ISSN No. 0976-5697

Volume 7, No. 4, July-August 2016
International Journal of Advanced Research in Computer Science
RESEARCH PAPER
Available Online at www.ijarcs.info

A Novel Approach for Prediction of Type 2 Diabetes

Alby S
Research Scholar
Research and Development Centre
Bharathiar University, Coimbatore – 44,Indai

Dr. B L Shivakumar
Professor
Sri Ramakrishna Engineering College
Coimbatore – 22, Indai

Abstract: This paper proposes a new approach for the prediction of type2 diabetes. Many different techniques have been used for the prediction of chronic diseases by different researchers. Among them Adaptive Neuro Fuzzy Inference system (ANFIS) is very popular and already used for the prediction of type 2 diabetes. In this paper, the proposed system is optimization of ANFIS using Mine Blasting Algorithm(MBA) which reduces the complexity of ANFIS and increases the accuracy of prediction.

Keywords: Diabetes ,Adaptive Neuro Fuzzy Inference system, Data Mining, Mine Blasting Algorithm, Prediction.

I. INTRODUCTION

In the present world, as the population increases the number of diseases has also increased drastically. A new trend in lifestyle which is resulting in overweight and obesity is one of the main causes behind the steep increase in health issues. People suffering from chronic diseases are also increasing worldwide. Prevention is better than cure; hence if we can predict the occurrence of ill health and chronic diseases, it can be prevented at the starting stage or before the occurrence.

Many researches have been done and many are still going on in this field, because of the high level of importance. Of the many chronic diseases, diabetes comes on the top level and this is a chronic, progressive non communicable disease which is if not treated properly becomes dangerous. According to WHO report 2016, the number of people living with diabetes has almost quadrupled since 1980 to 422 million adults[1].

For the prediction of any disease, a huge amount of data has to be analyzed. For this analysis Data Mining(DM) techniques is the most reliable one .Many researchers have used different DM techniques.

A. Data Mining

Data Mining (DM) can be viewed as a result of the natural evolution of information technology.[2] In recent years the information industry deals with a huge amount of data and needs turning such data into useful information and knowledge. The information and knowledge gained can be used for a wide range of applications. Data mining refers to extracting knowledge from large amounts of data. It is also treated as knowledge discovery in database (KDD). Knowledge discovery as a process is depicted in Fig. 1 and consists of an iterative sequence of the following steps:

1. Data cleaning (to remove noise and inconsistent data).
2. Data integration (where multiple data sources may be combined).
3. Data selection (where data relevant to the analysis task are retrieved from the database).
4. Data transformation (where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance).
5. Data mining (an essential process where intelligent methods are applied in order to extract data patterns).
6. Pattern evaluation (to identify the truly interesting patterns representing knowledge based on some interestingness measures).
7. Knowledge presentation (where visualization and knowledge representation techniques are used to present the mined knowledge to the user).

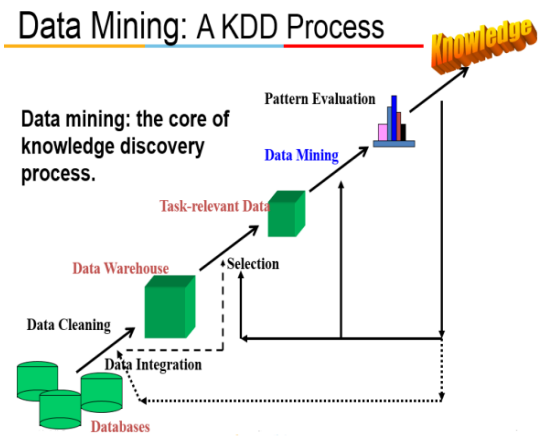


Fig.1: KDD process

B. Classification of Data Mining Systems

Data mining is an interdisciplinary field, the confluence of a set of disciplines, including database systems, statistics, machine learning, visualization, and information science (Figure 1.2).Moreover, depending on the data mining approach used, techniques from other disciplines may be applied, such as neural networks, fuzzy and/or rough set theory, knowledge representation, inductive logic programming, or high-performance computing. Depending on the kinds of data to be mined or on the given data mining application, the data mining system may also integrate techniques from spatial data analysis, information retrieval, pattern recognition, image analysis, signal processing, computer graphics, Web technology, economics, business, bioinformatics, or psychology. Because of the diversity of disciplines contributing to data mining, data mining research

is expected to generate a large variety of data mining systems.

