Int Cardiovasc Res J.2017;**11**(2):71-76.icrj.10657

International Cardiovascular Research Journal
www.ircrj.com

Simple Prediction of Type 2 Diabetes Mellitus via Decision Tree Modeling

Mehrab Sayadi [1], Mohammadjavad Zibaeenezhad [2], Seyyed Mohammad Taghi Ayatollahi [1,*]

[1] Department of Biostatistics, School of Medicine, Shiraz University of Medical Sciences, Shiraz, IR Iran
[2] Cardiovascular Research Center, Shiraz University of Medical Sciences, Shiraz, IR Iran

ARTICLE INFO

Article Type:
Research Article

Article History:
Received: 02 Nov 2016
Revised: 07 Dec 2016
Accepted: 23 Jan 2017

Keywords:
Heart Disease
Decision Tree
Risk Factors
Screening Test

ABSTRACT

Background: Type 2 Diabetes Mellitus (T2DM) is one of the most important risk factors in cardiovascular disorders considered as a common clinical and public health problem. Early diagnosis can reduce the burden of the disease. Decision tree, as an advanced data mining method, can be used as a reliable tool to predict T2DM.

Objectives: This study aimed to present a simple model for predicting T2DM using decision tree modeling.

Materials and Methods: This analytical model-based study used a part of the cohort data obtained from a database in Healthy Heart House of Shiraz, Iran. The data included routine information, such as age, gender, Body Mass Index (BMI), family history of diabetes, and systolic and diastolic blood pressure, which were obtained from the individuals referred for gathering baseline data in Shiraz cohort study from 2014 to 2015. Diabetes diagnosis was used as binary datum. Decision tree technique and J48 algorithm were applied using the WEKA software (version 3.7.5, New Zealand). Additionally, Receiver Operator Characteristic (ROC) curve and Area Under Curve (AUC) were used for checking the goodness of fit.

Results: The age of the 11302 cases obtained after data preparation ranged from 18 to 89 years with the mean age of 48.1 ± 11.4 years. Additionally, 51.1% of the cases were male. In the tree structure, blood pressure and age were placed where most information was gained. In our model, however, gender was not important and was placed on the final branch of the tree. Total precision and AUC were 87% and 89%, respectively. This indicated that the model had good accuracy for distinguishing patients from normal individuals.

Conclusions: The results showed that T2DM could be predicted via decision tree model without laboratory tests. Thus, this model can be used in pre-clinical and public health screening programs.

1. Background

Diabetes causes a great risk for cardiovascular disorders and is associated with high rates of myocardial infarction and stroke (1, 2), such a way that mortality rate due to these disorders has reached up to 80% (3). Furthermore, diabetic individuals may develop heart disease 10 to 15 years earlier compared to normal ones, which accounts for premature mortality in diabetic individuals (4, 5).

Most diabetic patients (90 - 95%) suffer from type 2 diabetes, which is considered to be a controllable condition among other non-communicable diseases (6). Early diagnosis is the key factor for disease control and can reduce the related complications and high expenses (7). Studies have suggested a 5-year lag for diagnosis of Type 2 Diabetes Mellitus (T2DM) (8), which is associated with grave cardiovascular outcomes (7, 9), kidney diseases, renal failure, and other long-term vascular diseases (10, 11). Late diagnosis can also lead to lack of self-care behaviors since most T2DM patients are not aware of their disease (3, 12). Self-care programs are in fact essential strategies to reduce the disease burden in the society by means of early recognition of asymptomatic T2DM (3).

Data mining is one of the practical branches of artificial

*Corresponding author: Seyyed Mohammad Taghi Ayatollahi, Department of Biostatistics, School of Medicine, Shiraz University of Medical Sciences, Zand Blvd, Shiraz, Iran, Postal Code: 7134845794, Tel: +98-7132305884, *E-mail:* ayatolahim@sums.ac.ir

intelligence, discovers latent patterns by looking for relationships between features in large databases, and serves for diagnostic purposes as well as cost reduction in clinical contexts (13). Decision tree, as one of the data mining methods in advanced statistical contexts, can predict T2DM. This method is a simple reliable tool for diagnosing the disease before emergence of clinical symptoms (14). Decision trees in clinical practice lead to clear presentation of complex data and facilitate data interpretation and application. As a result, decision trees have been proven to be feasible for diagnostic purposes (15).

