Hormozgan Medical Journal

doi 10.34172/hmj.2023.3026

Hormozgan Med J. 2023; 27(3): 133-140

Original Article

Investigating Some Biological Parameters in Patients with Diabetes to Diagnose the Disease Using a Machine Learning Approach

Parvin Norozi, Kahin Shahanipour*, Ali Asghar Rastegari

Department of Biochemistry, Falavarjan Branch, Islamic Azad University, Isfahan, Iran

***Correspondence to**
Kahin Shahanipour,
Email: Shahanipur_k@yahoo.com

Abstract

Background: Diabetes has several complications and late diagnosis of this disease leads to an increase in the complications. The present study aimed to investigate the possibility of predicting diabetes using machine learning techniques.

Methods: This study was a cross-sectional descriptive-analytical study. The population included the people referred to Falavarjan Social Security Center in Isfahan province in Iran in 2020 for diabetes screening. Blood samples were collected from 250 diabetic patients and 100 healthy non-diabetic samples. Then, glucose, cholesterol, triglyceride, high-density lipoprotein (HDL), low-density lipoprotein (LDL), and very low-density lipoprotein (VLDL) were measured and some characteristics such as height, weight, age and gender were collected from patients' records. Finally, the data were analyzed and compared using the k-nearest neighbor (KNN) algorithm, artificial neural networks (ANNs), support vector machine (SVM), Naive Bayes, and decision tree (DT). All analyses and modeling were performed in Python programming environment.

Results: In all criteria, the best results were obtained by SVM with an accuracy of 0.98, followed by ANNs with an accuracy of 0.96, respectively. Then, the K-NN algorithm with an accuracy of 0.87, Naive Bayes with an accuracy of 0.87, and DT with an accuracy of 0.76 were considered.

Conclusion: Both ANNs and linear SVMs are recommended as superior final models for the diagnosis of diabetes due to their higher performance (accuracy) in final decision-making.

Keywords: Diabetes, Machine learning, Support vector machines, Artificial neural networks

Received: April 16, 2021, **Accepted:** May 14, 2022, **Published Online:** September 28, 2023

Background

Diabetes is one of the most common metabolic diseases. Its main symptom is blood sugar higher than 120 mg/dL in fasting state and other symptoms are overeating and overdrinking. Diabetes is caused by a reduction in insulin level or a reduction in the effect of insulin in the body. Its prevalence has increased drastically during the recent decades due to lifestyle changes and obesity (1, 2). Late diagnosis or non-diagnosis of diabetes leads to the development of different chronic vascular complications. Generally, diabetes has multiple irreversible complications (3). Type 2 diabetes is not often diagnosed until complications appear (4). Early diagnosis and prevention can reduce mortality, prevent and decrease the complications of diabetes, and improve quality of life (5). Data mining methods have been extensively used in recent years in medicine and health care in the field of disease diagnosis and prevention, treatment method selection, mortality prediction, and treatment cost prediction (6). Data mining and machine learning can be used for automating diagnostic procedures in medicine (7). Different studies on diabetes mellitus have been conducted using data mining methods with different approaches (8, 9). Machine learning (ML) technology is highly effective for analyzing medical data. Today, the use of diagnostic data and automatic analysis of such data, as well as the detection of the patterns which help to identify and predict diseases quickly and inexpensively, have been considered. In recent years, different data mining methods such as decision tree (DT), artificial neural network (ANN), support vector machine (SVM) have been used for diagnosing and predicting diabetes (10-12). Given the dramatic progression of this disease and the lack of a way to detect this disease early and prevent its complications, this study aimed to diagnose it early and prevent its complications using biological parameters and machine learning methods.

The following parameters were used in the diagnosis of diabetes:

Glucose: Glucose is a type of carbohydrate that is consumed by the body cells and produces energy. It is the most important free sugar in the bloodstream.

Cholesterol: Cholesterol is a steroid and an important membrane material. Blood cholesterol comes from two

main sources, including diet and production in the liver.
TG: A type of lipid that is made up of three fatty acids
VLDL: Very low-density lipoprotein
LDL: Low-density lipoprotein, bad cholesterol
HDL: High-density lipoprotein, good cholesterol
BMI: Body mass index.

