BMJ Open

Predictive model and risk analysis for diabetic retinopathy using machine learning: a retrospective cohort study in China

Wanyue Li ,[1,2] Yanan Song ,[3] Kang Chen,[4] Jun Ying,[5] Zhong Zheng,[6] Shen Qiao,[3] Ming Yang,[3] Maonian Zhang,[1,2] Ying Zhang[2]

To cite: 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. doi:10.1136/bmjopen-2021-050989

► Prepublication history for this paper is available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2021-050989).

WL and YS contributed equally.

Received 08 March 2021
Accepted 19 October 2021

For numbered affiliations see end of article.

Correspondence to
Dr Ying Zhang;
zhangyingdoc@126.com

ABSTRACT

Objective Aiming to investigate diabetic retinopathy (DR) risk factors and predictive models by machine learning using a large sample dataset.

Design Retrospective study based on a large sample and a high dimensional database.

Setting A Chinese central tertiary hospital in Beijing.

Participants Information on 32 452 inpatients with type-2 diabetes mellitus (T2DM) were retrieved from the electronic medical record system from 1 January 2013 to 31 December 2017.

Methods Sixty variables (including demography information, physical and laboratory measurements, system diseases and insulin treatments) were retained for baseline analysis. The optimal 17 variables were selected by recursive feature elimination. The prediction model was built based on XGBoost algorithm, and it was compared with three other popular machine learning techniques: logistic regression, random forest and support vector machine. In order to explain the results of XGBoost model more visually, the Shapley Additive exPlanation (SHAP) method was used.

Results DR occurred in 2038 (6.28%) T2DM patients. The XGBoost model was identified as the best prediction model with the highest AUC (area under the curve value, 0.90) and showed that an HbA1c value greater than 8%, nephropathy, a serum creatinine value greater than 100 µmol/L, insulin treatment and diabetic lower extremity arterial disease were associated with an increased risk of DR. A patient's age over 65 was associated with a decreased risk of DR.

Conclusions With better comprehensive performance, XGBoost model had high reliability to assess risk indicators of DR. The most critical risk factors of DR and the cut-off of risk factors can be found by SHAP method to render the output of the XGBoost model clinically interpretable.

Strengths and limitations of this study

- This study is based on a large sample and a high dimensional database.
- XGBoost algorithm supports multi-threaded calculations, is less time-consuming, and has high model accuracy and good robustness.
- The Shapley Additive exPlanation value is a good method to render the output of the XGBoost model clinically interpretable, so as to provide more targeted suggestions for the treatment and management of type 2 diabetes inpatients.
- This was a single-centre retrospective study with only internal validation.

INTRODUCTION

Diabetic retinopathy (DR) is the leading cause of permanent and irreversible blindness in working-age adults globally.[1] DR is one of the common microvascular complications, and it not only affects a large population (25%, 95% CI 19% to 31%), but also presents more severe conditions, such as proliferative diabetic retinopathy (PDR) (15%, 95% CI 10% to 20%) in China.[2] Thus, controlling or reducing DR and its related vision loss is essential. Exploring the predictive and clinically significant factors influencing the occurrence of DR has garnered significant research interest.

The pathogenesis of DR is complex and multi-factorial.[3–5] Many experimental and clinical studies have explored the influencing factors related to the occurrence of DR.[6–8] However, the ordinarily used statistical methods, including logistic regression, show the over-fitting and instability of coefficients when a number of intercorrelated biomarkers are used and thus many practically significant factors are not supported by statistical results due to the limitations.[9]

Machine learning algorithms that have better generalisability and discrimination in high-dimensional data can prevent the samples from following the strict inclusion and exclusion criteria, thus reflecting the real health status of all patients. Many machine learning algorithms have been widely used in diabetes mellitus diagnosis, management and other related clinical administration aspects,