C. Data Mining Techniques

Several core techniques that are used in data mining describe the type of mining and data recovery operation.

Association is one of the best known data mining technique. In association, a pattern is discovered based on a relationship between items in the same transaction. That's the reason why association technique is also known as relation technique. The association technique is used in market basket analysis to identify a set of products that customers frequently purchase together.

Classification is a classic data mining technique based on machine learning. Basically classification is used to classify each item in a set of data into one of predefined set of classes or groups. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics.

Clustering is a data mining technique that makes meaningful or useful cluster of objects which have similar characteristics using automatic technique. The clustering technique defines the classes and puts objects in each class, while in the classification techniques, objects are assigned into predefined classes.

Prediction, as it name implied, is one of a data mining techniques that discovers relationship between independent variables and relationship between dependent and independent variables.

Sequential patterns analysis is one of data mining technique that seeks to discover or identify similar patterns, regular events or trends in transaction data over a business period.

Decision tree is one of the most used data mining techniques because its model is easy to understand for users. In decision tree technique, the root of the decision tree is a simple question or condition that has multiple answers. Each answer then leads to a set of questions or conditions that help us determine the data so that we can make the final decision based on it.

D. Data Mining Tasks

Data mining tasks are mainly classified into two broad categories:

- Predictive model
- Descriptive mode

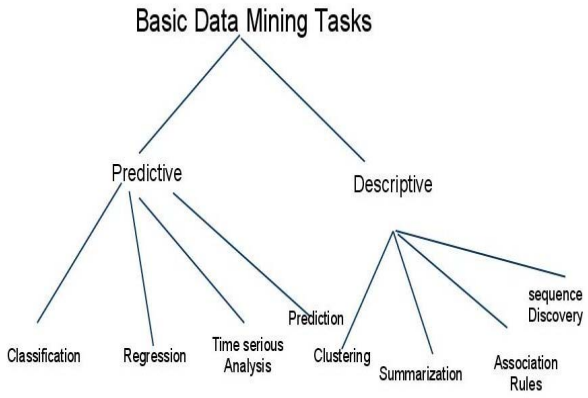


Fig.2: Data Mining Tasks

E. Applications of Data Mining

Data mining is widely used in diverse areas.

- Financial Data Analysis
- Retail Industry
- Telecommunication Industry
- Biological Data Analysis
- Other Scientific Applications
- Intrusion Detection

Data mining applications *are continuously developing in* various industries to provide more hidden knowledge that increases business efficiency and grows businesses. In recent times, there is a tremendous growth in the field of biology such as genomics, proteomics, functional Genomics and biomedical research. Biological data mining is a very important part of Bioinformatics.

F. Data Mining in Bioinformatics

Bioinformatics can be defined as the application of computer technology to the management of biological information[3]. Bioinformatics is the science of storing, extracting, organizing, analyzing, interpreting and utilizing information from biological sequences and molecules[4]. A particular active area of research in bioinformatics is the application and development of data mining techniques to solve biological problems. Analyzing large biological data sets requires making sense of the data by inferring structure or generalizations from the data. Applications of data mining to bioinformatics include gene finding, protein function domain detection, function motif detection, protein function inference, disease diagnosis, disease prognosis, disease treatment optimization, protein and gene interaction network reconstruction, data cleansing, and protein sub-cellular location prediction.

In the future of medicine, there exists a universal bioinformatics-based healthcare system in which physicians are knowledgeable in computer science technology and patient records (for example, genomes and proteomes) are stored on huge online data warehouses. Bioinformatics, more specifically translational bioinformatics, is the missing link between futuristic healthcare and modern computational technology. Over the last few decades, life-threatening

diseases, such as cancer, diabetes and the escalating cost of drug development have combined to increase patient suffering. The growing healthcare burden can be significantly reduced by the design and development of novel methods in translational bioinformatics and allied health information-driven sciences.