Materials and Methods

This study was performed to evaluate diabetic individuals in Falavarjan city in Isfahan province in 2020. Criteria for including patients in the research included age range of 25 to 70 years, having type 2 diabetes, willingness to participate in research, duration of at least 6 months from the time of diagnosis, having no underlying disease, and not being pregnant.

In this study, five different machine learning models such as k-nearest neighbors (KNN), SVM, ANN, Naive Bayes, and DT methods were developed and implemented on the data. Finally, the best model was selected as the recommended model for predicting the disease or the health of the sample based on the measured characteristics by evaluating each model and comparing their performance with each other.

Artificial Neural Networks

An ANN includes input variables, output variables, and weight. Network weights depend on the relationship between input and output variables. In general, there are three types of layers, each one made up of some processing units called neurons (13).

Decision Tree

DT is one of the most powerful and commonly used methods for classification and prediction and it is a model of machine learning that provides a tree-like structure like a flowchart for decision-making, group determination, and labeling of a specific datum.

Naive Bayes Algorithm

The method aims to classify the phenomena based on the probability of occurrence or non-occurrence of a phenomenon. The possibility of estimating model parameters with a small sample size as a training data set is one of the significant advantages of Naive Bayes.

Support Vector Machine

It is a pattern recognition model using supervised learning which uses a linear function for determining the maximum margin with the largest margin relative to specific points belonging to each group of the training sample.

The k-Nearest Neighbor Algorithm

It is one of the most well-known training algorithms used for pattern classification. In addition, it is a non-parametric method used for classification and regression. In this method, KNNs (samples) are examined for each new sample based on the defined distance criterion. Machine learning techniques are divided into many different categories, including supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning (14). In the present study, since the true label for each case was available, the supervised learning technique was used.

This study was performed in several steps including data gathering, data preparation, data labeling, initial data processing, reducing the dimensions of data by principal component analysis (PCA), classification of training and testing data, model evaluation, and comparison of results. In this study, it was attempted to classify laboratory data into two classes of healthy individuals and patients using machine learning algorithms. In addition, laboratory data were collected from 350 patients, including 250 diabetic samples and 100 healthy samples. For this purpose, blood samples were taken from diabetic patients and healthy individuals and then the serum was separated from the blood cells immediately. Afterwards, the analytes were measured with an auto-analyzer with the help of appropriate laboratory kits. Data on anthropometric indices including height, weight, gender, and age were collected from patients' records after gaining permission

Of these data, about 80% were considered for model training and 20% for testing and evaluating the model performance. Then, the samples were labeled and the data were pre-processed. Afterwards, the dimensions of the database were reduced from 11 to 7 to increase the speed and performance of the model.

First, the following steps were conducted to pre-process the data:

Sorting the Database

After determining the classes, all features should be fully available for each sample. In case of any defect in each sample, the sample should be completely removed from the data set. In case of the lack of any defect in the samples or the values of features, all of the data can be used for creating the model.

Converting Qualitative Variables to Quantitative Variables

Since most of the modeling procedures only deal with numerical values, it is necessary to attribute numbers to qualitative features to turn them into comparable and measurable values. For example, the gender of individuals was a qualitative feature with the values of men and women. In order to map this feature to a quantitative attribute, the numerical value 1 was assigned to the female feature and the numerical value 2 was assigned to the male feature.

Standardizing the Features

Since the range of changes in each feature is different, different data at different scales interfere with most of the

modeling methods including the KNN method and SVM. This defect is due to the assignment of higher weights to variables with larger scales or units. One of the common methods for solving this problem is using the feature standardization. Standardization of a dataset is to obtain the values which have a mean of zero and a standard variance or deviation of 1. Thus, if the mean of the main data equals μ and their standard deviation is σ, the Z value for the initial value of x can be obtained based on the following equation. In this study, Python programming environment was used for standardizing the features.