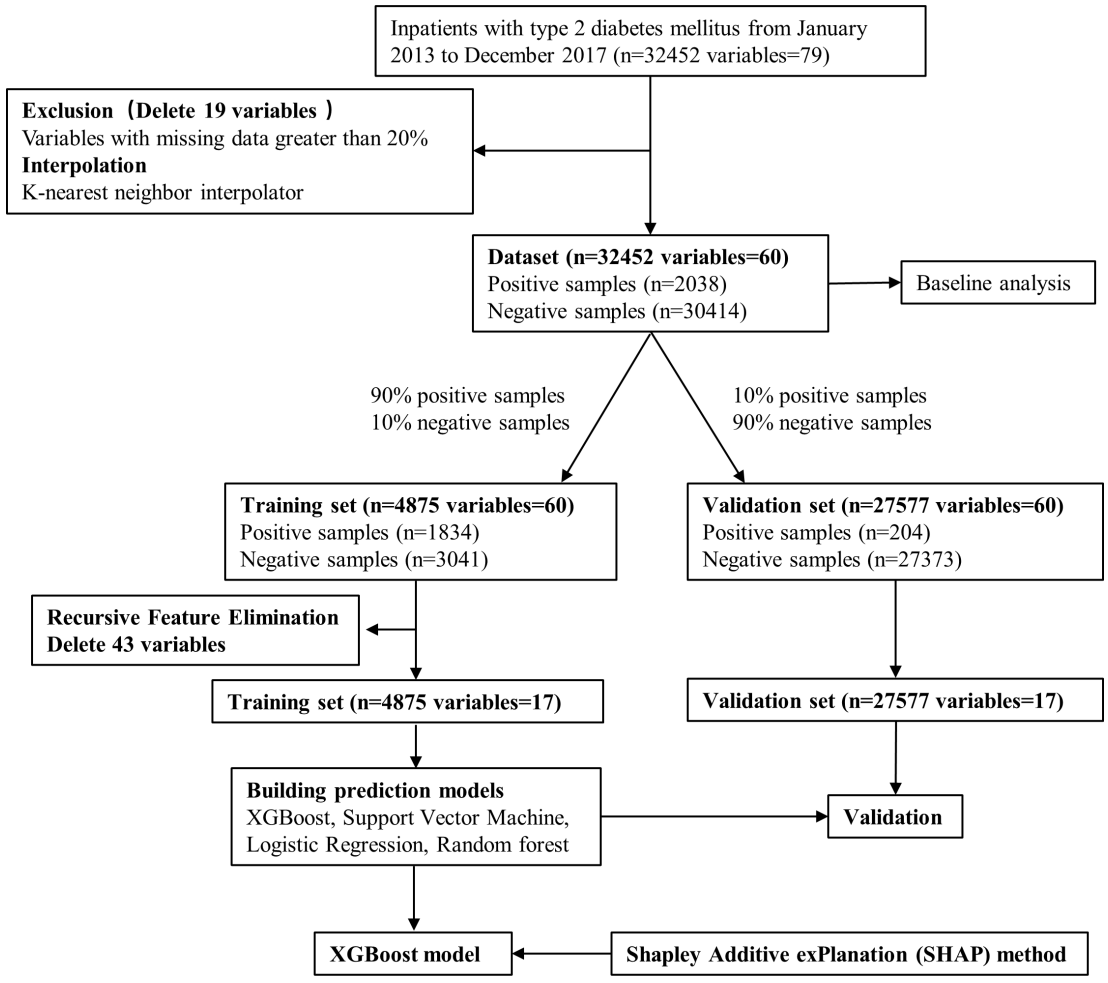

Figure 1 General schema for prediction model building and evaluation. The positive samples were defined as patients with diabetic retinopathy (DR), and negative samples were patients without DR.

particularly in the occurrence and progression of complications.[10] Therefore, machine learning algorithms are an effective means to use abundant available diabetes-related data to extract information. Thus far, there have been only a few reports of machine learning analysis of electronic health record data to assess the risks of DR.[11] However, these studies have mainly compared different models without specific explanation of variables retained in the model.[12]

In this study, we built an extreme gradient boosting (XGBoost) model to predict the risk of DR. In addition, the Shapley Additive exPlanation (SHAP) method is used to explain the XGBoost model to quantify the influence of risk factors of DR.

MATERIALS AND METHODS

Data collection

In this study, the clinical data of inpatients with type-2 diabetes mellitus (T2DM) were retrieved from the Chinese PLA general hospital electronic medical record system from 1 January 2013 to 31 December 2017.

Inclusion criteria

Patients with a discharge diagnosis of T2DM. Diagnostic information of the diseases was extracted from the discharge diagnosis records. The first record of measurement on the first admission for each variable was extracted.