G. Data Mining in Healthcare

In present era, Data Mining is becoming popular in healthcare field because there is a need of efficient analytical methodology for detecting unknown and valuable information in health data. In health industry, Data Mining provides several benefits such as detection of the fraud
in health insurance, availability of medical solution to the patients at lower cost, detection of causes of diseases and identification of medical treatment methods. It also helps the healthcare researchers for making efficient healthcare policies, constructing drug recommendation systems, developing health profiles of individuals etc. The data generated by the health organizations is very vast and complex due to which it is difficult to analyze the data in order to make important decision regarding patient health. This data contains details regarding hospitals, patients, medical claims, treatment cost etc. So, there is a need to generate a powerful tool for analyzing and extracting important information from this complex data. The analysis of health data improves the healthcare by enhancing the performance of patient management tasks. The outcome of Data Mining technologies is to provide benefits to healthcare organization for grouping the patients having similar type of diseases or health issues so that healthcare organization provides them effective treatments. It can also useful for predicting the length of stay of patients in hospital, for medical diagnosis and making plan for effective information system management. Recent technologies are used in medical field to enhance the medical services in cost effective manner. Data Mining techniques are also used to analyze the various factors that are responsible for diseases for example type of food, different working environment, education level, living conditions, availability of pure water, health care services, cultural ,environmental and agricultural factors .

The Healthcare industry is generally “information rich”, which is not feasible to handle manually. These large amounts of data are very important in the field of Data Mining to extract useful information and generate relationships amongst the attributes. The doctors and experts available are not in proportion with the population. Also, symptoms often be neglected. Heart disease diagnosis is a complex task which requires much experience and knowledge. Heart disease is a single largest cause of death in developed countries and one of the main contributors to disease burden in developing countries. In the health care industry the data mining is mainly used for predicting the diseases from the datasets. The Data Mining techniques, namely Decision Trees, Naive Bayes, Neural Networks, Associative classification, Genetic Algorithm are analyzed on Heart disease database. Cancer is the another most important cause of death for both men and women. The early detection of cancer can be helpful in curing the disease completely. So the requirement of techniques to detect the occurrence of cancer nodule in early stage is increasing. Predicting outcome of a disease is a challenging task. Data mining techniques tends to simplify the prediction segment. Automated tools have made it possible to collect large volumes of medical data, which are made available to the medical research groups. The results being an increasing popularity of data mining techniques to identify patterns and relationship among large number of variables, which make it possible to predict the outcome of the disease using pre-existential datasets.

In today’s world, one of the major public health challenges is Diabetic Mellitus. Diabetes is nearly four times as common as all types of cancer combined and causes more deaths than breast and prostate cancer combined. It is fast becoming the 21st century's major public-health concern. Diabetes is unlike other diseases, there are a lot of other components to diabetes, such as: the diabetes disease process, nutritional management, physical activity, medications, glucose monitoring, and psychosocial adjustment. Diabetes is a serious, sometimes life-threatening disease. Over time it can affect every body part and may cause kidney damage, nerve damage, amputations and blindness. It also raises your risks for heart and blood vessel disease and stroke. In fact, at least 65% of people with diabetes die from heart disease or stroke, according to the National Institutes of Health. Pregnant women with diabetes have a higher risk of delivering babies with birth defects than women who don't have diabetes.

According to WHO report, 347 million people worldwide have diabetes. WHO projects that diabetes will be the 7th leading cause of death in 2030. In 2012, an estimated 1.5 million deaths were directly caused by diabetes. More than 80% of diabetes deaths occur in low- and middle-income countries.

In 2000, India (31.7 million) topped the world with the highest number of people with diabetes mellitus followed by China (25.8 million) with the United States (17.7 million) in second and third place respectively. It is predicted that by 2030 diabetes mellitus may afflict up to 79.4 million individuals in India. It is estimated that 61.3 million people aged 20-79 years live with diabetes in India (2011 estimates). This number is expected to increase to 101.2 million by 2030.

