A Comparative Study on Risk Prediction Model of type 2 Diabetes based on Machine Learning Theory

Shu Wang
Department of Epidemiology and Medical Statistics, Guangdong Medical University, Dongguan, Guangdong, 523808
Rong Chen
Department of Epidemiology and Medical Statistics, Guangdong Medical University, Dongguan, Guangdong, 523808
Chunwen Lin
Department of Epidemiology and Medical Statistics, Guangdong Medical University, Dongguan, Guangdong, 523808
Ling Luo
Department of Epidemiology and Medical Statistics, Guangdong Medical University, Dongguan, Guangdong, 523808
Jialu Huang
Department of Epidemiology and Medical Statistics, Guangdong Medical University, Dongguan, Guangdong, 523808
Hao Liu
Department of Epidemiology and Medical Statistics, Guangdong Medical University, Dongguan, Guangdong, 523808
Weiying Chen
Department of Epidemiology and Medical Statistics, Guangdong Medical University, Dongguan, Guangdong, 523808
Jian Xu
Department of Epidemiology and Medical Statistics, Guangdong Medical University, Dongguan, Guangdong, 523808
Qiaoli Zhang
Preventive Medicine and Hygienics, Dongguan Center for Disease Control and Prevention, Dongguan, Guangdong, 523808
Haibing Yu
Department of Epidemiology and Medical Statistics, Guangdong Medical University, Dongguan, Guangdong, 523808
Yuanlin Ding (gdmusbd@gdmu.edu.cn)
Department of Epidemiology and Medical Statistics, Guangdong Medical University, Dongguan, Guangdong, 523808

Research Article

Keywords: Machine learning, Classification, Type 2 diabetes, Disease predictions

Posted Date: May 6th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-457165/v1

Abstract

In this study, the risk prediction model of type 2 diabetes was established by Logistic regression, decision tree, BP neural network, support vector machine and deep neural network methods based on the survey data of residents of Dongguan City, Guangdong Province during 2016-2018 and its risk factors. The prediction effect of the model was evaluated based on the accuracy rate, recall rate, AUC value of the area under the curve and other indicators. DeLong test was used to statistically analyze the difference in AUC value of each model, and the prediction results of each model were compared and analyzed. The results showed that, based on the selected data set, the prediction effect of the backpropagation neural network model was the best, the accuracy was as high as 93.7%, the recall rate was 92.8%, and the AUC was 0.977. This study could provide a methodical reference for the prediction of the disease risk of type 2 diabetes.

1. Introduction

Diabetes mellitus (DM) is a metabolic disease characterized by disorder of blood glucose metabolism, which is one of the major public health problems in the 21st century[1].In 2017, the number of diabetes patients in the world has been 451 million, and this number is expected to increase to 693 million by 2045, which will bring a great burden to the health care system[2].Type 2 diabetes mellitus (T2DM) is the most common form of diabetes[3].Early lifestyle changes or pharmacological interventions have been shown to be effective in delaying or preventing type 2 diabetes and its complications [4].Therefore, it is very important to accurately predict the risk of T2DM in advance. However, the onset of T2DM is slow and the clinical incubation period is long. Related detection and diagnosis methods are improving. Resulting in a possible delay of more than 10 years from the onset to the diagnosis of T2DM[5].Therefore, timely screening and management of diabetes high-risk groups is of great significance to reduce the incidence of diabetes[6].

In recent years, it has been widely used to predict the risk of diabetes by using machine learning method and mathematical model based on the basic situation of people and routine physical examination and other indicators[7].Machine learning represents a powerful set of methods for characterizing, adjusting, learning, predicting, and analyzing data. These methods use large amounts of data input and output to recognize patterns and learn efficiently to train machines to make autonomous recommendations or decisions. After sufficient repetition and modification, the machine can receive the input and predict the output[8,9].The output results were compared with known results to judge their accuracy, and then iteratively adjusted to improve their ability to predict disease[10].Machine learning can be divided into three types: supervised learning, unsupervised learning and reinforcement learning. Some of the most common supervised learning methods include Logistic Regression, K-Nearest Neighbor (KNN), Naive Bayes (NB), Decision Tree, Artificial Neural Network (ANN) and Support Vector Machine (SVM).At present, the supervised learning method is widely used in T2DM related research.

At present, in the research on the risk prediction model of T2DM, the commonly used machine learning methods include logistic regression, CART decision tree, C4.5 decision tree, support vector machine, Back Propagation (BP) neural network and deep neural network. At present, there is no report on the systematic comparison of the prediction effects of the above six models. Based on the above background, this study uses the survey data of T2DM among Dongguan residents from 2016 to 2018.The prediction effects of six risk prediction models of T2DM based on machine learning theory including Logistic regression, CART, C4.5, back-propagation neural network, support vector machine and deep neural network were further compared to provide methodological reference for risk prediction of T2DM.

2. Material And Methods

2.1 Study design

This design is a cross-sectional study.

2.2 The research objects

The respondents were residents aged 18 or older who had lived in the monitoring area for six months or more. Patients with T2DM were diagnosed using the 1999 World Health Organization criteria. The sample size required for calculation using a dedicated formula was 1340. At each monitoring point, the objects were selected by multi-stage cluster random sampling. The sampling methods of each stage were as follows: Phase 1 sampling: at each monitoring site, 3 communities are randomly selected according to a sampling method proportional to the population size. The second stage sampling: within each selected community, 2 administrative villages are randomly selected according to the proportion of population size. The third stage sampling: in each selected administrative village, a simple random sampling method is adopted to select more than 75 households. The fourth stage sampling: in each selected household, according to the Kish table method, randomly select 1 qualified permanent resident.

