






Value of machine learning algorithms for predicting diabetes risk: A subset analysis from a real-world retrospective cohort study

Yaqian Mao^{1†} , Zheng Zhu^{2†} , Shuyao Pan^{2†}, Wei Lin² , Jixing Liang², Huibin Huang², Liantao Li², Junping Wen² , Gang Chen^{2,3*} 

¹Department of Internal Medicine, Fujian Provincial Hospital South Branch, Shengli Clinical Medical College of Fujian Medical University, Fuzhou, China, ²Department of Endocrinology, Fujian Provincial Hospital, Shengli Clinical Medical College of Fujian Medical University, Fuzhou, China, and ³Fujian Provincial Key Laboratory of Medical Analysis, Fujian Academy of Medical, Fuzhou, China

Keywords

Diabetes, Machine learning algorithms, Predictive model

*Correspondence

Gang Chen
Tel: +86-135-0933-7027
Fax: +86-591-8755-7768
E-mail address:
chengangfj@163.com

J Diabetes Investig 2023; 14: 309–320

doi: [10.1111/jdi.13937](https://doi.org/10.1111/jdi.13937)

ABSTRACT

Aims/Introduction: To compare the application value of different machine learning (ML) algorithms for diabetes risk prediction.

Materials and Methods: This is a 3-year retrospective cohort study with a total of 3,687 participants being included in the data analysis. Modeling variable screening and predictive model building were carried out using logistic regression (LR) analysis and 10-fold cross-validation, respectively. In total, six different ML algorithms, including random forests, light gradient boosting machine, extreme gradient boosting, adaptive boosting (AdaBoost), multi-layer perceptrons and gaussian naive bayes were used for model construction. Model performance was mainly evaluated by the area under the receiver operating characteristic curve. The best performing ML model was selected for comparison with the traditional LR model and visualized using Shapley additive explanations.

Results: A total of eight risk factors most associated with the development of diabetes were identified by univariate and multivariate LR analysis, and they were visualized in the form of a nomogram. Among the six different ML models, the random forests model had the best predictive performance. After 10-fold cross-validation, its optimal model has an area under the receiver operating characteristic value of 0.855 (95% confidence interval [CI] 0.823–0.886) in the training set and 0.835 (95% CI 0.779–0.892) in the test set. In the traditional LR model, its area under the receiver operating characteristic value is 0.840 (95% CI 0.814–0.866) in the training set and 0.834 (95% CI 0.785–0.884) in the test set.

Conclusions: In the real-world epidemiological research, the combination of traditional variable screening and ML algorithm to construct a diabetes risk prediction model has satisfactory clinical application value.

INTRODUCTION

With the continuous changes in lifestyle and eating habits, the incidence of diabetes increases year by year. Meanwhile, diabetes has become one of the most important chronic non-communicable diseases in the world, with huge impacts on the health of humans. According to the latest statistics of the

International Diabetes Federation¹, there were approximately 536.6 million people (aged 20–79 year) with diabetes worldwide in 2021, and this number will rise to 783.2 million in 2045. Global diabetes-related health spending was estimated at \$US966 billion in 2021, and it is projected to reach \$US1,054 billion by 2045¹. Unfortunately, more than four out of five people with diabetes (80.6%, 432.7 million) live in low- and middle-income countries. According to International Diabetes Federation statistics, in 2021, a total of 141 million adults

†These authors contributed equally to this work.

Received 1 February 2022; revised 4 October 2022; accepted 16 October 2022

(aged 20–79 years) in China had diabetes, and more than half of them were undiagnosed². It can be seen that the current situation of diabetes is very serious. Therefore, early identification and intervention of diabetes risk factors plays a crucial role in preventing the occurrence and development of diabetes.

Modern medicine is faced with a large amount of data collection and analysis, as well as the clinical challenge of applying the acquired knowledge to solve complex problems^{3,4}. Machine learning (ML) methods can use recognized patterns to predict new data, which is conducive to finding difficult-to-recognize patterns from a complex combination of multiple clinical markers⁵. In the medical field, the most commonly used ML methods are random forests (RF), light gradient boosting machine, extreme gradient boosting, adaptive boosting, multi-layer perceptrons and gaussian naive bayes. The advantages of these technologies are that they can capture the non-linear relationship in the data, and improve the accuracy and effectiveness of the model. Many studies have proven that they have good performance in disease prediction and diagnosis^{6–17}.

It should be noted that there was controversy about ML models and traditional regression models in predicting the risk of disease. Studies have shown that ML models have better predictive performance than traditional regression models^{16–19}. There are also studies showing that, compared with traditional methods, the predictive performance of ML is not better than that of conventional regression models^{20,21}. There is still disagreement on whether to choose ML algorithm or conventional regression analysis for chronic disease prediction model construction in clinical work. The reason for the difference in predicting risk between the two methods is related to the size of the dataset, the type of variables and the incidence of positive events. Compared with classical modeling techniques, such as logistic regression (LR), modern modeling techniques, such as support vector machines, neural networks and RF, might require more than 10-fold the number of events per variable to achieve a stable area under the receiver operating characteristic (AUROC) curve and predictive advantages²². This means that this modern technique is only suitable for medical prediction problems when there are very large datasets.

In fact, in real-world studies, there are many data with complex variables and few positive events. Therefore, we urgently need to compare the predictive performance of ML algorithms and traditional regression analysis on these data, so as to select the best performing predictive model. The purpose of the present study was to evaluate the performance of several commonly used ML algorithms in diabetes risk prediction and compare them with traditional LR models. Our hypothesis was that ML classifiers have equally strong predictive power and clinical utility in dealing with real-world epidemiological studies.

MATERIALS AND METHODS

Participants

The data came from the chronic disease research database of Wuyishan City, Fujian Province, China. This is a retrospective

cohort study that belongs to the Fujian subcohort of the Chinese Diabetic Tumor Risk Assessment Study (REACTION study)²³. The study period was from March 2011 to January 2015. The study used a cluster sampling method to randomly select approximately 4,314 residents in the area, which has a typical representative. From March to December 2011, through a baseline survey, the distribution of diabetes among residents in this area was roughly understood. To further explore risk factors associated with the development of diabetes, we carried out a retrospective analysis of all residents who were followed up for approximately 36 months in the region. All participants met the following inclusion criteria: (i) there was no limit to men and women, with age >40 years; and (ii) the baseline and follow-up data were complete and not missing. Exclusion criteria were: (i) patients with known diabetes at baseline ($n = 268$); (ii) patients with newly diagnosed diabetes at baseline ($n = 359$); (iii) taking drugs that affect glucose metabolism before the examination²⁴, such as long-term antipsychotics, antidepressant treatments, statins or glucocorticoid treatments and so on; ($n = 0$); and (iv) suffering from diseases that affect glucose metabolism²⁴, such as polycystic ovary syndrome and so on ($n = 0$). A total of 3,687 participants were finally included, and the average follow-up time was approximately 36 months. At the end of the follow-up period, anthropometric measurements, blood tests and auxiliary tests for each patient were carried out again. This study was approved by the Ethics Committee of Fujian Provincial Hospital (ID: K2021-01-026) and it conforms to the provisions of the Declaration of Helsinki. The detailed research flowchart is shown in Figure 1.

