Review Article

BIOMEDICAL Journal of Scientific & Technical Research

ISSN: 2574 -1241

DOI: 10.26717/BJSTR.2026.65.010222

A Predictive Model for the Risk of Chronic Complications in Type 2 Diabetes

Xiaoyuan Ma[1]*, Maiming Yang[1]*, Weijia Wei[1], Gao Wei[1] and Qinpeng Zhe[2]

[1]*China Unicom Digital Intelligence Medical Technology Co., Ltd., Guangzhou, China*

[2]*Guangzhou Center for Disease Control and Prevention Guangzhou, China*

***Corresponding author:** Xiaoyuan Ma and Maiming Yang, China Unicom Digital Intelligence Medical Technology Co., Ltd.,Guangzhou, China

ARTICLE INFO

Received: April 22, 2026

Published: May 13, 2026

Citation: Xiaoyuan Ma, Maiming Yang, Weijia Wei, Gao Wei and Qinpeng Zhe. A Predictive Model for the Risk of Chronic Complications in Type 2 Diabetes. Biomed J Sci & Tech Res 65(4)-2026. BJSTR. MS.ID.010222.

ABSTRACT

Objective: To evaluate the effectiveness of the combined model of XG-BOOST and logistic regression model for predicting the incidence of diabetes mellitus.

Methods: Using diabetes management data from January 2017 to December 2021 in Guangzhou city as fitting data and selecting 25480 chronic disease management data from 2017-2021 from Guangzhou city's universal health management platform as prediction data, a combined model of XG-BOOST, logistic regression model was established for the chronic disease management characteristic fields, including demographic information, follow-up information, the laboratory test data, the characteristic variables were initially screened by the XG-BOOST model, and then the fitted risk factors of the logistic regression model were used.

Results: Fifty-two characteristic variables were screened by the XG-BOOST model for the 436 fields of chronic disease management characteristics, and 33 risk factors for diabetes screening were fitted using the logistic regression model. ROC curves, AUC, and PR curves were fitted for the logistic regression model.

Conclusion: The combined model fit was better than the single model, and the predictive effect was better than or equivalent to the single model.

Abbreviations: LR: Logistic Regression; DT: Decision Tree; NN: Neural Network; LVQ: Learning Vector Quantization; SVM: Support Vector Machine; ECG: Electrocardiogram; AL: Axial Length

Introduction

The high prevalence of type 2 diabetes (T2D) and the serious consequences of complications have led many experts to focus on the research of factors related to T2D complications. Because of the marked heterogeneity in the clinical presentation of diabetic patients, it is difficult to achieve an effective early diagnosis and differential diagnosis of diabetic complications by simple laboratory indices. The rapid development of machine learning techniques has made prospective prediction possible, and several machine learning methods have been applied to the medical field, such as decision trees, logistic regression, support vector machines, and artificial neural networks have been used to make predictions about diabetes or its complications. In this study, we build a risk model for predicting T2D complications based on a novel two-step method using data from about 130,000 diabetic patients in Guangzhou from 2017-2021, analyze the role of characteristic data of T2D complications. The two-step method consists of the following steps, first being building an XGBoost model, which aims at feature selection, and second being building a logistic regression model for prediction.

Methods

Data Source

The data for this project was obtained from the Guangzhou All-Ming Health Information Platform, which was launched in 2011. As of December 2021, the platform has continuously collected 11 years of medical, public health and physical examination data from 290 medical and health institutions in the city. According to the survival analysis sample size calculation method, about 130,000 diabetic patients in the city in 2017 were selected as study subjects, and their baseline and year-by-year physiological and biochemical indicators, lifestyle,

and treatment intervention-related variables from 2017 to 2021 were obtained and organized to prepare for the construction of statistical models. The 25,480 records obtained according to the data entry criteria were used to train the model.