In this study, a total of 4157 subjects were selected from the survey data of T2DM in Dongguan residents for 3 years from 2016 to 2018. After screening and exclusion, a total of 4106 subjects were included, including 149 patients with T2DM in the case group, and the rest were classified as the control group with 3957 subjects. This study was approved by the Medical Ethics Committee of the Affiliated Hospital of Guangdong Medical University, and was carried out in accordance with relevant guidelines and regulations. All studies were conducted with the consent and informed consent of the subjects.

2.3 Data collection

2.3.1 The questionnaire surveys

The items of the questionnaire are as follows: Gender, Age, Education level, Smoking, Drinking, Cereals, Potatoes, Beans, Eggs, Milk, Livestock, Poultry, Fish, Vegetables, fruits per week consumption frequency.

2.3.2 Body measurements

That included Body mass index (BMI), Waistline, Systolic blood pressure (SBP) and Diastolic blood pressure (DBP). All indicators are measured twice in succession before being averaged.

2.3.3 Laboratory testing

After at least 8 hours of fasting, about 5 mL of peripheral blood was collected by professional nurses in the morning for the detection of clinical biochemical indexes. A Total of five detection measures, including Fasting plasma glucose (FPG), Total cholesterol（TC), Triglyceride (TG), High-density lipoprotein cholesterol (HDL-C), Low-density lipoprotein cholesterol (LDL-C).

2.4 Data analysis

2.4.1 Data cleaning

R Studio was used to check the data distribution. Samples with age below 18 years old, excessive missing data and outliers were removed, and KNN method was used to fill in the remaining missing values.

2.4.2 Process unbalanced data

The proportion of type 2 diabetes patients to normal population in the original data was 1:27, and the data was unbalanced. Synthetic Minority Over-Sampling Technique (SMOTE) is used to process the unbalanced data by R Studio software.70% of the data processed by SMOTE were randomly selected as the training set and 30% as the test set.

2.4.3 Variable selection

The equilibrium data were tested for normality, and then univariate analysis was performed using Mann-Whitney U rank sum test and chi-square test. Screen out the target variables.

2.4.4 10-fold cross validation

The 10-flod cross validation is when a data set is randomly divided into 10 equally sized subsets. One subset is used as the test set, and the other nine are used as the training set. In this study, the 10-fold cross-validation method was applied to 70% of the training sets, and the included variables were statistically significant variables in univariate analysis. By continuously adjusting the model parameters and comparing the corresponding 10-fold cross-validation results, the optimal parameters were selected to construct the risk prediction model of type 2 diabetes.

2.4.5 Model construction

70% of the processed data were randomly selected as the training set and the remaining 30% as the test set. Logistic regression, CART, C4.5, BP neural network, support vector machine and deep neural network were applied to the training set by R Studio software. The risk prediction model of type 2 diabetes was established, and the prediction effects of each model were evaluated and compared according to the accuracy, recall rate, AUC and other indicators.

3. Results

3.1 Statistical description of baseline data

Subjects with many missing values and obvious data errors were deleted, and the final sample size was 4106, including 149 patients with type 2 diabetes and 395 people in the normal population. The difference in sample size between the two groups was huge, so SMOTE method was adopted to process the unbalanced data. The parameters were perc.over=2600 and perc.under=103 (sampling ratio was 2600% and 103%, respectively). After treatment, 4023 patients with type 2 diabetes were enrolled, and 3990 were in the control group. 70% data were randomly selected as the model training set, and the remaining 30% data were used as the test set. The specific results are shown in Table 1.

3.2 Screening results of variables in univariate analysis

Univariate analysis was performed on the balanced samples, α=0.1. The normal test found that the distribution of each characteristic attribute in the two groups of samples was mostly skewed distribution, so the Mann-Whitney U rank sum test and chi-square test in SPSS were used to analyze the quantitative

and qualitative data respectively. The results of univariate analysis were shown in Table 2. The results showed that except the educational level, the distribution of the other 23 characteristic variables between the case group and the control group was statistically different.

3.3 Parameter tuning results

In this study, the 10-fold cross validation method was applied to 70% of the training sets, and the included characteristic variables were statistically significant variables in univariate analysis. The corresponding 10-fold cross validation results were compared by continuously adjusting the model parameters. For SVM, linear kernel function, radial basis function, polynomial kernel function and Sigmoid kernel function are used for 10-fold cross validation. The results show that the linear kernel function is the best predictor. For BP neural network, the maximum number of iterations is set to 3000, and the number of hidden layer neurons within the range of 5-20 is respectively cross-verified by ten times. The results show that when the number of hidden layer neurons is 2, the prediction effect is the best. For deep neural network(DNN), the range of hidden layers was 8-12, and the number of neurons in hidden layers was 25-35. The number of neurons in each hidden layer was set to be equal in this study. The results showed that the prediction effect was best when the number of hidden layers was 9 and the number of neurons in each hidden layer was 33.

3.4 Model construction results

3.4.1 Logistics regression model

Fit all data except education level, build logistics regression model, and then use stepwise regression method to screen variables based on AIC information criterion. A total of 16 variables were finally screened, as shown in Table 3, which were age, alcohol consumption, consumption frequency of cereals, potatoes, beans, fruits, eggs, dairy, poultry and fish, DBP, FPG, TC, TG, HDL-C, and LDL-C. Variables screened by stepwise regression were applied to the training set to build a Logistic model, as shown in Table 4. In this model, the factors that had a greater influence on T2DM included potato consumption frequency, fish consumption frequency, TC, FPG, HDL-C. In addition, the frequency of cereal consumption and TC were negatively correlated with the incidence of T2DM, while the other variables were positively correlated with the incidence of type 2 diabetes.

The logistics model equation is: $\text{logit}(P) = -17.486 + 0.027\text{Age} + 0.173\text{Drinking} - 0.236\text{Cereals} + 0.442\text{Potatoes} + 0.176\text{Beans} + 0.199\text{Fruits} + 0.294\text{Eggs} + 0.154\text{Milk} + 0.373\text{Poultry} + 0.491\text{Fish} + 0.026\text{DBP} + 2.112\text{FPG} - 0.724\text{TC} + 0.249\text{TG} + 0.573\text{HDLC} + 0.303\text{LDL-C}$.