Data collection

Questionnaire investigation

All participants completed a standard questionnaire to collect the information, including age, sex, family history of diabetes, education level, marital status, comorbidity (impaired fasting glucose [IFG], impaired glucose tolerance [IGT], hypertension, dyslipidemia, fatty liver, abdominal obesity, overweight, obesity, osteopenia and osteoporosis), personal history (drinking history, smoking history, tea drinking history and load exercise) and eating habits (seafood, fruits, eggs, dairy products and soy products). The history of smoking or alcohol was classified into three levels, namely current (smoking or drinking alcohol in the past 6 months), past (smoking or drinking alcohol in the past for > 6 months) and never. The first two levels were defined as having a history of smoking or alcohol. Tea drinking history was divided into three levels: never, occasionally and often. Load exercise includes playing basketball, swimming, running and so on for more than three times a week, at least 30 min each time. Past medical history should be definitively diagnosed by a clinician in a secondary or tertiary hospital.

Anthropometric measuring

Anthropometric measurements included height, weight, waist circumference, hip circumference and neck circumference.

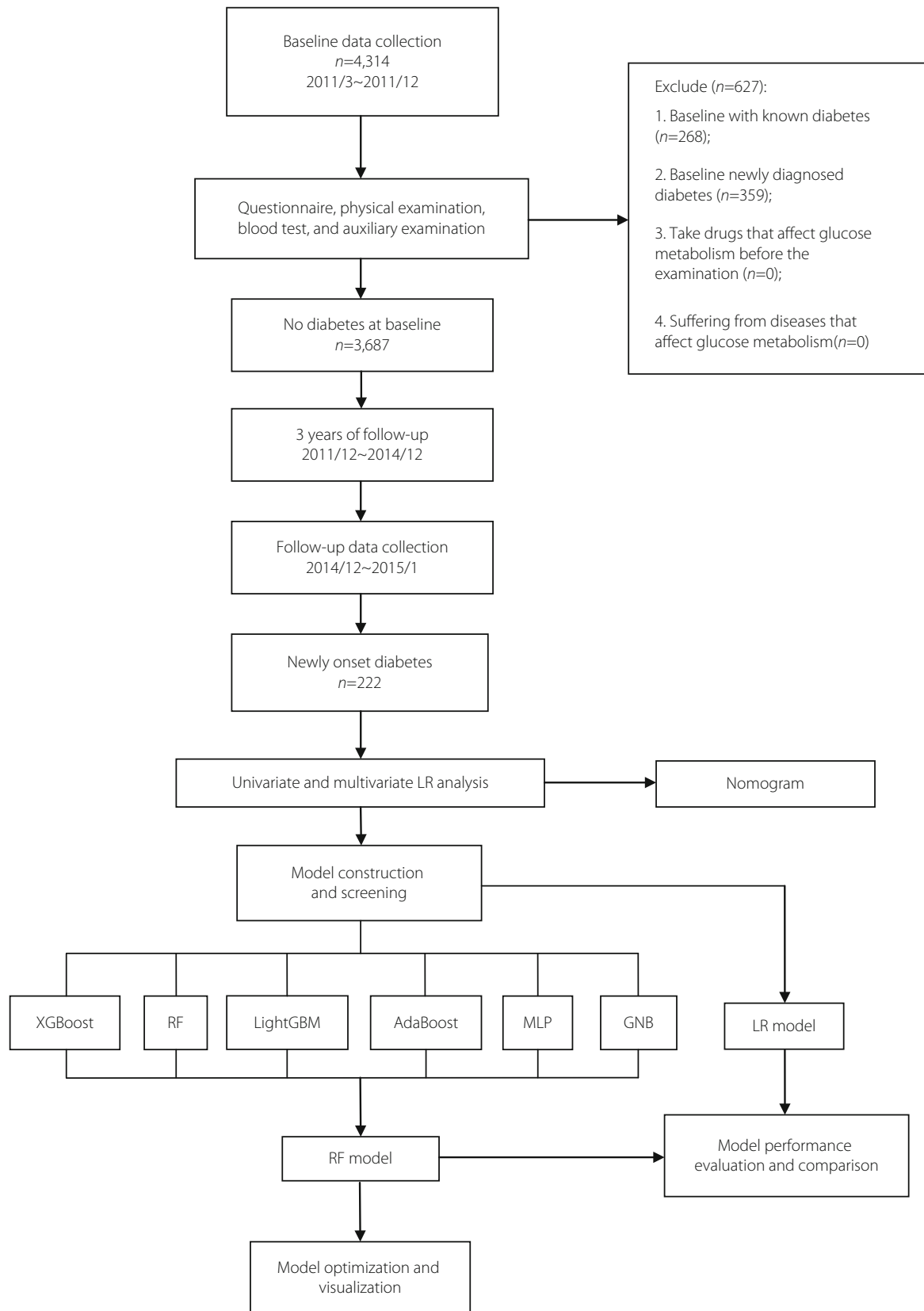


Figure 1 | Flowchart of the cohort study. AdaBoost, adaptive boosting; GNB, Gaussian naive bayes; LightGBM, light gradient boosting machine; LR, logistic regression; MLP, multi-layer perceptrons; RF, random forests; XGBoost, extreme gradient boosting.

Participants wore light clothing and no shoes. The researchers measured the participants' height and weight using calibrated height-weight scales. Waist circumference was measured at the waist level transumbilical point; neck circumference was measured at the junction of the superior border of the seventh cervical vertebra at the back of the neck and the inferior junction of the anterior Adam's apple; hip circumference was measured at the most convex point of the pubic symphysis and gluteus maximus. Auxiliary examinations included blood pressure, bone mineral density and brachial-ankle pulse wave velocity. The participants received the blood pressure measure of the right upper arm in a sitting position after resting for at least 10 min. The blood pressure was measured three times with an Omron sphygmomanometer (Kyoto, Japan), and the average values were taken for statistical analysis. Brachial-ankle pulse wave velocity was measured by Japan Omron BP-203RPEIII arteriosclerosis doppler ultrasound automatic analyzer. According to the automatic display of the instrument, the brachial-ankle pulse wave velocity values of the left and right sides were obtained, and the average value of both sides was taken for statistical analysis. The left heel bone was selected for bone mineral density measurement, and the Achilles Express ultrasonic bone density analyzer (GE Lunar Corp., Madison, WI, USA) was used for bone mineral density measurement.

Blood detection

All participants fasted overnight for at least 10 h, the blood samples were collected early in the morning and the standard 75-g oral glucose tolerance test was carried out next. Blood biochemical indicators included fasting plasma glucose (FPG), 2-h plasma glucose (2hPG) after the 75-g oral glucose tolerance test, fasting serum insulin, high-density lipoprotein cholesterol, low-density lipoprotein cholesterol, total cholesterol, triacylglycerol, alanine aminotransferase, aspartate aminotransferase, gamma glutamyl transpeptidase, alkaline phosphatase and uric acid. According to the $(FPG \times \text{fasting serum insulin}) / 22.5$, the insulin resistance index was calculated. The blood glucose was determined by the glucose oxidase method. The remaining blood samples were detected by chemiluminescence methods with an autoanalyzer. Data collection and analysis were completed by two collaborators (Y Mao and Z Zhu), and disputes were resolved by discussion.

