INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

DATA MINING TECHNIQUES IN PREVENTION AND DIAGNOSIS OF NON COMMUNICABLE DISEASES

Nimna Jeewandara[1], PPG Dinesh Asanka[2]

[1] *Faculty of Graduate Studies and Research*
Sri Lanka Institute of Information Technology, Sri Lanka.
nimnajeewandara@gmail.com

[2] *Faculty of Graduate Studies and Research*
Sri Lanka Institute of Information Technology, Sri Lanka.
dineshasanka@gmail.com

Abstract: - The complex heterogeneous and voluminous data are generated through health care industry which is difficult to analyse by traditional methods. Data mining in non-communicable disease area is the support to identify the pattern in the different perspective and summarize to obtain useful information in the medical domain. This paper summarizes and evaluates the current state of knowledge in data mining techniques which helps to prevent and diagnosis of non-communicable diseases. Some of the recent research papers which were published by addressing Diabetes Mellitus, Heart Disease, Hypertension, Cancer and Hyperlipidaemia were used for this purpose. In early stages, one data mining technique was used for analysis the diagnosis of one disease but in recent years many data mining techniques were used to analyse prevention and diagnosis of the diseases. For the purpose of analysing the multiple diseases the researchers have used multiple stage analysis with many algorithms.

Keywords: *NCD; Data mining; Diabetes Mellitus; Heart Disease; Hypertension; Cancer.*

1. Introduction

1.1 Non-communicable Diseases (NCDs)

Cardiovascular diseases (e.g.: hypertension, heart attacks, stroke), diabetes mellitus, cancers, chronic respiratory disease (e.g.: asthma, chronic obstructive pulmonary disease (COPD)) are main non-communicable diseases which are now global disease burden [1].

The main risk factors for NCDs are demographical risk factors (age, gender, race and family history) physiological risk factors and risky behaviors. Among them, risky behaviors which are alcohol consumption, tobacco use, beetle chewing, unhealthy diet, and physical inactivity can be reduced or controlled by intervention. Unless, these behavioral risk factors can lead to metabolic or physiologic changes in human being.

Physiological risk factors are raised blood pressure, overweight, obesity, raised blood glucose and raised cholesterol.

Most of the food-related workings are handled by the female at home. Unhealthy diet pattern of the main character of a family will affect the whole family as a unit. Other behavioral risk factors such as alcohol consumption, tobacco use, beetle chewing, and physical inactivity will also deviate the family members (male person of that family and children) into the risk.

1.2 Data Mining in Medical Domain

Data Mining is a process of analyzing the available large amount of data in the different perspective to discover the useful information [2]. It combines statistics, machine learning, artificial intelligence and database technology. The data integration, selection, cleaning, transformation, mining, pattern evaluation, and knowledge presentation are the steps involved in data mining process. There is no difference the health sector is also a major data generating area which is voluminous, heterogeneous and valuable. Data mining methods are used to predicts and prevent the diseases. Medical diagnosis is subjective and depends on the available data and the experience of the physician and sometimes differ from one person from another. Healthcare related data mining is a challenging field. Medical data mining explores the hidden pattern and knowledge discovery in a database to predict diseases [2]. Descriptive and predictive are two data mining models in big data. In descriptive data analysis, it uses historical data to identify the pattern in data and analyze the relationship between variables or samples. Descriptive models are clustering, association rule, and summarization. These models are developed by using whole data set. In predictive data analysis, it uses historical data and current data to predict the future or used to diagnose the disease. Decision tree, random forecast, artificial neural network, support vector machine, regression are predictive data models.

A Cluster is a collection of the group of similar elements or objects which are unsupervised. It doesn't have pre-defined classes. K-means, Gaussian mixture models, kernel K-means, special clustering, nearest neighbor are few examples for clustering algorithms [3].

The association rule is a data mining procedure to find the frequent relationships/patterns or associations between variables within a larger database. These rules are useful for analyzing and predicting the behaviors in a given scenario. Support indicates the how frequently appear the item in given database and confidence indicates the number of times of true statements are occurred. Positive rule mining (classical rule), negative rule mining and Interestingness measures are the 3 type of association rules. Positive rule mining is based on frequent item sets in a dataset (positive relationship). Examples are Apriori algorithm and FP-growth algorithm. Negative rule mining is based on the infrequent item sets in a database.