The Logistic model confusion matrix table and ROC curve were obtained by applying the model to test set for verification. As shown in Table 5, Table 6 and Figure 3(A), it can be concluded that the accuracy rate of this model is 89.4%, the recall rate is 86.0%, the accuracy rate is 93.0%, and the area under the ROC curve AUC is 0.962.

3.4.2 Support Vector Machine Model

By using the linear kernel function, the 23 characteristic variables that are significant in the single factor analysis in the training set were substituted into the SVM model, the constructed SVM model was applied to the test set for verification, and the confusion matrix table and ROC curve of the SVM model were obtained. As shown in Table 5, Table 6 and Figure 3(B), it can be concluded that the accuracy rate of this model is 91.2%, the recall rate is 89.0%, the accuracy rate is 93.3%, and the AUC of the area under the ROC curve is 0.911.

3.4.3 BP neural network

The three-layer neural network structure is adopted. The hidden layer has 20 neurons and the maximum number of iterations is 3000. Twenty-three significant characteristic variables from univariate analysis in the training set were substituted into the BP neural network model. The final model constructed was applied to the test set for verification, and the correlation confusion matrix table and ROC curve were obtained. As shown in Table 5, Table 6 and Figure 3 (C), it can be concluded that the accuracy rate of this model is 93.7%, the recall rate is 92.8%, the accuracy rate is 94.6%, and the area under the ROC curve AUC is 0.977.

3.4.4 Decision tree model

(1) CART decision tree

The 23 characteristic variables that were significant in the single factor analysis in the training set were substituting into the CART decision tree model, and the output model of the CART decision tree was shown in Figure 1. When FPG≥5.6mmol/ L, type 2 diabetes was diagnosed directly; When FPG<5.6mmol/ L, Potatoes= 0,1, the patient was diagnosed as non-type 2 diabetes mellitus; When FPG<5.6mmol/ L, Potatoes≠ 0,1 and AGE <34, the patients were diagnosed as non-type 2 diabetes mellitus; When FPG<5.6mmol/ L, Potatoes≠ 0,1 and AGE ≥34, the patient was diagnosed as having type 2 diabetes. The CART decision tree model constructed was applied to test set for verification, and the correlation confusion matrix table and ROC curve were obtained. As shown in Table 5, Table 6 and Figure 3 (D), we can conclude that the accuracy rate of this model is 88.7%, the recall rate is 84.8%, the accuracy rate is 93.3%, and the area under the ROC curve AUC is 0.906.

(2) C4.5 decision tree

Twenty-three significant characteristic variables from univariate analysis in the training set were put into the C4.5 decision tree model. As shown in Figure 2, the decision tree model output by C4.5 algorithm includes 6 root nodes and 9 leaf nodes. According to the model, type 2 diabetes was diagnosed when FPG> was 5.61mmol/ L; When FPG≤5.61mmol/ L: Potatoes=0 was diagnosed as non-type 2 diabetes; Potatoes=1, Age≤54 was diagnosed as non-type 2 diabetes; Potatoes=1, Age>54, TC≤5.11mmol/ L was diagnosed as type 2 diabetes. Potatoes=1, Age>54, TC BBB>11 mmol/ L was diagnosed as non-type 2 diabetes mellitus. Potatoes=2, DBP≤81mmHg was diagnosed as non-type 2 diabetes; Potatoes=2, DBP>81mmHg was diagnosed as type 2 diabetes; Potatoes=3, Age≤34 was diagnosed as non-type 2 diabetes; Potatoes=3, Age>34 was diagnosed as type 2 diabetes. The C4.5 model was applied to the test set for verification, and the correlation confusion matrix table and ROC curve were obtained. As shown in Table 5, Table 6 and Figure 3 (E), it can be concluded that the accuracy rate of this model is 88.6%, the recall rate is 84.9%, the accuracy rate is 92.7%, and the area under the ROC curve AUC is 0.888.

3.4.5 Deep neural network model construction

The 23 characteristic variables that were significant for univariate analysis in the training set were substituted into the DNN model. The number of hidden layers was 9, with 33 neurons in each layer, and the correlation confusion matrix table and ROC curve were obtained. As shown in Table 5, Table 6 and Figure 3(F), the accuracy rate of this model was 84.5%, the recall rate was 82.9%, the accuracy rate was 86.1%, and the AUC of the area under the ROC curve was 0.845.

3.5 Comparison of model performance

DeLong test in R Studio was used to compare the AUC values of each model, as shown in Table 7 and Figure 4. Based on the data set and incorporating the robustness of the model and the prediction effect of type 2 diabetes, BP neural network model is the best, the accuracy is as high as 93.7%, the recall rate is 92.8%, accurate rate was 94.6%, the AUC value is 0.977, followed by logistic regression model, the SVM model, CART decision tree model, C4.5 decision tree model, depth of neural network model. The prediction effect of SVM and CART was similar, and the difference was not statistically significant.

4. Discussion

With rapid economic growth and changes in people's lifestyle, the prevalence of diabetes in China has increased significantly. According to the 2017 survey results of the International Diabetes Association, the number of diabetes patients in China has reached 114 million, making China the country with the largest number of diabetes patients in the world. T2DM is a chronic disease characterized by the body's inability to metabolize glucose effectively, raising blood sugar levels and leading to hyperglycemia. Chronic high blood glucose levels can affect the kidneys, nervous system, heart and vascular systems, leading to a range of serious complications, and have a significant impact on the health and medical costs of the population. It is estimated that approximately 5 million people aged 20 to 99 died of diabetes in 2017, accounting for 9.9% of all cause mortality in this age group globally, and more than one-third of all diabetes deaths occurred in people under 60 years of age [2]. At present, more and more researchers are committed to using machine learning methods to explore the risk factors related to T2DM and the construction of prediction models for T2DM. The early diagnosis or prediction of diabetes through the model is of great significance for the prevention of T2DM, the improvement of the quality of life of T2DM patients and the prevention of related complications. In this study, the risk prediction model of T2DM was established by using logistic regression, decision tree, BP neural network, support vector machine and deep neural network methods based on the survey data of residents of Dongguan during 2016-2018 and its risk factors. The prediction results of each model were compared and analyzed to provide methodological reference for the risk prediction of T2DM.