Definition and classification of variables

According to the guideline for the prevention and treatment of type 2 diabetes in China (2020 edition)²⁴, diabetes is defined as $FPG \geq 7.0$ mmol/L and/or $2hPG \geq 11.1$ mmol/L; IFG is defined as 6.1 mmol/L $\leq FPG < 7.0$ mmol/L and $2hPG < 7.8$ mmol/L; and IGT is defined as $FPG < 7.0$ mmol/L and 7.8 mmol/L $\leq 2hPG < 11.1$ mmol/L; IFG and IGT are known as prediabetes. See Table S1 for classification and definitions of other variables.

Variable screening and traditional regression model construction

Logistic regression, a commonly used statistical model, uses logistic functions to model binary dependent variables. It not only provides the probability of occurrence of the predicted outcome, but also additional information on the prediction results, such as odds ratio (OR), 95% confidence interval (CI) and so on²⁵. In the present study, univariate and multivariate LR analysis were used to screen independent risk factors for diabetes, and visualized risk factors with $P < 0.05$ in multivariate analysis in the form of a nomogram. The statistical definition of a nomogram is the geometric expression of a mathematical formula, which graphically shows the interaction and superposition of predictors, and provides patients with an individualized disease risk assessment²⁶.

Significant ($P < 0.05$) risk factors in multivariate analysis were used as modeling variables to construct a traditional LR model, and 25% of the internal data were randomly selected as the test set for validation.

ML model

In this study, first, a 10-fold cross-validation approach was used to build the model with the best predictive performance. Second, the data were split into training set (validation set) and test set (25%). The best ML model was further optimized by 10-fold cross-validation for better performance; 10-fold cross-validation is currently a preferred technique in computer science^{27,28}, and it randomly divided all samples in the data into 10 mutually exclusive subsets of similar size and number of events. The optimal model should be trained and tested 10 times, nine subsets were selected as the training set and validation set each time, and the remaining subsets were used as the test set. Afterwards, Shapley additive explanation (SHAP) was used to visually explain the effects of important variables on the model. SHAP is a game theory-based framework²⁹ that has recently been shown to be effective in explaining various ML models^{30–32}. SHAP actually attributes the model output value to the Shapley value of each feature. Intuitively, the contribution of a feature to an outcome can be explained by estimating the Shapley value of each feature. The Shapley value can intuitively reflect the influence of the feature in each sample, and further understand whether the feature is a protective factor or a risk factor of the model. Finally, our algorithm ranked eight risk factors based on RF by variable importance and was compared with the SHAP method.

Model performance evaluation was mainly carried out by the AUROC, sensitivity, specificity and negative predictive value. The ML classifier with the largest AUROC value was selected as the best model and compared with the traditional nomogram model. As an indicator of comprehensive evaluation sensitivity and specificity, AUROC provides a more intuitive standard for judging the accuracy of prediction models³³. The larger the AUROC is, the higher the accuracy of its prediction is.

Statistical analysis

Statistical analysis was carried out with IBM SPSS software (version 25.0 for windows; SPSS Inc., Chicago, IL, USA), R software (version 3.6.3; The R Foundation for Statistical Computing, Vienna, Austria) and Python software (version 3.7.0; Beaverton, OR, USA). Categorical variables were expressed as frequency (n/N , %). Baseline data analysis was carried out using the χ^2 -test, and two-sided $P < 0.05$ was considered statistically significant.

RESULTS

Demographics features

A total of 3,687 participants were included in the present study, ranging in age from 41 to 79 years. In total, 222 patients developed diabetes after 3 years of follow up. Combined with literature reading and expert knowledge, 45 risk factors associated with the occurrence of diabetes were finally identified (Table S1). Table 1 shows a comparison of baseline characteristics between diabetes and non-diabetes patients.

Variable screening and model construction

Univariate and multivariate LR analyses were used for feature variable screening (Table S2). According to the LR analysis results, eight factors, including age, family history, IFG, IGT, hypertension, triacylglycerol, alanine aminotransferase and

gamma glutamyl transpeptidase, were finally selected as modeling variables (see Figures 2a,b for details).

Based on the 10-fold cross-validation, a diabetes prediction model was built using six different ML classifiers (Figure 2c,d). According to AUROC value ranking, in the training set, the ML classifier with the best model performance was RF (AUROC 0.848, 95% CI 0.820–0.876). Also in the validation set, the ML classifier with the best model performance was RF (AUROC 0.826, 95% CI 0.741–0.912). Therefore, we hold the view that among all ML classifiers, the RF model has the best predictive performance. Table 2 presents the prediction performance comparison of six different ML classifiers. In the model construction, the definitions and parameter settings of six different ML classifiers are detailed in Table S3. Furthermore, based on the univariate and multivariate LR analyses results, a traditional diabetes prediction model was also constructed. In the training set (full data), the model has an AUROC value of 0.840 (95% CI 0.814–0.866), and in the test set (random 25% data), the model has an AUROC value of 0.834 (95% CI 0.785–0.884; see Figure S1 for details). The performance comparison of the best ML model (RF model) and the traditional LR model is shown in Table 3.

To further improve the performance and accuracy of the model, 10-fold cross-validation was used to train and test the RF model. Figure 3a–c show the training process of the RF

Table 1 | Baseline characteristic analysis of 3,687 patients with or without type 2 diabetes mellitus

Baseline characteristics	Non-diabetes <i>n</i> = 3,465	Diabetes <i>n</i> = 222	<i>P</i>	Baseline characteristics	Non-diabetes <i>n</i> = 3,465	Diabetes <i>n</i> = 222	<i>P</i>
Demographic characteristics				Laboratory examination			
Sex				TG (mmol/L)			
Female	1880 (54.26)	119 (53.60)	0.850	<1.7	2,155 (62.19)	87 (39.19)	<0.001
Male	1,585 (45.74)	103 (46.40)		≥1.7	1,310 (37.81)	135 (60.81)	
Age (years)				TC (mmol/L)			
<50	1,168 (33.71)	36 (16.22)	<0.001	<5.2	1,731 (49.96)	97 (43.69)	0.070
50–60	1,420 (40.98)	88 (39.64)		≥5.2	1,734 (50.04)	125 (56.31)	
60–70	659 (19.02)	72 (32.43)		HDL-C (mmol/L)			
≥70	218 (6.29)	26 (11.71)		≥1.0	3,118 (89.99)	196 (88.29)	0.416
Family history				<1.0	347 (10.01)	26 (11.71)	
No	3,242 (93.56)	200 (90.09)	0.044	LDL-C (mmol/L)			
Yes	223 (6.44)	22 (9.91)		<3.4	2,432 (70.19)	136 (61.26)	0.005
Education level				≥3.4	1,033 (29.81)	86 (38.74)	
Below junior high school	1,079 (31.14)	78 (35.14)	0.421	UA (μmol/L)			
Junior high school	950 (27.42)	60 (27.03)		≤420	2,708 (78.15)	163 (73.42)	0.100
Above junior high school	1,436 (41.44)	84 (37.84)		>420	757 (21.85)	59 (26.58)	
Marital status				FINS			
Married	3,296 (95.12)	209 (94.14)	0.514	≤5	1,562 (45.08)	59 (26.58)	<0.001
Others	169 (4.88)	13 (5.86)		5–10	1,519 (43.84)	103 (46.40)	
Personal history				>10	384 (11.08)	60 (27.03)	
Smoking history				IR			
No	2,350 (67.82)	161 (72.52)	0.145	<2.69	3,148 (90.85)	162(72.97)	<0.001
Yes	1,115 (32.18)	61 (27.48)		≥2.69	317 (9.15)	60(27.03)	