Diabetes Prediction Model

Using cohort data containing diabetic patients and healthy populations, train machine learning models to predict the risk of developing diabetes in healthy populations over 5 years and identify risk factors for diabetes and its complications. Determining the scope of data. Baseline and annual follow-up data including physical examination and behavioral factors are included, and demographic, physiological, biochemical indicators, lifestyle and treatment intervention-related variables are selected from a predefined pool of predictors to create training data for the prediction model of chronic complications of diabetes in the adult population, and parallel sub-data from the cohort are used as the model validation set. Data preprocessing. It includes data cleaning, outlier processing, data statute, data transformation, data sampling, etc. Model development and selection. Our primary goal is to obtain accurate predictions with high interpretability. To achieve this, we consider logistic regression for its linearity in features. As to feature selection, we build an auxiliary model of XGBoost which we train with the data of diabetic patient in Guangzhou from 2017-2021 and obtain future importance to get the desired features.

Model Evaluation Validation

The internal validity and external validity of the model were validated. Internal validity reflects the reproducibility of the model and external validity reflects the generalizability of the model and needs to be tested with data outside the research project itself (temporally and geographically independent, or completely independent data). The internal validity was tested using the half-fold cross-validation method, which involves dividing the original data into two parts, one for model building and the other for model validation. The training and validation sets were randomly divided to obtain a training set containing 20480 data points and a validation set with 5000 data points.

Results

XGBoost

We used the data to train the XGBoost model after processing the data for unique thermal coding and missing values to help us derive the importance of features from the perspective of model computation. The AUC of the Roc curve for the model was 0.8878, indicating that the model performed well by which the feature importance was scored as follows (Figure 1). In order to obtain a set of features with highest significance among all features, we conduct a thorough experiment where we first feed all features to an XGBoost model which return us a list of scores of importance for each feature. Then we feed features to another XGBoost model one at a time with the feature importance from the most to the least and discover the model performance corresponding to each added feature. The graph below shows that as the number of features increases, the performance has an upward trend different number of features (Figure 2). A further observation reveals that as the number of features is reaching eight, the performance has a relatively obvious improvement, whereas after the number reaches eight, the improvement of model performance becomes less noticeable until it reaches around 45. The options for the number of features thus remain to be eight and 45. Since the performance of the model with 45 features is barely significantly better than that of the model with eight features, besides, taking into account the computational cost and the interpretability of features, we choose the eight most important features for our logistic regression model.

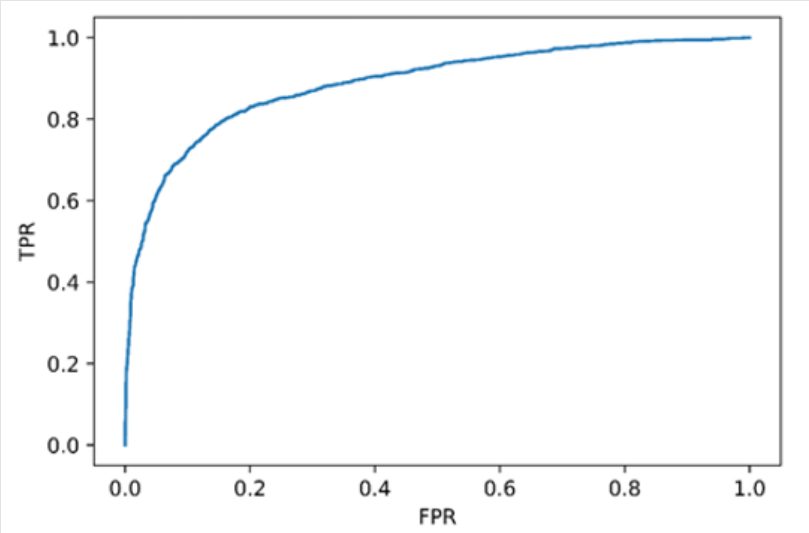

Figure 1: The ROC of XGBoost model trained with all features.