Exclusion criteria

Variables with more than 20% of missing data. Patients with cataract, keratitis, corneal speckles and other eye diseases that affect fundus examination. Patients with fundus diseases other than DR.

Diagnostic criteria

The diagnosis criteria for T2DM followed the criteria of the 2003 American Diabetes Association.[13] DR was diagnosed according to the International Clinical Diabetic Retinopathy Severity Scale[14] using the macula-centred 45° fundus photograph and indirect ophthalmoscopy when pupils were dilated. Fundus photograph reading and examinations were performed by two experienced ophthalmologists. All patients with diabetic fundus lesions, including mild non-proliferative DR, were

Table 1 Baseline analysis results of 60 variables of 32 452 patients with T2DM

Variables		Total (n=32 452)	Non-DR (n=30 414)	DR (n=2038)	P value
Age		59.71±12.64	59.86±12.69	57.43±11.67	<0.001**
Sex (Female)	Female	10 962 (33.78)	10 217 (33.59)	745 (36.56)	0.007**
Nationality	Han	30 461 (93.86)	28 550 (93.87)	1911 (93.77)	0.834
	Others	1806 (5.57)	1689 (5.55)	117 (5.74)	
	Unknown	185 (0.57)	175 (0.58)	10 (0.49)	
Marital status	Married	31 526 (97.15)	29 544 (97.14)	1982 (97.25)	0.820
	Others	926 (2.85)	870 (2.86)	56 (2.75)	
Permanent residence	Urban	27 484 (84.69)	25 830 (84.93)	1654 (81.16)	<0.001**
	Rural	4968 (15.31)	4584 (15.07)	384 (18.84)	
Occupation	Stable	14 404 (44.39)	13 570 (44.62)	834 (40.92)	0.001**
	Unstable	18 048 (55.61)	16 844 (55.38)	1204 (59.08)	
Hypertension	Yes	20 834 (64.20)	19 328 (63.55)	1506 (73.90)	<0.001**
Hyperlipidaemia	Yes	9567 (29.48)	9164 (30.13)	403 (19.77)	<0.001**
Atherosclerosis	Yes	17 083 (52.64)	16 022 (52.68)	1061 (52.06)	0.604
Stroke	Yes	2264 (6.98)	2050 (6.74)	214 (10.50)	<0.001**
Fatty liver	Yes	9849 (30.35)	9165 (30.13)	684 (33.56)	0.001**
Liver cirrhosis	Yes	550 (1.69)	525 (1.73)	25 (1.23)	0.109
Other chronic liver disease	Yes	4605 (14.19)	4311 (14.17)	294 (14.43)	0.778
Pancreatic disease	Yes	726 (2.24)	691 (2.27)	35 (1.72)	0.118
Biliary tract diseases	Yes	4613 (14.21)	4291 (14.11)	322 (15.80)	0.037*
Nephropathy	Yes	8611 (26.53)	7383 (24.28)	1228 (60.26)	<0.001**
Kidney failure	Yes	817 (2.52)	608 (2.00)	209 (10.26)	<0.001**
Nervous system disease	Yes	2362 (7.28)	2238 (7.36)	124 (6.08)	0.036*
Coronary heart disease	Yes	13 114 (40.41)	12 553 (41.27)	561 (27.53)	<0.001**
Myocardial infarction	Yes	3026 (9.32)	2919 (9.60)	107 (5.25)	<0.001**
Arrhythmias	Yes	2790 (8.60)	2648 (8.71)	142 (6.97)	0.008**
Respiratory system diseases	Yes	5545 (17.09)	5202 (17.10)	343 (16.83)	0.774
Diabetic lower extremity arterial disease	Yes	2963 (9.13)	2456 (8.08)	507 (24.88)	<0.001**
Hemopathy	Yes	2556 (7.88)	2122 (6.98)	434 (21.30)	<0.001**
Rheumatic immune disease	Yes	1252 (3.86)	1194 (3.93)	58 (2.85)	0.017*
Endocrine disease	Yes	8855 (27.29)	7992 (26.28)	863 (42.35)	<0.001**
Digestive system neoplasms	Yes	2593 (7.99)	2532 (8.33)	61 (2.99)	<0.001**
Urinary neoplasms	Yes	458 (1.41)	438 (1.44)	20 (0.98)	0.109
Gynaecological neoplasms	Yes	1149 (3.54)	1103 (3.63)	46 (2.26)	0.001*
Lung neoplasms	Yes	855 (2.63)	838 (2.76)	17 (0.83)	<0.001**
Other neoplasms	Yes	3327 (10.25)	3202 (10.53)	125 (6.13)	<0.001**
Insulin treatment	Yes	20 037 (61.74)	18 249 (60.00)	1788 (87.73)	<0.001**
SBP, mm Hg		135±19	135±19	142±21	<0.001**
DBP, mm Hg		79±11	79±11	82±12	<0.001**
FBG, mmol/L		7.25 (5.93, 9.51)	7.23 (5.94, 9.44)	7.83 (5.78, 10.73)	<0.001**
HbA1c, %		7.1 (6.4, 8.3)	7.1 (6.4, 8.2)	7.9 (6.7, 9.4)	<0.001**
TG, mg/day		1.55 (1.10, 2.28)	1.55 (1.10, 2.27)	1.53 (1.11, 2.34)	0.621
TC, mg/dL		4.34 (3.62, 5.10)	4.32 (3.61, 5.09)	4.52 (3.81, 5.37)	<0.001**
HDL, mg/dL		1.02 (0.86, 1.23)	1.02 (0.85, 1.23)	1.03 (0.87, 1.24)	0.044*