Table 1. (Continued)

Baseline characteristics	Non-diabetes <i>n</i> = 3,465	Diabetes <i>n</i> = 222	<i>P</i>	Baseline characteristics	Non-diabetes <i>n</i> = 3,465	Diabetes <i>n</i> = 222	<i>P</i>
Drinking history				VAI			
No	1,500 (43.29)	102 (45.95)	0.439	<1	783 (22.60)	30 (13.51)	<0.001
Yes	1,965 (56.71)	120 (54.05)		1–2	1,264 (36.48)	59 (26.58)	
Tea drinking history				2–3	665 (19.19)	48 (21.62)	
Never	1,323 (38.18)	85 (38.29)	0.920	3–4	347 (10.01)	34 (15.32)	
Occasionally	1,103 (31.83)	73 (32.88)		≥4	406 (11.72)	51 (22.97)	
Often	1,039 (29.99)	64 (28.83)		Anthropometric characteristics			
Load exercise				BMI (kg/m ²)			
No	2,905 (83.84)	188 (84.69)	0.740	<24	1,765 (50.94)	70 (31.53)	<0.001
Yes	560 (16.16)	34 (15.32)		24–28	1,327 (38.30)	97 (43.69)	
Comorbidity (yes)				≥28	373 (10.77)	55 (24.78)	
IFG	285 (8.23)	38 (17.12)	<0.001	HC (cm)			
IGT	536 (15.47)	131 (59.01)	<0.001	<95	2,059 (59.42)	98 (44.14)	<0.001
Hypertension	423 (12.21)	66 (29.73)	<0.001	95–105	1,302 (37.58)	102 (45.95)	
Dyslipidemia	671 (19.37)	67 (30.18)	<0.001	≥105	104 (3.00)	22 (9.91)	
Fatty liver	82 (2.37)	14 (6.31)	<0.001	NC (cm)			
Abdominal obesity	653 (18.85)	94 (42.34)	<0.001	<35	2,221 (64.10)	111 (50.00)	<0.001
Overweight	1,262 (36.42)	102 (45.95)	0.004	≥35	1,244 (35.90)	111 (50.00)	
Obesity	343 (9.90)	49 (22.07)	<0.001	WHR			
Osteopenia	592 (17.09)	51 (22.97)	0.025	<0.5	1,164 (33.59)	36 (16.22)	<0.001
Osteoporosis	100 (2.89)	12 (5.41)	0.034	≥0.5	2,301 (66.41)	186 (83.78)	
Laboratory examination				WHR [†]			
ALT (U/L)				<0.85/0.9	1,508 (43.52)	56 (25.23)	<0.001
≤25	2,761 (79.68)	138 (62.16)	<0.001	≥0.85/0.9	1,957 (56.48)	166 (74.78)	
25–50	608 (17.55)	59 (26.58)		Auxiliary examination			
>50	96 (2.77)	25 (11.26)		SBP (mmhg)			
AST (U/L)				<140	2,343 (67.62)	101 (45.50)	<0.001
≤20	1,823 (52.61)	93 (41.89)	<0.001	≥140	1,122 (32.38)	121 (54.51)	
20–40	1,516 (43.75)	99 (44.60)		DBP (mmhg)			
>40	126 (3.64)	30 (13.51)		<90	2,916 (84.16)	178 (80.18)	0.118
GGT (U/L)				≥90	549 (15.84)	44 (19.82)	
≤30	2,265 (65.37)	96 (43.24)	<0.001	BaPWV (cm/s)			
30–60	771 (22.25)	71 (31.98)		≤1,400	1,710 (49.35)	54 (24.32)	<0.001
>60	429 (12.38)	55 (24.78)		1,400–1,800	1,268 (36.60)	93 (41.89)	
ALP (U/L)				>1,800	487 (14.06)	75 (33.78)	
≤100	2,849 (82.22)	156 (70.27)	<0.001	Eating habits (yes)			
100–125	454 (13.10)	46 (20.72)		Seafood	3,214 (92.76)	200 (90.09)	0.141
>125	162 (4.68)	20 (9.01)		Fruit	3,241 (93.54)	199 (89.64)	0.024
Non-HDL-C (mmol/L)				Egg	2,937 (84.76)	176 (79.28)	0.029
<4.1	2,131 (61.50)	108 (48.65)	<0.001	Soy products	2,735 (78.93)	173 (77.93)	0.722
≥4.1	1,334 (38.50)	114 (51.35)		Dairy products	1,706 (49.24)	94 (42.34)	0.046

[†]Waist-to-hip ratio (WHR) was calculated as waist circumference divided by hip circumference. We considered abdominal obesity as WHR ≥0.9 in male or WHR ≥0.85 in female. ALP, alkaline phosphatase; ALT, alanine aminotransferase; AST, aspartate aminotransferase; baPWV, brachial-ankle pulse wave velocity; BMI, body mass index; DBP, diastolic blood pressure; FINS, fasting serum insulin; GGT, gamma glutamyl transpeptidase; HC, hip circumference; HDL-C, high-density lipoprotein cholesterol; IFG, impaired fasting glucose; IGT, impaired glucose tolerance; IR, insulin resistance; LDL-C, low-density lipoprotein cholesterol; NC, neck circumference; Non-HDL-C, non-high-density lipoprotein cholesterol; SBP, systolic blood pressure; TC, total cholesterol; TG, total triglyceride; UA, uric acid; VAI, visceral adiposity index; WHtR, waist-to-height ratio.

model. When the training results of the training set and validation set tend to be consistent, the prediction performance of the RF model is the best (Figure 3d). Its optimal AUROC value is 0.855 (95% CI 0.823–0.886) in the training set, 0.821 (95% CI 0.721–0.921) in the validation set and 0.835 (95% CI 0.779–

0.892) in the test set. At the same time, its optimal parameters are set as: criterion gini, max depth 4, min impurity decrease 0 and *n* estimators 200. Figure 3e is the variable importance ranking based on the RF algorithm. Figure 3f is a visualization of the RF model. As shown in Figure 3f, the larger the Shapley