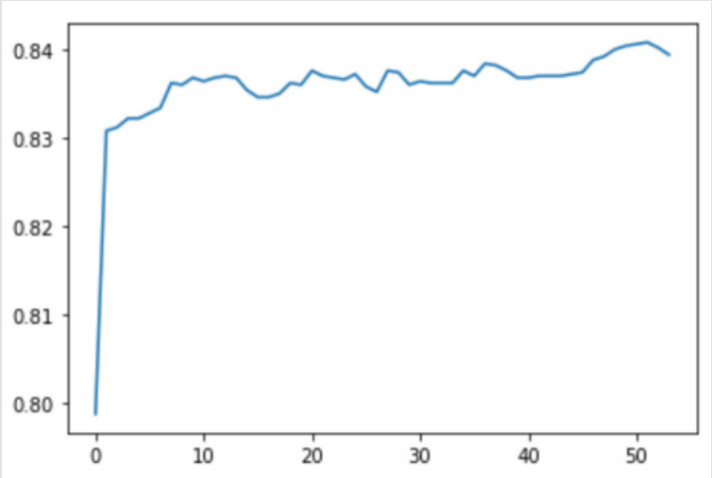

Figure 2: The curve of AUC of ROC for XGBoost models with different number of features.

Logistic Regression

Considering the trade-off between model performance and number of features, to ensure the performance, interpretability and training efficiency of the logistic regression model, we selected the top 8 features in terms of importance for logistic regression model training, which are: fasting blood glucose, urine routine urine glucose, glycated hemoglobin, blood routine hemoglobin, family history mother, electrocardiogram, blood type RH, and living environment - kitchen exhaust facilities. Logistic regression models were constructed using the above features and compared with XGBoost. The AUC of the logistic regression model ROC was 0.8536, which was slightly worse than the AUC of the XGBoost model trained with the full amount of data, but as seen from the comparison of the two ROC curves in the above, the logistic regression model performed slightly better than the XGBoost model trained with the same amount of data (Figure 3).

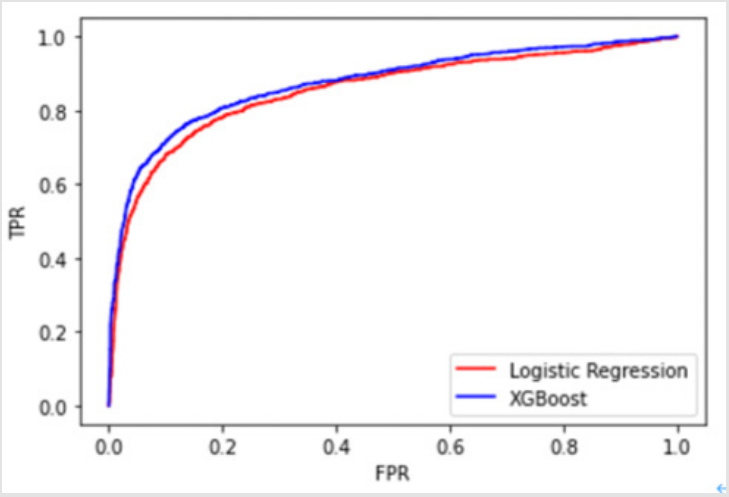

Figure 3: The ROC of logistic regression model and XGBoost model.

It is evident in other studies that XGBoost models often out perform logistic regression models hence it is generally held that XGBoost is more robust than logistic regression. However, in our study, the XGBoost model only out performs the logistic regression by a small margin. To maintain a reasonably high interpretability, we resort to selecting logistic regression and conduct further study (Table 1). Based on the logistic regression model, we obtained the ranking of the correlations between each characteristic variable and diabe-

tes mellitus. As shown in the Table 1, the correlation between fasting blood glucose level and glycosylated hemoglobin and diabetes was high and in accordance with medical common sense. Also, we can see that living environment-kitchen exhaust facilities is correlated with diabetes, and we consider that this variable reflects the economic level of that observer, and the risk of diabetes is relatively low for the observer with high economic level.

Table 1: Feature importance with scores obtained by the XGBoost model.