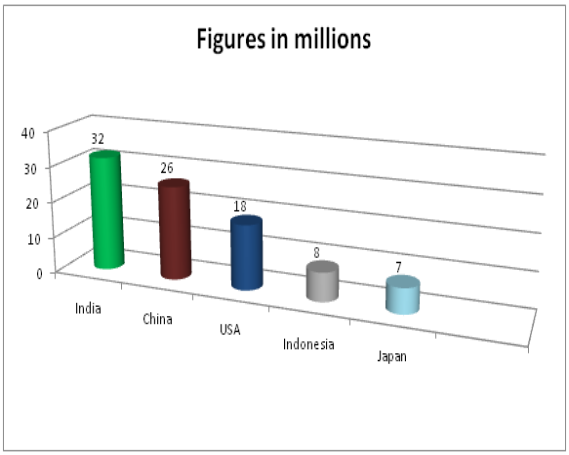


Fig.3. Statistics of people with diabetes

H. Diabetes

Diabetes is a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces. Insulin is a

hormone that regulates blood sugar. Hyperglycemia, or raised blood sugar, is a common effect of uncontrolled diabetes and over time leads to serious damage to many of the body's systems, especially the nerves and blood vessels.

Types of diabetes: There are three types of diabetes. Type 1 diabetes, Type 2 diabetes and Gestational diabetes.**Type 1 diabetes** (previously known as insulin-dependent, juvenile or childhood-onset) is characterized by deficient insulin production and requires daily administration of insulin. The cause of type 1 diabetes is not known and it is not preventable with current knowledge. Symptoms include excessive excretion of urine (polyuria), thirst (polydipsia), constant hunger, weight loss, vision changes and fatigue. These symptoms may occur suddenly. Type 2 diabetes (formerly called non-insulin-dependent or adult-onset) results from the body's ineffective use of insulin.

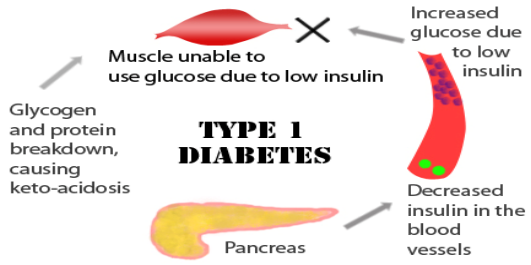


Fig.4: Type 1 Diabetes

Type 2 diabetes comprises 90% of people with diabetes around the world, and is largely the result of excess body weight and physical inactivity. Symptoms may be similar to those of Type 1 diabetes, but are often less marked. As a result, the disease may be diagnosed several years after onset, once complications have already arisen. Until recently, this type of diabetes was seen only in adults but it is now also occurring in children.

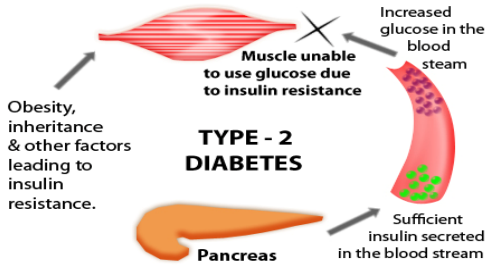


Fig.5: Type 2 Diabetes

Gestational diabetes is hyperglycemia with blood glucose values above normal but below those diagnostic of diabetes, occurring during pregnancy. Women with gestational diabetes are at an increased risk of complications during pregnancy and at delivery. They are also at increased risk of type 2 diabetes in the future.

Diabetes is a silent killer. If uncontrolled, it can lead to deadly complications.