Continued

Table 1 Continued

Variables	Total (n=32 452)	Non-DR (n=30 414)	DR (n=2038)	P value
LDL, mg/dL	2.71±0.99	2.70±0.97	2.93±1.19	<0.001**
Fbg, g/L	3.27 (2.80, 3.98)	3.26 (2.80, 3.94)	3.59 (2.96, 4.62)	<0.001**
BUN, mmol/L	5.41 (4.43, 6.69)	5.38 (4.40, 6.60)	6.30 (4.96, 8.70)	<0.001**
SCr, µmol/L	70.1 (59.0, 83.5)	69.9 (59.0, 82.6)	77.5 (59.8, 114.6)	<0.001**
SUA, umol/L	324.3±99.2	323.5±99.1	335.9±100.6	<0.001**
Hb, g/L	137±21	137±20	128±24	<0.001**
Hct, %	41 (37, 44)	41 (38, 44)	38 (34, 42)	<0.001**
PLT, 10^9/L	205 (170, 247)	205 (170, 247)	208 (172, 252)	0.023*
TBil, umol/L	10.4 (7.7, 14.0)	10.5 (7.8, 14.1)	8.9 (6.2, 12.6)	<0.001**
DBil, umol/L	3.2 (2.3, 4.5)	3.3 (2.4, 4.5)	2.5 (1.6, 3.6)	<0.001**
TP, g/L	67.34±6.68	67.55±6.55	64.15±7.77	<0.001**
ALB, g/L	41.5 (38.7, 44.1)	41.7 (38.9, 44.2)	39.7 (35.4, 42.3)	<0.001**
LDH, U/L	153.9 (134.9, 180.0)	153.3 (134.5, 179.3)	161.4 (140.9, 191.7)	<0.001**
ALT, U/L	19.6 (13.8, 29.9)	19.8 (13.9, 30.4)	16.3 (11.9, 23.4)	<0.001**
AST, U/L	17.2 (13.8, 22.8)	17.4 (13.9, 23.0)	15.6 (12.6, 20.1)	<0.001**
GGT, U/L	28.1 (18.8, 47.8)	28.6 (19.1, 48.7)	22.4 (15.7, 34.7)	<0.001**
ALP, U/L	68.2 (56.4, 83.2)	68.2 (56.4, 83.2)	67.9 (55.7, 82.9)	0.147
PT, s	13.1 (12.6, 13.7)	13.1 (12.6, 13.7)	12.9 (12.4, 13.5)	<0.001**
PTA, %	99 (90, 108)	99 (90, 108)	100 (91, 110)	<0.001**
APTT, s	35.8 (33.3, 38.7)	35.8 (33.3, 38.7)	35.7 (33.3, 38.58)	0.145
GLO, g/L	25.9 (22.9, 29.3)	25.9 (22.9, 29.3)	25.5 (22.5, 28.7)	<0.001**

