Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Development and Validation of a Machine Learning-Based Model for Predicting Diabetes Risk in the Health Screening Population of Lanzhou City

Yinfan Gou
Gansu University of Traditional Chinese Medicine
Ping Yang
The No. 2 People's Hospital of Lanzhou
Ling Tao
The No. 2 People's Hospital of Lanzhou
Haibin Luo
The No. 2 People's Hospital of Lanzhou
Xiaoqiang Zhang
The No. 2 People's Hospital of Lanzhou
Jinyu Wang
Lanzhou University
Sheng Li
2124025512@qq.com

The No. 2 People's Hospital of Lanzhou

Research Article

Keywords: diabetes, risk prediction model, machine learning, logistic regression, random forest, health checkup

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11274348/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Development and Validation of a Machine Learning-Based Model for Predicting Diabetes Risk in the Health Screening Population of Lanzhou City

Yinfan Gou[1] ¶, Ping Yang[2] ¶, Ling Tao[2] , Haibin Luo[2] , Xiaoqiang Zhang[2] , Jinyu Wang[3] *, Sheng Li[2] *

1 School of Public Health, Gansu University of Chinese Medicine, Lanzhou 730000, China

2 The No.2 People's Hospital of Lanzhou, Lanzhou 730046, China

3 School of Public Health, Lanzhou University, Lanzhou 730000, China

* Correspondence:
Corresponding Author:Jinyu Wang;Sheng Li
Wangjiny@lzu.edu.cn (Jinyu Wang);1178708407@qq.com(Sheng Li)

¶ These authors contributed equally to this work:Yinfan Gou and Ping Yang.

Abstract

Objective: To investigate the prevalence of diabetes and its associated risk factors among individuals undergoing health checkups at a hospital in Lanzhou, and to develop a diabetes risk prediction model based on traditional logistic regression and machine learning algorithms, thereby providing a scientific basis for the early screening and precise prevention and control of diabetes in the local population.

Methods: A total of 1,546 adults who underwent health checkups at the health examination center of a tertiary hospital in Lanzhou from January 2020 to December 2024 and met the inclusion and exclusion criteria were selected. Data collected included gender, age, body mass index (BMI), systolic blood pressure, diastolic blood pressure, fasting blood glucose, total cholesterol, triglycerides, low-density lipoprotein (LDL), and high-density lipoprotein (HDL). Intergroup comparisons were performed using the χ^2 test, t-test, or Mann–Whitney U test; variables found to be statistically significant in univariate analysis were included in forward stepwise logistic regression, and odds ratios (ORs) and 95% confidence intervals (CIs) were reported. The entire sample was randomly divided into a training set (1,082 cases) and a validation set (464 cases) in a 7:3 ratio. Category weight balancing was applied, and the classification threshold was optimized using the Joden index. The accuracy, precision, recall, F1 score, and AUC of the three models were compared.

Results: Among the 1,546 participants who underwent physical examinations, 89 cases of diabetes and 1,457 non-diabetic cases were identified, resulting in a diabetes detection rate of 5.76%. There were 973 males and 573 females. The diabetic group differed statistically from the non-diabetic group in terms of age, BMI, systolic blood pressure, diastolic blood pressure, total cholesterol, low-density lipoprotein (LDL), high-density lipoprotein (HDL), triglycerides, and fasting blood glucose levels ($P < 0.05$); specifically, the diabetic group had higher levels of age, BMI, blood pressure, total cholesterol, LDL, triglycerides, and fasting blood glucose, while HDL levels were lower. Multivariate logistic regression analysis showed that age (OR = 1.081, 95%

CI: 1.053–1.109), BMI (OR = 1.101, 95% CI: 1.022–1.185), systolic blood pressure (OR = 1.025, 95% CI: 1.004–1.047), and triglycerides (OR = 1.195, 95% CI: 1.043–1.369) were positively associated with diabetes status, while HDL-C was a protective factor (OR = 0.306, 95% CI: 0.097–0.973); Gender did not reach the conventional level of statistical significance (P = 0.065) but was retained in the predictive model. On the validation set, the logistic regression AUC was 0.770, accuracy 0.793, precision 0.165, recall 0.630, and F1 score 0.262; the AUC for the Random Forest model was 0.733, with an accuracy of 0.655, precision of 0.120, recall of 0.778, and an F1 score of 0.208; the AUC for the XGBoost model was 0.725, with an accuracy of 0.728, precision of 0.133, recall of 0.667, and an F1 score of 0.222.