Interestingness measures are also known as the constraint-based association rule. Instead of above rules, these rules are applied due to the interesting to users. These constraints can be knowledge based or data constraints [6].

Decision Tree is a structure with a root node, branches and leaf nodes. It can be handle both numerical and categorical data. Decision tree gives the simple illustration of collected knowledge. It is easy to transform leaf and nodes to a set of rules by mapping from the root node to the leaf nodes one by one. Decision trees have high classification accuracy and it is a reliable and effective decision-making technique for medical diagnostics. Decision trees are also used in medical and health care applications for more than 20 years. Classical decision tree algorithms are ID3, C4.5, CART, OCI, SADT, and ID5R. [9]. To reduce the drawback of the decision, tree a group of researches introduced a hybrid approach which is a combination of decision trees and artificial neural networks. Many evolutionary approaches were also introduced by researches (e.g. genTrees, VEDEC). All approaches were tested to the medical domain and were encountered with size, sensitivity and accuracy problems. They show that the evolutionary method has least problems [9]. AREX algorithm is an extended version of decision tree which is also known as decision programs [9].

Numerical outcome of the predictive model is regression which variables on one or more predictors. These predictors can be numerical or categorical variables [11]. A linear regression will predict the probabilities range between 1 and 0. A logistic regression produces a logistic curve. The values are limited to between 0 and 1. Logistic regression is similar to a linear regression. The curve of logistic regression is constructed using the natural logarithm of the "odds" of the target variable, rather than the probability [11]. If there is no linear relationship between dependent and independent variable are called nonlinear regression. Multiple linear regression is used to model the linear relationship between a dependent variable and one or more independent variables [12] [13]. Generalized Linear Models (GLM) and Support Vector Machines (SVM) are 2 algorithms for regression. GLM for linear modelling and SVM for linear, nonlinear and other mining functions [12]. To analyze the medical data regression is the important statistical method. It is used to identify the relationship between multiple factors. As an example, if a patient has high fasting blood sugar and it is influenced by the age,

weight, sex. The dependent variable is FBS and independent variables are age, weight and sex. Age can be a single variable (single linear regression) multiple variable (multiple linear regression) or logistic regression. The strength of the relationship can also be calculated [14].

ANN are computational models which are based on the biological model of the brain. They are nonlinear data modelling tools. ANN are flexible to apply incomplete, missing and noisy data. When under lying data, relationship is unknown the ANN are the powerful tool for data modelling. It can be approximate complex nonlinear mapping. It can enhance the performance by combining ANN with genetic algorithms and neuro-fuzzy systems [17].

2. Literature Review

The literature review indicates that most of the times these researches used data mining algorithms for diagnosis the diseases than prevention. These algorithms show different accuracy, sensitivity and specificity while diagnosing one disease in different methods which helps to evaluate each method. There are no single methods has found superior than other methods and every method have advantages and disadvantages in different conditions.

2.1 Algorithm for Diabetes Millets

2.1.1) Clustering:

Razan Paul, Abu Sayed Md. Latiful Hoque have used K-means and k- mode algorithms to predict the likelihood of diseases. They have developed algorithm for both continues and discrete data [5]. This research had shown that combining algorithm and background knowledge had produce excellent results [5]. However, it is not indicated that what was the base to analyze the diabetic for each cluster and there was less information about the data base.

2.1.2) Association rule:

Azra Ramezankhani, Omid Pournik, Jamal Shahrabi, Fereidoun Azizi, and Farzad Hadaegh have proceed a research using ARM to analyze the risk patterns for type 2 diabetes occurrence. They found that the risk factors for male are IGT, IFG, levels, the length of stay in the city, central obesity, the ratio between total cholesterol and high-density lipoprotein ≥ 5.3, low physical activity, chronic kidney disease and wrist circumference. The female risk factors are impaired fasting glucose (IFG) and impaired glucose tolerance (IGT), with body mass index (BMI) ≥ 30 kg/m2, wrist circumference > 16.5 cm and waist to height ratio ≥ 0.5 and family history of diabetes [8]. They had not taken the equal number of male and female to compare. Sociological factors and nutritional factors were not considered in this study.