2. Objectives

The present study aims to present a simple model for predicting T2DM using decision tree thorium. Introducing this model as a screening tool for diabetes would express its benefit to public health programming.

3. Materials and Methods

2.1. Patients and Their Measurements

This analytical model-based study used a part of the data related to a cohort of 11302 cases from the database of Healthy Heart House of Shiraz, Iran. Shiraz is the fifth most populous city of Iran and the capital of Fars province with a population of over 2 million individuals. Shiraz is located in the Southwest of Iran, and is the center of patient referring in the south of the country. Participants were volunteer individuals aging above 18 years who were enrolled in Shiraz cardiovascular cohort study from 2014 to 2015. Demographic, clinical, and anthropometric data, such as age, gender, Body Mass Index (BMI), family history of diabetes, and systolic and diastolic blood pressure, were extracted from the database. Then, conventional risk factors for T2DM and cardiovascular disease were selected for analyzing and decision tree modeling. Participants with missing data were excluded from the study. All measurements were in line with the study protocol and were prepared before analysis. Fasting Blood Sugar (FBS) > 126 was considered to be diabetes and this type of diagnosis of T2DM was used as the binary datum. Among the 14321 records, 11302 fulfilled all the required data regarding the six candidate variables for entering the decision tree. These cases were subsequently subdivided to individuals with or without diabetes. It should be noted that inputs of the decision tree were age, gender, systolic and diastolic blood pressure, family history of diabetes, and BMI.

2.2. Decision Tree Model

There are many methods for classification in multivariate approach, including discriminate analysis, artificial networks, and regression models, especially logistic regression and fuzzy logistic regression (16, 17). Regression model is a suitable statistical method for modeling with respect to controlling confounders. However, this does not imply that controlling all confounders is essential in disease prediction. In addition to confounders, important variables may be excluded from the final model. Decision tree, as one of data mining and artificial intelligence methods, is a tree structure in form of a diagram. One advantage of decision tree modeling is that it is a reliable and simple variable selection tool for clinical practice. Another advantage of using decision trees is providing clear results from sophisticated data that allows its simple application in clinical practice. Studies have shown that decision tree is a reasonable method for diagnostic plans (18). Decision tree is used as a method for classification in advanced statistical methods context by presenting a tree model including some nodes. Root and internal nodes are test cases that are used to divide samples to different groups. Besides, internal nodes are the result of variable test cases and leaf nodes denote the class variable. Various decision tree algorithms are available to classify data, including C4.5, C5, J48, CART, and CHAID. In this paper, J48 decision tree algorithm (19) was chosen to run the model. Each node for the decision tree is found by considering the highest information gain for all variables. If a variable gives a clear end product, the branch of this variable is terminated and the target value is assigned to it. Overall, decision tree provides a powerful technique for prediction of diabetes diagnosis problem. Considering these advantages, this technique was used for diabetes classification in the present study. It was used for predicting T2DM, which is simple to be used before the onset of subclinical symptoms. To build decision trees, WEKA software (3.7.5 version, New Zealand) was applied (20). In doing so, 10-fold cross-validation as well as 30-70% train and test method were used for model checking. The results of the decision tree model were reported as accuracy and precision indices as well as True Positive (TP), False Positive (FP), and Positive Predictive Value (PPV) (Recall) indices.

2.3. Other Statistical Methods

Receiver Operator Characteristic (ROC) curve, recall, and Area Under Curve (AUC) were utilized for assessment of goodness of fit. Moreover, logistic regression model was used using the SPSS statistical software, version 22 (SPSS Inc, Chicago, IL, USA) to identify the association between diabetes and study parameters. Categorical variables were expressed as numbers (percentages) and continuous ones as means ± standard deviation (SD). Normally distributed parametric variables were compared using unpaired student t-test, while nonparametric ones were compared using chi-square test. P-values less than 5% were considered to be statistically significant.