In recent years, machine learning techniques have been widely used to predict the risk of developing T2DM. Ye Hong [11] built a diabetes prediction model based on BP neural network, SVM and integrated learning, and the results showed that the prediction effect of BP neural network was better than that of SVM. Jing Gao [12] used BP neural network and Logistic regression to construct the prediction model of T2DM complications, and the study found that the prediction effect of BP neural network was higher than that of Logistic regression. Liu et al. [13] constructed Logistic model, BP neural network model and decision tree model to analyze the risk factors of T2DM, and the results showed that the prediction effects of the three models from high to low were BP neural network, logistic regression model and decision tree model, respectively. Dwivedi et al. [14] used six algorithms of Classification trees, SVM, ANN, NB, Logistic and KNN to predict diabetes, and the results showed that the prediction effect of Logistic regression model was better than that of support vector machine. These are consistent with the results of this study. The results of this study showed that, considering the robustness of the model and the prediction effect of the model for T2DM, the prediction effect of the BP neural network model was the best among the six models, including Logistic regression, SVM, CART decision tree, C4.5 decision tree, BP neural network and deep neural network, which were constructed by the training set containing 70% samples. The accuracy rate was as high as 93.7%, the recall rate was 92.8%, and the AUC value was 0.977. followed by Logistic model, SVM model, CART decision tree model, C4.5 decision tree model, and deep neural network model. The prediction effect of SVM and CART was similar, and the difference was not statistically significant. Logistic model AUC value is 0.962, the prediction effect is better than SVM. Although the Logistic model is weaker than BP neural network, it has a strong explainability to the results and can reflect the relationship between various factors and T2DM.

Faruque [15] and Kandhasamy [16] found that the prediction effect of C4.5 decision tree model was significantly higher than that of SVM model. The results of this study showed that the prediction effects of SVM and CART models were similar, and the difference in AUC value was not statistically significant, and the prediction effects of the two models were slightly better than that of C4.5 decision tree model. In practical application, compared with SVM model, CART decision tree model and C4.5 decision tree model can present the variables included in the model more intuitively. Cheruku et al. [17] built a prediction model for T2DM based on PIDD data set by using a variety of machine learning methods. The results showed that the prediction effect of C4.5 decision tree model was

better than that of CART decision tree model, and the accuracy rates were 74.2% and 70.7%, respectively. Althunayan et al. [18] compared the performance of nine algorithms such as Naive Bayes, C4.5, CART and random forest in the prediction of T2DM, and the results showed that the accuracy of random forest was the highest, up to 100%, and the accuracy of C4.5 algorithm was much higher than that of CART algorithm. Meng et al. [19] used ANN, logistic regression and C4.5 data mining technology to predict diabetes, and finally concluded that C4.5 machine learning technology is more effective and accurate than other methods. The results of this study show that the prediction effect of CART decision tree model is slightly better than that of C4.5 decision tree model. The CART decision tree model is a binary tree, and compared with C4.5 decision tree model, its operation is faster and the model formed is more concise. Therefore, CART is more suitable for the processing of large sample data. Ayon et al. [20] used deep neural network method to build a prediction model for T2DM based on PIDD data set, and the accuracy of the model reached 98.35%. Mohapatra [21] applied deep neural network to the prediction of T2DM, and the accuracy of the algorithm was as high as 97.11%. The results of this study showed that compared with the other 5 T2DM prediction models, DNN had the worst prediction effect, with an accuracy rate of only 84.5% and an AUC value of 0.845. The results of this study are different from those of previous related studies, which may be caused by the differences in sample size, data quality, included characteristic variables, definition of related variables and construction techniques of different data sets.

In this paper, SMOTE algorithm is used to process unbalanced data, but this algorithm cannot overcome the problem of data distribution of unbalanced data sets, and it is easy to produce the problem of distribution marginalization. If a sample of a few categories is at the distribution edge of the sample set of a few categories, then the sample generated by it and its adjacent samples will also be at this edge and will become more and more marginalized, thus blurring the boundary of the two sample types and increasing the difficulty of classification algorithm. In addition, the model constructed in this paper has only been validated internally, lacking external validation, and the extrapolation is limited. In the later stage of consideration, it will be verified in larger population samples from different regions. In conclusion, the BP neural network model in this study has the best prediction effect, while the deep neural network, different from previous studies, has the worst prediction effect in this study. BP neural network model to predict the effect is best, but the results can be interpreted is not strong, which means that the BP neural network model can efficiently identify patients with T2DM, but due to its principle depends on the internal control mechanism, unable to understand the characteristics of the factors influence on T2DM, so cannot to more accurately the relevant risk factors intervention. Although the decision tree can visually present the classification process, it cannot understand the influence of the factors included in the model on T2DM. Logistic regression model is second only to BP neural network in predicting effect, but it has strong explainability to the results, and its principle is simple and easy to understand. Therefore, suitable classifiers can be selected according to research purposes when applied in clinical decision-making.

5. Conclusion

Machine learning technology has high accuracy, low error rate and low cost in the early prediction of various diseases. Early diagnosis of T2DM is of great significance to improve the quality of life of patients with T2DM and prevent complications. This study is based on a variety of indicators, such as accuracy, recall rate, AUC, etc. The prediction effects of six risk prediction models of Type 2 glucose and urine disease were compared, including Logistic regression, CART, C4.5, back propagation neural network, support vector machine and deep neural network. The results showed that the prediction effect of the back propagation neural network model based on the selected data set was the best.