II. LITERATURE REVIEW

Milan Zormana, in literature [5] addressed the problem of mining rules from the diabetes database using a combination of decision trees and association rules. About 1251 different cases from original database with selected attributes were considered and with the help of association rule approaches, different trees are built and converted them into different set of rules and these rules where further reduced and filtered. The main objective of this paper was to analyze the number of rules which are generated and how many rules will be balanced after performing filtering and reduction. It also analyze how many rules will be generated employing association rule approach on the same database and the conclusion is that, the sets of rules built by decision trees were much smaller than results of association rules. The literature [6] analyzed the Pima Indian diabetes data sets using the tool Rapid Miner. Discovered the hidden relationships between Plasma Glucose and Class attributes which finds that the patients with higher Plasma- Glucose values are having more chance to develop diabetes and with low Plasma – Glucose values have less chance to develop diabetes nearby future. The paper [7] discussed the potential of applying the apriori – Gen algorithm to the association study for the type – 2 diabetes. The relative risk (RR), which is the risk of developing a disease relative to exposure and odds ratio (OR) , which is the ratio of the odds of an event occurring in control group, are used to prove that interaction of Multi SNPs is associated with the disease. B. M. Patil, in literature [8] introduced a new approach to generate association rules on numeric data. They used pre-processing to improve the quality of data by handling the missing values and applied equal interval binning with approximate values based on medical expert's advice to Pima Indian diabetes data. Lastly apriori association rule algorithm is applied to generate the rules. Only type- 2 diabetic patients those who are pregnant woman below 21 years are included in their study. . S.M.Nuwangi, in paper [9] used advanced and reliable data mining techniques to identify different risk factors behind the diabetes and the relationship between the diabetes and the other diseases. Using association rule generation, the relationship between edema and diabetes and wheezes and diabetes has been identified. The result shows, the females aged between 39 – 75 years with normal BMI range, systolic BP range and diastolic BP range and having wheezes will have a high risk towards developing high FBS (fasting blood sugar) level. Palivela Hemant at [10] combines K-means clustering with various different classification algorithms like SMO, Naive Bayes, Bagging, AdaBoost, J48, Rotation Forest and Random Forest to predict the positive and negative of disease. The data consists of 768 different entries in accordance with attributes like Skin , Mass , Age , Insulin , pregnant etc are

used . SMO implements the sequential minimal optimization algorithms for training a support vector classifier. Missing values are replaced globally, nominal attributes are transformed into binary ones and attributes are normalised. Bagging bags a classifier to reduce variance. By using various classifiers, the authors propose a hybrid model for the prediction of the positivity and negativity of the diabetes. Bum Ju Lee, in paper [11], did a study among a total of 4870 subjects to predict the fasting plasma glucose status that is used in the diagnosis of type -2 diabetes by a combination of various anthropometric measures that are measured in a greater no of specific sites in the body can improve the predictive power of diagnosing type 2 diabetes. According to their findings it is indicated that a combination of anthropometric measures can clearly improve the predictive power for normal and high FPG status when compared with individual measures and prediction experiments using balanced data of normal and high FPG subject can improve the prediction performance and reduce the intrinsic bias of the model towards the majority class. The research [12] was based on three techniques of Expectations – Maximisation (EM) algorithm, H – means clustering and Genetic algorithm (GA). Pima Indian Diabetes Datasets were used on WEKA software tool. About 35 % of the total of 768 test samples was found with diabetes presence.

Sonukumari and Archana Singh, in paper [13], tried to propose an intelligent and effective methodology for the automated detection of Diabetes Mellitus based on neural network. A survey has been done among 100 data sets which include people from different age groups, gender and life style. Around 13 parameters like gender, age, weight, height, thirst increase, hunger increase, appetite increase, vomiting etc along with the possible valued are fed in the neural network system. The output is in the binary form. The value zero means the person is not affected from DM and if it is one, it reveals that the person is suffering from DM. As a result of this study the authors proved that their neural network system is having an accuracy rate of 92.8 %.In paper [14] Adaptive Neuro Fuzzy Inference system (ANFIS) is used for the diagnosis of diabetes. The input nodes in neural network are constructed based on the input attribute. The hidden nodes are used to classify given input based on the training dataset with the help of AGKNN. The paper [15] presents an intelligent expert based system ANFISGA for the dosage planning for type-2 diabetes male patients. Two artificial intelligence techniques ANFIS and GA were combined. The study [16] assessed the association between the HW phenotype and type-2 diabetes and evaluated the predictive powers of combined anthropometric measurements and TG levels based on machine learning.

III. PROPOSED SYSTEM FOR PREDICTION OF DIABETES

For the prediction of diabetes many researches have been conducted. The propsed approach for the prediction of diabetes is optimization of ANFIS using Mine Blast Algorithm(MBA). ANFIS has been popular and widely used in medical field. Use of optimization mechanism reduces the complexity of ANFIS and increases the accuracy of ANFIS.