4. Results

The participants' ages ranged from 18 to 89 years, with the mean age of 48.1 ± 11.4 years. Additionally, more than half of the participants (51.1%) were male. The prevalence of diabetes in the whole sample was 12.1%. Demographic and clinical routine data have been presented in Table 1. The results of univariate analysis revealed that all continuous variables were higher in diabetic patients in comparison to healthy individuals. Moreover, the results of multiple logistic regressions indicated that diabetes was predictable by all variables, except for gender (Table 2). Then, WEKA software was used to run the decision tree model. A total of 11302 records were used for data mining, and it took 0.23 seconds to build the model.

The rate of correctly classified samples was 89% and the model was more precise in identifying healthy individuals than patients. The total precision, recall, and accuracy of the model were 87%, 89%, and 88%, respectively. Details of the model measures have been shown in Table 3.

Table 1. Demographic and Clinical Routine Characteristics of the Study Participants

Variable	Diabetics (n = 1363)	Healthy Individuals (n = 9939)	P value
Age (year), Mean ± SD	56.6 ± 9.4	46.9 ± 11.1	< 0.001
Gender (male), n (%)	784(57.5)	4996(50.3)	< 0.001
BMI, Mean ± SD	28.2 ± 4.2	26.2 ± 4.1	< 0.001
BP (diastolic), Mean ± SD	89.0 ± 9.4	75.0 ± 11.0	< 0.001
BP (systolic), Mean ± SD	143.9 ± 17.7	122.5 ± 15.6	< 0.001
Family history, n (%)	519(38.1)	1662(16.7)	< 0.001

Abbreviations: BMI, body mass index; BP, blood pressure

Table 2. The Results of Logistic Regression Analysis on Diabetics

Variable	Coefficient	SE	OR	95%CI for OR		P value
Age (year)	0.061	0.004	1.063	1.056	1.071	< 0.001
Gender (female = ref)	-0.024	0.072	0.976	0.849	1.123	0.739
BMI	0.060	0.008	1.062	1.045	1.079	< 0.001
BP (diastolic)	0.085	0.004	1.089	1.079	1.098	< 0.001
BP (systolic)	0.016	0.003	1.016	1.010	1.021	< 0.001
Family history	0.695	0.074	2.004	1.734	2.316	< 0.001

Abbreviations: SE, standard error of mean; OR, odds ratio; CI, confidence interval; BMI, body mass index; BP, blood pressure

Table 3. Detailed Accuracy for Healthy Individuals and Diabetics Patients Obtained from the Decision Tree Model

Samples	TP Rate	FP Rate	Precision	Recall	F-measure	ROC Area
Healthy individuals	0.95	0.58	0.92	0.95	0.93	0.89
Diabetic patients	0.41	0.04	0.56	0.41	0.47	0.89
Average	0.89	0.52	0.87	0.89	0.88	0.89

Abbreviations: TP, true positive; FP, false positive; ROC, receiver operator characteristic

The area under the ROC curve reached 0.890 (Figure 1). As the figure depicts, the capability of the model was high, especially in identification of healthy individuals. The tree included 36 leaves and the size of the tree was 71. Diastolic blood pleasure was placed at the root node of the tree due to higher information gain, followed by age. At first, the tree was split into two branches; i.e., ≤ 80 and > 80 mmHg for diastolic blood pressure. Then, it was divided into two branches; i.e., ≤ 47 and > 47 years old. Family history of diabetes and systolic blood pressure were located at the next level of the tree (Figure 2). In this tree, gender was not as important as other variables and it was located in the last level.

