Hindawi
Journal of Diabetes Research
Volume 2020, Article ID 6873891, 12 pages
https://doi.org/10.1155/2020/6873891

Research Article

Identification of Potential Type II Diabetes in a Large-Scale Chinese Population Using a Systematic Machine Learning Framework

Mingyue Xue,[1,2] Yinxia Su,[2] Chen Li,[3] Shuxia Wang ,[4] and Hua Yao [4]

[1] *Hospital of Traditional Chinese Medicine Affiliated to the Fourth Clinical Medical College of Xinjiang Medical University, Urumqi, China*
[2] *College of Public Health, Xinjiang Medical University, Urumqi, China*
[3] *The First Affiliated Hospital of Xinjiang Medical University, Urumqi, China*
[4] *Center of Health Management, The First Affiliated Hospital, Xinjiang Medical University, Urumqi, China*

Correspondence should be addressed to Shuxia Wang; 2724443591@qq.com and Hua Yao; yaohua01@sina.com

Received 12 March 2020; Revised 1 August 2020; Accepted 2 September 2020; Published 24 September 2020

Academic Editor: Janet H. Southerland

Copyright © 2020 Mingyue Xue et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Background. An estimated 425 million people globally have diabetes, accounting for 12% of the world's health expenditures, and the number continues to grow, placing a huge burden on the healthcare system, especially in those remote, underserved areas. *Methods*. A total of 584,168 adult subjects who have participated in the national physical examination were enrolled in this study. The risk factors for type II diabetes mellitus (T2DM) were identified by p values and odds ratio, using logistic regression (LR) based on variables of physical measurement and a questionnaire. Combined with the risk factors selected by LR, we used a decision tree, a random forest, AdaBoost with a decision tree (AdaBoost), and an extreme gradient boosting decision tree (XGBoost) to identify individuals with T2DM, compared the performance of the four machine learning classifiers, and used the best-performing classifier to output the degree of variables' importance scores of T2DM. *Results*. The results indicated that XGBoost had the best performance (accuracy = 0.906, precision = 0.910, recall = 0.902, F-1 = 0.906, and AUC = 0.968). The degree of variables' importance scores in XGBoost showed that BMI was the most significant feature, followed by age, waist circumference, systolic pressure, ethnicity, smoking amount, fatty liver, hypertension, physical activity, drinking status, dietary ratio (meat to vegetables), drink amount, smoking status, and diet habit (oil loving). *Conclusions*. We proposed a classifier based on LR-XGBoost which used fourteen variables of patients which are easily obtained and noninvasive as predictor variables to identify potential incidents of T2DM. The classifier can accurately screen the risk of diabetes in the early phrase, and the degree of variables' importance scores gives a clue to prevent diabetes occurrence.

1. Introduction

Diabetes, as a group of metabolic disorders, is characterized by hyperglycemia, which can lead to many serious conditions such as heart disease, kidney disease, vision loss, and lower limb amputation [1]. According to the data from the World Health Organization (WHO), the global epidemic of diabetes currently affects more than 422 million people in 2014 and increased notably in recent decades [2, 3]. In China, the incidence rate of diabetes (100 million of adult patients) was the highest in the world. About 52.7% of diabetes patients have no awareness, and this proposition remains upward [4]. Research has proven that a healthy lifestyle and a reasonable diet structure can effectively delay and prevent the occurrence of type II diabetes mellitus (T2DM) [5]. The American Diabetes Association recommends annual diabetes screening for people over 45 years of age and with major risk factors [6]. China's national plan for the prevention and control of noncommunicable diseases (2012-2015) listed diabetes as one of the key diseases in China and proposed diabetes prediction suggestions based on a blood glucose test and routine physical examination [7].

However, the traditional diabetes screening method needs an expensive blood test and extra manpower, which is a big challenge for the backward remote areas [8]. A diabetes screening model built by easily available indicators, without expensive examinations, is crucial to the occurrence and development of diseases [9, 10].

The analysis of diabetes data is a challenging issue because most of the medical data are nonlinear, nonnormal, correlation structured, and complex in nature [11]. The machine learning (ML) algorithms have dominated in the field of medical healthcare [11–15] and medical imaging for diseases such as stroke, coronary artery disease, and cancer [16–20]. A decision tree (DT) is one of the classical algorithms of ML. This simple and sensitive tree algorithm provides a unique ability to build disease prediction for large datasets [21–23]. Tree embedding algorithms aggregate the results from multiple trees, which usually have better accuracy and generalization ability than a single tree. This includes combining stumps with an enhancement program [24]. The random forest (RF) of a boosting procedure to combine stumps of trees belongs to a "bagging" algorithm [25], which has already been widely used in biological medicine researches [26, 27], especially in the diagnosis of diabetes [11, 12]; AdaBoost with a decision tree (AdaBoost) [28] and an extreme gradient boosting decision tree (XGBoost) [29] belong to "boosting" algorithms, and they had better performance than a decision tree in the prediction and classification [30–32]. In this study, LR- and tree-based models were used. Some studies have confirmed that this method can accurately classify diabetes mellitus [33]. Previous studies have used ML models to classify diabetes. To the best of our knowledge, this is the first diabetes screening model established by comparing four tree-based ML algorithms.

2. Methods

2.1. Study Population. The national physical examination (NPE) is a free physical examination provided by the Chinese government for all Xinjiang people. The data came from the physical examination data of Urumqi in 2018. A total of 643,439 subjects participated in the examination and signed a written informed consent form. The exclusion criteria of potential participants are the following: (1) pregnancy, (2) people with type I diabetes mellitus (T1DM), and (3) age less than 20 years. Finally, a total of 584,168 subjects from the eligible participants were included in the final analysis. This study was performed in accordance with the principles outlined in the Declaration of Helsinki and approved by the Xinjiang Uygur Autonomous Region CDC ethical committee and the institutional review board.

2.2. Diagnosis of T2DM. Subjects with the following criteria were classified as having T2DM: blood glucose 2 hours after meal ≥ 11.1 mmol/l, fasting blood glucose ≥ 7.0 mmol/l, or the main complaint of diabetes and taking hypoglycemic drugs; the final incidence of diabetes was 12.4%.

2.3. Baseline Survey. NPE investigates a wide range of lifestyle, dietary, psychosocial, occupational, and biochemical and genetic factors related to the development of chronic diseases. Therefore, the epidemiologists and medical professionals from the CDC in the Uygur Autonomous Region referred to a previous study [34] to design a standard medical examination form, which included 3 parts: a questionnaire, physical examination, and laboratory testing. The examination of all the participants was done by the medical and health teams in the administrative regions, which were made up of full-time employees with medical qualifications and fieldwork experience. In order to ensure the accuracy of the results, all participants were asked to bring their unique national identity (ID) cards and take them as the unique identification. After the investigation, all the data were summarized into the Health Management Hospital of Xinjiang Medical University.

Trained interviewers administered questionnaires during face-to-face interviews. The questionnaires included demographic information, occupational history, socioeconomic status, family and personal disease histories, smoking history, alcohol use, diet, physical activity, and contact history of occupational disease-inductive factors. The physical examinations were performed by trained physicians, nurses, and technicians, in which items included standing height, body weight, waist circumference, heart rate, blood pressure, and abdominal ultrasound. Abdominal ultrasound can observe the shape and size of the abdominal organs; also, it can determine whether these organs have tumors, cysts, or stones, including the liver, kidney, gallbladder, and other organs. For each participant, a 10 ml nonfasting blood sample was collected into three vacuum tubes. The samples were then kept in a portable, insulated cool box with ice packs for up to a few hours before being taken to the local study laboratory for immediate processing. Blood test indicators include blood glucose and blood biochemistry. In this study, we wanted to establish a simple model that can predict the risk of T2DM without blood sampling. We selected 18 variables from the questionnaire and physical examination based on the previous studies [35–37] (Table 1).