$$z = \frac{x - \mu}{\sigma}$$

Reducing the Dataset Dimensions

The main and independent features can be identified and dependent features can be removed using the current mathematical models. The database dimensions which are as many as the measurement features were reduced for the better performance of the model. In this study, dimensionality reduction was conducted on the database in Python programming environment using the PCA method. This method first mapped 11 features of data including gender, age, height, weight, BMI, sugar, cholesterol, triglyceride, HDL, LDL, and VLDL into a new 11-dimensional space. Then, key components were identified among the new features and a series of more valuable features were included in the analysis instead of evaluating all the features. Based on the results using 6 and 7 features, the model reached 100% performance and using more features as input of the model did not outperform the result. Thus, in this study, the dimensions of the database were reduced from 11 to 7. Since the PCA method transforms the feature space into another space, it is not easy to understand the value of each feature in the current space intuitively. For this reason, another method called the chi-square method was used for better understanding the relationship between each feature and the final labeling (diagnosis of diabetes). The chi-square value of each feature was extracted using this method. The high value of this index in each feature indicated the degree of dependence of this feature on the final labels.

Results

In this study, it was attempted to classify laboratory data into two groups of healthy and diabetic individuals using machine learning algorithms. In addition, laboratory data were collected from 350 patients, including 250 diabetic samples and 100 healthy samples. Of these samples, 170 were female and 180 were male. Of these data, about 80% were used for model training and 20% for testing and evaluating the model performance. Then, the samples were labeled and the data were pre-processed. In order to increase the speed and performance of the model, the dimensions of the database were reduced from 11 factors, including gender, age, height, weight, BMI, sugar, cholesterol, triglyceride, HDL, LDL, and VLDL, to 7 factors. Five different machine learning models such as KNN, SVM, ANN, Naive Bayes, and DT methods were constructed and implemented on the data. Ultimately, the best model was selected as the superior model for predicting the disease or the health of the sample based on the measured features by evaluating each of the models and comparing their performance with each other.

Figure 1 illustrates the value of the chi-square for each feature. As can be observed, the higher the index, the stronger its effect on labeling.

Thus, this feature with its high share in describing this space cannot be eliminated from the set of features, and on the contrary, the features which have a lower index can be eliminated from the database or measurement of new samples because they have little effect on the conclusion.

- In this section, the evaluation of the performance of the proposed models for grouping the data related to diabetic patients in the form of the criteria of training accuracy, test accuracy, test sensitivity, and test specificity can be observed. Training and test accuracy indicate the proportion of the total number of predictions that were correct for training and test data, respectively. Specificity describes the proportion of true negative cases which are correctly identified using the model. Sensitivity shows the proportion of true positive cases which are correctly identified. All these metrics can be used to evaluate the model performance. To summarize the performance of each classification model, a confusion matrix is used. In the present study, this matrix was 2x2 since there were two groups (healthy/unhealthy). As can be seen below, the matrix consists of true positive (correctly predicted healthy cases), false positive (incorrectly predicted healthy cases), true negative (correctly predicted unhealthy cases), and false negative (incorrectly predicted unhealthy cases) cases.

$$Confusion\ Matrix = \begin{bmatrix} true\ negative & \text{false negative} \\ \text{false positive} & \text{true positive} \end{bmatrix}$$

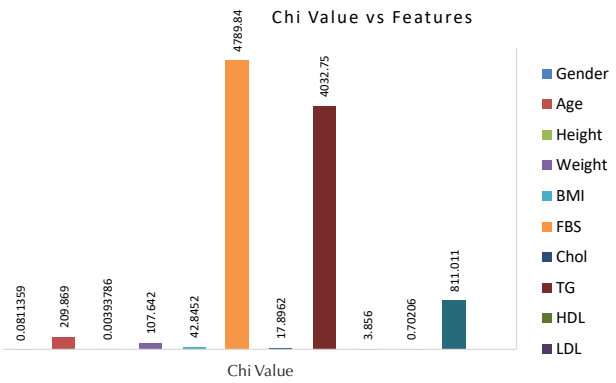

Figure 1. Chi-square value for each feature

Results of Modeling by KNN Method

Table 1 presents the modeling and labeling performance of the samples based on the KNN method considering the optimal value (k = 3). Based on the results obtained from the first to the fifth series, as well as the average of different criteria, this method can be regarded as an acceptable method for classifying diabetic patients. The average index related to the test specificity in this method was about 0.69 and the average sensitivity index was about 0.968, indicating that this model had a better performance on the data of the individuals with diabetes. The unbalanced frequency of healthy and patient groups is one of the reasons for this issue.