Declarations

Acknowledgements

We appreciate all authors for their contributions and physicians and participants. The study was funded by the Medical Scientific Research Foundation of Guangdong Province(Grant No. A2021395), Natural Science Foundation of Basic and Applied Basic Research Foundation of Guangdong Province (Grant No.2021A1515010061), Undergraduate Innovation Experiment Project of Guangdong Medical University (Grant No. ZZDG003), Natural Science Key Cultivation Project of Scientific Research Fund of Guangdong Medical University (Grant No. GDMUZ2020008), Zhanjiang Science and Technology Development Special Fund Competitive Allocation Project (Grant No. 2020A01031), Guangdong Provincial Colleges and Universities Characteristic Innovation Project (Grant No. 2019KTSCX047), Basic and Applied Basic Research Foundation of Guangdong Province Regional Joint Fund Project(Grant No.2020B1515120021) and Young Innovative Talents Project of Guangdong Universities(Grant No.2018KQNCX088).

Author Contributions

S.W. and R.C. have the same contribution. S.W. and R.C. conceived of the study and wrote the manuscript. C.L., L.L, J.H., H.L., W.C. and J.X. contributed to study design, data collection and review of the manuscript. Y.D., H.Y. and Q.Z. critically reviewed the manuscript and put forward modification opinions. All authors approved the final version.

Competing interests

The authors declare no competing interests.

References

1. Zinman, B. The international diabetes federation world diabetes congress 2015. *Eur Endocrinol.* **11,** 66 (2015).

2. Cho, N. H., Shaw, J. E., Karuranga, S., Huang, Y., Fernandes, J. R. & Ohlrogge, A. W., et al. IDF diabetes atlas: global estimates of diabetes prevalence for 2017 and projections for 2045. *Diabetes Res Clin Pract*. **138,** 271-281 (2018).
3. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N. & Chouvarda, I. Machine learning and data mining methods in diabetes research. *Comput Struct Biotechnol J*. **15,** 104-116 (2017).
4. Knowler, W.C. et al. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *New Engl J Med*. **346,** 393-403 (2002).
5. Harris, M.I., Klein, R., Welborn, TA. & Knuiman MW. Onset of NIDDM occurs at least 4–7 yr before clinical diagnosis. *Diabetes care*. **15,** 815-819 (1992).
6. Li, G. et al. The long-term effect of lifestyle interventions to prevent diabetes in the China Da Qing diabetes prevention Study: a 20-year follow-up study. *The Lancet*. **371,** 1783-1789 (2008).
7. Wan, X.Y., Zou, L.Z. & Fu, S.H. Discussion on the management of chronic diseases in hospital. *J Aerospace Med*. **26,** 1260-1261 (2015).
8. Bini, S.A. Artificial intelligence, machine learning, deep learning, and cognitive computing: what do these terms mean and how will they impact health care? *J Arthroplasty*. **33,** 2358-2361 (2018).
9. Naylor, C.D. On the prospects for a (deep) learning health care system. *Jama*. **320,** 1099-1100 (2018).
10. Haeberle, H. S., Helm, J. M., Navarro, S. M., Karnuta, J. M. & Ramkumar, P. N. Artificial intelligence and machine learning in lower extremity arthroplasty: a review. *J Arthroplasty*. **34,** 2201-2203 (2019).
11. Hong, Y. Research on diabetes prediction models based on machine learning algorithm. *Harbin Institute of Technology.* 20-46 (2016).
12. Gao, J. Prediction of different stages of type 2 diabetes mellitus by machine learning. *Xi 'an Medical College*. 14-34 (2019).
13. Liu, S.et al. Application of three statistical models for predicting the risk of diabetes. *BMC Endocr Disord*. **19,** 126-126 (2019).
14. Dwivedi, A.K. Analysis of computational intelligence techniques for diabetes mellitus prediction. *Neural Comput Appl*. **30,** 3837-3845 (2018).
15. Faruque, M. F., Asaduzzaman, & Sarker, I. H. Performance analysis of machine learning techniques to predict diabetes mellitus. (2019).
16. Kandhasamy, J.P. & Balamurali, S. Performance analysis of classifier models to predict diabetes mellitus. *Procedia Computer Science*. **47,** 45-51 (2015).
17. Cheruku, R., Edla, D.R. & Kuppili, V. SM-RuleMiner: Spider monkey based rule miner using novel fitness function for diabetes classification. *Comput Biol Med*. **81,** 79-92 (2017).
18. Althunayan, L., Alsahdi N, & Syed, L. Comparative analysis of different classification algorithms for prediction of diabetes disease. 144:1-144:6 (2017).
19. Meng, X. H., Huang, Y. X., Rao, D. P., Zhang, Q. & Liu, Q. Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. *Kaohsiung J Med Sci*. **29,** 93-99 (2013).
20. yon, S. I. & Islam, M. M. Diabetes prediction: a deep learning approach. *International Journal of Information Engineering and Electronic Business*. **1,** 21-27 (2019).
21. Mohapatra, S.K., Nanda. S. & Mohanty, M.N. Diabetes detection using deep neural network. *Soft Computing Systems*. 225-231 (2018).