2.4. Variable Definitions. The potential risk factors in this study to assess T2DM included the following: age, gender, ethnicity, body mass index (BMI), physical activity, smoking, drinking, eating habits, waist circumference, blood pressure, and some comorbidities.

Sociodemographic information included age (years), gender including "male" and "female," and ethnic groups which were divided into six categories: "Han," "Uygur," "Kazak," "Hui," "Mongolian," and "other nationalities"; the baseline comorbidities considered in this study were fatty liver and hypertension (yes or no).

Lifestyle information included smoking, drinking, physical activity, and eating habits. Physical activity was defined as regularly doing at least 20 min per day of physical activity during leisure time over the previous 6 months (yes or no) [38]. Individuals who had been smoking at least one cigarette per day for at least 6 months were defined as smokers, and those who had been drinking alcohol at least once per week for at least 6 months were considered drinkers [39]. We also included daily smoking amount (cigarettes) and weekly

Table 1: Characteristics of variables.

Variables	Diabetes ($N = 72,027$)	Nondiabetes ($N = 510,411$)	p value
Age (years)	66.43 ± 13.43	52.41 ± 16.06	<0.001
BMI (kg/m^2)	25.92 ± 3.65	24.37 ± 3.42	<0.001
Waist circumference (cm)	90.20 ± 10.75	84.95 ± 10.71	<0.001
Systolic pressure (mmHg)	130.20 ± 16.52	121.30 ± 14.27	<0.001
Diastolic pressure (mmHg)	77.80 ± 10.56	75.14 ± 9.65	<0.001
Ethnicity, *n* (%)			<0.001
Han	50,691 (70.38)	331,413 (64.93)	
Uygur	10,864 (15.08)	95,913 (18.79)	
Kazak	1147 (1.59)	18,893 (3.70)	
Hui	8126 (11.28)	52,838 (10.35)	
Mongolian	76 (0.11)	1214 (0.24)	
Other nationalities	1123 (1.56)	10,140 (1.99)	
Gender, *n* (%)			<0.001
Male	34,641 (48.09)	239,875 (47.00)	
Female	37,386 (51.91)	270,536 (53.00)	
Physical activity, *n* (%)			<0.001
Yes	26,239 (36.43)	154,585 (30.29)	
No	45,788 (63.57)	355,826 (69.71)	
Drinking status, *n* (%)			<0.001
Yes	15,944 (22.14)	102,852 (20.15)	
No	56,083 (77.86)	407,559 (79.85)	
Drinking amount (g)			<0.001
≥170	6687 (9.30)	39,479 (7.73)	
<170	65,240 (90.70)	470,932 (92.27)	
Smoking amount (cigarettes)	10 (8-20)*	10 (7-20)*	<0.001
Smoking status, *n* (%)			<0.001
Yes	10,683 (14.83)	63,920 (12.52)	
No	61,344 (85.17)	446,491 (87.48)	
Dietary ratio, *n* (%)			<0.001
Meat based	2849 (3.96)	13,554 (2.66)	
Meat balanced	66,603 (92.47)	482,864 (94.60)	
Vegetarian based	2575 (3.58)	13,993 (2.74)	
Sugar loving, *n* (%)			<0.001
Yes	940 (1.31)	4560 (0.89)	
No	71,087 (98.69)	505,851 (99.11)	
Oil loving, *n* (%)			<0.001
Yes	2722 (3.78)	13,068 (2.56)	
No	69,305 (96.22)	497,343 (97.44)	
Salt loving, *n* (%)			<0.001
Yes	4261 (5.92)	20,896 (4.09)	
No	67,766 (94.08)	489,515 (95.91)	
Fatty liver, *n* (%)			<0.001
Yes	22,331 (31.00)	52,800 (10.34)	
No	49,696 (69.00)	457,611 (89.66)	
Hypertension, *n* (%)			<0.001
Yes	29,937 (41.56)	112,348 (22.01)	
No	42,090 (58.44)	398,063 (77.99)	

*Median (IQR). Abbreviation: BMI: body mass index.

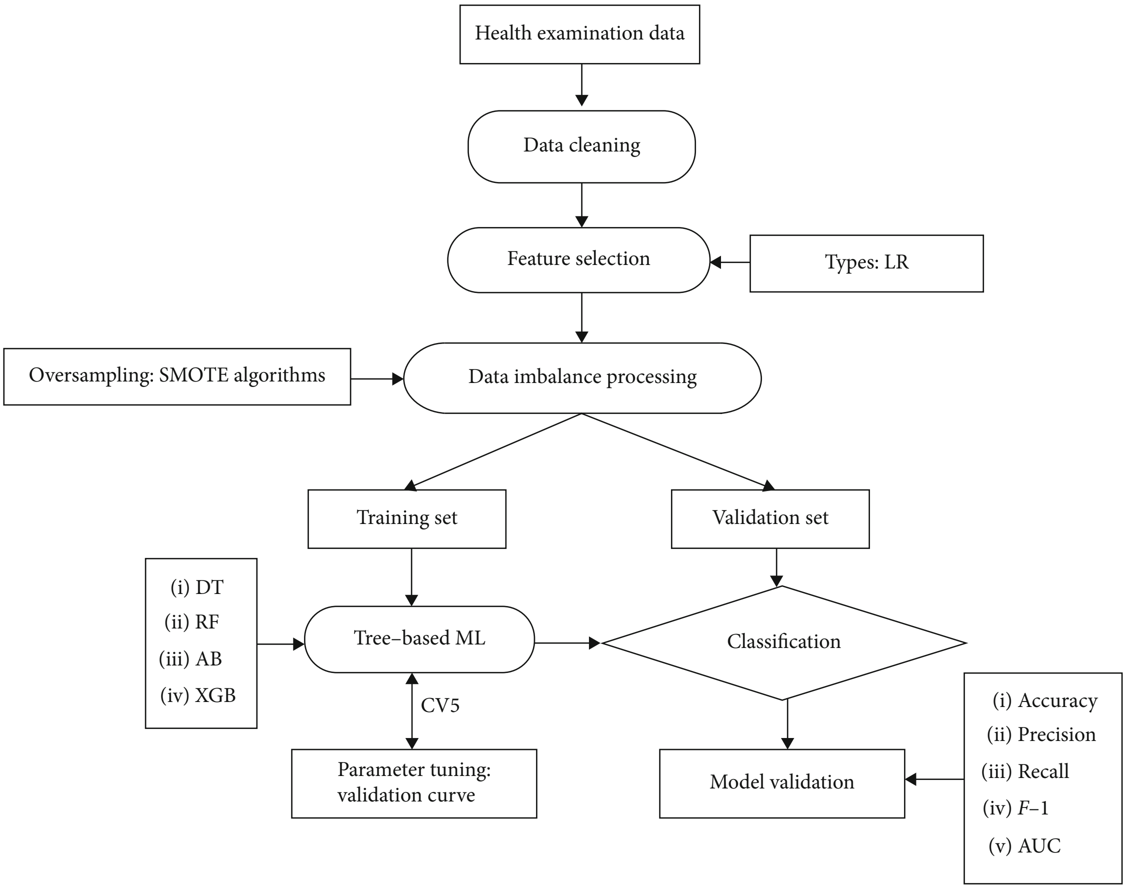


Figure 1: Machine learning flowchart of this study. Abbreviations: LR: logistic regression; DT: decision tree; RF: random forest; AB: AdaBoost; XGB: XGBoost; ML: machine learning.

drinking amount ("≥170 g" or "<170 g"). Diet habits included 6 options: "meat based," "meat balanced," "vegetarian based," "oil loving," "sugar loving," and "salt loving"; participants can choose one or more of them.

