining techniques. She has published 2 articles in international well reputed journals.

S.Prem Kumar: He received Ph.D. degree in Computer Science and Technology from Sri Krishnadevaraya University, Anantapur, Andhra Pradesh, in 2010. He is Professor of computer science and engineering, Department of CSE&IT, G.Pullaiah college of Engineering & Technology, Nandikotkur Road, Kurnool, Andhra Pradesh, India. His teaching and research areas include Data mining techniques, mobile computing and Internet frame works. He has published 10 articles in national and international well reputed journals.