The continuous variables were expressed as mean±SD or the median (IQR) after the normality distribution test. The categorical variables were expressed as number (percentage).
*P value <0.05; **p value <0.01.
ALB, albumin; ALP, alkaline phosphatase transferase; ALT, alanine aminotransferase; APTT, activated partial thromboplastin time; APTT, activated partial thromboplastin time; AST, aspartate aminotransferases; BUN, blood urea nitrogen; DBil, direct bilirubin; DBP, diastolic blood pressure; DR, diabetic retinopathy; FBG, fasting blood glucose; Fbg, fibrinogen; GGT, glutamine; GLO, globulin; Hb, haemoglobin; Hct, haematocrit; HDL-C, high density lipoprotein; LDH, lactate dehydrogenase; LDL-C, low density lipoprotein; Marital status, others (single, divorced, widow); non-DR, diabetics without diabetic retinopathy; PLT, platelet count; PT, prothrombin time; PTA, prothrombin activity; SBP, systolic blood pressure; SCr, serum creatinine; SUA, serum uric acid; TBil, total bilirubin; TC, total cholesterol; TG, triglyceride; TP, total protein.

defined in the DR group (microaneurysms, more than 20 intraretinal haemorrhages in each of the four quadrants, definite venous beading in 2+ quadrants, prominent intraretinal microvascular abnormalities in 1+ quadrant, neovascularisation or vitreous/preretinal haemorrhage). The positive samples were defined as patients with DR, and negative samples were patients without DR.

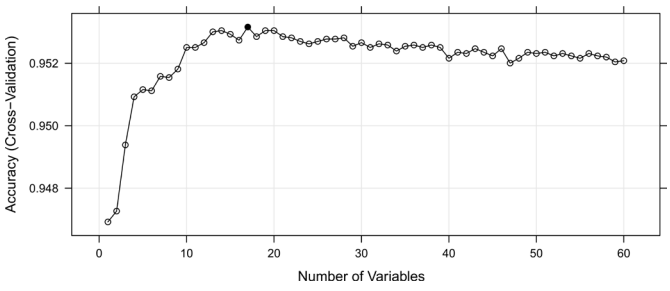

Figure 2 Feature selection accuracy curve. The accuracy got the highest value when the number of variables was 17 (represented as a solid point).

Statistical analysis

Data interpolation

In order to improve the data utilisation, the missing data needed to be interpolated. The k-nearest neighbour interpolator (KNNI) method was used to interpolate the individual missing data. Based on the available variables of

Table 2 Performance of prediction models in the validation set

Method	Accuracy	Sensitivity	Specificity	ROC-AUC
XGBoost	**0.90**	**0.70**	**0.90**	**0.90**
SVM	0.89	0.45	0.90	0.79
LR	0.86	0.59	0.86	0.83
RF	0.92	0.63	0.92	0.87

LR, logistic regression; RF, random forest; ROC-AUC, areas under receiver operator characteristic curves; SVM, support vector machine; XGBoost, Extreme Gradient Boosting.

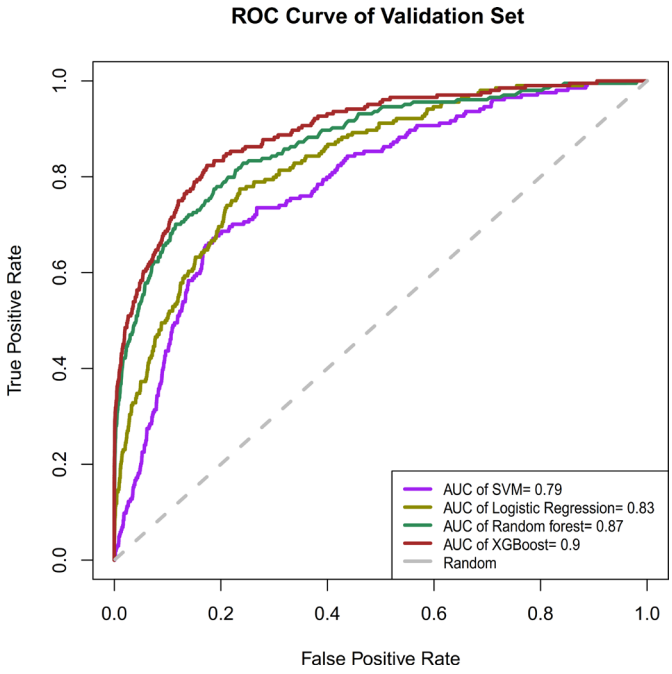

Figure 3 ROC curve of validation set. LR, logistic regression; RF, random forest; ROC-AUC, areas under receiver operator characteristic curves; SVM, support vector machine; XGBoost, Extreme Gradient Boosting.