Results of Modeling by SVM

Table 2 indicates the grouping results using the linear SVM method. Due to the good results of this method and the good performance of this method on different data, this method can be recommended as an effective method for the classification of diabetic data. High accuracy results on new data (test data) as the main criterion for evaluation indicate the good performance of this model. Furthermore, the high and close mean value in the two criteria of specificity (0.97) and sensitivity (0.98) reveal the good and uniform performance of the model in both groups. The mean above 0.90 is desirable and high in both groups.

Results of Modeling by ANN

This method is one of the best proposed methods with its high scalability and adaptability (increasing the number of layers and neurons in each layer). In this study, a three-layer model including an input layer with seven neurons (number of independent features), a hidden layer with 2000 neurons, and an output layer with two neurons (number of groups) was used. Based on the results of accuracy in this model, this method can be proposed as an effective method. As can be observed in Table 3, the sensitivity and specificity results, indicating the performance of the model in each group, are acceptably high and close to each other.

Results of Modeling by Naïve Bayes Method

The results of accuracy on test data, indicating the performance of the model on new data, show the relatively good performance of this model. It should be stated that a difference of 0.08 between the sensitivity and specificity metrics indicates a uniform model performance in both groups (healthy and unhealthy) (Table 4).

Results of Modeling by DT Method

This model is regarded as a powerful tool due to high intuition in the modeling process. The results of test accuracy in this method indicate a value of 0.84, which is a good performance. Nevertheless, a difference of 0.12 between specificity and sensitivity indicates that the model fails to provide a relatively acceptable performance in the diagnosis of the healthy group (Table 5).

By considering the accuracy metric in testing and training data as a key indicator for model performance, it can be stated that all five methods (the KNN algorithm, linear SVM, ANN, Naive Bayes, and DT) are acceptable for predicting the group of diabetic patients.

Based on Figure 2, the SVM and ANN methods with average test accuracies of 0.9800 and 0.9685, respectively, can be considered as priorities for selecting the optimal method.

Figure 3 displays the performance of each model for healthy and unhealthy individuals. As indicated in the figure below, the SVM and ANN methods in both groups had higher performance compared to the other models.

The performance of the SVM method was 0.97 in the

Table 1. Performance Modeling and Sample Prediction Based on kNN

	First Series	Second Series	Third Series	Fourth Series	Fifth Series	Mean
Training accuracy	0.9321	0.9536	0.9393	0.9393	0.9464	0.9421
Test accuracy	0.9000	0.9143	0.8114	0.8714	0.8857	0.8765
Test specificity	0.7000	0.7500	0.6500	0.7000	0.6500	0.6900
Test sensitivity	0.9800	0.9800	0.9600	0.9400	0.9800	0.9680
Confusion matrix (Test data)	14 6 1 49	15 5 1 49	13 7 2 48	14 6 3 47	13 7 1 49	

Table 2. Modeling Performance and Sample Prediction Based on SVM Method

	First Series	Second Series	Third Series	Fourth Series	Fifth Series	Mean
Training accuracy	1	1	1	1	1	1
Test accuracy	0.9857	0.9571	0.9714	0.9571	0.9714	0.9685
Test specificity	0.9500	0.9000	1	0.9000	1	0.9500
Test sensitivity	1	0.9800	0.9600	0.9800	0.9600	0.9760
Confusion matrix (Test data)	19 1 0 50	18 2 1 49	20 0 2 48	18 2 1 49	20 0 2 48	

Table 3. Modeling Performance and Sample Prediction Based on ANN Method

	First Series	Second Series	Third Series	Fourth Series	Fifth Series	Mean
Training accuracy	1	1	1	1	1	1
Test accuracy	0.9857	0.9571	0.9714	0.9571	0.9714	0.9685
Test specificity	0.9500	0.9000	1	0.9000	1	0.9500
Test sensitivity	1	0.9800	0.9600	0.9800	0.9600	0.9760
Confusion matrix (Test data)	19 1 0 50	18 2 1 49	20 0 2 48	18 2 1 49	20 0 2 48	