Tables

Table 1

Comparison of baseline data between case group and normal group

	Before Balance		After Balance	
	NC (n=3957)	T2DM(n=149)	NC (n=3990)	T2DM(n=4023)
Age (year) *	33(14)	45(24)	33(13)	43(23)
Waistline (cm) *	77.5(14.90)	87.0(15.50)	77.4(14.60)	84.9(12.67)
BMI (kg/m^2) *	22.41(4.87)	24.91(4.69)	22.48(4.9)	24.53(3.71)
SBP(mmHg) *	118(19.00)	127(20.00)	118(20.00)	126(17.00)
DBP(mmHg) *	77(12.00)	82(14.00)	77(13.00)	81(12.00)
FPG(mmol/l) *	4.52(0.70)	7.30(3.79)	4.51(0.71)	7.41(3.66)
TC(mmol/l) *	4.75(1.17)	5.16(1.67)	4.76(1.18)	4.96(1.39)
TG(mmol/l) *	1.03(0.84)	1.57(1.45)	1.05(0.85)	1.65(1.45)
HDL-C(mmol/l) *	1.40(0.48)	1.31(0.59)	1.40(0.49)	1.29(0.43)
LDL-C(mmol/l) *	2.60(1.02)	2.89(0.61)	2.59(0.99)	2.89(1.05)
Gender (%)				
male	1880(47.5)	70(47.0)	1910(47.0)	2131(53.0)
female	2077(52.5)	79(53.0)	2080(52.1)	1892(47.0)
Education level (%)				
below junior college	2900(73.3)	124(83.2)	2861(71.7)	2915(72.5)
junior college or above	1025(26.7)	25(16.8)	1129(28.3)	1108(27.5)
Smoking (%)				
yes	906(22.9)	33(22.1)	905(22.7)	1164(28.9)
no	3051(77.1)	116(77.9)	3085(77.3)	2859(71.1)
Drinking (%)				
yes	2210(55.9)	55(36.9)	2256(56.5)	1856(46.1)
no	1747(44.1)	94(63.1)	1734(43.5)	2167(53.9)
Cereal consumption frequency (%)				
not eat	24(0.6)	1(0.7)	24(0.6)	8(0.2)
1~3 times per week	292(7.4)	8(5.4)	325(8.1)	169(4.2)
4~6 times per week	1266(32.0)	62(41.6)	1285(32.2)	1617(40.2)
everyday	2375(60.0)	78(52.3)	2356(59.0)	2229(55.4)
Potato consumption frequency (%)				
not eat	854(21.6)	37(24.8)	835(20.9)	1046(26.0)
1~3 times per week	2760(69.7)	95(63.8)	2823(70.8)	1914(47.6)
4~6 times per week	241(6.1)	9(6.0)	228(5.7)	536(13.3)
everyday	102(2.6)	8(5.4)	104(2.6)	527(13.1)
Bean consumption frequency (%)				
not eat	626(15.8)	35(23.5)	585(14.7)	955(23.7)
1~3 times per week	2870(72.5)	93(62.4)	2916(73.1)	1841(45.8)
4~6 times per week	321(8.1)	10(6.7)	335(84)	650(16.2)
everyday	140(3.5)	11(7.4)	154(3.9)	577(14.3)
Vegetable consumption frequency (%)				
not eat	23(0.6)	2(1.3)	22(0.6)	44(1.1)
1~3 times per week	505(12.8)	13(8.7)	512(12.8)	259(6.4)
4~6 times per week	1291(32.6)	60(40.3)	1348(33.8)	1624(40.4)
everyday	2138(54.0)	74(49.7)	2108(52.8)	2096(52.1)

Fruit consumption frequency (%)				
not eat	260(6.6)	5(3.4)	272(6.8)	88(2.2)
1~3 times per week	1895(47.9)	70(47.0)	1954(49.0)	1449(36.0)
4~6 times per week	916(23.1)	38(25.5)	920(23.1)	1254(31.2)
everyday	886(22.4)	36(24.2)	844(21.2)	1232(30.6)
Egg consumption frequency (%)				
not eat	388(9.8)	14(9.4)	362(9.1)	413(10.3)
1~3 times per week	2709(68.5)	94(63.1)	2755(69.0)	1809(45.0)
4~6 times per week	496(12.5)	29(19.5)	520(13.0)	1100(27.3)
everyday	364(9.2)	12(8.1)	353(8.8)	701(17.4)
Milk consumption frequency (%)				
not eat	1050(26.5)	54(36.2)	1072(26.9)	1330(33.1)
1~3 times per week	2138(54.0)	64(43.0)	2137(53.6)	1327(33.0)
4~6 times per week	423(10.7)	19(12.8)	436(10.9)	795(19.8)
everyday	346(8.7)	12(8.1)	345(8.6)	571(14.2)
Livestock consumption frequency (%)				
not eat	904(22.8)	25(16.8)	924(23.2)	596(14.8)
1~3 times per week	1537(38.8)	53(35.6)	1566(39.2)	1031(25.6)
4~6 times per week	553(14.0)	28(18.8)	548(13.7)	866(21.5)
everyday	963(24.3)	43(28.9)	952(23.9)	1530(38.0)
Poultry consumption frequency (%)				
not eat	406(10.3)	10(6.7)	438(11.0)	422(10.5)
1~3 times per week	2755(69.6)	99(66.4)	2745(68.8)	1831(45.5)
4~6 times per week	509(12.9)	29(19.5)	502(12.6)	1121(27.9)
everyday	287(7.3)	11(7.4)	305(7.6)	649(16.1)
Fish consumption frequency (%)				
not eat	493(12.5)	12(8.1)	480(12.0)	424(10.5)
1~3 times per week	2830(71.5)	95(63.8)	2832(71.0)	1915(47.6)
4~6 times per week	452(11.4)	28(18.8)	486(12.2)	1004(25.0)
everyday	182(4.6)	14(9.4)	192(4.8)	680(16.9)

Table 1. NC, normal population; * indicates that the distribution of each feature attribute in the two groups of samples is mostly skewed distribution, so the median and interquartile range are used to describe the mean value and variation degree.