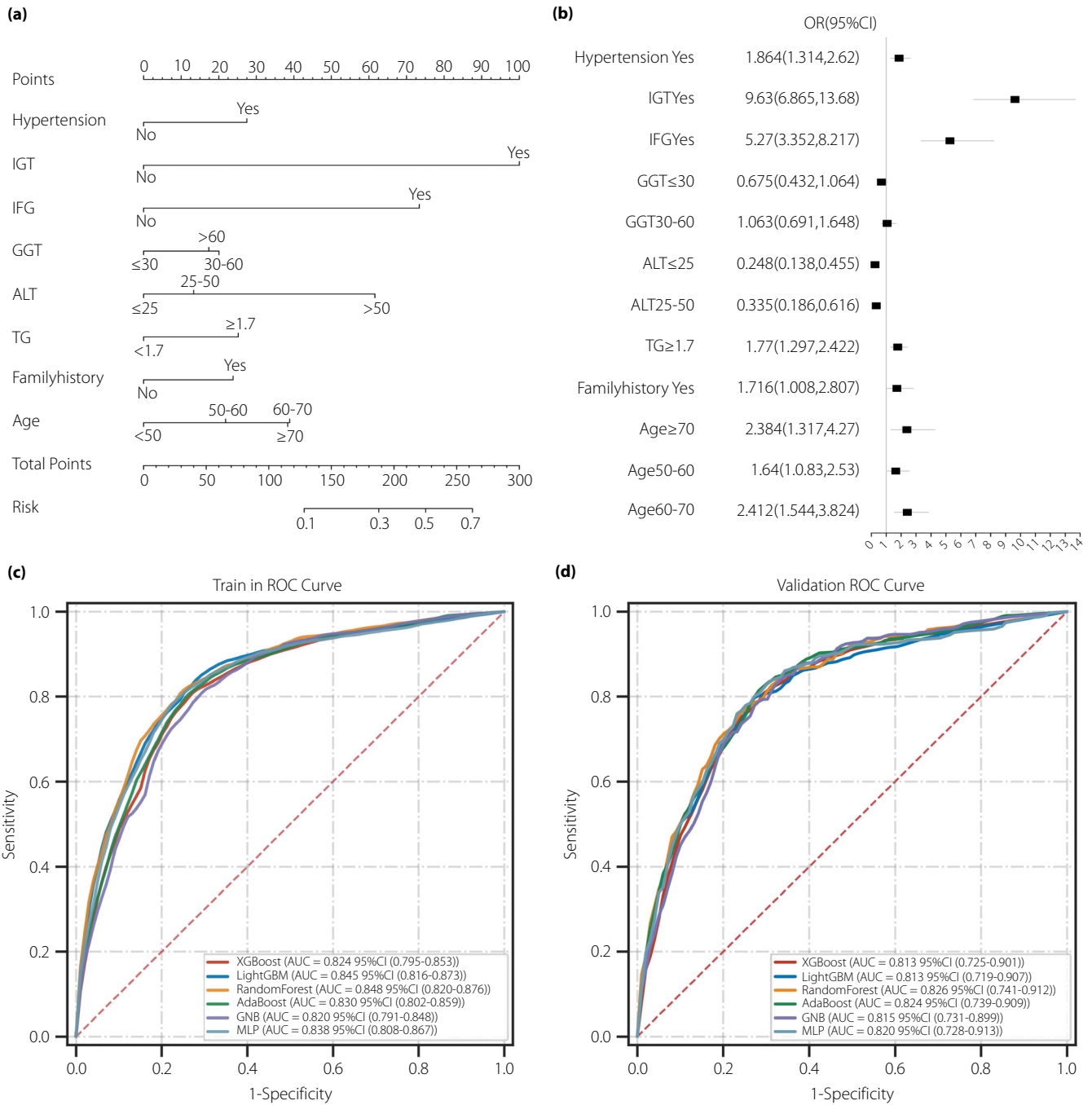


Figure 2 | Logistic regression analysis and model construction. (a,b) Feature variable screening by logistic regression (LR) analysis. (a) Nomogram of diabetes risk factors. (b) Forest plot of diabetes risk factors. According to the LR analysis results, eight factors, including age, family history, impaired fasting glucose (IFG), impaired glucose tolerance (IGT), hypertension, triacylglycerol, alanine aminotransferase (ALT) and gamma glutamyl transpeptidase (GGT) were finally selected as modeling variables. (c,d) Area under the receiver operating characteristic values of six different machine learning models in training set and validation set. AdaBoost, adaptive boosting; AUC, area under the curve; CI, confidence interval; GNB, Gaussian naive bayes; LightGBM, light gradient boosting machine; MLP, multi-layer perceptrons; OR, odds ratio; RF, random forests; ROC, receiver operating characteristic; TG, total triglyceride; XGBoost, extreme gradient boosting.

Table 2 | Prediction performance comparison of six different machine learning classifiers on training and validation sets

ML classifiers		Performance (95% CI)			
		AUROC	Accuracy	Sensitivity	Specificity
Training set	XGBoost	0.824 (0.795–0.853)	0.761 (0.748–0.774)	0.801 (0.780–0.822)	0.743 (0.726–0.759)
	LightGBM	0.845 (0.816–0.873)	0.754 (0.735–0.773)	0.824 (0.806–0.843)	0.741 (0.721–0.761)
	RF [†]	0.848 (0.820–0.876)	0.768 (0.759–0.777)	0.807 (0.793–0.821)	0.764 (0.753–0.775)
	AdaBoost	0.830 (0.802–0.859)	0.762 (0.752–0.772)	0.795 (0.784–0.805)	0.754 (0.742–0.766)
	GNB	0.820 (0.791–0.848)	0.730 (0.722–0.738)	0.805 (0.797–0.813)	0.719 (0.710–0.728)
	MLP	0.838 (0.808–0.867)	0.764 (0.749–0.780)	0.807 (0.791–0.824)	0.761 (0.743–0.779)
Validation set	XGBoost	0.813 (0.725–0.901)	0.756 (0.738–0.773)	0.820 (0.788–0.852)	0.738 (0.688–0.787)
	LightGBM	0.813 (0.719–0.907)	0.746 (0.722–0.770)	0.797 (0.744–0.851)	0.754 (0.712–0.796)
	RF [‡]	0.826 (0.741–0.912)	0.761 (0.742–0.780)	0.820 (0.774–0.866)	0.759 (0.715–0.804)
	AdaBoost	0.824 (0.739–0.909)	0.756 (0.740–0.772)	0.829 (0.794–0.864)	0.749 (0.701–0.797)
	GNB	0.815 (0.731–0.899)	0.728 (0.715–0.741)	0.834 (0.776–0.893)	0.724 (0.673–0.774)
	MLP	0.820 (0.728–0.913)	0.762 (0.744–0.780)	0.842 (0.804–0.881)	0.739 (0.681–0.797)

[†]Machine learning (ML) model with the best prediction performance in the training set is random forests (RF; sorted according to the size of the area under the receiver operating characteristic [AUC] value). [‡]ML model with the best prediction performance in the validation set is RF (sorted according to the size of the AUC value). AdaBoost, adaptive boosting; CI, confidence interval; GNB, Gaussian naive bayes; LightGBM, light gradient boosting machine; MLP, multi-layer perceptrons; RF, random forests; XGBoost, extreme gradient boosting.