Conclusion: The diabetes detection rate among the health examination population in Lanzhou City was 5.76%. Age, BMI, systolic blood pressure, triglycerides, and HDL cholesterol were independently associated with diabetes status. Among the three models, logistic regression demonstrated the best overall predictive performance, while Random Forest achieved the highest recall rate and can serve as an auxiliary tool for initial screening. Fasting blood glucose is a core indicator for the diagnosis and screening of diabetes; however, due to overlap with the definition of the outcome, it was not included in the multivariate predictive models.

Keywords: diabetes; risk prediction model; machine learning; logistic regression; random forest; health checkup

1 Introduction

Diabetes is a metabolic disorder primarily characterized by chronic hyperglycemia. It can cause long-term damage to multiple organ systems, including the cardiovascular system, kidneys, retina, and peripheral nerves, and has become a major global public health issue [1]. According to a World Health Organization report, the number of people with diabetes worldwide rose from approximately 200 million in 1990 to 830 million in 2022, and the disease burden continues to rise. The 11th edition of the International Diabetes Federation's *Diabetes Atlas* estimates that in 2024, approximately 589 million adults aged 20–79 worldwide will have diabetes, of whom about 252 million are unaware of their condition [2].

China also faces significant pressure in diabetes prevention and control. National surveillance data show that, based on a comprehensive assessment of fasting blood glucose, 2-hour postprandial glucose levels after an oral glucose tolerance test, or glycated hemoglobin, the prevalence of diabetes among Chinese adults was approximately 12.4% in 2018 [3][4], and there is still room for improvement in awareness, treatment rates, and control levels. The latest Chinese Guidelines for the Prevention and Treatment of Diabetes (2024 Edition) emphasize the need to strengthen proactive screening, risk factor management, and lifelong health management for populations at high risk of diabetes.

Previous studies have shown that advancing age, overweight and obesity, unhealthy lifestyles, and dyslipidemia are all closely associated with the risk of developing type 2 diabetes [5][6][7]. Among these, obesity, as reflected by an elevated BMI, can promote the development of diabetes through mechanisms such as insulin resistance and chronic inflammatory responses[8]; elevated triglycerides and reduced high-density lipoprotein (HDL) cholesterol are also considered important metabolic risk factors [9].

Traditional logistic regression offers advantages such as a clear model structure and interpretable effect sizes, but it makes certain assumptions regarding the form of variables and the structure of relationships. In recent years, machine learning methods have been widely adopted for disease prediction [10][11], as they can capture nonlinear relationships and variable interactions through tree-based models and ensemble learning. However, existing research suggests that machine

learning models for diabetes exhibit significant heterogeneity in terms of feature selection, data balance, validation strategies, and interpretability, and their transferability across different regional populations is limited [13].

Compared to the national population, the sociodemographic structure, lifestyle, and healthcare utilization patterns of individuals undergoing health checkups in the Northwest region exhibit certain regional characteristics. Therefore, based on health checkup data from a tertiary hospital in Lanzhou, this study compares the performance of logistic regression, random forest, and XGBoost models in identifying diabetes status and analyzes the applicability of traditional statistical models and machine learning methods in the local health checkup context.

2 Materials and Methods

2.1 Study Population

The study population consisted of individuals who underwent health examinations at the health examination center of a tertiary hospital in Lanzhou City between January 2020 and December 2024. Inclusion criteria: (1) Age ≥ 18 years; (2) Complete health examination records with no missing key variables; (3) Continuous residence in Lanzhou for at least six months. Exclusion criteria: (1) Individuals with a prior confirmed diagnosis of malignant tumors or severe mental disorders; (2) Pregnant or lactating women; (3) Outliers with missing key variables or obvious logical errors in their health examination records. A total of 1,546 examinees were ultimately included in the study.

2.2 Data Sources and Variable Definitions

General demographic indicators (gender, age), physical examination indicators (height, body weight, BMI, systolic blood pressure, diastolic blood pressure), and laboratory indicators (fasting blood glucose, total cholesterol, triglycerides, low-density lipoprotein, high-density lipoprotein) were collected. BMI = body weight (kg) / height² (m^2).