A. Adaptive network based fuzzy inference system (ANFIS)

Concept and Structure:ANFIS combines the advantages of two intelligent approaches neural network and fuzzy logic to allow good reasoning in quantity and quality. A network obtained has an excellent ability of training by means of neural networks and linguistic interpretation of variables via fuzzy logic. The both of them encode the information in parallel and distribute architecture in a numerical framework.

Rule: if x is A1 and y is B1 then f (x) = px + qy + r

Where x and y are the inputs, A and B are the fuzzy sets, f are the output, p, q and r are the design parameters that determined during the training process. ANFIS is composed of two parts is the first part is the antecedent and the second part is the conclusion, which are connected to each other by rules in network form. Five layers are used to construct this network. Each layer contains several node sits structure shows in figure 1.

layer1: executes a fuzzification process which denotes membership functions (MFs) to each input. In this paper we choose Gaussian functions as membership functions:

$$O_i^1 = \mu_{Ai} = \exp\left(\frac{-(x-c)^2}{\sigma^2}\right)$$

layer2: executes the fuzzy AND of antecedents part of the fuzzy rules

$$O_i^2 = w_i = \mu_{Ai}(x_1) \times \mu_{Bi}(x_2), i = 1,2,3,4$$

layer3: normalizes the MFs

$$O_i^3 = \overline{w_i} = \frac{w_i}{\sum_{j=1}^{4} w_j}, i = 1,2,3,4$$

layer4: executes the conclusion part of fuzzy rules

$$O_i^4 = \overline{w_i} y_i = \overline{w_i}\left(\alpha_1^i x_i + \alpha_2^i x_2 + \alpha_3^i\right), i = 1,2,3,4.$$

layer5: computes the output of fuzzy system by summing up the outputs of the fourth layer which is the defuzzification process.

$$O_i^5 = overll\_output = \sum_{i=1}^{4} \overline{w}_i y_i = \frac{\sum_{i=1} w_i y_i}{\sum_{i=1}^{4} w_i}$$

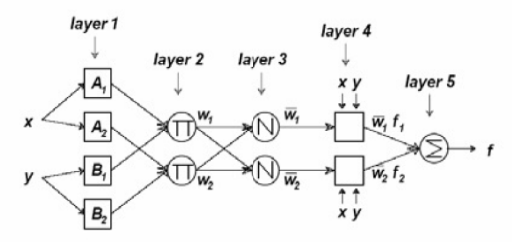


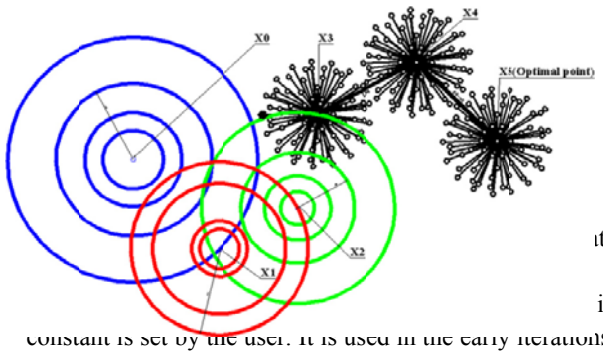
Fig.6: ANFIS architecture

Circles in ANFIS represent fixed nodes that predefined operators to their inputs and no other parameters but the input participate in their calculations. While square are the representative for adaptive nodes that affected by internal parameters.

B. Mine blast algorithm

Basic Concepts : MBA is a recently developed optimization method used for handling complex optimization problems introduced by Sadollah [17]. The fundamental concepts and ideas of MBA are derived by the explosion of mines where the thrown shrapnel pieces collide with other landmines near the explosion area resulting in their explosion. Thus, the goal is to find the most explosive landmine (min or max) located at the optimal point(Xf).

.



ıtion for is a constant is set by the user. It is used in the early iterations of the algorithm as below:

If μ = I then [I is the iteration number index]
Exploration
Else
Exploitation

The proposed technique for optimizing rule-base of ANFIS and tuning its parameters used MBA as an optimizer.

IV.CONCLUSION

This paper proposes optimization of ANFIS using MBA as an optimizer for the prediction of type 2 diabetes. In ANFIS, all the rules are not potential rules. So for the efficiency, it is very important to optimize the rules. This paper suggests a better optimization algorithm MBA which can be used as an optimizer and this will reduce the network's complexity and computational cost.