5. Discussion

T2DM can be controlled by medication and life style modifications, such as special diets, if it is diagnosed before becoming complicated. Diabetes is known as a silent disease because clinical manifestations are absent until appearance of complications (21). Neuropathy, cataracts,

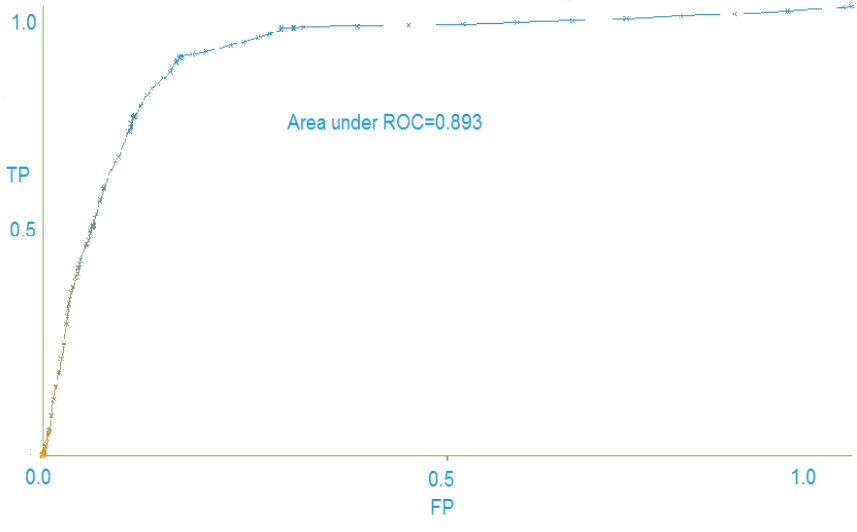

Figure 1. Receiver Operator Characteristic (ROC) Curve for Separating the Patients and Healthy Individuals Considering True Positive (TF) and False Positive (FP)