the sample to be imputed, the k closest complete samples were found. Thereafter, the distance function was used to calculate the distances between these k complete samples and the sample to be interpolated. Finally, we weighted the variables of k samples according to their distances and generated the estimated value.

Baseline analysis of the complete data set was conducted in the interpolated data set. The continuous variables were expressed as mean±SD or the median (IQR) after the normality distribution test. The categorical variables were expressed as number and percentage. χ^2 test in categorical variables and t-test in continuous variables were performed. The value of $p<0.05$ was considered statistically significant.

Data set division

Because of the imbalance in the distribution of positive and negative samples, the random under-sampling method was used to generate the training and validation sets. The training set, containing 90% positive samples and 10% negative samples, was used to train the prediction models. The validation set, which comprised the rest of samples, was used to assess the ability of the machine learning models to predict DR in diabetic patients.

Feature selection

Feature selection was aimed to exclude redundant factors without losing key information and to determine a factor set of lower dimensions, improve the accuracy and reduce the complexity of the model. Recursive feature elimination (RFE) method was used to determine the optimal variables for feature selection. The RFE method is a greedy algorithm, which is the representative of the wrapper model algorithm. With the whole data set as the starting point and the prediction accuracy as the evaluation criterion, the least relevant variable is eliminated through each iteration; furthermore, the feature ranking is performed based on this. The more relevant the variable, the higher the ranking. The RFE method will generate some feature subsets according to the above evaluation criteria and finally select the optimal feature subset. In this study, random forest was determined as the basic classifier for RFE, and the feature selection was performed on the training set. The criterion of feature screening was model optimisation. Therefore, the multicollinearity between variables was not considered in this study. For machine learning algorithms, the multicollinearity between variables had little impact on the predictive performance of the model, thus it was more important to select the best combination of variables.

Prediction model training and validation

In this study, XGBoost was used to develop the predictive model. XGBoost was proposed by Chen 2016,[15] using the negative gradient of the loss function as the residual value of the current fitting to achieve an accurate classification effect. XGBoost performs a second-order Taylor expansion of the loss function and adds a regular term outside the loss function to balance the decline of the loss function and the complexity of the model, thereby reducing the possibility of overfitting.

To make the model more convincing, we also compared the performance of XGBoost with three other popular machine learning techniques: logistic regression (LR), random forest (RF) and support vector machine (SVM). In this study, accuracy, sensitivity, specificity and the areas under the receiver operator characteristic curves (ROC-AUC) were used as the criteria to compare the performance of the model. A 10-fold cross validation was performed to compare the AUC of XGBoost and random forest models and to determine the overall best performance. Given the values of true negative (TN), the values of true positive (TP), false negative (FN) and false positive (FP) were calculated from the confusion matrix; the formulas of the afore-mentioned measures are detailed in the following text.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

$$\text{Specificity} = \frac{TN}{TN+FP}$$

With traditional XGBoost output, only the importance of variables is sorted; however, it is impossible to measure the direction and level of influence of the variables on outcomes. To better explain the results of machine learning models, the SHAP method was used for visualisation analysis. SHAP is a framework based on additive feature attribution methods, which was first proposed by Lloyd Shapley in game theory.[16] Intuitively, a SHAP value is the contribution of the feature to the outcome

value. A positive SHAP value indicates that the feature improves the outcome value and has a positive effect; on the contrary, a negative SHAP value indicates that this feature reduces the outcome value and has a negative effect. This method can output the importance ranking of the features, as well as the relationship between the features and the outcome.

In this study, data were retrieved by suing Procedural Language/SQL on Oracle Database (a database management system). R programming language (V.3.6.1) and Python (V.3.7.7) were used for statistical analysis. The general schema for the prediction model building is shown in figure 1.