2.1.3) Decision tree:

Asma A. AlJarullah has used decision tree algorithm to predict the developing diabetes using Pima Indians Diabetes Data Set [24]. She has used 724 females with 6 attribute which are number of time patient pregnant, plasma-glucose, diastolic BP, BMI, Diabetes pedigree function(DPF) and age. Weka software was used to construct the decision tree. This research has 2 phases which are data pre-processing at the first stage and decision tree construction at the second stage. The accuracy of the decision tree model was 78.177%. The other important factors are not considered in this research which are family history, gestational DM, dietary pattern and life style[24].

2.1.4) Regression:

Abdullah A. Aljumah, Mohammed Gulam Ahamad, and Mohammad Khubeb Siddiqui have completed the predictive analysis for diabetic treatment in Saudi Arabia. It was used regression-based data mining technique. They have used a software to analyze the data which has Predictive data mining algorithms such as Regression and support vector machine (SVM). This research showed the drug treatment is effective in both age groups but more effective for old age group than young the young group. Apart from medication, for the effective diabetic treatment, exercise, weight reduction and smoke cessation are also important factors. In this paper, they show the education level did not impact on the HbA1c changes. Some patients follow single drug treatment but some patients need combination drug treatments to control the diabetic. Young patients are advised to focus more on

diet control, weight reduction, exercise, and smoke cessation than medication. They also show the dietary treatment is effective for both groups but more effective for old age group people than young. Weight reduction is strongly suggested for old age patients but it is also effective for both groups. The research shows the smoking cessation is effective who suffer diabetic. Even though the exercise is effective for both groups the research shows it is more effective for old age diabetic patient. The blood glucose level normally rises with age shows in researches [16] [15]. When they had reclassified 5 age group in to 2 age groups the age group 35-44 were common to both groups. If they had considered the gender the research will be more valuable.

2.1.5) Logistic regression, artificial neural networks (ANNs) and decision tree models for DM:

The combination of many data mining algorithms is very useful for better decision making than using one algorithm for one disease in medical domain. There is 2 type of diabetes which is type 1 DM and type 2DM. Among them, the genetic basis of type 2 diabetes has yet to be identified. Xue-Hui Meng, Yi-Xiang Huang, Dong-Ping Rao, Qiu Zhang, and Qing Liu have used logistic regression, artificial neural networks (ANNs) and decision tree models to predict diabetes or pre-diabetes using common risk factors in China [19]. They evaluated the three models for accuracy, sensitivity and specificity. The decision tree model had the best classification accuracy and next the logistic regression model was the next. The ANN gave the lowest accuracy. The risk factors for diabetes are genetic predisposition and behavior and an environment. Such modifiable risk factors as obesity and physical inactivity are the main non-genetic determinants of the disease. Physical activity level, dietary habits, adiposity, alcohol consumption, smoking, and duration of sleep are the lifestyle factors related to developing diabetes. Previous studies have proposed that anthropometric measurement and adipocyte size help to predict diabetes incidence using traditional statistical methods. The positively related factors were older age, family history of diabetes, BMI, and preference for salty food for diabetes. The negatively related factors were the education level and drinking coffee for diabetes. Previous researches also have shown the similar results for risk factors for diabetes [19]. Several trials have demonstrated that the lifestyle intervention is benefits for the prevention of diabetes.

This study also showed that physical activity and extent of stress related to work were also positively associated with diabetes. The age is itself a risk factor for diabetes. This study has shown that Smoking, drinking alcohol, and eating vegetables and fruits daily were not significantly associated with diabetes. But previous studies have shown that there was relationship with above mention factors with diabetes [19].

The logistic regression model is commonly used for the prediction of disease or health status is of interest. In some researches it had been shown that The ANN model was identified as better performance than decision tree model. The decision tree model was stated to perform better than logistic regression model. In this study, it was indicated that decision tree is the best predictor. The logistic regression model was second and the ANNs model was poorest of the three models [19]. They had not considered the equal percentage of male and female as gender. The number of normal participation and the disease participation is also differed. The attribute of the number of cigarette for a life is not a correct measure. The work stress was also very difficult to measure as correctly.

2.2 Algorithm for Heart Diseases

2.2.1) Clustering:

Shantakumar B.Patil and Dr. Y.S.Kumaraswamy have presented a research on K-Means clustering which is simple and easily implemented was used to predict the heart attack in pre-processed data warehouse [4]. They have used MAFIA algorithm to find the frequent items. They have used only one algorithm to find the frequent item set. It did not provide the accuracy or sensitivity and compare only with predefined threshold to predict the heart disease.