Table 3 | Model performance comparison

Model	RF model	LR model
Variable	Age, family history, IFG, IGT, hypertension, TG, ALT, GGT	Age, family history, IFG, IGT, hypertension, TG, ALT, GGT
Training set	AUROC: 0.855 (95% CI: 0.823–0.886)	AUROC: 0.840 (95% CI, 0.814–0.866)
Test set	AUROC: 0.835 (95% CI: 0.779–0.892)	AUROC: 0.834 (95% CI, 0.785–0.884)

ALT, alanine aminotransferase; AUROC, area under the receiver operating characteristic; CI, confidence interval; GGT, gamma glutamyl transpeptidase; IFG, impaired fasting glucose; IGT, impaired glucose tolerance; LR, logistic regression; RF, random forests; TG, total triglyceride.

value of a variable is, the larger the SHAP value is and the greater the probability of diabetes is. The red point in the figure represents that the variable of the corresponding sample plays a positive role in probability prediction, and the blue point plays a negative role.

DISCUSSION

Artificial intelligence-based ML technology has been widely used in the area of tumors and chronic diseases^{11,19,34–36}. Obviously, this innovative method is an important tool in the field of precision medicine, which can facilitate the selection of the best diagnosis and treatment strategies. However, currently there is little known about the value of ML algorithms in predicting the risk of developing chronic diseases. How to intervene risk factors in the early stage of diabetes is the fundamental way to achieve individualized optimal medical care. In the present study, traditional variable screening methods and novel ML algorithms were combined to assess the risk of developing diabetes after 3 years in the general population. We found that the diabetes prediction model built based on the ML algorithm has the same good prediction performance as the traditional LR model.

In recent years, relevant literature has reported the positive application of ML algorithms in diabetes risk prediction^{37–39}. In 2019, Xiong *et al.*³⁷ compared the ability of five different ML methods in predicting diabetes risk factors, and the results showed that combining ML methods could provide an accurate diabetes risk prediction assessment model. Wang *et al.*³⁹ developed and validated a predictive model without biochemical parameters to identify individuals at high risk for diabetes. The findings suggested that the artificial neural network classifier based on demographic, lifestyle and anthropometric data was an effective predictive model³⁹. Birk *et al.*⁴⁰ applied several ML and statistical methods, combined with survey data from the Food Frequency Questionnaire to calculate a global diet quality score, and developed a predictive tool for screening for prediabetes. Findings showed that the model had a positive predictive effect. Furthermore, Zhang *et al.*⁴¹ developed a data mining approach to characterize the risk of diabetes using a series of ML models. The results show that ML has superior capabilities in identifying risk factors and predicting outcomes over a large range of data, and an increasing number of variables⁴¹. However, the aforementioned studies all have certain limitations. First, these studies are cross-sectional surveys, so the model

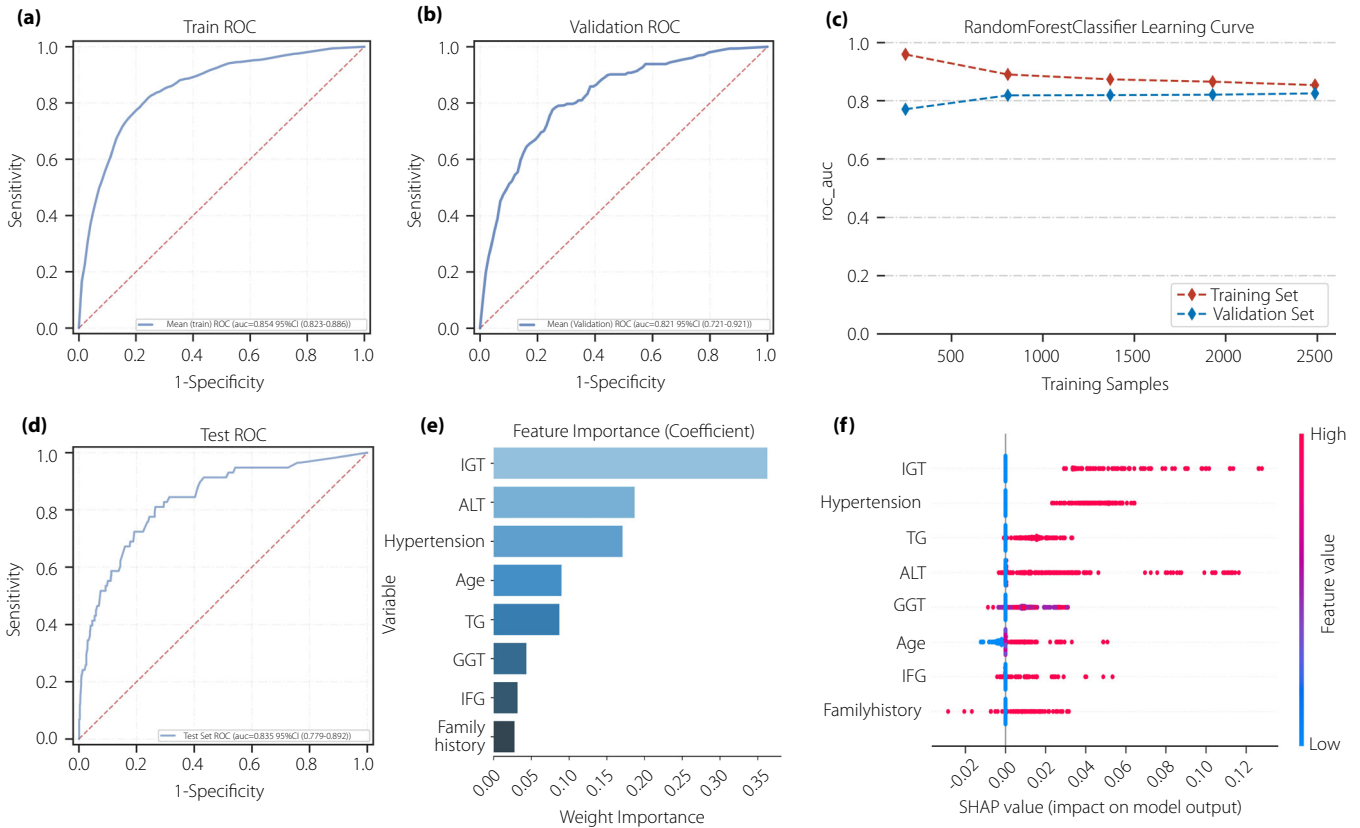


Figure 3 | Optimization and visualization of random forests (RF) model. (a–c) Fitting optimization process of RF model by 10-fold cross-validation. (d) Area under the receiver operating characteristic value of RF model in the test set. (e) Variable importance ranking based on the RF algorithm. (f) Shapley additive explanations (SHAP) of the RF model. Model parameters in (e,f) are set as follows: Criterion: gini, max depth: 3, min impurity decrease: 0; n estimators: 10. ALT, alanine aminotransferase; AUC, area under the curve; GGT, gamma glutamyl transpeptidase; IFG, impaired fasting glucose; IGT, impaired glucose tolerance; ROC, receiver operating characteristic; TG, total triglyceride.