2.5. Statistical Analysis. The baseline characteristics of the study population were presented as mean ± SD (standard deviation) for continuous normal distribution variables, median (IQR) for continuous nonnormally distributed variables, and number (percentage) for the categorical variables. Differences in variables between diabetes and nondiabetes patients are analyzed by the independent t-test for continuous normal distribution variables, the Mann-Whitney test for nonnormally distributed variables, and the chi-square test for categorical variables. All of the tests were two-tailed and considered significant factors whose p values were less than 0.05.

2.6. Machine Learning System. The major objective of the tree-based ML algorithms is to classify the T2DM. The overview of the proposed tree-based ML algorithms has been shown in Figure 1.

2.6.1. Data Cleaning. NPE data are large and with jumbled variables, with many missing and abnormal values. So data preprocessing is a very important step, and the quality of preprocessing will directly affect the performance of the later prediction model [40]. Firstly, we deleted nearly 200 variables that were not meaningful to this study. Secondly, we filled in outliers and nulls, classification variables were filled with the most frequent value, and continuous variables were filled with a mean value.

2.6.2. Feature Selection. There were some commonly used feature selection techniques in ML/statistics, namely: RF [12, 41], LR [42, 43], mutual information [12, 44], principal component analysis [12, 44, 45], analysis of variance [12, 46], and Fisher's discriminant ratio [12, 44, 47]. In this study, we have used the LR model to identify the risk factor for diabetic disease based on a p value ($p < 0.05$) and OR.

2.6.3. Data Imbalance Processing. The number of nondiabetes subjects was greater than the number of subjects with diabetes (an unbalanced-class problem). Generally, classes with few subjects are more difficult to predict than those with numerous subjects [48–51]. We used the SMOTE algorithm to solve the negative impact of class imbalance, which belonged to the method of oversampling; the principle of the method is to increase the number of a few classes of samples in classification to achieve sample balance, and it is widely used because of its ability to preserve important information in samples.

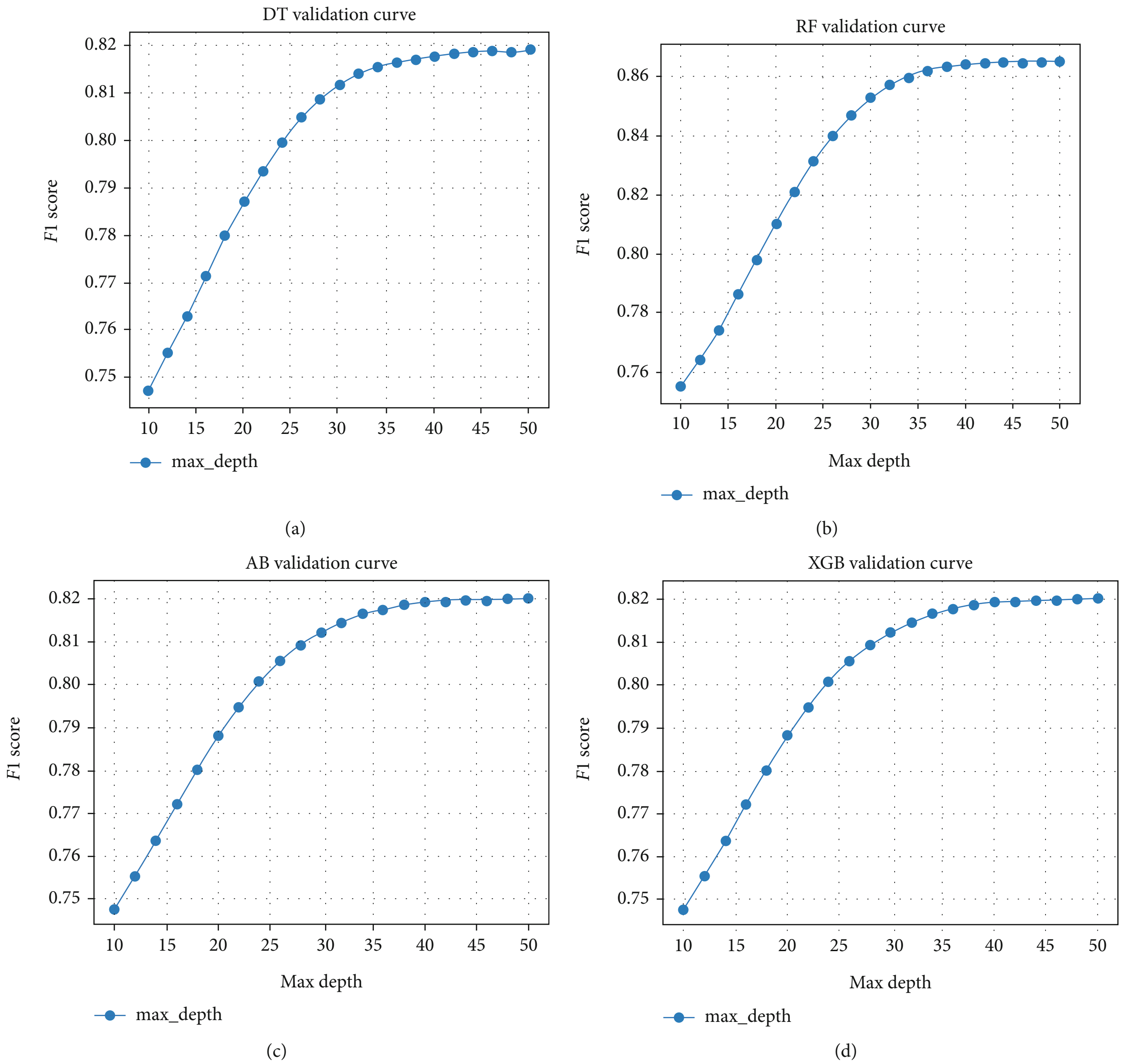


Figure 2: Parameter selection process of the prediction model constructed by four classification tree models: (a) decision tree, (b) random forest, (c) AdaBoost, and (d) XGBoost. Note: the score of F-1 has been tested when the max depth parameter of the model is between 10 and 50.

2.6.4. Classifier Comparison. In this study, we used four tree-based ML algorithms: DT, RF, AdaBoost, and XGBoost, all of which were supervised ML methods. DT is a tree structure-based model which describes the classification process based on input features [52]; the advantage of DT is that it is simple and easy to implement, but it often exhibits high variance and overfitting problems, which limits its utility as an independent prediction model. However, it is possible to improve the overall prediction by aggregating the results from multiple trees, which is called the embedding method. RF is one of the common tree embedding methods [53], which uses the bagging method to combine multiple trees. Another ensemble approaches, AdaBoost and XGBoost algorithms [24], use the boosting procedure to combine stumps of trees. These ensemble methods can be loosely conceptualized as forming a robust overall prediction by aggregating the predictions of many simpler predictive models. This is similar to the process of drawing on the advice of many experts to arrive at a clinical diagnosis for a patient, each of whom views the patient in a slightly different way.