2.2.2) Association Rule:

Ibrahim Umar Said, Abdullahi Haruna Adam, Dr. Ahmed Baita Garko together have conducted the research to predict the heart diseases using Apriori algorithm [7]. They have categorized the data based on the gender and the risk factors. They have found that females are more chance to free from coronary heart diseases than males. It was supported by previous researchers as well [7]. This research has more than 80% of confidence level with more than 99% accuracy levels. The fasting blood sugar attribute value should be > 126 mg /dl to consider as

diabetic and it is a major drawback. They had used only one algorithms instead of using many to compare the results.

2.2.3) Decision Tree

Vikas Chaurasia and Saurabh Pal have used three decision tree algorithms which were CART, ID3 and DT (decision table) to predict the heart disease early. They have used three evaluation criteria to find the best algorithm. They found that CART had the best accuracy (83.49%) than other 2 algorithms [10]. Fasting blood sugar is a main attribute to diagnose the diabetic to predict the heart disease. Hence it should be 126 mg/dl instead of 120 mg/dl.

2.2.4) Regression

Irfana P. Bhatti, Heman D. Lohano , Zafar A. Pirzado and Imran A. Jafri have conducted the research to investigate factors that contribute significantly to enhancing the risk of ischemic heart disease[23]. The data set has 585 individuals. Among them 484 individuals had IHD and 101 individuals had no IHD (control). The factors that used for this research were cholesterol level, age, cooking oil, profession, locality, sex, education, blood pressure, smoke, narcotic use, DM, history of IHD, Apo Protein A, Apo Protein B, high density Lipo protein, low density Lipo protein, and phospholipids, Total Lipid, uric acid, weight, glucose and triglycerides. Among these factors this research showed that the use of banaspati ghee, living in urban area, high cholesterol level, age group of 51 to 60 year are significantly contribute to enhancing the risk of ischemic heart disease. This research has shown 98.63 % of overall correctly prediction with this model [23].

2.2.5) K-means clustering with decision tree, WAC (Weighted Associative Classifier) with Apriori algorithm and Naive Bayes:

The chance of getting heart disease was predicted by Aswathy Wilson, Jismi Simon, Liya Thomas, and Soniya Joseph using data mining algorithms. The algorithms were K-Means Clustering with Decision Tree, WAC (Weighted Associative Classifier) with Apriori algorithm and Naive Bayes [20]. They have used many attributes such as age, sex, blood pressure and blood sugar for this research. WAC was used to classify the data set and Apriori algorithm was used to find out the frequent item set. Naive Bayes rule was used to create models with predictive capabilities. The decision tree algorithm was used to make the decision. The study was shown that K-means clustering with decision tree offers high accuracy (83.24%) and Naive Bayes (81%) [20].

2.3 Algorithm for Breast Cancer

2.3.1) Naïve Bayes, Back-propagated Neural Network and C4.5 Decision Tree:

Abdelghani Bellaachia and Erhan Guven have conducted the research using three data mining algorithms to predict the survivability rate of breast cancer patients [26]. They have used Weka toolkit with Naïve Bayes, the back-propagated neural network, and the C4.5 decision tree algorithms. They have used SEER database which include patient-level information on cancer site, tumor pathology, stage, and cause of death. They have used a pre-classification approach with three variables which were Survival Time Recode (STR), Vital Status Recode (VSR) and the Cause of Death (COD). They have found that Decision Tree algorithm has more accuracy than other two [26].

2.3.2) Artificial Neural Networks, Decision trees and Logistic regression

Dursun Delen, Glenn Walker and Amit Kadam have completed the research using three data mining algorithms which are ANN, Decision tree and Logistic regression [25]. They compared these algorithms accuracy, sensitivity and specificity to predict breast cancer survivability. They have used SEER Public data with 72 variables which provide the socio-demographic and cancer specific information concerning an incidence of cancer. This research has shown that the decision tree (C5) is the best predictor with 93.6% accuracy and ANN was the second highest accuracy (91.2%). Logistic regression has shown that 89.2% accuracy. The grade was the most important variable. The second was the stage of the cancer in the prediction of survival [25].