based on this data might not be able to accurately predict the risk of diabetes or prediabetes in the future. Second, in terms of algorithms, the algorithms used in some studies are relatively old, so the latest method needs to be further refined. The present study was a retrospective cohort study of approximately 36 months, with a total of 45 risk factors, which has strong clinical reference significance for exploring disease risk factors and building predictive models. We know that in exploring the occurrence and development of chronic diseases, there are strong correlations among many variables. Especially, as the number of variables increases, the multicollinearity between explanatory variables might cause various problems. In the present study, a diabetes risk prediction model was constructed using a novel ML algorithm. It was found that the prediction models constructed by these six novel ML algorithms have strong predictive value, among which the RF model has the best prediction performance.

Random forests is a common ensemble learning model, which can further improve the performance of the model by synthesizing the classification results of multiple weak

classifiers^{42,43}. Compared with other ML classifiers, RF has the following advantages. First, the RF model has high accuracy, strong generalization ability, fast operation speed and easy implementation. Second, the RF algorithm is good at identifying important relevant variables from high-dimensional feature variables, eliminating redundant and irrelevant features, and improving the predictive ability of the model. The two randomization processes (training sample randomization and feature randomization) in the RF model make RF more advantageous when dealing with high-dimensional data problems, and also provide stronger generalization capabilities⁴⁴. From predicting postoperative complications in elderly head and neck squamous cell carcinoma patients to identifying malignant pulmonary nodules^{45,46}, the ability of RF models to solve clinical problems has received increasing attention.

In the present study, the feature importance ranking based on the RF algorithm was used to screen the variables. By calculating the weight (the number of times the feature is used to split nodes) of each feature in the model, the cover (the average number of times a feature is covered by each split), gain (the

average gain of each split) and other indicators, we can roughly observe the variable dimension that features play an important role in the model. The eight most influential variables were screened from the 3,687 samples as variables to construct the model. However, feature importance only provides the important variable, rather than a judgment on the positive or negative effects of the results. Therefore, to better show the interpretability of the model, we introduced SHAP.

By showing the performance of the model on the full dataset, the SHAP method makes up for this shortcoming, giving both the importance of the variables and the positive and negative effects of the results. SHAP is a model post-exposure approach that interprets complex ML models and quantifies the contribution of each feature to the predictions made by the model. In the present study, the bar chart of the absolute value of the Shapley value of each feature was shown on the whole dataset after averaging, and the dispersion of the Shapley value on different features can also be plotted for each sample on the full dataset. Dot plots imply the importance of each feature to better help clinical researchers understand the role of each feature in the model.

The significance of the present study is that it is a real-world risk assessment cohort study based on 3,687 samples. By comparing the performance of six classic ML algorithms, the ML model with the best predictive performance is screened and compared with the traditional LR model. In a traditional ML model, algorithms only rank the importance of variables. However, the extent to which these variables affected outcome events could not be measured, which means that we were unable to assess whether the risk factor had positive or negative effects on the outcome event. To better interpret the results of the ML model, in the present study, the SHAP method was used to carry out a visual analysis of the RF model. The SHAP method is a reliable method to make the output of RF models clinically interpretable. In addition, this study combines traditional variable selection methods and ML algorithms. On the one hand, it can well explain the relationship between variables and outcome events. On the other hand, ML algorithms can easily absorb new clinical data and continuously update and optimization.

In fact, performance comparisons between ML algorithms and traditional statistical analysis are highly dependent on the nature of a given dataset. In the absence of a given dataset, we cannot conclude whether one method is better than the other. In real-event research, one method cannot be the best choice in all situations, and the choice of which method can be used will largely depend on the nature of the given dataset and the goals of the researcher. When the primary goal is to analyze the relationship between risk factors and outcome events, traditional statistical methods might be particularly useful. ML methods can be particularly useful when there are significant correlations between variables in a dataset or when the amount of data is particularly large.

The present study also has the following limitations. First, due to the epidemiological nature of this study, the possibility of confounding cannot be completely eliminated. Second, the risk factors included in this study are limited, and need to be further improved and supplemented in future studies. Third, all study participants were from the same center, so some cautions should be exercised in summarizing our findings. In future work, opportunities for multicenter collaboration and improved data mining capabilities will provide more opportunities for multicenter prospective studies.

The present study proved that ML models have predictive performance as good as traditional statistical analysis. Most importantly, this study can provide effective key information in predicting the risk of developing diabetes, which guides the research on other chronic diseases.

ACKNOWLEDGMENT

The authors thank the Department of Endocrinology, Shengli Clinical Medical College of Fujian Medical University.

FUNDING

This work was supported by the Chinese Medical Association Foundation and Chinese Endocrine Society (Grant number: 12020240314), the National Natural Science Foundation of China (Grant number: 81270874), the Natural Science Foundation of Fujian Province (Grant number: 2011J06012) and the Startup Fund for Scientific Research, Fujian Medical University (Grant number: 2020QH2049).

DISCLOSURE

The authors declare no conflict of interest.

Approval of the research protocol: N/A.

Informed consent: This study has been approved by the Ethics Committee of Fujian Provincial Hospital (ID: K2021-01-026).

Registry and the registration no. of the study/trial: N/A.

Animal studies: N/A.

REFERENCES

1. Sun H, Saeedi P, Karuranga S, *et al.* IDF diabetes atlas: global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes Res Clin Pract* 2022; 183: 109119.
2. Ogurtsova K, Guariguata L, Barengo NC, *et al.* IDF diabetes atlas: global estimates of undiagnosed diabetes in adults for 2021. *Diabetes Res Clin Pract* 2022; 183: 109118.
3. Abhari S, Niakan Kalhori SR, Ebrahimi M, *et al.* Artificial intelligence applications in type 2 diabetes mellitus care: focus on machine learning methods. *Healthc Inform Res* 2019; 25: 248–261.
4. Fallah M, Niakan Kalhori SR. Systematic review of data mining applications in patient-centered mobile-based information systems. *Healthc Inform Res* 2017; 23: 262–270.