Table 4. Modeling Performance and Sample Prediction Based on Naive Bayes Method

	First Series	Second Series	Third Series	Fourth Series	Fifth Series	Mean
Training accuracy	0.9036	0.9170	0.8964	0.8964	0.9107	0.9036
Test accuracy	0.8857	0.8857	0.9000	0.8286	0.8714	0.8743
Test specificity	0.8500	0.8500	0.9000	0.8000	0.7000	0.8200
Test sensitivity	0.9000	0.9000	0.9000	0.8400	0.9400	0.8960
Confusion matrix (Test data)	17 3 5 45	17 3 5 45	18 2 5 45	16 4 8 42	6 47 3	

Table 5. Modeling Performance and Sample Prediction Based on the DT Method

	First Series	Second Series	Third Series	Fourth Series	Fifth Series	Mean
Training accuracy	1	1	1	1	1	1
Test accuracy	0.8857	0.8143	0.8286	0.8000	0.9000	0.8457
Test specificity	0.9000	0.6000	0.8000	0.6500	0.8500	0.7600
Test sensitivity	0.8800	0.9000	0.8400	0.8600	0.9200	0.8800
Confusion matrix (Test data)	18 2 6 44	12 8 5 45	16 14 8 42	13 7 7 43	17 3 4 46	

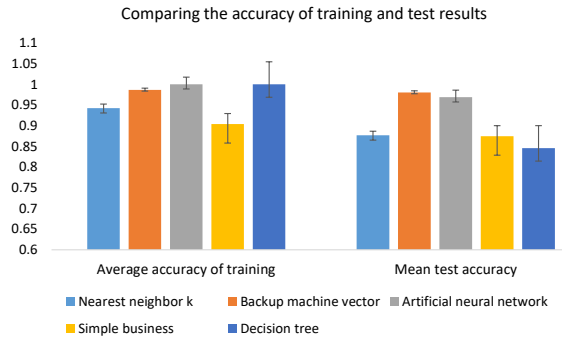

Figure 2. Comparison of the Accuracy of Training and Testing Results

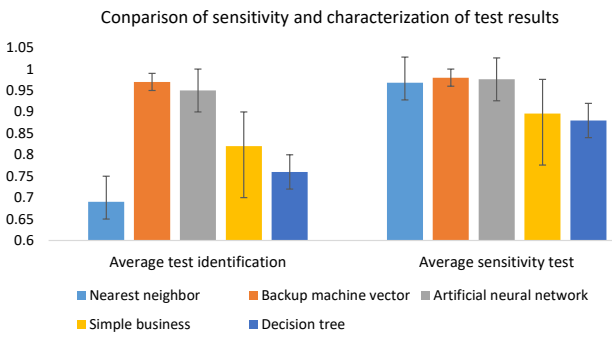

Figure 3. Comparison of the Sensitivity and Diagnosis of Test Results

healthy group and 0.98 in the patient group. In addition, the performance of ANN method in the healthy group was 0.95 and it was 0.976 in the patient group.

Discussion

In all criteria, the best results were obtained by the SVM with an accuracy of 0.98 and the ANN with an accuracy of 0.96.

Mirzakhani et al compared the ANN and DT in identifying and predicting the factors related to type 2 diabetes using the records of 901 individuals referring to health centers in Mashhad during 2012-2013. The results indicated that the multilayer ANN model had a better performance in predicting type 2 diabetes compared to the CART DT (15). In our study, the SVM and ANN methods in both groups had higher performance than the other models.

Yaghoubzadeh et al conducted a review on data mining methods for diagnosing diabetes and stated that one of the most significant factors in predicting diabetes is the age of individuals and the risk of developing diabetes in women is much stronger than in men. Another significant factor in predicting diabetes is blood pressure. In addition, body mass index is one of the essential factors studied

in predicting diabetes (16). Yaghoubzadeh et al studied different machine learning algorithms and stated that the SVM algorithm is the most accurate and useful method in predicting diabetes and its only limitation is the computational complexity which can be solved by users.