2.6.5. Model Evaluation. Balanced datasets were randomly divided into two parts: the training set accounted for 70% of the data and the test set accounted for 30% of the data [21, 54]. In order to improve the accuracy of the classification tree, we have drawn a "verification curve" based on 5-fold cross-validation of four classification trees, and the optimal hyperparameter has been obtained (Figure 2). The algorithms were compared based on a confusion matrix and some indicators including accuracy, precision, recall, F-1, and receiver operating characteristic (ROC) curve. Several important measures, such as accuracy, precision, recall, and F-1, could be calculated by using the confusion matrix.

$$\begin{aligned}\text{Accuracy} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}},\\ \text{Precision} &= \frac{\text{TP}}{\text{TP} + \text{FN}},\\ \text{Recall} &= \frac{\text{TP}}{\text{TP} + \text{FP}},\\ F\text{-}1 &= \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.\end{aligned} \tag{1}$$

2.7. Feature Importance Ranking. Tree-based models can provide measures of variable importance. Unlike the OR values of regression models, ML algorithms cannot estimate an easy explanation number because the relationships that ML algorithms fit are more complex than those of regression models. Therefore, it is not usual to generalize this relationship directly into any one parameter, nor is there a causal relationship or even a statistical explanation [55]. Instead, the measure can often be thought of as rank ordering of which variables are most "important" to the fitted model [56]. Although the variable importance ranking cannot replace the target hypothesis test for a given parameter, it can be used as a means of generating hypotheses to help identify factors that warrant further study, allowing some insight into the factors that most influence the predictions [57].

The software used in this study was Python software version 3.7.2. The "Pandas" library, "NumPy" library, and "Matplotlib" library were used for null and outlier determination and interpolation, the "Imlearn" library was used to solve data imbalance, and the "Sklearn" library was used to establish machine learning models and verify the validation.

3. Results

3.1. Patients and Variables. A total of 72,027 patients (12.4%) from the pool of 582,438 subjects had T2DM. Each subject was composed of 18 variables (Table 1), including age, BMI, gender, waist circumference, ethnicity, drinking, physical activity, smoking, eating habits, blood pressure, fatty liver, and hypertension. It is observed that all attributes are highly statistically ($p < 0.001$) associated with diabetes.

3.2. Feature Extraction Using Logistic Regression. Table 2 shows the effect of the selected factors on T2DM by logistic regression. It was shown that age, BMI, waist circumference, systolic pressure, ethnicity, physical activity, drinking status, weekly drinking amount (g), daily smoking amount (cigarettes), smoking status, dietary ratio (meat to vegetable), diet habit (oil loving), fatty liver, and hypertension are statistically significant factors for T2DM at a 5% level of significance and the rest of the factors are insignificant. These 14 variables were used for tree-based ML algorithms to classify T2DM. Among these statistically significant variables, variables with OR > 1 were the risk factors for T2DM, including age, BMI, waist circumference, systolic pressure, ethnicity (Hui), weekly drinking amount ≥ 170 g, daily smoking amount (cigarettes), smoking status, diet habit (oil loving), fatty liver, and hypertension; variables with OR < 1 were the protective factors, including ethnicity (Kazak and Mongolian), physical activity, drinking status, and diet habit (meat balanced).

3.3. Tuning of Parameters. Finally, we got 1,020,822 samples by the SMOTE algorithm (Table 3): 714,575 subjects as the training set and 306,247 subjects as the validation set. The average F-1 score for different models and their parameter are listed in the validation set (Figure 2). When the "maximum depth" of DT takes 44 and that of RF, XGBoost, and AdaBoost takes 40, we got a relatively economical and accurate classification tree model.

3.4. Validation of the Training Set. Our study has built four tree-based ML algorithms. Table 4 shows the performance of all classifiers. The confusion matrix has been displayed by heatmap; the larger the number, the darker the color of the region, that is, the closer the color of TN and TP regions to orange. On the contrary, the lighter the color of FN and FP regions, the higher the accuracy of the classification model. We got that the result of XGBoost was better than that of the others (accuracy = 0.906, precision = 0.910, recall = 0.902, F-1 = 0.906, and AUC = 0.968). Figure 3 presents the ROC of all classifiers.

3.5. Variable Importance Ranking by XGBoost. In this study, XGBoost was used to rank the LR-selected variables because of its best classification performance. XGBoost provided the importance score of each variable, attributing the predictive risk in 3 ways. Specifically, we chose the default method, which represented the relative number of times a variable is used to distribute the data across all trees. There were only very small differences among the importance scores through the three methods, which did not influence the rank of the variable's impact. The important measurement scores of 14 variables have been shown in Figure 4. BMI is the most significant feature, followed by age, waist circumference, systolic pressure, ethnicity, smoking amount, fatty liver, hypertension, physical activity, drinking status, dietary ratio (meat to vegetable), drink amount, smoking status, and diet habit (oil loving).

4. Discussion

In this paper, cases were recruited and consisted of easily acquired variables to establish a screening model for T2DM. LR models were used for selecting the risk factors. Then, we compared the performance of four tree-based ML algorithms (DT, RF, AdaBoost, and XGBoost), and XGBoost got the best performance, which had accuracy = 0.906, precision = 0.910, recall = 0.902, F-1 = 0.906, and AUC = 0.968. Finally, through the best classifier to establish the most important ranking of factors affecting the incidence of diabetes, the results indicate that this strategy successfully achieves accurate and rapid diabetes screening.

The order of feature importance (Figure 3) showed that age, BMI, and waist circumference were the top three influencing factors of diabetes, which was consistent with

Table 2: Screening the risk factors for T2DM by multiple logistic regression (CI = confidence interval).