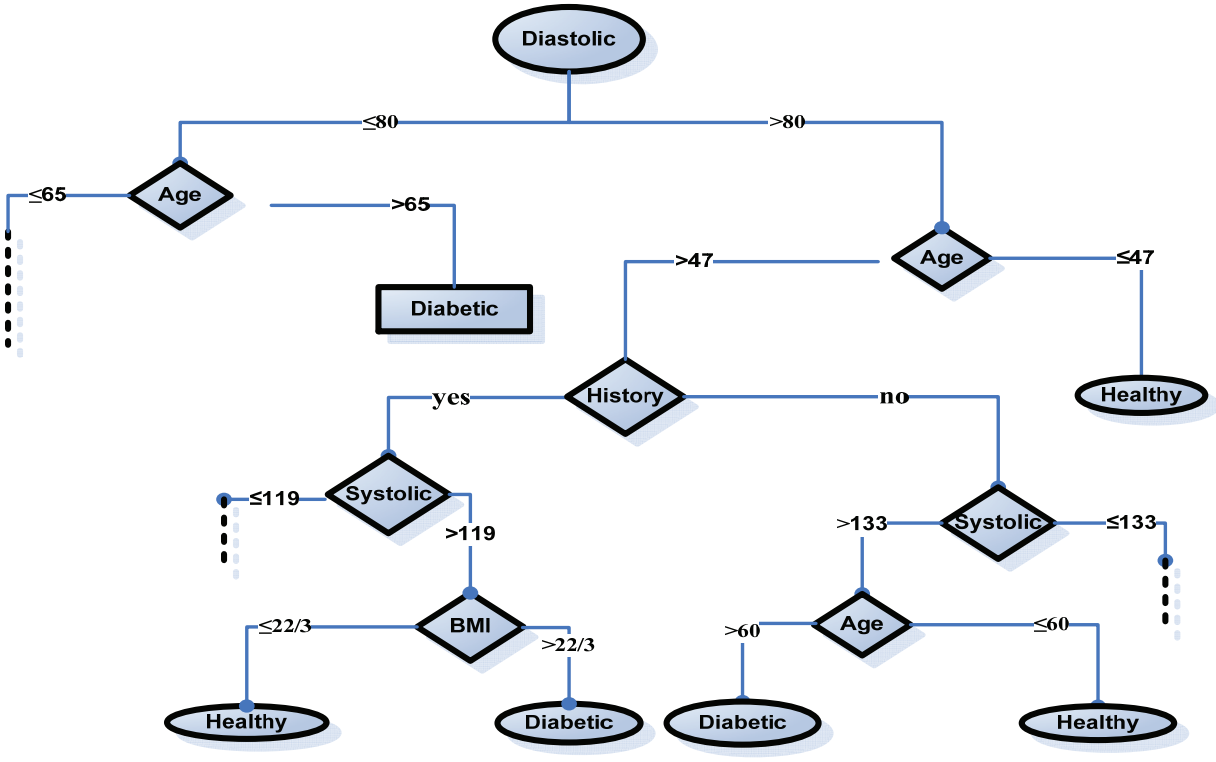

Figure 2. Decision tree results. Diastolic blood pleasure was placed at the root node of the tree due to higher information gain, which was followed by age. Family history of diabetes and systolic blood pressures were placed at the next level of the tree. In case of large number of leaves, they have been indicated by dashes and only nodes at higher levels have been displayed.

increased risk of kidney diseases, heart attack, and stroke (21, 22), which are associated with T2DM, necessitate a comprehensive population approach toward early diagnosis and management. Measurement of FBS to detect diabetes (23) is only performed after occurrence of symptoms and may be a late presentation of abnormal glucose metabolism. Although molecular tests capable of very early detection of deranged glucose metabolism have been introduced, they are not generally available as a routine diagnostic tool (22, 24). Besides, individuals need to be rechecked frequently because there will be no guarantee for those whose tests have been negative recently. Hence, a simple model of prediction that can easily be used by health providers can be potentially very useful in terms of individual patient care and population based surveillance. This was the main focus of the present study, which introduced such a model in a cohort of Iranian population.

Similar to this study, some studies provided models for early detection of T2DM and prevention of the potential complications related to late diagnosis. The results of our study were in line with those of the previous ones, and the six variables mentioned in our study were a replication of the previous reports (25, 26). Most of these studies have demonstrated that age, blood pressure, BMI, and family history of DM, as the most important risk factors, as well as laboratory data could be used for diagnosis of diabetes (14, 22). One advantage of our study was utilization of available variables instead of lab data in the prediction model.

Some studies have investigated other risk factors, such as eating habits and physical activity, for T2DM prediction (20). However, some of these studies have used data mining for predicting T2DM and have reported higher accuracy and precision values (27, 28). In the current study, we reported a highly accurate and precise model, but the precision and recall index was lower in patients than in healthy individuals. These results were in agreement with those of other studies (14, 26). However, the precision of our model was higher compared to the recall index, which is similar to most studies conducted on the issue (26, 29). Hence, our model can be used both as a screening tool in public health and as a diagnostic clinical test. The variables used in our study were indeed the same as those used in diabetes screening and recommended by most health policymakers and providers (30). Similar to our study, some studies have confirmed that these variables are important as prediction variables (31, 32).

Contemporary practice includes multiple diagnostic tests, such as HbA1C, Fasting Plasma Glucose (FPG), and Oral Glucose Tolerance Test (OGTT), for diabetic patients care. Therefore, recent studies, especially those that had used the variables related to the main diagnosis of diabetes, have compared the capability of the decision tree model and indicated that they were not to be considered in diabetes prediction or diagnosis. Nonetheless, the present study applied the decision tree using real data and the variables used in screening programs in primary healthcare and surveillance system, except for waist circumferences, without the need for any lab data that was a strength of our study. However, waist circumference could not be used because it was not available, which was one our study limitations. Overall, the results showed that this model

could be used in screening programs. It could also be designed as a computer-based program for automatically screening of the work field.

5.1. Conclusion

The present study used the decision tree model for screening of T2DM without laboratory tests. This model had some advantages, such as using big real data that were obtained from a cohort study. It should also be noted that this model was different from other screening or diagnostic tests. Moreover, instead of the lab data applied in primary healthcare or surveillance system, some routine variables were employed. Therefore, this study is a key step towards early diagnostic screening test of diabetes without using diagnostic laboratory tests and can consequently be used in pre-clinical and public health screening programs.

Acknowledgements

The present paper was extracted from the cohort data in Shiraz Healthy Heart House database. Hereby, the authors would like to acknowledge Ms. Houri Mosavinezhad for her assistance in improving the use of English.

Authors' Contribution

Mehrab sayadi, Mohammad Javad Zibaeenezhad, and Seyyed Mohammad Taghi Ayatollahi: providing the final draft, study supervision; Mehrab Sayadi: preparing the data, data analysis, and providing the first draft.