Criteria for determining diabetic status: Fasting blood glucose ≥ 7.0 mmol/L, or individuals with a prior confirmed diagnosis of diabetes who are currently receiving antidiabetic treatment, were classified as having diabetes. Current guidelines also list fasting blood glucose ≥ 7.0 mmol/L as one of the diagnostic criteria for diabetes; however, in the absence of a clear hyperglycemic state, confirmation typically requires repeat testing or the use of other diagnostic methods [1][2]. Therefore, this study primarily uses this definition for identifying diabetic status and predicting risk; a single health examination value is not directly interpreted as new-onset diabetes.

Overweight is defined as 24.0 ≤ BMI < 28.0 kg/m^2, and obesity is defined as BMI ≥ 28.0 kg/m^2; dyslipidemia is defined in accordance with the *Chinese Guidelines for Dyslipidemia Management (2023) [5][7].

2.3 Statistical Analysis

Data analysis was performed using SPSS 27.0 and Python. Continuous variables that followed a normal distribution were expressed as $\bar{x} \pm s$, and comparisons between groups were performed using the independent samples t-test; non-normal variables were expressed as M (P25, P75) and analyzed using the Mann–Whitney U test; categorical variables were expressed as n (%) and analyzed using the χ^2 test. The two-sided significance level was set at $\alpha = 0.05$.

For the analysis of influencing factors, diabetic status was used as the dependent variable (0 = non-diabetic, 1 = diabetic). Variables that were statistically significant in univariate analysis

were included in forward stepwise logistic regression, and the results are reported as β, standard error, Wald χ^2, P-value, OR, and 95% CI .

2.4 Machine Learning Model Development

Three types of models were developed using logistic regression, random forests, and XGBoost. The entire dataset was randomly split into a training set and a validation set in a 7:3 ratio; the training set was used for model parameter estimation and tuning, while the validation set was used for internal validation. Model development, validation, and results reporting were conducted in accordance with relevant reporting guidelines for clinical prediction models [14–15]. Given that there were only 89 positive diabetes cases in the tabular data, resulting in a significantly imbalanced class distribution, class weight balancing (class_weight=balanced) was applied during the training phase, and classification thresholds were selected based on the principle of maximizing the Jorden index. The Random Forest model was built as an ensemble classifier based on Bootstrap aggregation and random feature selection [10]; XGBoost is a gradient-boosting tree framework that improves classification performance by progressively fitting residuals [11]. Model performance was evaluated using Accuracy, Precision, Recall, F1 score, and AUC [19].

3 Results

3.1 Baseline Characteristics

A total of 1,546 participants aged 19 to 77 years were included in the study, with a mean age of (41.71 ± 10.32) years. There were 973 males (63.0%) and 573 females (37.0%); 89 participants had diabetes, while 1,457 did not, resulting in a diabetes detection rate of 5.76%. The mean age in the diabetes group was 49.46 years, higher than the 41.23 years in the non-diabetes group; BMI, systolic blood pressure, and diastolic blood pressure were also higher than those in the non-diabetes group.

Table 1. Comparison of General Characteristics Between the Diabetes and Non-Diabetes Groups Among Health Examination Participants at a Hospital in Lanzhou

Variables	Study Population (n=1,546)	Non-diabetic group (n=1,457)	Diabetes Group (n=89)	Statistics	P-value
Gender (Male/Female)	973/573	897/560	76/13	20.49	<0.001
Age	41.71±10.32	41.23±10.14	49.46±10.27	- 7.43	<0.001
BMI	23.68±3.45	23.55±3.40	25.82±3.58	- 6.10	<0.001
Systolic blood pressure	121.74±16.42	120.92±15.95	135.03±18.29	- 8.03	<0.001
Diastolic blood pressure	79.01±11.15	78.53±10.92	86.46±11.82	- 6.59	<0.001
Total Cholesterol	4.53±0.85	4.52±0.83	4.83±1.07	- 3.40	<0.001
Low-Density Lipoprotein	2.66±0.71	2.65±0.70	2.82±0.71	- 2.16	0.031
High-Density	1.21±0.32	1.22±0.32	1.05±0.29	- 4.88	<0.001