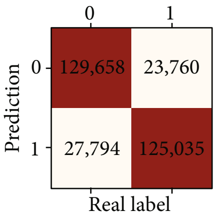
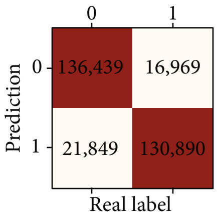
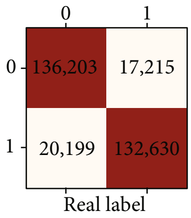
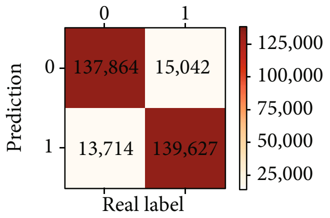
Intercept and variable	Odds ratio	95% CI	*Z* value	*p* value
Age (years)	1.047	(1.046-1.048)	113.625	<0.001
BMI (kg/m^2)	1.016	(1.012-1.020)	7.894	<0.001
Waist circumference (cm)	1.016	(1.015-1.018)	23.905	<0.001
Systolic pressure (mmHg)	1.002	(1.001-1.003)	5.304	<0.001
Diastolic pressure (mmHg)	1.001	(0.999-1.002)	1.650	0.099
Ethnicity, *n* (%)				
Han	1	Ref	—	—
Uygur	1.011	(0.981-1.043)	0.734	0.463
Kazak	0.460	(0.426-0.497)	-19.669	<0.001
Hui	1.075	(1.040-1.111)	4.269	<0.001
Mongolian	0.464	(0.342-0.616)	-5.127	<0.001
Other nationalities	0.989	(0.912-1.072)	-0.263	0.793
Gender, *n* (%)				
Male	1	Ref	—	—
Female	1.017	(0.994-1.041)	1.444	0.149
Physical activity, *n* (%)				
No	1		—	—
Yes	0.715	(0.699-0.731)	-29.179	<0.001
Drinking status, *n* (%)				
No	1	Ref	—	—
Yes	0.891	(0.864-0.918)	-7.424	<0.001
Drinking amount (g)				
<170	1	Ref	—	—
≥170	1.239	(1.185-1.296)	9.432	<0.001
Smoking amount (cigarettes)	1.005	(1.002-1.007)	3.921	<0.001
Smoking status, *n* (%)				
No	1	Ref	—	—
Yes	1.137	(1.086-1.191)	5.452	<0.001
Dietary ratio, *n* (%)				
Meat based	1	Ref	—	—
Meat balanced	0.917	(0.869-0.969)	-3.105	0.002
Vegetarian based	1.019	(0.941-1.103)	0.455	0.649
Sugar loving, *n* (%)				
No	1	Ref	—	—
Yes	0.994	(0.896-1.101)	-0.119	0.906
Oil loving, *n* (%)				
No	1	Ref	—	—
Yes	1.157	(1.072-1.249)	3.730	<0.001
Salt loving, *n* (%)				
No	1	Ref	—	—
Yes	0.989	(0.932-1.049)	-0.362	0.718
Fatty liver, *n* (%)				
No	1	Ref	—	—
Yes	2.224	(2.168-2.280)	62.430	<0.001
Hypertension, *n* (%)				
No	1	Ref	—	—
Yes	2.373	(2.312-2.435)	65.334	<0.001

Abbreviation: BMI: body mass index.

Table 3: Dataset description.

Dataset	Sample distribution	Ratio	Description
Original data	510,411/72,027	7 : 1	Original data with full instances
SMOTE data	510,411/510,411	1 : 1	Dataset is balanced utilizing SMOTE oversampling

Table 4: The results of classification algorithms.

Testing criteria	DT	RF	AB	XGB
Confusion matrix				
Accuracy	0.832	0.873	0.878	0.906
Precision	0.823	0.862	0.871	0.910
Recall	0.845	0.889	0.888	0.902
F-1	0.834	0.875	0.879	0.906
AUC	0.832	0.947	0.948	0.968

Abbreviations: AUC: the area under the receiver operating characteristic (ROC) curve; DT: decision tree; RF: random forest.

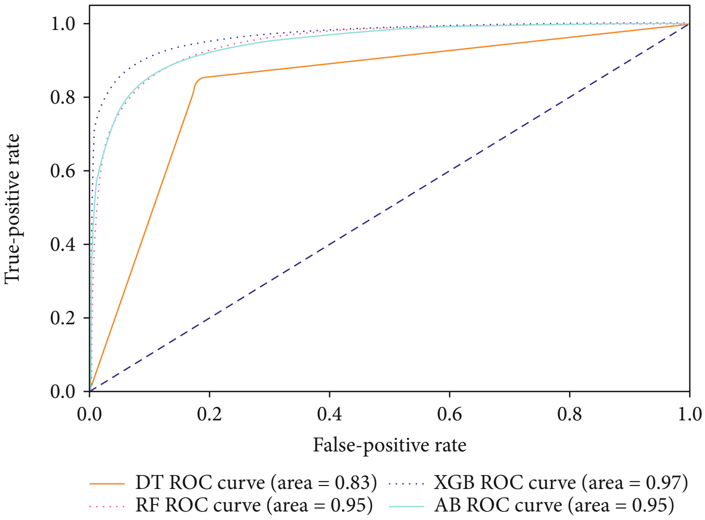


Figure 3: ROC curve of all algorithms. Abbreviations: DT: decision tree; RF: random forest; AB: AdaBoost; XGB: XGBoost.

Pei et al.'s T2MD screening model based on a j48 decision tree [35]. The variables whose OR > 1 are risk factors for the disease, including age, BMI, waist circumference, systolic pressure, hypertension, ethnicity (Hui), daily smoking amount (cigarettes), fatty liver, weekly drinking amount ≥ 170 g, smoking status, and diet habit (oil loving). Xu et al. [36] used the data of the national cross-sectional survey in 2010 for study and found that the risk factors for diabetes were age, smoking, overweight, obesity, dyslipidemia, elevated triacylglycerol, and high systolic blood pressure. Other countries had developed diabetes screening tools, and the American Diabetes Association (ADA) provides a simple "T2DM risk test" that used age, gender, family history of diabetes, hypertension, physical activity, and weight status to assess diabetes risk in the general population [37]. The above conclusions were consistent with the conclusions of this study. Variables with OR < 1 are protective factors, including ethnicity (Kazak and Mongolian), physical activity, weekly alcohol consumption < 170 g, and diet habit (diet balanced). The protective factors include three adjustable indicators, which suggested that people could control the occurrence of the disease through a good lifestyle. Several large-scale trials have demonstrated the benefits of targeted lifestyle interventions to prevent diabetes [58–60].

There are several strengths of our study. First, all the variables come from noninvasive and easily available

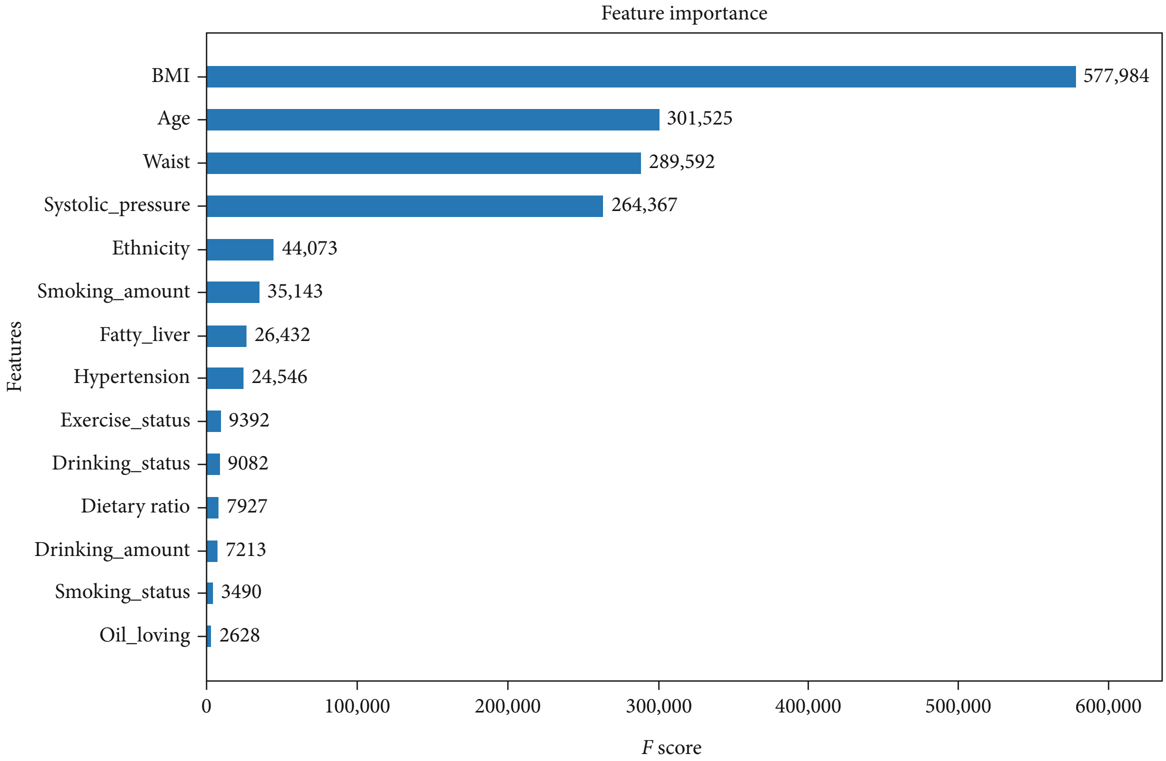


Figure 4: Feature importance contributed to the XGBoost model measured by the *F* score.