The results of the research of Yaghoubzadeh et al are consistent with the results of the present study. According to the results of the present study, ANN and especially linear SVM due to high performance in the final decision were suggested as final models.

Kurdish conducted a study to detect and diagnose diabetes using the k-means algorithm and the SVM using the UCI database (14). These data included 9 features, each of which indicated the situation of a healthy person and a patient with diabetes. The results indicated a higher accuracy for the SVM method (94.2%). The results of the study by Kurdish are consistent with those of the present study. In this study, the SVM method has been selected as an effective method. Teimouri et al compared different methods of machine learning in the diagnosis of diabetic patients and hypertension in terms of gender, age, body mass index, family history of diabetes, pregnancy history, history of gestational diabetes, history of abortion, systolic blood pressure, diastolic blood pressure, fasting blood sugar, cholesterol, triglyceride, high-density lipoprotein, creatinine, urea, and albumin. In addition, in the non-cost-sensitive scenario, logistic regression, SVM, and DT with accuracies of 69.70, 69.61, and 68.62%, respectively, showed the best performance (17).

The results of the study by Teimouri et al on the comparison of different methods of machine learning in the diagnosis of diabetic patients are not consistent with the present study. The DT method did not work best in the present study, while the SVM and the ANN had the best performance. Ture et al classified 9 characteristics including age, gender, family history of hypertension, smoking, blood triglyceride, uric acid, fatty lipoprotein, body mass index, and blood cholesterol and stated that the DT with 83% accuracy had the best result (18). They showed that DT with an accuracy of 0.83 showed the best result, while in the present study, the results with an accuracy of 0.84 showed that the performance was acceptable. Barakat et al used the SVM method to diagnose type 2 diabetes with an accuracy of 94%. A total of 4682 individuals were included in the study and the variables included gender, BMI, blood pressure, cholesterol, and blood sugar (19). The results of the study by Barakat et al are consistent with the results of the present study. In this study, the SVM was also recognized as an effective method. Barfei et al concluded that since the artificial network model does not require the usual assumptions of classical statistical methods, the accuracy of prediction of neural networks is higher compared to statistical methods (20). Based on the results of this study, the ANN was selected as the superior method. Zabbah et al stated that a faster and more accurate diagnosis of diabetes can be achieved using modern data mining methods (21). Whatever has been observed in similar studies is the power of using the k-fold method in separating the test and training samples. Kalhor et al designed an intelligent system for the diagnosis of diabetes using a data mining approach and stated that DT, KNN algorithm, and SVM had the best results, respectively (22).

The results of the study conducted by Kalhor et al contradict the present study because, in the present study, the best method was the linear SVM method, followed by the ANN.

Arena et al conducted a comparative study on machine learning techniques to diagnose diabetes and reported the highest accuracy and sensitivity for the SVM, Naive Bayes, and the DT, respectively (23). The results of the study by Arena et al are consistent with our study. The SVM method was more accurate compared to other methods.

Faraji Gavgani et al used the SVM and regression methods to diagnose and predict diabetes, indicating that the SVM algorithm was better than other methods such as logistics in diagnosing and predicting the disease (24). The results of the study by Faraji Gavgani et al are consistent with those of the present study. Due to the high accuracy of the linear SVM method in the present study, it can be used as an effective method for classifying diabetic data.

Rafeh and Arbabi used data mining techniques and lipid parameters to diagnose diabetes. In this study, some models were developed in RapidMiner software to diagnose various types of diabetes and predict and classify patients as diabetic and non-diabetic. The best accuracy of the evaluation model belonged to the DT model. This model can be used for individuals with high levels of blood lipids to predict blood sugar levels and diagnose diabetes (25). The results of the study by Rafeh and Arbabi are not consistent with the present studies. In the present study, SVM and neural network methods have been selected as the superior methods, respectively. Additionally, a high triglyceride level can be used as a key factor after glucose for the early diagnosis of diabetes using data mining methods. In the mentioned studies, the difference between the studied populations and the considered factors can change the results.