Lipoprotein					
Triglycerides*	1.43(0.99,2.18)	1.39(0.97,2.10)	2.35(1.48,3.60)	- 6.73	<0.001
Fasting Blood Glucose*	5.20(4.86,5.67)	5.16(4.84,5.57)	8.92(7.70,11.53)	- 15.86	<0.001

Note: Variables marked with an asterisk (*) do not follow a normal distribution and are presented as median (P25, P75); comparisons between groups were performed using the Mann-Whitney U test. Other continuous variables are presented as mean ± standard deviation, and comparisons between groups were performed using the independent samples t-test. Comparisons of categorical variables between groups were performed using the chi-square (χ^2) test.

3.2 Univariate Analysis of Diabetes Prevalence

As shown in Table 1, there were statistically significant differences ($P < 0.05$) between the diabetes group and the non-diabetes group in terms of gender, age, BMI, systolic blood pressure, diastolic blood pressure, total cholesterol, triglycerides, low-density lipoprotein (LDL), high-density lipoprotein (HDL), and fasting blood glucose. Specifically, triglycerides and fasting blood glucose levels in the diabetes group tended to be significantly higher than those in the non-diabetes group, while HDL cholesterol levels were lower than in the non-diabetes group.

3.3 Multivariate Analysis of Diabetes Prevalence

Variables found to be statistically significant in the univariate analysis were included in a forward stepwise logistic regression. As shown in Table 2, age, BMI, systolic blood pressure, triglycerides, and HDL cholesterol were included in the final model and reached statistical significance. For every 1-year increase in age, the odds ratio (OR) for diabetes prevalence was 1.081 (95% CI: 1.053–1.109, $P < 0.001$); for every 1 kg/m² increase in BMI, the OR was 1.101 (95% CI: 1.022–1.185, $P = 0.011$); for every 1 mmHg increase in systolic blood pressure, the OR was 1.025 (95% CI: 1.004–1.047, $P = 0.020$); and for every 1 mmol/L increase in triglycerides, the OR was 1.195 (95% CI: 1.043-1.369, $P = 0.010$). Elevated high-density lipoprotein cholesterol was associated with a lower risk of diabetes (OR = 0.306, 95% CI: 0.097–0.973, $P = 0.045$). The OR for gender was 1.892 (95% CI: 0.962–3.721, $P = 0.065$), which did not reach the conventional level of significance but was retained in the predictive model as a clinical adjustment variable.

Table 2. Multivariate Logistic Regression Analysis of Factors Influencing Diabetes in a Population Undergoing Health Screenings at a Hospital in Lanzhou

Variables	β value	Standard error	Wald χ^2 value	P-value	OR value	95%CI
Age	0.078	0.013	34.478	<0.001	1.081	（1.053，1.109）
BMI	0.096	0.038	6.495	0.011	1.101	（1.022，1.185）
Systolic blood pressure	0.025	0.011	5.427	0.020	1.025	（1.004,1.047）
Diastolic blood pressure	- 0.006	0.016	0.200	0.655	0.993	（0.962,1.025）
Total Cholesterol	0.180	0.267	0.457	0.499	1.198	（0.710,2.019）

Triglycerides	0.178	0.070	6.549	0.010	1.195	（1.043,1.369）
Low-Density Lipoprotein	0.095	0.290	0.290	0.743	1.100	（0.622,1.943）
High-Density Lipoprotein	－1.183	0.589	0.589	0.045	0.306	（0.097,0.973）
Gender (Male)	0.637	0.345	0.345	0.065	1.892	（0.962,3.721）
Constants	－12.006	1.578	1.578	<0.001	-	-

Note: Gender was retained as a clinical adjustment variable.

3.4 Development and Validation of a Diabetes Risk Prediction Model

Based on the six independent risk factors identified through multivariate logistic regression (age, BMI, systolic blood pressure, triglycerides, HDL, and gender), diabetes risk prediction models were constructed using three algorithms: logistic regression, random forest, and XGBoost. The dataset was randomly split in a 7:3 ratio into a training set (1,082 cases, 70%) and a validation set (464 cases, 30%). To address the class imbalance—with a ratio of diabetic to non-diabetic samples of approximately 1:16—model training employed class weight adjustment (class_weight='balanced') and optimized classification thresholds (based on maximizing the Jorden index) to enhance the models' ability to identify the minority class (diabetic patients). The performance evaluation results of each model on the validation set are shown in Table 3.