measurement indexes and questionnaire indexes. The model can be applied to the prediabetes and noninvasive prediction of diabetes without the need for expensive laboratory testing, which is useful, particularly in areas of high epidemiological risk and low socioeconomic status [2, 61].

Second, this study was based on a large Chinese population, with a wide range of population choices and high extrapolation and representativeness. Moreover, our dataset included many major ethnic groups in China, which better evaluated the characteristics of the Chinese population.

Third, in most previous diabetes screening models, smoking and drinking were only divided into two categories (have and have not), so they failed to reflect the impact of frequency and quantity on the disease. Through Figure 3, we knew that compared with the smoking status, the daily smoking amount was more important to the disease. Furthermore, our studies have shown that alcohol was a protective factor for T2DM, but alcohol consumption > 170 g a week increased the risk of diabetes. Previous studies have also confirmed that light-to-moderate alcohol consumption could reduce the risk of T2DM [62, 63]; however, there was a strong dose-response relationship between smoking number, alcohol consumption, and diabetes and cardiovascular disease [64–66], suggesting that while quitting smoking completely and controlling alcohol consumption were our goals, even smoking fewer cigarettes and drinking less alcohol can reduce the risk of the disease.

Fourth, we compared the performance of four tree-based classification models, and XGBoost achieved the best performance. XGBoost used in this study has received extensive attention in recent years due to its excellent learning effect and efficient training speed. XGBoost has more advantages than LR in predicting the occurrence of results rather than measuring the relationship between specific risk factors and events, but its disadvantage is the poor interpretation of risk factors [55]. LR provides a clear interpretation of its coefficients as the odds ratios of the risk factors. We know that the former could get higher prediction accuracy and the latter could get better explanation among variables. In this study, we have first used LR to screen variables and then used XGBoost to classify diseases, which not only improves the accuracy of classification but also gets the risk factors and protective factors for diseases, enlightening us which characteristics may lead to T2DM and which characteristics can prevent T2DM.

Surprisingly, previous studies have found that the course of diabetes is closely related to diet. For example, the Diabetes Prevention Program (DPP) reported that a reasonable diet and exercise can reduce the incidence of type 2 diabetes by 58% [67]. But in this study, we only got the weak effects of meat and vegetable matching and oil preference on T2DM (Figure 3) and did not find that halophilia or sugar addiction is associated with diabetes. However, the effects of these factors on diabetes have been confirmed in previous studies [68, 69]. Eating habits are the main influencing factors of waist circumference and BMI, so we think that diabetes and eating habits are closely related; the possible reasons for the irrelevance might be that the diet survey of Xinjiang national health examination was a cross-sectional study and there was no professional person to evaluate the diet of the physical examination population. The main reason for the error was that the self-reported eating habits of the physical examination population were subjective and professional evaluation indicators are lacking; for this, in the future research, more accurate results can be obtained through the follow-up of people's living habits.

There are several limitations in this study: firstly, since this was a cross-sectional study, we could not assess the causal relationship between T2DM and other comorbidities. Secondly, the data used in this study was the physical examination data of China, which might limit the extrapolation of the results. It is generally believed that there are some differences in the pathophysiology of diabetes between Asians and Caucasians and there are similar differences between Asian countries. Thirdly, previous studies have confirmed that education and family history are also important determinants of diabetes. However, our physical examination data failed to obtain the education and family history of participants. Fourthly, this study only optimizes the "maximum depth" parameter of the classification trees. The machine learning model can improve the performance of the model by tuning multiple parameters, which needs to be further implemented in the future. Finally, some indicators do not have objective and unified evaluation criteria, such as eating habits, which may reduce the accuracy of the prediction model.

5. Conclusion

We have proposed a classifier combining tree-based ML algorithms and LR to build a diabetes screening model using 582,438 subjects in China. The ranking of disease risk factors and protective factors provided us with inspiration to prevent diabetes. We also got the dose relationship between smoking and drinking and the disease. In a word, our model can help China's health system to improve the level of early diagnosis of diabetes, suggesting the significance of lifestyle change in the prevention and delay of the disease.

Abbreviations

AUC:	Area under the receiver operating characteristic curve
ROC:	Receiver operating characteristic curve
NPE:	National physical examination
BMI:	Body mass index
ML:	Machine learning
OR:	Odds ratio
DT:	Decision tree
LR:	Logistic regression
RF:	Random forests
WHO:	World Health Organization
T2DM:	Type II diabetes mellitus
T1DM:	Type I diabetes mellitus
ADA:	American Diabetes Association.

Data Availability

Data supporting the results of this study can be available by requesting the first author or corresponding author.

Conflicts of Interest

The authors declare no conflict of interest.

Authors' Contributions

Mingyue Xue is responsible for the conceptualization. Yinxia Su, Chen Li, and Shuxia Wang are responsible for the data curation. Mingyue Xue is responsible for the formal analysis. Hua Yao is responsible for the funding acquisition. Shuxia Wang is responsible for the investigation. Mingyue Xue and Hua Yao are responsible for the project administration. Mingyue Xue is responsible for the software. Hua Yao is responsible for the supervision. Mingyue Xue is responsible for the validation. Mingyue Xue is responsible for the visualization. Mingyue Xue is responsible for the writing of the original draft. Mingyue Xue and Yinxia Su contributed to this work equally.

Acknowledgments

The project was supported by the National Natural Science Foundation of China (81960608), and the Xinjiang Science and Technology Department set up the project (2018E02057). We would like to thank Dr. Ma Jixuan of Huazhong University of Science and Technology and Ms. Liu Li of the Fourth Affiliated Hospital of Xinjiang Medical University for their technical guidance.

References

[1] World Health Organization, *Global Reportion Diabetes*, World Health Organization,, 2016, http://www.who.int/iris/handle/10665/204871.

[2] M. Xue, Y. Su, Z. Feng et al., "A nomogram model for screening the risk of diabetes in a large-scale Chinese population: an observational study from 345,718 participants," *Scientific Reports*, vol. 10, no. 1, p. 11600, 2020.

[3] A. Ramachandran, R. C. Wan Ma, and C. Snehalatha, "Diabetes in Asia," *The Lancet*, vol. 375, no. 9712, pp. 408–418, 2010.

[4] M. Jahani and M. Mahdavi, "Comparison of predictive models for the early diagnosis of diabetes," *Healthcare Informatics Research*, vol. 22, no. 2, pp. 95–100, 2016.