2.4 Algorithm for Multiple Diseases- Hypertension and Hyperlipidemia

2.4.1) Logistic regression analysis, Classification and Regression Tree (CART), decision tree, Chi-squared Automatic Interaction Detector (CHAID), exhaustive CHAID, and discriminant analysis:

Most of the previous studies had used many predictive models for one disease prediction. But in case of suffering more than one diseases for a single person, the behavior of the common risk factors may be change due to reciprocal effect between diseases. Cheng-Ding Chang, Chien-Chih Wang and Bernard C. Jiang have suggested a two-phase analysis technique predict hypertension and hyperlipidemia simultaneously. This approach offers a more effective way for preventive medicine. They have suggested six data mining approaches to select the individual risk factors of these two diseases firstly. The 6 data mining approaches were logistic regression analysis, Classification and Regression Tree (CART), decision tree, Chi-squared Automatic Interaction Detector (CHAID), exhaustive CHAID, and discriminant analysis [21]. Then they have used voting principle to determine the common risk factors. Next, they have used the Multivariate Adaptive Regression Splines (MARS) method to construct multiple predictive models for two diseases which are based on common risk factors for both diseases. The systolic blood pressure is more than 140 hg mm and /or diastolic blood pressure is more than 90 is considered as the Hyperion. Hypertension can also be associated with hyperlipidemia. The previous research showed that the risk factors for hypertension are the family history of HT, old age, sex, BMI, high sodium intake, high total fat intake, smoking, excess alcohol intake and physical inactivity. Previous studies also showed that incidence of the major cardiovascular event can be reduced by 22 -24 % when introducing lipid-lowering therapy for type 2 diabetic patients' medication. The proposed analysis procedure showed that Systolic Blood Pressure (SBP), Triglycerides, Uric Acid (UA), Glutamate Pyruvate Transaminase (GPT), and gender are common risk factors for both diseases. The proposed multi-diseases predictor method had a classification accuracy rate of 93.07%. This research showed that the effective and appropriate methodology for simultaneously predicting hypertension and hyperlipidemia [21]. Some important risk factors for both diseases were not included for both diseases such as total cholesterol, age, BMI.

.

3. Conclusion

As the non-communicable disease burden is a challenge many programs and research are conducted to prevent and control of these diseases. Most of the developed countries has very successive programs than others.

All above researches used data mining approaches to predict the non-communicable diseases and the states by using many demographical, physiological, behavioral factors. They have used one or many data mining algorithms for one or more diseases to understand the hidden pattern and relationships of these diseases. They have been produced different accuracy, sensitivity and specificity according to their algorithms and the variables in large complex data sets. The results can be more accurate and validated using experienced medical professionals' evaluation which explore novel clinical research direction to the world.

Acknowledgment

The authors are utmost grateful for the support and kind assistance given by Prof. Koliya Pulasinghe, Dean, Faculty of Computing, Sri Lanka Institute of Information Technology.

REFERENCES

[1] V. Poznyak, L. Riley, V.D. Costa, S. Mendis, T. Armstrong, D. Bettcher, F. Branca, J. Lauer, C. Mace, Gretch, "Global Status Report on noncommunicable diseases 2014," WHO, Geneva, 2014.

[2] N.S. Nithya, K. Duraiswamy, P. Gomathy, "A Survey on Clustering Techniques in Medical Diagnosis," International Journal of Computer Science Trends and Technology (IJCST), vol. 1, no. 2, pp. 17-22, 2013.

[3] A.K. Jain ,R. Chitta, R. Jin, "Clustering Big Data," 29/11/2012. [Online]. Available: http://www.cse.nd.edu/Fu_Prize_Seminars/jain/

[4] slides.pdf. [Accessed: 1/4/2017].

[5] S.B. Patil, Y.S. Kumaraswamy, "Extraction of Significant Patterns from Heart Disease Warehouses for Heart Attack Prediction," IJCSNS International Journal of Computer Science and Network Security, vol. 9, no. 2, pp. 228-235, 2009.

[6] R. Paul, A.S.Md.L. Hoque, "Clustering Medical Data to Predict the Likelihood of Diseases," 2010.

[7] K.P. Kumar, S. Arumugaperumal, "Association Rule Mining and Medical Application: A Detailed Survey," International Journal of Computer Applications (0975 – 8887), vol. 80, no. 17, pp. 10-19, 2013.

[8] I.U. Said, A.H. Adam, A.B. Garko, "Association Rule Mining on Medical Data to Predict Heart Disease," International journal of science technology and management, vol. 4, no. 8, pp. 26-35, 2015.