The results show that the AUC values for the three models ranged from 0.725 to 0.770, indicating moderate overall discriminatory ability. Among them, the logistic regression model had the highest AUC (0.770), and its accuracy (0.793) and F1 score (0.262) were also superior to those of the other two models, indicating that it has a relative advantage in terms of a balanced consideration of sensitivity and precision, while also featuring a simple model structure and strong clinical interpretability. The Random Forest model had the highest recall (0.778), suggesting that it can identify nearly 80% of patients with diabetes and has a low rate of missed diagnoses, making it suitable for large-scale population screening; however, its precision was relatively low (0.120), and its positive predictive value needs improvement, so individuals with positive screening results require further confirmation in conjunction with other clinical indicators. The XGBoost model's metrics were moderate, with an AUC of 0.725 and a recall of 0.667.

Given the data imbalance, the precision of all three models was generally low (0.120–0.165), which is a systemic characteristic of classification tasks involving imbalanced data. Considering the AUC, accuracy, and F1 score collectively, the logistic regression model demonstrated the best overall performance in terms of predictive capability and clinical interpretability. Model performance could be further improved in the future by increasing the sample size, introducing resampling techniques such as SMOTE, or incorporating additional potential predictors. The ROC curves for each model are shown in Figure 1.

Table 3. Results of the Evaluation of the Predictive Performance of the Diabetes Prediction Model

Model	Accuracy	Accuracy Rate	Recall Rate	F1 Score	AUC value
Logistic Regression	0.793	0.165	0.630	0.262	0.770

Random Forest	0.655	0.120	0.778	0.208	0.733
XGBoost	0.728	0.133	0.667	0.222	0.725