Features	Score
Fasting Blood Glucose Level	0.159079
Routine urine glucose	0.094123
Glycosylated Hemoglobin	0.065557
Mother with Diabetes	0.024763
Electrocardiogram	0.02286
Blood Group RH	0.018091
Living Environment-kitchen Exhaust Facilities	0.017928

In addition, we can see some correlation between routine blood hemoglobin, blood group RH, and electrocardiogram (ECG) and diabetes. Among them, the higher the level of routine hemoglobin, the lower the risk of developing diabetes. Since there is no evidence in the current medical research to prove a significant correlation between routine hemoglobin, blood group RH, and ECG and diabetes, the next study will further verify the correlation between these three characteristics and diabetes from a modeling perspective. In addition, the correlation between family history mother and diabetes was low, and family history mother here includes the history of disease of the observed mother, including diabetes, hypertension, and cardiovascular disease. Considering the significant effect of maternal diabetes on whether the offspring develop diabetes, we will further distinguish the offspring of mothers with diabetes from the offspring of mothers without diabetes in the data preprocessing stage in the next study to verify the effect of maternal diabetes on the risk of diabetes in the offspring.

Discussion

There are many domestic and international studies applied to diabetic complications, focusing on the predictive effect of different machine learning models on diabetic complications. Li, et al. [1] Used Logistic Regression (LR), Decision Tree (DT) and Neural Network (NN) to predict the occurrence of diabetic neuropathy based on personal lifestyle variables such as smoking index, whether or not to exercise, etc. Li, et al. [2] applied Logistic screening index and used Learning Vector Quantization (LVQ) neural network to build a prediction model with more satisfactory results in other than diabetic neuropathy prediction. Cho, et al. [3] used 39 features screened to build a model of diabetic nephropathy based on Support Vector Machine (SVM) and developed software for visualization. Huang, et al. [4] used artificial neural networks to model serum protein mass spectrometry for type II diabetic nephropathy and achieved good results. Li, et al. [5] used artificial neural networks to develop predictive models for diabetic complications based on clinical tests and TCM symptoms. Liu SY [6] established three risk prediction models based on optimized logistic regression models for diabetic nephropathy, diabetic retinopathy, and diabetic foot, and the models were validated with good results. Lv, et al. [7] collected 630 eyes of 315 patients with T2DM into their study, and logistic regression analysis was used to establish risk prediction models, and found that axial length (AL), age, duration of diabetes, glycosylated hemoglobin (HbAlc), and urine protein were significantly associated with the occurrence of diabetic eye lesions. Pal, et al. [8] used different machine learning models to validate the prediction of diabetic retinal complications. Ananthi, et al. [9] developed a fuzzy classification model using six clinical data variables to predict diabetic nephropathy and diabetic heart disease complications. Liu, et al. [10] screened several variables related to diabetic complications based on multilayer perceptual neural networks, respectively.

In addition, several studies have used endostatin [11,12], microRNA (miRNA) [13] and erythrocyte deformation index [14] to predict the progression of T2D complications and specific populations. Other studies have focused on the impact of glycated hemoglobin [15] and blood uric acid [16,17] on diabetic complications. In these studies, the focus was on the impact of clinical data and medical indicators on diabetic complications, but the impact of individual or common variables on the prognosis of diabetes and the determination of disease risk was missing. In this study, we explore the role of correlations between warning indicators, between warning indicators and complications, and between complications from the perspective of diabetes patients performing complication self-prevention, which is more informative for patients in the form of probabilities to let them know the likelihood of their disease. It can enable patients with type 2 diabetes to carry out more targeted self-management, enhance prevention awareness, control diet and rest, and improve lifestyle habits, thus reducing the probability of complications and alleviating patients' pain and economic burden.

References

1. Li Changping (2009) Comparison of the performance of Logistic Regression, Decision Tree and Neural Network in predicting peripheral neuropathy in type 2 diabetes mellitus [D]. Beijing: Academy of Military Medical Sciences of the Chinese People's Liberation Army.
2. Li Ge (2004) Predicting Chronic Complications of Type 2 Diabetes Based on Data Mining Technology [D]. Tianjin: Tianjin Medical University.
3. Cho B H, Yu H, Kim K W, Tae Hyun Kim, In Young Kim, et al. (2008) Application of irregular and unbalanced data to predict diabetic nephropathy using visualization and feature selection methods[J]. Artificial Intelligence in Medicine 42(1): 37-53.
4. Huang Bo (2013) Study on the diagnostic model of serum protein profile in type 2 diabetic nephropathy using artificial neural networks [J]. Laboratory Medicine and Clinical Practice 10(13): 1686-1687.