The strength of this study in comparison to previous studies was the use of 5 different data mining methods to carry out this project. The major limitation of this study was the limited number of samples to collect and perform experiments.

Conclusion

Among the selected methods, DT and KNN methods could not perform well in diagnosing healthy individuals due to the imbalance in the frequency of categories; therefore, they were labeled as patients. Another limitation of the KNN method was the high number of

samples. The difference in performance between the two different groups of healthy individuals and patients in the Naive Bayes method was lower compared to DT and KNN; however, this method did not have good credibility in this issue due to the relatively low accuracy of test and training data. Based on the results mentioned above, the ANN method and especially the linear SVM method have been proposed as the final models due to their higher performance in the final decision. Based on the presented results, the accuracies of the linear SVM model and ANN methods were 0.98 and 0.96, respectively, showing the best results in the classification of experimental samples.

Authors' Contribution

Conceptualization: Kahin Shahanipour.
Data curation: Parvin Norozi, kahin Shahanipour.
Formal analysis: Parvin Norozi, Ali Asghar Rastegari.
Funding acquisition: Parvin Norozi.
Investigation: Parvin Norozi, Ali Asghar Rastegari, Kahin Shahanipour.
Methodology: Kahin Shahanipour.
Project administration: Parvin Norozi.
Resources: Parvin Norozi, Ali Asghar Rastegari, Kahin Shahanipour.
Software: Parvin Norozi, Ali Asghar Rastegari.
Supervision: Kahin Shahanipour.
Validation: Ali Asghar Rastegari.
Visualization: Parvin Norozi, Ali Asghar Rastegari.
Writing–original draft: Parvin Norozi.
Writing–review & editing: Kahin Shahanipour.

Competing Interests

The authors declare that they have no conflict of interests.

Ethical Approval

This study was approved by the Ethics Committee of Islamic Azad University, Falavarjan Branch (IR.IAU.FALA.REC.1399,034).

Funding

It is not declared by the authors.