Funding/Support

This study was supported by Shiraz University of Medical Sciences, Shiraz, Iran.

Financial Disclosure

There is no financial disclosure.

References

1. Ansar H, Mazloom Z, Kazemi F, Hejazi N. Effect of alpha-lipoic acid on blood glucose, insulin resistance and glutathione peroxidase of type 2 diabetic patients. *Saudi medical journal*. 2011;**32**(6):584-8.
2. Sattar N, Gaw A, Scherbakova O, Ford I, O'Reilly DS, Haffner SM, *et al.* Metabolic syndrome with and without C-reactive protein as a predictor of coronary heart disease and diabetes in the West of Scotland Coronary Prevention Study. *Circulation*. 2003;**108**(4):414-9.
3. Zibaeenezhad MJ, Aghasadeghi K, Bagheri FZ, Khalesi E, Zamirian M, Moaref AR, *et al.* The Effect of Educational Interventions on Glycemic Control in Patients with Type 2 Diabetes Mellitus. *Int Cardiovasc Res J*. 2015;**9**(1):17-21.
4. Gazzaruso C, Solerte SB, Pujia A, Coppola A, Vezzoli M, Salvucci F, *et al.* Erectile dysfunction as a predictor of cardiovascular events and death in diabetic patients with angiographically proven asymptomatic coronary artery disease: a potential protective role for statins and 5-phosphodiesterase inhibitors. *Journal of the American College of Cardiology*. 2008;**51**(21):2040-4.
5. Zinman B, Wanner C, Lachin JM, Fitchett D, Bluhmki E, Hantel S, *et al.* Empagliflozin, Cardiovascular Outcomes, and Mortality in Type 2 Diabetes. *The New England journal of medicine*. 2015;**373**(22):2117-28.
6. American Diabetes Association. Diagnosis and classification of diabetes mellitus. *Diabetes care*. 2010;**33**(Supplement 1):S62-S9.
7. Heydari I, Radi V, Razmjou S, Amiri A. Chronic complications of diabetes mellitus in newly diagnosed patients. *International journal of diabetes mellitus*. 2010;**2**(1):61-3.
8. Harris MI, Eastman RC. Early detection of undiagnosed diabetes mellitus: a US perspective. *Diabetes/metabolism research and reviews*. 2000;**16**(4):230-6.
9. Karter AJ, Stevens MR, Herman WH, Ettner S, Marrero DG, Safford MM, *et al.* Out-of-pocket costs and diabetes preventive services: the Translating Research Into Action for Diabetes (TRIAD) study. *Diabetes Care*. 2003;**26**(8):2294-9.
10. Dreyer G, Hull S, Aitken Z, Chesser A, Yaqoob MM. The effect of ethnicity on the prevalence of diabetes and associated chronic kidney disease. *QJM : monthly journal of the Association of Physicians*. 2009;**102**(4):261-9.
11. Khajehdehi P, Malekmakan L, Pakfetrat M, Roozbeh J, Sayadi M. Prevalence of chronic kidney disease and its contributing risk factors in southern Iran: a cross-sectional adult population-based study. *Iranian journal of kidney diseases*. 2014;**8**(2):109-15.
12. Saudek CD, Herman WH, Sacks DB, Bergenstal RM, Edelman D, Davidson MB. A new look at screening and diagnosing diabetes mellitus. *The Journal of Clinical Endocrinology & Metabolism*. 2008;**93**(7):2447-53.
13. Witten IH, Frank E, Hall MA, Pal CJ. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann; 2016.
14. Hische M, Luis-Dominguez O, Pfeiffer AF, Schwarz PE, Selbig J, Spranger J. Decision trees as a simple-to-use and reliable tool to identify individuals with impaired glucose metabolism or type 2 diabetes mellitus. *European journal of endocrinology*. 2010;**163**(4):565-71.
15. Stern SE, Williams K, Ferrannini E, DeFronzo RA, Bogardus C, Stern MP. Identification of individuals with insulin resistance using routine clinical measurements. *Diabetes*. 2005;**54**(2):333-9.