[5] Y. Zheng, S. H. Ley, and F. B. Hu, "Global aetiology and epidemiology of type 2 diabetes mellitus and its complications," *Nature Reviews. Endocrinology*, vol. 14, no. 2, pp. 88–98, 2018.

[6] K. Pippitt, M. Li, and H. E. Gurgle, "Diabetes mellitus: screening and diagnosis," *American Family Physician*, vol. 93, no. 2, pp. 103–109, 2016.

[7] M. Hong and D. O. Finance, "Analysis of existing problems and solutions in the management and application of funds for basic public health services," *China Health Industry*, vol. 15, no. 6, pp. 102-103, 2018.

[8] V.-M. Lélis, E. Guzmán, and M.-V. Belmonte, "A statistical classifier to support diagnose meningitis in less developed areas of Brazil," *Journal of Medical Systems*, vol. 41, no. 9, pp. 145–145, 2017.

[9] B. Xi, S. Li, Z. Liu et al., "Intake of fruit juice and incidence of type 2 diabetes: a systematic review and meta-analysis," *PLoS One*, vol. 9, no. 3, pp. e93471–e93471, 2014.

[10] C. L. Gillies, K. R. Abrams, P. C. Lambert et al., “Pharmacological and lifestyle interventions to prevent or delay type 2 diabetes in people with impaired glucose tolerance: systematic review and meta-analysis,” *BMJ*, vol. 334, no. 7588, pp. 299–299, 2007.

[11] M. Maniruzzaman, N. Kumar, M. Menhazul Abedin et al., “Comparative approaches for classification of diabetes mellitus data: machine learning paradigm,” *Computer Methods and Programs in Biomedicine*, vol. 152, pp. 23–34, 2017.

[12] M. Maniruzzaman, M. J. Rahman, M. Al-MehediHasan et al., “Accurate diabetes risk stratification using machine learning: role of missing value and outliers,” *Journal of Medical Systems*, vol. 42, no. 5, pp. 92–92, 2018.

[13] S. K. Srivastava, S. K. Singh, and J. S. Suri, “Healthcare text classification system and its performance evaluation: a source of better intelligence by characterizing healthcare text,” *Journal of Medical Systems*, vol. 42, no. 5, pp. 97–97, 2018.

[14] G. Luo, “Automatically explaining machine learning prediction results: a demonstration on type 2 diabetes risk prediction,” *Health Information Science and Systems*, vol. 4, no. 1, 2016.

[15] G. Luo, “MLBCD: a machine learning tool for big clinical data,” *Health Information Science and Systems*, vol. 3, no. 1, 2015.

[16] E. Deniz, A. Şengür, Z. Kadiroğlu, Y. Guo, V. Bajaj, and Ü. Budak, “Transfer learning based histopathologic image classification for breast cancer detection,” *Health information science and systems*, vol. 6, no. 1, pp. 18–18, 2018.

[17] A. S. Ashour, A. R. Hawas, and Y. Guo, “Comparative study of multiclass classification methods on light microscopic images for hepatic schistosomiasis fibrosis diagnosis,” *Health information science and systems*, vol. 6, no. 1, pp. 7–7, 2018.

[18] S. K. Banchhor, N. D. Londhe, T. Araki et al., “Wall-based measurement features provides an improved IVUS coronary artery risk assessment when fused with plaque texture-based features during machine learning paradigm,” *Computers in Biology and Medicine*, vol. 91, pp. 198–212, 2017.

[19] V. Kuppili, M. Biswas, A. Sreekumar et al., “Extreme learning machine framework for risk stratification of fatty liver disease using ultrasound tissue characterization,” *Journal of Medical Systems*, vol. 41, no. 10, pp. 152–152, 2017.

[20] S. K. Banchhor, N. D. Lond0he, T. Araki et al., “Calcium detection, its quantification, and grayscale morphology-based risk stratification using machine learning in multimodality big data coronary and carotid scans: a review,” *Computers in Biology and Medicine*, vol. 101, pp. 184–198, 2018.

[21] A. Ramezankhani, O. Pournik, J. Shahrabi, D. Khalili, F. Azizi, and F. Hadaegh, “Applying decision tree for identification of a low risk population for type 2 diabetes. Tehran Lipid and Glucose Study,” *Diabetes Research and Clinical Practice*, vol. 105, no. 3, pp. 391–398, 2014.

[22] J. S. Kammerer, S. J. N. McNabb, J. E. Becerra et al., “Tuberculosis transmission in nontraditional settings: a decision-tree approach,” *American Journal of Preventive Medicine*, vol. 28, no. 2, pp. 201–207, 2005.

[23] M. Marinov, A. S. M. Mosa, I. Yoo, and S. A. Boren, “Data-mining technologies for diabetes: a systematic review,” *Journal of Diabetes Science and Technology*, vol. 5, no. 6, pp. 1549–1556, 2011.

[24] Y. Freund and R. E. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting,” *Journal of Computer & System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[25] L. Breiman, “Bagging predictors,” *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[26] S. Shah, X. Luo, S. Kanakasabai, R. Tuason, and G. Klopper, “Neural networks for mining the associations between diseases and symptoms in clinical notes,” *Health Information Science and Systems*, vol. 7, 1 pages, 2019.

[27] Z. Liao, Y. Ju, and Q. Zou, “Prediction of G protein-coupled receptors with SVM-Prot features and random forest,” *Scientifica*, vol. 2016, Article ID 8309253, 10 pages, 2016.

[28] Y. Freund and R. E. Schapire, *Experiments with a new boosting algorithm*, ICML, 1996.

[29] T. Chen and C. Guestrin, “XGBoost: a scalable tree boosting system,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[30] D. Lam, X. Zhang, H. Li et al., “Predicting gamma passing rates for portal dosimetry-based IMRT QA using machine learning,” *Medical Physics*, vol. 46, no. 10, pp. 4666–4675, 2019.

[31] Y. Zhang, Y. Wang, W. Zhou et al., “A combined drug discovery strategy based on machine learning and molecular docking,” *Chemical Biology & Drug Design*, vol. 93, no. 5, pp. 685–699, 2019.

[32] C. Wang, L. Liu, C. Xu, and W. Lv, “Predicting future driving risk of crash-involved drivers based on a systematic machine learning framework,” *International Journal of Environmental Research and Public Health*, vol. 16, no. 3, p. 334, 2019.

[33] M. Maniruzzaman, M. J. Rahman, B. Ahammed, and M. M. Abedin, “Classification and prediction of diabetes disease using machine learning paradigm,” *Health Information Science and Systems*, vol. 8, no. 1, pp. 7–7, 2020.

[34] Z. Chen, J. Chen, R. Collins et al., “China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up,” *International Journal of Epidemiology*, vol. 40, no. 6, pp. 1652–1666, 2011.

[35] D. Pei, C. Zhang, Y. Quan, and Q. Guo, “Identification of potential type II diabetes in a Chinese population with a sensitive decision tree approach,” *Journal of Diabetes Research*, vol. 2019, Article ID 4248218, 7 pages, 2019.

[36] Y. Xu, L. Wang, J. He et al., “Prevalence and control of diabetes in Chinese adults,” *JAMA*, vol. 310, no. 9, pp. 948–959, 2013.

[37] B. D. Pollock, T. Hu, W. Chen et al., “Utility of existing diabetes risk prediction tools for young black and white adults: evidence from the Bogalusa Heart Study,” *Journal of Diabetes and its Complications*, vol. 31, no. 1, pp. 86–93, 2017.