Table 2

Results of univariate analysis

Variable	*P-value*	Variable	*P-value*	Variable	*P-value*	Variable	*P-value*
Age	<0.1	TC	<0.1	Beans	<0.1	Vegetables	<0.1
Waistline	<0.1	TG	<0.1	Fruits	<0.1	Milk	<0.1
BMI	<0.1	HDL-C	<0.1	Gender	<0.1	Livestock	<0.1
SBP	<0.1	LDL-C	<0.1	Education level	0.452	Poultry	<0.1
DBP	<0.1	Fish	<0.1	Smoking	<0.1	Cereals	<0.1
FPG	<0.1	Eggs	<0.1	Drinking	<0.1	Potatoes	<0.1

Table 3

Logistic model variable screening results

	Estimate	Std.Error	*Z-value*	*Pr(>|z|)*
(Intercept)	-17.689	0.650	-27.198	<0.001
Age	0.029	0.003	8.436	<0.001
DBP	0.025	0.004	5.693	<0.001
FPG	2.171	0.064	34.039	<0.001
TC	-0.712	0.091	-7.863	<0.001
TG	0.250	0.031	8.143	<0.001
HDL-C	0.589	0.135	4.375	<0.001
LDL-C	0.294	0.106	2.770	<0.01
Drinking	0.204	0.082	2.492	<0.05
Cereals	-0.218	0.065	-3.367	<0.001
Potatoes	0.422	0.057	7.351	<0.001
Beans	0.177	0.057	3.092	<0.01
Fruits	0.135	0.050	2.713	<0.01
Eggs	0.246	0.053	4.613	<0.001
Milk	0.213	0.046	4.661	<0.001
Poultry	0.374	0.054	6.890	<0.001
Fish	0.500	0.056	8.883	<0.001

Table 4

Logistic model

	Estimate	Std.Error	*Z-value*	*Pr(>|z|)*
(Intercept)	-17.468	0.768	-22.757	<0.001
Age	0.027	0.004	6.880	<0.001
DBP	0.026	0.005	5.041	<0.001
FPG	2.112	0.075	28.271	<0.001
TC	-0.724	0.111	-6.525	<0.001
TG	0.249	0.035	7.143	<0.001
HDL-C	0.573	0.159	3.610	<0.001
LDL-C	0.303	0.129	2.355	<0.05
Drinking	0.173	0.097	1.776	0.076
Cereals	-0.236	0.078	-3.033	<0.01
Potatoes	0.442	0.068	6.481	<0.001
Beans	0.176	0.068	2.579	<0.01
Fruits	0.199	0.059	3.384	<0.001
Eggs	0.294	0.063	4.656	<0.001
Milk	0.154	0.054	2.832	<0.01
Poultry	0.373	0.064	5.814	<0.001
Fish	0.491	0.067	7.416	<0.001

Table 5

Confusion matrix of each model

		Forecast classification											
		logistic		SVM		BP		CART		C4.5		DNN	
		1	0	1	0	1	0	1	0	1	0	1	0
True	1	1068	174	1105	137	1153	89	1048	194	1054	188	1030	212
classification	0	81	1081	78	1084	63	1099	78	1084	85	1077	161	1001

Table 6

Performance indexes of each model

Model	Accuracy	Precision	Recall	AUC
logistic	0.894	0.930	0.860	0.962
SVM	0.912	0.933	0.890	0.911
BP	0.937	0.946	0.928	0.977
CART	0.887	0.933	0.848	0.906
C4.5	0.886	0.927	0.849	0.888
DNN	0.845	0.861	0.829	0.845

Table 7

Difference test of AUC values

	Z -value	*P-value*
BP vs logistic	-12.25	<2.2e-16
logistic vs SVM	12.423	<2.2e-16
SVM vs CART	0.70627	0.355
CART vs C4.5	4.9083	9.188e-07
C4.5 vs DNN	5.6613	1.502e-08