5. Vokinger KN, Feuerriegel S, Kesselheim AS. Continual learning in medical devices: FDA's action plan and beyond. *Lancet Digit Health* 2021; 3: e337–e338.
6. He J, Baxter SL, Xu J, *et al.* The practical implementation of artificial intelligence technologies in medicine. *Nat Med* 2019; 25: 30–36.
7. Jiang F, Jiang Y, Zhi H, *et al.* Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol* 2017; 2: 230–243.
8. Hsieh MH, Sun LM, Lin CL, *et al.* The performance of different artificial intelligence models in predicting breast cancer among individuals having type 2 diabetes mellitus. *Cancers (Basel)* 2019; 11: 1751.
9. Rau HH, Hsu CY, Lin YA, *et al.* Development of a web-based liver cancer prediction model for type II diabetes patients by using an artificial neural network. *Comput Methods Programs Biomed* 2016; 125: 58–65.
10. Hsieh MH, Sun LM, Lin CL, *et al.* Development of a prediction model for colorectal cancer among patients with type 2 diabetes mellitus using a deep neural network. *J Clin Med* 2018; 7: 277.
11. Yoo TK, Kim SK, Kim DW, *et al.* Osteoporosis risk prediction for bone mineral density assessment of postmenopausal women using machine learning. *Yonsei Med J* 2013; 54: 1321–1330.
12. Weng SF, Reys J, Kai J, *et al.* Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS One* 2017; 12: e0174944.
13. Senders JT, Staples PC, Karhade AV, *et al.* Machine learning and neurosurgical outcome prediction: a systematic review. *World Neurosurg* 2018; 109: 476–486.e1.
14. Kruppa J, Liu Y, Diener HC, *et al.* Probability estimation with machine learning methods for dichotomous and multicategory outcome: applications. *Biom J* 2014; 56: 564–583.
15. Taylor RA, Pare JR, Venkatesh AK, *et al.* Prediction of in-hospital mortality in emergency department patients with sepsis: a local big data-driven, machine learning approach. *Acad Emerg Med* 2016; 23: 269–278.
16. Singal AG, Mukherjee A, Elmunzer BJ, *et al.* Machine learning algorithms outperform conventional regression models in predicting development of hepatocellular carcinoma. *Am J Gastroenterol* 2013; 108: 1723–1730.
17. Churpek MM, Yuen TC, Winslow C, *et al.* Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. *Crit Care Med* 2016; 44: 368–374.
18. Wu X, Yuan X, Wang W, *et al.* Value of a machine learning approach for predicting clinical outcomes in young patients with hypertension. *Hypertension* 2020; 75: 1271–1278.
19. Xu Y, Yang X, Huang H, *et al.* Extreme gradient boosting model has a better performance in predicting the risk of 90-day readmissions in patients with ischaemic stroke. *J Stroke Cerebrovasc Dis* 2019; 28: 104441.
20. Nusinovi S, Tham YC, Chak Yan MY, *et al.* Logistic regression was as good as machine learning for predicting major chronic diseases. *Clin Epidemiol* 2020; 122: 56–69.
21. 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.
22. van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC Med Res Methodol* 2014; 14: 137.
23. Bi Y, Lu J, Wang W, *et al.* Cohort profile: risk evaluation of cancers in Chinese diabetic individuals: a longitudinal (REACTION) study. *J Diabetes* 2014; 6: 147–157.
24. Chinese Diabetes Society. Guideline for the prevention and treatment of type 2 diabetes mellitus in China (2020 edition). *Chin J Diabetes Mellitus* 2021; 13: 95 (Chinese).
25. Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. *J Biomed Inform* 2002; 35: 352–359.
26. Balachandran VP, Gonen M, Smith JJ, *et al.* Nomograms in oncology: more than meets the eye. *Lancet Oncol* 2015; 16: e173–e180.
27. Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: a comparison of resampling methods. *Bioinformatics* 2005; 21: 3301–3307.
28. Witten IH, Frank E. Data mining: practical machine learning tools and techniques. *ACM Sigmod Record* 2011; 31: 76–77.
29. Roth AE. Lloyd shapley (1923–2016). *Nature* 2016; 532: 178.
30. Li W, Song Y, Chen K, *et al.* Predictive model and risk analysis for diabetic retinopathy using machine learning: a retrospective cohort study in China. *BMJ Open* 2021; 11: e050989.
31. Ogami C, Tsuji Y, Seki H, *et al.* An artificial neural network-pharmacokinetic model and its interpretation using Shapley additive explanations. *CPT Pharmacomet Syst Pharmacol* 2021; 10: 760–768.
32. Zheng P, Yu Z, Li L, *et al.* Predicting blood concentration of tacrolimus in patients with autoimmune diseases using machine learning techniques based on real-world evidence. *Front Pharmacol* 2021; 12: 727245.
33. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin Chem* 1993; 39: 561–577.
34. Kawakami E, Tabata J, Yanaihara N, *et al.* Application of artificial intelligence for preoperative diagnostic and prognostic prediction in epithelial ovarian cancer based on blood biomarkers. *Clin Cancer Res* 2019; 25: 3006–3015.
35. Kourou K, Exarchos TP, Exarchos KP, *et al.* Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J* 2015; 13: 8–17.
36. Rakha EA, Reis-Filho JS, Ellis IO. Combinatorial biomarker expression in breast cancer. *Breast Cancer Res Treat* 2010; 120: 293–308.

37. Xiong XL, Zhang RX, Bi Y, *et al.* Machine learning models in type 2 diabetes risk prediction: results from a cross-sectional retrospective study in Chinese adults. *Curr Med Sci* 2019; 39: 582–588.
38. Choi SB, Kim WJ, Yoo TK, *et al.* Screening for prediabetes using machine learning models. *Comput Math Methods Med* 2014; 2014: 618976.
39. Wang C, Li L, Wang L, *et al.* Evaluating the risk of type 2 diabetes mellitus using artificial neural network: an effective classification approach. *Diabetes Res Clin Pract* 2013; 100: 111–118.
40. Birk N, Matsuzaki M, Fung TT, *et al.* Exploration of machine learning and statistical techniques in development of a low-cost screening method featuring the global diet quality score for detecting prediabetes in rural India. *J Nutr* 2021; 151: 110 S–118 S.
41. Zhang L, Wang Y, Niu M, *et al.* Machine learning for characterizing risk of type 2 diabetes mellitus in a rural Chinese population: the Henan Rural Cohort Study. *Sci Rep* 2020; 10: 4406.
42. Breiman L. Random forests. *Mach Learn* 2001; 45: 5–32.
43. Byeon H. Can the random forests model improve the power to predict the intention of the elderly in a community to participate in a cognitive health promotion program? *Iran J Public Health* 2021; 50: 315–324.
44. Touw WG, Bayjanov JR, Overmars L, *et al.* Data mining in the life sciences with random forest: a walk in the park or lost in the jungle? *Brief Bioinform* 2013; 14: 315–326.
45. Chen Y, Cao W, Gao X, *et al.* Predicting postoperative complications of head and neck squamous cell carcinoma in elderly patients using random forest algorithm model. *BMC Med Inform Decis Mak* 2015; 15: 44.
46. Ding M, Pan SY, Huang J, *et al.* Optical coherence tomography for identification of malignant pulmonary nodules based on random forest machine learning algorithm. *PLoS One* 2021; 16: e0260600.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Figure S1 | Traditional regression model construction.

Table S1 | Variables used for feature screening.

Table S2 | Univariate and multivariate logistic regression analysis in predicting risk factors of diabetes in the whole cohort.

Table S3 | Parameter settings and definitions for six different machine learning algorithms.