[38] L. Yang, K. Yan, D. Zeng et al., “Association of polycyclic aromatic hydrocarbons metabolites and risk of diabetes in coke oven workers,” *Environmental Pollution*, vol. 223, pp. 305–310, 2017.

[39] L. Yang, Y. Zhou, H. Sun et al., “Dose-response relationship between polycyclic aromatic hydrocarbon metabolites and risk of diabetes in the general Chinese population,” *Environmental Pollution*, vol. 195, pp. 24–30, 2014.

[40] K. Y. Ngiam and I. W. Khor, “Big data and machine learning algorithms for health-care delivery,” *The Lancet Oncology*, vol. 20, no. 5, pp. e262–e273, 2019.

[41] F. Degenhardt, S. Seifert, and S. Szymczak, “Evaluation of variable selection methods for random forests and omics data sets,” *Briefings in Bioinformatics*, vol. 20, no. 2, pp. 492–503, 2019.

[42] P. C. Austin and J. V. Tu, "Automated variable selection methods for logistic regression produced unstable models for predicting acute myocardial infarction mortality," *Journal of Clinical Epidemiology*, vol. 57, no. 11, pp. 1138–1146, 2004.

[43] M. Maniruzzaman, H. S. Suri, N. Kumar et al., "Risk factors of neonatal mortality and child mortality in Bangladesh," *Journal of Global Health*, vol. 8, no. 1, pp. 010417–010417, 2018.

[44] V. K. Shrivastava, N. D. Londhe, R. S. Sonawane, and J. S. Suri, "A novel and robust Bayesian approach for segmentation of psoriasis lesions and its risk stratification," *Computer Methods and Programs in Biomedicine*, vol. 150, pp. 9–22, 2017.

[45] V. K. Shrivastava, N. D. Londhe, R. S. Sonawane, and J. S. Suri, "Computer-aided diagnosis of psoriasis skin images with HOS, texture and color features: a first comparative study of its kind," *Computer Methods and Programs in Biomedicine*, vol. 126, pp. 98–109, 2016.

[46] M. Bejani, D. Gharavian, and N. M. Charkari, "Audiovisual emotion recognition using ANOVA feature selection method and multi-classifier neural networks," *Neural Computing & Applications*, vol. 24, no. 2, pp. 399–412, 2014.

[47] Y. Wang, J. Ji, and P. Liang, "Feature selection of fMRI data based on normalized mutual information and fisher discriminant ratio," *Journal of X-Ray Science and Technology*, vol. 24, no. 3, pp. 467–475, 2016.

[48] B. J. Lee, B. Ku, J. Nam, D. D. Pham, and J. Y. Kim, "Prediction of fasting plasma glucose status using anthropometric measures for diagnosing type 2 diabetes," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 2, pp. 555–561, 2014.

[49] B. J. Lee and J. Y. Kim, "A comparison of the predictive power of anthropometric indices for hypertension and hypotension risk," *PLoS One*, vol. 9, no. 1, article e84897, 2014.

[50] H. Yu, X. Yang, S. Zheng, and C. Sun, "Active learning from imbalanced data: a solution of online weighted extreme learning machine," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 4, pp. 1088–1103, 2019.

[51] Y. Tang, Y. Zhang, N. V. Chawla, and S. Krasser, "SVMs modeling for highly imbalanced classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 39, no. 1, pp. 281–288, 2009.

[52] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.

[53] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[54] C.-p. Li, X.-y. Zhi, J. Ma et al., "Performance comparison between logistic regression, decision trees, and multilayer perceptron in predicting peripheral neuropathy in type 2 diabetes mellitus," *Chinese Medical Journal*, vol. 125, no. 5, pp. 851–857, 2012.

[55] B. A. Goldstein, A. M. Navar, and R. E. Carter, "Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges," *European Heart Journal*, vol. 38, no. 23, pp. 1805–1814, 2017.

[56] B. A. Goldstein, E. C. Polley, and F. B. Briggs, "Random forests for genetic association studies," *Statistical Applications in Genetics and Molecular Biology*, vol. 10, no. 1, pp. 32–32, 2011.

[57] J. Taylor and R. J. Tibshirani, "Statistical learning and selective inference," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 112, no. 25, pp. 7629–7634, 2015.

[58] G. Li, P. Zhang, J. Wang et al., "The long-term effect of lifestyle interventions to prevent diabetes in the China Da Qing Diabetes Prevention Study: a 20-year follow-up study," *The Lancet*, vol. 371, no. 9626, pp. 1783–1789, 2008.

[59] T. Saaristo, L. Moilanen, E. Korpi-Hyövälti et al., "Lifestyle intervention for prevention of type 2 diabetes in primary health care: one-year follow-up of the Finnish National Diabetes Prevention Program (FIN-D2D)," *Diabetes Care*, vol. 33, no. 10, pp. 2146–2151, 2010.

[60] J. Tuomilehto, J. Lindström, J. G. Eriksson et al., "Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance," *The New England Journal of Medicine*, vol. 344, no. 18, pp. 1343–1350, 2001.

[61] M. Xue, L. Liu, S. Wang et al., "A simple nomogram score for screening patients with type 2 diabetes to detect those with hypertension: a cross-sectional study based on a large community survey in China," *PLoS One*, vol. 15, no. 8, article e0236957, 2020.

[62] C. Holst, U. Becker, M. E. Jørgensen, M. Grønbæk, and J. S. Tolstrup, "Alcohol drinking patterns and risk of diabetes: a cohort study of 70,551 men and women from the general Danish population," *Diabetologia*, vol. 60, no. 10, pp. 1941–1950, 2017.

[63] C. Knott, S. Bell, and A. Britton, "Alcohol consumption and the risk of type 2 diabetes: a systematic review and dose-response meta-analysis of more than 1.9 million individuals from 38 observational studies," *Diabetes Care*, vol. 38, no. 9, pp. 1804–1812, 2015.

[64] A. Pan, Y. Wang, M. Talaei, F. B. Hu, and T. Wu, "Relation of active, passive, and quitting smoking with incident type 2 diabetes: a systematic review and meta-analysis," *The lancet Diabetes & Endocrinology*, vol. 3, no. 12, pp. 958–967, 2015.

[65] S. Akter, A. Goto, and T. Mizoue, "Smoking and the risk of type 2 diabetes in Japan: a systematic review and meta-analysis," *Journal of Epidemiology*, vol. 27, no. 12, pp. 553–561, 2017.

[66] S. Polsky and H. K. Akturk, "Alcohol consumption, diabetes risk, and cardiovascular disease within diabetes," *Current Diabetes Reports*, vol. 17, no. 12, pp. 136–136, 2017.

[67] The Diabetes Prevention Program Research, "The Diabetes Prevention Program. Design and methods for a clinical trial in the prevention of type 2 diabetes," *Diabetes Care*, vol. 22, no. 4, pp. 623–634, 1999.

[68] G. Asaad and C. B. Chan, "Food sources of sodium, saturated fat, and added sugar in the Physical Activity and Nutrition for Diabetes in Alberta (PANDA) trial," *Applied Physiology, Nutrition, and Metabolism*, vol. 42, no. 12, pp. 1270–1276, 2017.

[69] S. A. Sullivan and L. L. Birch, "Pass the sugar, pass the salt: experience dictates preference," *Developmental Psychology*, vol. 26, no. 4, pp. 546–551, 1990.