Figures

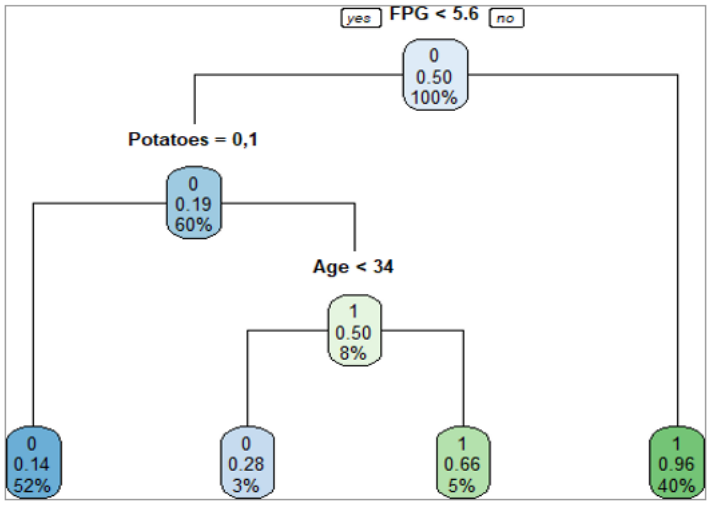

Figure 1

CART decision tree model

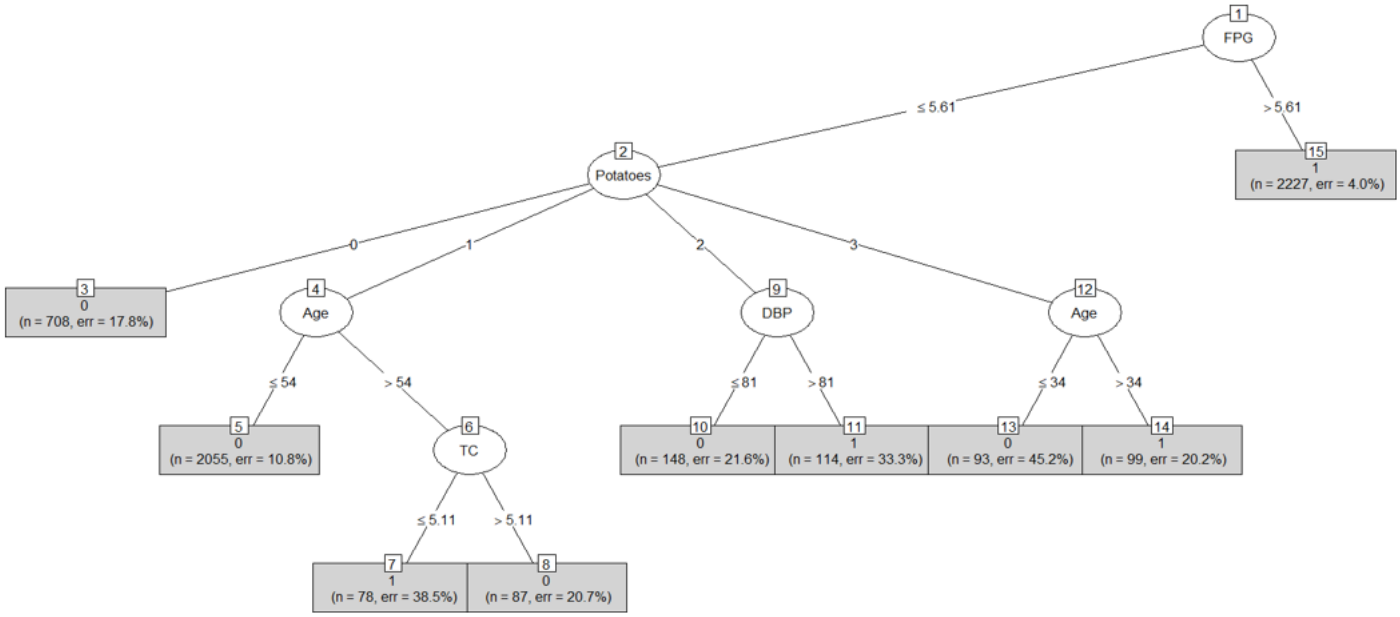

Figure 2

C4.5 decision tree model

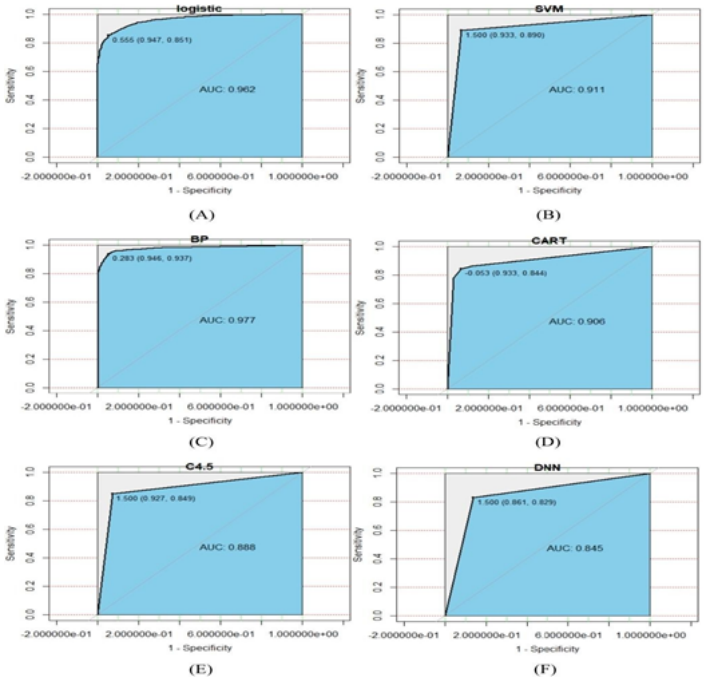

Figure 3

ROC curves of each model. (A) represents the ROC curve of the Logistic regression model; (B)ROC curve representing the support vector machine model; (C)Represents the ROC curve of the back-propagation neural network model; (D)ROC curve representing CART decision tree model; (E)ROC curve representing C4.5 decision tree model; (F)Represents ROC curve of deep neural network model.

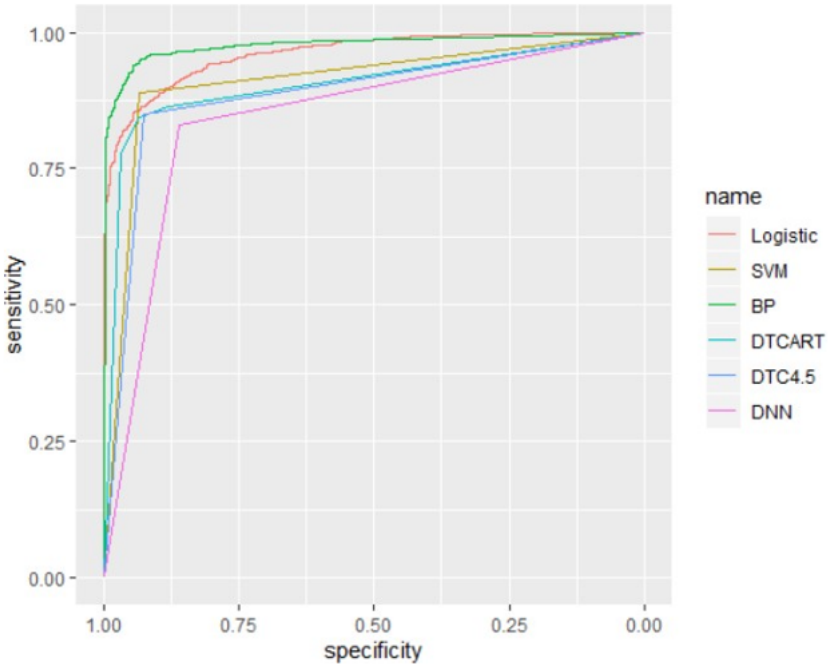

Figure 4

Comprehensive ROC curve of six models