References

1. Tuomilehto J, Lindström J, Eriksson JG, Valle TT, Hämäläinen H, Ilanne-Parikka P, et al. Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance. N Engl J Med. 2001;344(18):1343-50. doi: 10.1056/nejm200105033441801.
2. Heydari I, Radi V, Razmjou S, Amiri A. Chronic complications of diabetes mellitus in newly diagnosed patients. Int J Diabetes Mellit. 2010;2(1):61-3. doi: 10.1016/j.ijdm.2009.08.001.
3. Luijks H, Schermer T, Bor H, van Weel C, Lagro-Janssen T, Biermans M, et al. Prevalence and incidence density rates of chronic comorbidity in type 2 diabetes patients: an exploratory cohort study. BMC Med. 2012;10:128. doi: 10.1186/1741-7015-10-128.
4. Beagley J, Guariguata L, Weil C, Motala AA. Global estimates of undiagnosed diabetes in adults. Diabetes Res Clin Pract. 2014;103(2):150-60. doi: 10.1016/j.diabres.2013.11.001.
5. Zhuo X, Zhang P, Hoerger TJ. Lifetime direct medical costs of treating type 2 diabetes and diabetic complications. Am J Prev Med. 2013;45(3):253-61. doi: 10.1016/j.amepre.2013.04.017.
6. Iyer A, Jeyalatha S, Sumbaly R. Diagnosis of diabetes using classification mining techniques. International Journal of Data Mining & Knowledge Management Process IJDKP. 2015; 5(1):1-14. doi: 10.5121/ijdkp.2015.5101.
7. Soni J, Ansari U, Sharma D, Soni S. Predictive data mining for medical diagnosis: an overview of heart disease prediction. Int J Comput Appl. 2011;17(8):43-8.
8. Temurtas H, Yumusak N, Temurtas F. A comparative study on diabetes disease diagnosis using neural networks. Expert Syst Appl. 2009;36(4):8610-5. doi: 10.1016/j.eswa.2008.10.032.
9. Habibi S, Ahmadi M, Alizadeh S. Type 2 diabetes mellitus screening and risk factors using decision tree: results of data mining. Glob J Health Sci. 2015;7(5):304-10. doi: 10.5539/gjhs.v7n5p304.
10. Fang X. Are you becoming a diabetic? A data mining approach. In: 2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery. Tianjin, China: IEEE; 2009. p. 18-22. doi: 10.1109/fskd.2009.807.
11. Huang Y, McCullagh P, Black N, Harper R. Feature selection and classification model construction on type 2 diabetic patients' data. Artif Intell Med. 2007;41(3):251-62. doi: 10.1016/j.artmed.2007.07.002.
12. Jaimes F, Farbiarz J, Alvarez D, Martínez C. Comparison between logistic regression and neural networks to predict death in patients with suspected sepsis in the emergency room. Crit Care. 2005;9(2):R150-6. doi: 10.1186/cc3054.
13. Maniruzzaman M, Kumar N, Menhazul Abedin M, Shaykhul Islam M, Suri HS, El-Baz AS, et al. Comparative approaches for classification of diabetes mellitus data: machine learning paradigm. Comput Methods Programs Biomed. 2017;152:23-34. doi: 10.1016/j.cmpb.2017.09.004.
14. Kurdish F. Detection and Diagnosis of Diabetes Using Clustering and Data Mining Techniques [thesis]. Islamic Azad University, Boroujerd Branch; 2017. [Persian].
15. Mirzakhani F, Kazemi A, Rasoulian-Kasrineh M, Javad Moosavi SY, Amirabadi Zadeh AR. Comparison of artificial neural network and decision tree to identify and predict factors associated with type 2 diabetes. J Paramed Sci Rehabil. 2018;7(4):19-32. doi: 10.22038/jpsr.2018.26264.1695. [Persian].
16. Yaghoubzadeh R, Kamel R, Khairabladi M. A review of data mining methods for the diagnosis of diabetes. The First National Conference on the Application of New Technologies in Electrical Science and Engineering. 2017.
17. Teimouri M, Ebrahimi E, Alavinia S. Comparison of various machine learning methods in diagnosis of hypertension in diabetics with/without consideration of costs. Iran J Epidemiol. 2016;11(4):46-54. [Persian].
18. Ture M, Kurt I, Turhan Kurum A, Ozdamar K. Comparing classification techniques for predicting essential hypertension. Expert Syst Appl. 2005;29(3):583-8. doi: 10.1016/j.eswa.2005.04.014.
19. Barakat NH, Bradley AP, Barakat MN. Intelligible support vector machines for diagnosis of diabetes mellitus. IEEE Trans Inf Technol Biomed. 2010;14(4):1114-20. doi: 10.1109/titb.2009.2039485.
20. Barfei F, Salehi M, Najafi I. Predicting diabetes using artificial neural network. Razi J Med Sci. 2015;22(135):28-37. [Persian].
21. Zabbah I, Eskandari A, Sardari Z, Noghandi A. Diagnosis of diabetes using artificial neural network and neuro-fuzzy approach. J Torbat Heydariyeh Univ Med Sci. 2018;6(2):10-20. [Persian].
22. Kalhor R, Mortezagholi A, Naji F, Shahsavari S, Zakaria Kiaei M. Designing an intelligent system for diagnosing type 2 diabetes using the data mining approach: brief report. Tehran Univ Med J. 2019;76(12):827-31. [Persian].
23. Aruna S, Rajagopalan SP, Nandakishore LV. An empirical comparison of supervised learning algorithms in disease detection. Int J Inf Technol Converg Serv. 2011;1(4):81-92. doi: 10.5121/ijitcs.2011.1408.
24. Faraji Gavgani L, Sarbakhsh P, Asghari Jafarabadi M, Shamshirgaran M. Application of support vector machine for detection of functional limitations in the diabetic patients of the northwest of Iran in 2017: a descriptive study. J Rafsanjan Univ Med Sci. 2020;18(12):1270-86. [Persian].
25. Rafeh R, Arbabi M. Data mining techniques to diagnose diabetes using blood lipids. J Ilam Univ Med Sci. 2015;23(4):239-47. [Persian].