16. Heydari ST, Ayatollahi SMT, Zare N. Comparison of artificial neural networks with logistic regression for detection of obesity. *Journal of medical systems*. 2012;**36**(4):2449-54.
17. Pourahmad S, Ayatollahi SMT, Taheri SM, Agahi ZH. Fuzzy logistic regression based on the least squares approach with application in clinical studies. *Computers & Mathematics with Applications*. 2011;**62**(9):3353-65.
18. Mohlig M, Floter A, Spranger J, Weickert MO, Schill T, Schlosser HW, *et al.* Predicting impaired glucose metabolism in women with polycystic ovary syndrome by decision tree modelling. *Diabetologia*. 2006;**49**(11):2572-9.
19. Bhargava N, Sharma G, Bhargava R, Mathuria M. Decision tree analysis on j48 algorithm for data mining. *Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering*. 2013;**3**(6).
20. Meng XH, Huang YX, Rao DP, Zhang Q, Liu Q. Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. *The Kaohsiung journal of medical sciences*. 2013;**29**(2):93-9.
21. Nesto RW, Bell D, Bonow RO, Fonseca V, Grundy SM, Horton ES, *et al.* Thiazolidinedione use, fluid retention, and congestive heart failure: a consensus statement from the American Heart Association and American Diabetes Association. *Diabetes Care*. 2004;**27**(1):256-63.
22. Iyer A, Jeyalatha S, Sumbaly R. Diagnosis of diabetes using classification mining techniques. *arXiv preprint arXiv:150203774*. 2015.
23. World Health Organization. Use of glycated haemoglobin (HbA1c) in diagnosis of diabetes mellitus: abbreviated report of a WHO consultation. *Use of glycated haemoglobin (HbA1c) in diagnosis of diabetes mellitus: abbreviated report of a WHO consultation*; 2011.
24. Grundy SM, Hansen B, Smith SC, Jr., Cleeman JI, Kahn RA, American Heart A, *et al.* Clinical management of metabolic syndrome: report of the American Heart Association/National Heart, Lung, and Blood Institute/American Diabetes Association conference on scientific issues related to management. *Circulation*. 2004;**109**(4):551-6.
25. Brown N, Critchley J, Bogowicz P, Mayige M, Unwin N. Risk scores based on self-reported or available clinical data to detect undiagnosed type 2 diabetes: a systematic review. *Diabetes research and clinical practice*. 2012;**98**(3):369-85.
26. Habibi S, Ahmadi M, Alizadeh S. Type 2 Diabetes Mellitus Screening and Risk Factors Using Decision Tree: Results of Data Mining. *Global journal of health science*. 2015;**7**(5):304-10.
27. Kahramanli H, Allahverdi N. Design of a hybrid system for the diabetes and heart diseases. *Expert systems with applications*. 2008;**35**(1):82-9.
28. Lekkas S, Mikhailov L. Evolving fuzzy medical diagnosis of Pima Indians diabetes and of dermatological diseases. *Artificial intelligence in medicine*. 2010;**50**(2):117-26.

29. Kaur G, Chhabra A. Improved J48 classification algorithm for the prediction of diabetes. *International Journal of Computer Applications*. 2014;**98**(22).
30. Haring HU, Merker L, Seewaldt-Becker E, Weimer M, Meinicke T, Broedl UC, *et al*. Empagliflozin as add-on to metformin in patients with type 2 diabetes: a 24-week, randomized, double-blind, placebo-controlled trial. *Diabetes Care*. 2014;**37**(6):1650-9.
31. Dong JJ, Lou NJ, Zhao JJ, Zhang ZW, Qiu LL, Zhou Y, *et al*. Evaluation of a risk factor scoring model in screening for undiagnosed diabetes in China population. *Journal of Zhejiang University Science B*. 2011;**12**(10):846-52.
32. Robinson CA, Agarwal G, Nerenberg K. Validating the CANRISK prognostic model for assessing diabetes risk in Canada's multi-ethnic population. *Chronic diseases and injuries in Canada*. 2011;**32**(1):19-31.