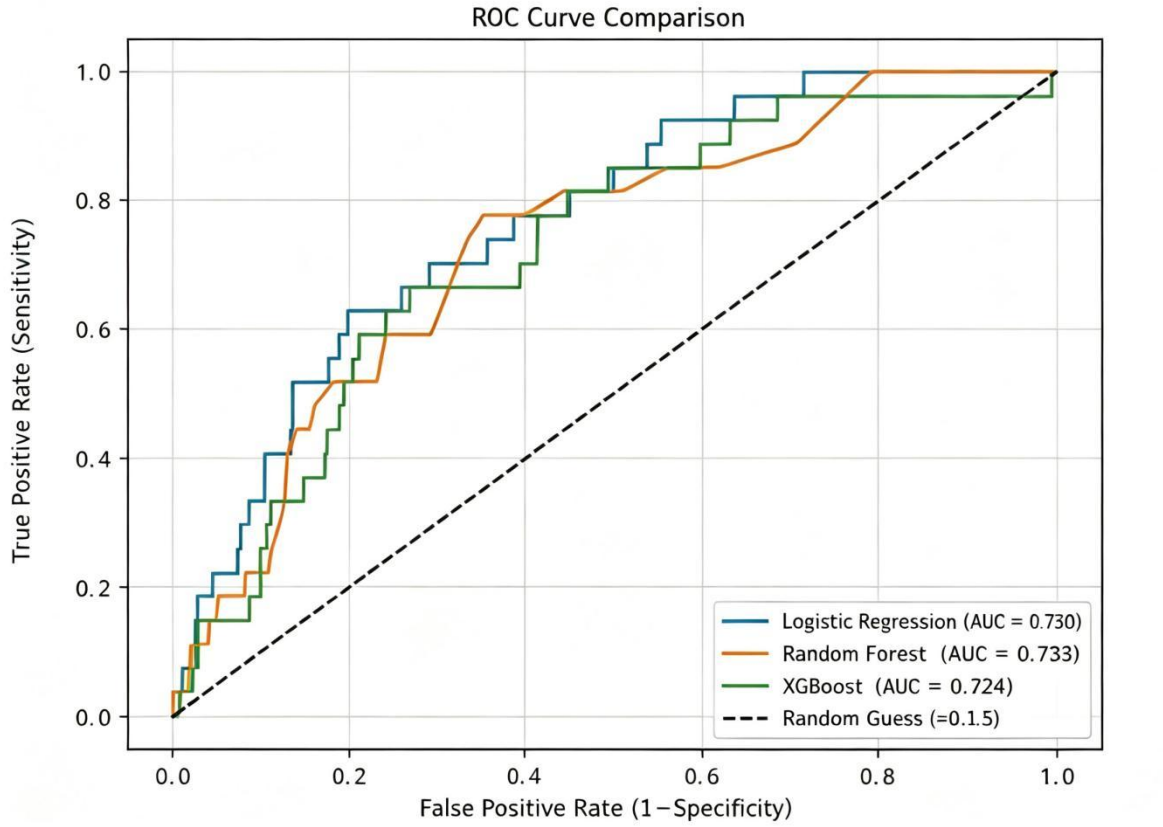


Figure 1. ROC Curve for the Diabetes Risk Prediction Model

4 Discussion

This study was based on data from a population undergoing health examinations at a tertiary hospital in Lanzhou City. A total of 1,546 examinees were included, of whom 89 had diabetes and 1,457 did not; the diabetes detection rate was 5.76%. This detection rate is higher than the previously reported prevalence of 4.2% among residents of Lanzhou, suggesting that the burden of diabetes among the local health screening population should not be overlooked; however, it is lower than the 2018 national prevalence of 12.4% among adults and the 10.6% prevalence among adults in Gansu Province [5], which may be related to the fact that the study population consisted primarily of middle-aged and young employed individuals undergoing health screenings, resulting in a younger age distribution [6]. Therefore, proactive screening and stratified management of high-risk populations for diabetes should still be strengthened in health examination settings, with particular attention given to metabolic abnormalities already present in young and middle-aged adults [7][8].

The results of the univariate analysis showed that there were statistically significant differences ($P < 0.05$) between the diabetes group and the non-diabetes group in terms of gender, age, BMI, blood pressure, and lipid profile. Overall, the diabetes group was characterized by higher age, BMI, and blood pressure levels, accompanied by abnormal lipid metabolism.

Multivariate logistic regression results showed that age, BMI, systolic blood pressure, triglycerides, and HDL cholesterol were included in the final model and reached statistical significance. For every 1-year increase in age, the odds ratio (OR) for diabetes was 1.081 (95% CI: 1.053–1.109, $P < 0.001$), representing an approximately 8.1% increase in risk; For every 1 kg/m^2 increase in BMI, the odds ratio (OR) was 1.101 (95% CI: 1.022–1.185, $P = 0.011$), representing an approximately 10.1% increase in risk; For every 1 mmHg increase in systolic blood pressure, the OR was 1.025 (95% CI: 1.004–1.047, $P = 0.020$), representing an approximately 2.5% increase in risk; For every 1 mmol/L increase in triglycerides, the OR was 1.195 (95% CI: 1.043–1.369, $P = 0.010$), representing an approximately 19.5% increase in risk; Elevated high-density lipoprotein cholesterol was associated with a lower prevalence of diabetes [8][9], with an OR of 0.306 (95% CI: 0.097–0.973, $P = 0.045$), suggesting it may be a protective factor. The OR for gender was 1.892 (95% CI: 0.962–3.721, $P = 0.065$), which did not reach the conventional significance level of 0.05; however, it was retained in the predictive model as a clinical adjustment variable. Diastolic blood pressure, total cholesterol, and low-density lipoprotein cholesterol did not reach statistical significance in the multivariable model. Fasting blood glucose showed the most pronounced difference in univariate analysis; however, since the definition of diabetes in this study itself included fasting blood glucose ≥7.0 mmol/L or a history of diabetes treatment, using it as a predictor could lead to label leakage or circular reasoning. Therefore, it was not included as an independent predictor in the final multivariate predictive model; nevertheless, its status as a core indicator for diabetes screening remains clear.