V.REFERNCES

[1] www.who.int/mediacentre/news/releases/2016

[2] Jiawei Han and Micheline Kamber, "Data Mining : Concepts and Techniques", 2nd edition.

[3] Khalid Raza,"Application of data mining in bioinformatics", Indian Journal of Computer Science and Engineering, Vol 1 No 2, pp 114-118.

[4] Mohammed J. Zaki, Jason T. L. Wang,Hannu T.T. Toivonen, "BIOKDD01: Workshop on Data Mining in Bioinformatics",www.kdd.org/sites/default/files/issues/3-2-2002-01/zaki.pdf.

[5] Milan Zormana, Gou Masudab, Peter Kokola, Ryuichi Yamamotob, Bruno Stiglica, "Mining Diabetes Database With Decision Trees and Association Rules", CBMS,Proc. of the 15th IEEE symposium , pp 134-139,2002.

[6] Jianchao Han, Juan C. Rodriguze, Mohsen Beheshti, "Diabetes Data Analysis and Prediction Model Discovery Using RapidMiner", Proc. of the Second International Conference on Future Generation Communication and Networking, Vol. 3,pp 96-99, 2008.

[7] Weidong Mao, Jinghe Mao, "The Application of Apriori-Gen Algorithm in the Association Study in Type 2 Diabetes", Proc. of the 3rd International Conference Bioinformatics and Biomedical Engineering(ICBBE 2009), pp 1-4, 2009.

[8] B. M. Patil, R. C. Joshi, Durga Toshniwal, "Association rule for classification of type -2 diabetic patients", Proc. of the Second International Conference on Machine Learning and Computing, pp 330-334, 2010.

[9] S.M.Nuwangi, C. R. Oruthotaarachchi, J.M.P.P. Tilakaratna & H. A. Caldera, "Usage of Association rules and Classification Techniques in Knowledge Extraction of Diabetes", Proc of the 6th International Conference on Advanced Information Management and Service(IMS), pp 372-377, 2010 .

[10] Palivela Hemant, Thotadara Pushpavathi, "A novel approach to predict by cascading clustering and classification", Proc. of the 3rd International Conference on Computing Communication & Networking Technologies, pp 1-7. 2012.

[11] Bum Ju Lee, Boncho Ku, Jiho Nam, Duong Duc Pham, Jong Yeol Kim, "Prediction of Fasting Plasma Glucose Status using Anthropometric Measures for Diagnosing Type 2 Diabetes", IEEE Journal of Biomedical and Health Informatics, Vol. pp, Issue. 9,page 1,TITB-00020-2013.

[12] C M Velu, K R Kashwan, "Visual Data Mining Techniques for Classification of Diabetes Patients". Proc. of the IEEE 3rd International Advance Computing Conference, pp. 1070-1075, 2013.

[13] Sonukumari, Archana Singh, " A data mining approach for the Diagnosis of Diabetes Mellitus" Proc. of the 7th International Conference on Intelligent Systems and Control,pp. 373-375,2013.

[14] C. Kalaiselvi, G. M. Nasira, "A New Approach for Diagnosis of Diabetes and Prediction of Cancer Using ANFIS". World

Congress on Computing and Communication Technologies (WCCCT), 2014 ,Pages: 188 – 190.

[15] C. P. Ronald Reagan; S. Prasanna Devi ,"An Android App for intelligent dosage planning in Type2 diabetes using ANFISGA", International Conference on Recent Trends in Information Technology (ICRTIT), 2014, Pages: 1 – 4.

[16] Bum Ju Lee, Jong Yeol Kim, Identification of Type 2 Diabetes Risk Factors Using Phenotypes Consisting of Anthropometry and Triglycerides based on Machine Learning" , IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, VOL. 20, NO. 1, JANUARY 2016 39.

[17] Sadollah, A., A. Bahreininejad, H. Eskandar, and M. Hamdi, "Mine blast algorithm for optimization of truss structures with discrete variables". Computers & Structures, 2012. **102**: p. 49-63.