[9] A. Ramezankhani, O. Pournik, J. Shahrabi, F. Azizi, and F. Hadaegh, "n Application of Association Rule Mining to Extract Risk Pattern for Type 2 Diabetes Using Tehran Lipid and Glucose Study Database," Int J Endocrinol Metab. 2015 Apr; 13(2): e25389., vol. 13, no. 2, 2015.

[10] V. Podgorelec, P. Kokol, B. Stiglic, I. Rozman, "Decision Trees: An Overview and Their Use in medicine," Journal of Medical Systems, vol. 26, no. 5, pp. 445-463, 2002.

[11] V. Chaurasia, S. Pal, "Early Prediction of Heart Diseases Using Data Mining Techniques," Caribbean Journal of Science and Technology, vol. 1, pp. 208-217, 2013.

[12] V. Podgorelec, P. Kokol, B. Stiglic, I. Rozman, "Decision Trees: An Overview and Their Use in medicine," Journal of Medical Systems, vol. 26, no. 5, pp. 445-463, 2002.

[13] "Regression," dmg.org, [Online]. Available: http://dmg.org/pmml/v2-0/Regression.html. [Accessed 4 4 2017].

[14] "Regression," oracle, [Online]. Available: http://www.comp.dit.ie/btierney/oracle11gdoc/datamine.111/b28129/regress.htm. [Accessed 4 4 2017].

[15] A. Schneider, G. Hommel, M. Blettner, "Linear Regression Analysis," Deutsches Ärzteblatt International, vol. 107, no. 44, pp. 776-782, 2010.

[16] A.A. Aljumah, M.G. Ahamad, M.K. Siddiqui, "Application of data mining: Diabetes health care in young and old patients," Journal of King Saud University – Computer and Information Sciences, no. 25, pp. 127-136, 2012.

[17] "Old Age Solutions, "Ministry of Science & Technology (Govt. of India), 2015. [Online]. Available: http://www.oldagesolutions.org/Health/Diabetes.aspx. [Accessed 4 April 2017].

[18] Y. Singh, A.S. Chauhan, "Neural Networks in Data Mining," Journal of Theoretical and Applied Information Technology, pp. 37-42, 2005-2009.

[19] F. Shadabi, D. Sharma, "Artificial Intelligence and Data Mining Techniques in Medicine – Success Stories," in International Conference on BioMedical Engineering and Informatics, 208.

[20] X-.H. Meng, Y-.X. Huang, D-.P. Rao, Q. Zhang, Q. Liu, "Comparison of three data mining models for predicting diabetes or prediabetes by risk factors," Kaohsiung Journal of Medical Sciences, vol. 29, pp. 93-99, 2013.

[21] A. Wilson, J. Simon, L. Thomas, S. Joseph, "Data Mining Techniques For Heart Disease Prediction," International Journal of Advances in Computer Science and Technology, vol. 3, no. 2, pp. 113-116, 2014.

[22] C-.D. Chang, C-.C. Wang, B.C. Jiang, "Using data mining techniques for multi-diseases prediction modeling of hypertension and hyperlipidemia by common risk factors," Expert Systems with Applications, no. 38, pp. 5507-5513, 2011.

[23] D.S.V. Mallawaarachchi, S.C. Wickremasinghe, L.C. Somatunga, V.T.S.K. Siriwardena, N.S. Gunawardena, "Healthy Lifestyle Centres: a service for screening noncommunicable diseases through primary health-care institutions in Sri Lanka," WHO South-East Asia Journal of Public Health, vol. 5, no. 2, pp. 89-95, 2016.

[24] Irfana P. Bhatti, Heman D. Lohano , Zafar A. Pirzado and Imran A. Jafri , 2006. A Logistic Regression Analysis of the Ischemic Heart Disease Risk. Journal of Applied Sciences, 6: 785-788.

[25] Asma A. AlJarullah," Decision Tree Discovery for the Diagnosis of Type II Diabetes", International Conference on Innovations in Information Technology , 2011.

[26] Dursun Delen, Glenn Walker, Amit Kadam," Predicting breast cancer survivability: a comparison of three data mining methods ", Elsevier, 2004.

[27] Abdelghani Bellaachia, Erhan Guven, "Predicting Breast Cancer Survivability Using Data Mining Techniques", Citeseer, 2006.