In terms of model performance, among the three models developed in this study, the logistic regression model achieved an AUC of 0.770, a precision of 0.793, and an F1 score of 0.262 on the validation set, demonstrating the best overall performance [13]. Furthermore, with its simple structure and high interpretability, it is well-suited for risk stratification in clinical health screening scenarios. The Random Forest model had the highest recall rate at 0.778, indicating a strong ability to identify patients with diabetes and a low rate of missed diagnoses, making it suitable for initial screening; however, its precision was only 0.120, accuracy was 0.655, and F1 score was 0.208, suggesting a high number of false positives, and those with positive screening results require further confirmation. The XGBoost model had an AUC of 0.725, a precision of 0.133, a recall of 0.667, and an F1 score of 0.222, indicating average overall performance. The AUCs of all three models ranged between 0.725 and 0.770, reflecting moderate discriminatory power, suggesting that there is still room for improvement in predicting diabetes risk based solely on routine physical examination indicators.

In this study, the precision of the three models was generally low, ranging from 0.120 to 0.165, primarily due to the small number of positive diabetes cases (only 89) and a positive rate of just 5.76%. In severely imbalanced datasets, models tend to correctly identify the majority of non-diabetes cases, while the precision, recall, and F1 score for positive cases are highly sensitive to thresholds and sample distribution [16–17] . Therefore, future studies should expand the sample size—particularly by increasing the number of positive diabetes cases—and explore resampling methods such as SMOTE [12], cost-sensitive learning, or ensemble strategies. Additionally, they should include the PR-AUC and 95% CI to more comprehensively evaluate the models' practical value in positive case identification [18].

The innovative aspects of this study are as follows: First, by focusing on Lanzhou—a representative city in Northwest China—this study complements existing research on diabetes risk prediction based on local health examination populations [19]; second, it systematically compares the performance of traditional logistic regression with two machine learning algorithms—Random Forest and XGBoost—in identifying diabetes status; third, all model input variables are derived from routine health examination items, including age, BMI, blood pressure, and blood lipid levels, ensuring good accessibility and potential for widespread application. However, this study also has certain limitations: First, the data were collected from a single

center, and the sample size was relatively limited, with only 89 cases of diabetes; the model has not yet undergone external validation, and its generalizability requires further testing [20–21]. Second, this study employed a retrospective cross-sectional design, which does not allow for the inference of causal relationships; third, potential important variables such as diet, physical activity, smoking, alcohol consumption, and family history of diabetes were not included, which may affect the model's predictive accuracy; finally, there is an overlap between fasting blood glucose and the definition of the outcome, which may influence variable selection and model interpretation [22]. In the future, efforts should be made to expand the sample size through multicenter collaboration, incorporate a broader range of predictive factors, and conduct prospective cohort studies and external validation to further enhance the model's robustness and clinical applicability.

In summary, the diabetes detection rate among the health examination population in Lanzhou was 5.76%, and age, BMI, systolic blood pressure, triglycerides, and high-density lipoprotein cholesterol were independently associated with diabetic status. Among the three predictive models constructed based on routine health examination indicators, the logistic regression model demonstrated the best overall predictive performance, while the random forest model had the highest recall rate, making both suitable as auxiliary tools for initial screening. If the goal is to achieve a balance between overall discrimination and interpretability, logistic regression should be prioritized; if the goal is to reduce missed diagnoses, the random forest model may be considered, but positive cases should be further confirmed in conjunction with other clinical indicators [23–24]. The aforementioned models can serve as a reference for the early identification and health management of high-risk populations for diabetes among residents of Lanzhou City.

5 Ethics approval and consent to participate

This study was conducted in accordance with the ethical principles of the Declaration of Helsinki and was approved by the Ethics Committee of Lanzhou Second People's Hospital (Approval No. KY2026060025). Because the study was retrospective and used anonymized clinical data, the requirement for informed consent was waived by the Ethics Committee. All data were de-identified to protect participant confidentiality and privacy.

6 Consent for publication

Not applicable.

7 Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

8 Funding

This study was supported by the Natural Science Foundation of Gansu Province (26JRRA180), the Lanzhou Municipal Science and Technology Program (2025-3-066, 2025-3-070, and 2023-2-72), and the Lanzhou Municipal Health Industry Scientific Research Project (A2026011), all awarded to Sheng Li. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

9 Authors' contributions

All authors contributed to the study conception and design. Yinfan Gou and Ping Yang contributed equally to this work and share first authorship. They conceived and designed the

study, and drafted the initial manuscript. Ling Tao and Haibin Luo were responsible for data collection and material preparation. Xiaoqiang Zhang performed the formal data analysis and statistical modeling. Jinyu Wang and Sheng Li are the corresponding authors; they supervised the research, critically revised the manuscript for important intellectual content, and provided administrative support. All authors reviewed and approved the final version of the manuscript.