5. Li Pan (2016) Study on the prediction model of complications of type 2 diabetes based on neural networks [D]. Guangzhou: Guangzhou University of Chinese Medicine.
6. Liu Xiaoyu (2016) Study on the logistic regression model based on meta-analysis of the risk of complications of type 2 diabetes [D]. Chongqing: Third Military Medical University.
7. Lü Zhe Chen Yiqi, Shen Lijun, et al. (2017) Establishment and preliminary verification of the risk prediction model of diabetic retinopathy in patients with type 2 diabetes [J]. Chinese Journal of Fundus Diseases, p. 3.
8. Pal R, Poray J, Sen M (2017) Application of machine learning algorithms on diabetic retinopathy[C]. Sri Venkateshwara Coll Engn, Bangalore: 2nd IEEE International Conference on Recent Trends in Electronics, Information and Communication Technology (RTEICT), pp. 2046-2051.
9. Ananthi S, Bhuvaneswari V (2017) Prediction of heart and kidney risks in diabetic prone population using fuzzy classification[C]. Coimbatore: International Conference on Computer Communication and Informatics (ICCCI).
10. Liu, Mimi Cai, Yongming (2018) Research on the prediction of diabetic complications based on multilayer perceptron neural networks [J]. software 39(10): 30-35.
11. Chauhan K, Verghese D A, Rao V, Chan L, Parikh C R, et al. (2019) Plasma endostatin predicts kidney outcomes in patients with type 2 diabetes[J]. Kidney International 95(2): 439-446.
12. El-Ashmawy H M, Roshdy H S, Saad Z (2019) Serum endostatin level as a marker for coronary artery calcification in type 2 diabetic patients[J]. Journal of the Saudi Heart Association (Elsevier) 31(1): 24-31.
13. Jiménez Lucena R, Rangel-Zúniga O A, Alcalá Díaz J F, Lopez-Moreno J, Roncero-Ramos I, et al. (2018) Circulating mi RNAs as predictive biomarkers of Type 2 Diabetes Mellitus development in coronary heart disease patients from the CORDIOPREV study[J]. Molecular Therapy - Nucleic Acids 12: 146-157.
14. Lee S B, Kim Y S, Kim J H, et al. (2018) Use of RBC deformability index as an early marker of diabetic nephropathy[J]. Clinical Hemorheology and Microcirculation.
15. Tang M, Dzm A, Liu H A, Wu N, Si Y, et al. (2019) Performance of atherosclerotic cardiovascular risk prediction models in a rural northern Chinese population: Results from the Fang shan cohort study[J]. American Heart Journal 213: 34-44.
16. Meer TPVD, Wolffen buttel BHR, Patel CJ (2021) Data-driven assessment, contextualisation and implementation of 134 variables in the risk for type 2 diabetes: an analysis of Lifelines, a prospective cohort study in the Netherlands[J]. Diabetologia, p. 1-11.
17. Xla B, Ql A, Wc A, Peng S, Yexiang S, et al. (2021) A dynamic risk-based early warning monitoring system for population-based management of cardiovascular disease - ScienceDirect[J]. Fundamental Research 1(5): 534-542.

ISSN: 2574-1241

DOI: 10.26717/BJSTR.2026.65.010222

Xiaoyuan Ma and Maiming Yang. Biomed J Sci & Tech Res

This work is licensed under Creative *Commons* Attribution 4.0 License

Submission Link: https://biomedres.us/submit-manuscript.php

Assets of Publishing with us

- Global archiving of articles
- Immediate, unrestricted online access
- Rigorous Peer Review Process
- Authors Retain Copyrights
- Unique DOI for all articles

https://biomedres.us/