10 Acknowledgments

The authors are deeply grateful to all staff participating in this study.

11 Data Availability Statement

The datasets used and/or analyzed during the current study are available from the corresponding author upon reasonable request.

12 References

[1] World Health Organization. Diabetes. Geneva: World Health Organization; 2024.

[2] International Diabetes Federation. IDF Diabetes Atlas. 11th ed. Brussels: International Diabetes Federation; 2025.

[3] Li Y, Teng D, Shi X, et al. Prevalence of diabetes recorded in mainland China using 2018 diagnostic criteria from the American Diabetes Association: national cross sectional study. BMJ. 2020;369:m997. DOI:10.1136/bmj.m997.

[4] Wang L, Peng W, Zhao Z, et al. Prevalence and treatment of diabetes in China, 2013-2018. JAMA. 2021;326(24):2498-2506. DOI:10.1001/jama.2021.22208.

[5] Chan JCN, Malik V, Jia W, et al. Diabetes in Asia: epidemiology, risk factors, and pathophysiology. JAMA. 2009;301(20):2129-2140. DOI:10.1001/jama.2009.726.

[6] Hu FB. Globalization of diabetes: the role of diet, lifestyle, and genes. Diabetes Care. 2011;34(6):1249-1257. DOI:10.2337/dc11-0442.

[7] Yu D, Martin CB, Fryar CD, et al. Prevalence of diabetes by BMI: China Nutrition and Health Surveillance (2015-2017) and U.S. National Health and Nutrition Examination Survey (2015-2018). AJPM Focus. 2024;3(3):100215. DOI:10.1016/j.focus.2024.100215.

[8] Zhong HJ, Luo LM, Wang XM, et al. Association between triglyceride to HDL cholesterol ratio and a risk of diabetes mellitus: a systematic review and meta-analysis. Lab Med. 2025;56(1):1-6. DOI:10.1093/labmed/lmae052.

[9] Qiang S, Zhang Y, et al. Dose-response association between the triglycerides: high-density lipoprotein cholesterol ratio and type 2 diabetes mellitus risk: the rural Chinese cohort study and meta-analysis. Diabetes Metab Res Rev. 2019;35(3):e3096. DOI:10.1002/dmrr.3096.

[10] Breiman L. Random forests. Mach Learn. 2001;45:5-32. DOI:10.1023/A:1010933404324.

[11] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM; 2016. p. 785-794. DOI:10.1145/2939672.2939785.

[12] Chawla NV, Bowyer KW, Hall LO, et al. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321-357. DOI:10.1613/jair.953.

[13] Christodoulou E, Ma J, Collins GS, et al. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol. 2019;110:12-22. DOI:10.1016/j.jclinepi.2019.02.004.

[14] Collins GS, Reitsma JB, Altman DG, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ. 2015;350:g7594. DOI:10.1136/bmj.g7594.

[15] Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. DOI:10.1136/bmj-2023-078378.

[16] Wolff RF, Moons KGM, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med. 2019;170(1):51-58. DOI:10.7326/M18-1376.

[17] Riley RD, Snell KIE, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: part II—binary and time-to-event outcomes. Stat Med. 2019;38(7):1276-1296. DOI:10.1002/sim.7992.

[18] Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. BMJ. 2020;368:m441. DOI:10.1136/bmj.m441.

[19] Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21(1):128-138. DOI:10.1097/EDE.0b013e3181c30fb2.

[20] Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J. 2014;35(29):1925-1931. DOI:10.1093/eurheartj/ehu207.

[21] Van Calster B, McLernon DJ, Van Smeden M, et al. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17:230. DOI:10.1186/s12916-019-1466-7.

[22] Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26(6):565-574. DOI:10.1177/0272989X06295361.

[23] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44(3):837-845.

[24] Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996;15(4):361-387. DOI:10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4.

[25] Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. 2nd ed. Cham: Springer; 2019. DOI:10.1007/978-3-030-16399-0.