RESULTS

The data of 32 452 T2DM inpatients including 2038 DR patients and 30 414 non-DR patients, and 79 variables was extracted. Nineteen variables were deleted for data missing greater than 20%. So there reserved 60 variables. The following variables were obtained: demography, other diseases besides T2DM and DR such as nephropathy, laboratory measurements, physical indicators, and insulin treatment. After the interpolation with KNNI, baseline analysis of data sets is shown in table 1. The average age of 32 452 patients with T2DM was 59.71±12.64 years, including 21 490 males (66%) and 10 962 females (34%). A total of 2038 patients (6.3%) were diagnosed with DR among which 63% were males and 37% were females.

Feature selection

According to the results of RFE, 17 variables were selected to build the prediction model, they were age, fasting blood glucose, HbA1c, total cholesterol, triglyceride, serum creatinine, serum urea, direct bilirubin, total protein, albumin, glutamine transferase, lactate dehydrogenase, fibrinogen, prothrombin activity, nephropathy, diabetic lower extremity arterial disease (DLEAD) and insulin treatment. Figure 2 shows how the accuracy varies with the number of variables.

Model performance

The training set comprised 1834 positive samples and 3041 negative samples. The validation set comprised 204 positive samples and 27 373 negative samples. XGBoost, Logistic regression (LR), Random Forest (RF) and Support Vector Machine (SVM) were developed based on the training set with the above-mentioned 17 variables. The results of the performance assessment—accuracy, sensitivity, specificity and ROC-AUC—are detailed in table 2 and figure 3. In the validation set, the XGBoost model showed the highest AUC value (0.90), which is the key index for evaluating the function of the predictive model.

XGBoost and RF were selected to be further assessed by 10-fold cross-validation in the whole data set because of their well comprehensive performance. The results showed that AUC values were 0.86 (95% CI 0.85 to 0.86) and 0.89 (95%CI 0.88 to 0.90), respectively. XGBoost model delivered optimal performance across the four machine learning algorithms. It was identified as the best model in this study.

DR influencing factors assessment

To identify the importance of each feature to the prediction model, a SHAP summary plot of the XGBoost model was framed (figure 4). HbA1c, nephropathy, serum creatinine and insulin treatment were at the top of the ranking list. As is illustrated in the SHAP summary plot, the higher the SHAP value of a feature, the more likely the occurrence of DR. The red dots represent higher feature values, and the blue dots represent lower feature values. The high value of HbA1c, nephropathy, serum creatinine and insulin treatment correspond to a SHAP value greater than zero. This suggests that these features are important risk factors for DR.

The SHAP dependence plot shows the effect of a single feature on the output of the XGBoost model (figure 5). When the SHAP value of each feature exceeds zero, this indicates an increased risk of DR. An HbA1c value greater than 8%, nephropathy, a serum creatinine value greater than 100 µmol/L, insulin treatment and DLEAD were associated with an increased risk of DR. A patient's age over 65 was associated with a decreased risk of DR. The actual application form of the model is shown in figure 6.

The red area implies that the feature value increases the probability of DR and the blue area indicates that the feature value decreases the probability of DR; f(x) indicates the comprehensive SHAP value of each patient. The base value indicates the average SHAP value of all samples. If the value of f(x) is greater than the base value, the model will predict that the patient has DR. The panel above shows that a DR patient was accurately predicted to suffer from DR. The panel below shows that a patient with a normal fundus was accurately predicted as not suffering from DR. The XGBoost model provides a good distinction between DR and non-DR patients and can indicate different risk probabilities according to the individualised circumstances of each patient.

DISCUSSION

DR, as one of the most common microvascular complications, harms the visual function of 14.77%~22.43% people with diabetes in China.[17] In our study, only 6.3% of T2DM patients suffered DR, which is similar to that mentioned in another report from Beijing (8.1%).[18] Most potential asymptomatic patients of diabetes are not aware of the illness until they start suffering from vision loss or even blindness caused by the deficiency of routine physical examination. The urgent need to provide diabetes patients with targeted guidance on the prevention and management of DR reflects the necessity of analysing the DR influencing factors.

Many studies have investigated the risk factors of DR among different populations or clinical samples.[2 8 12 19]

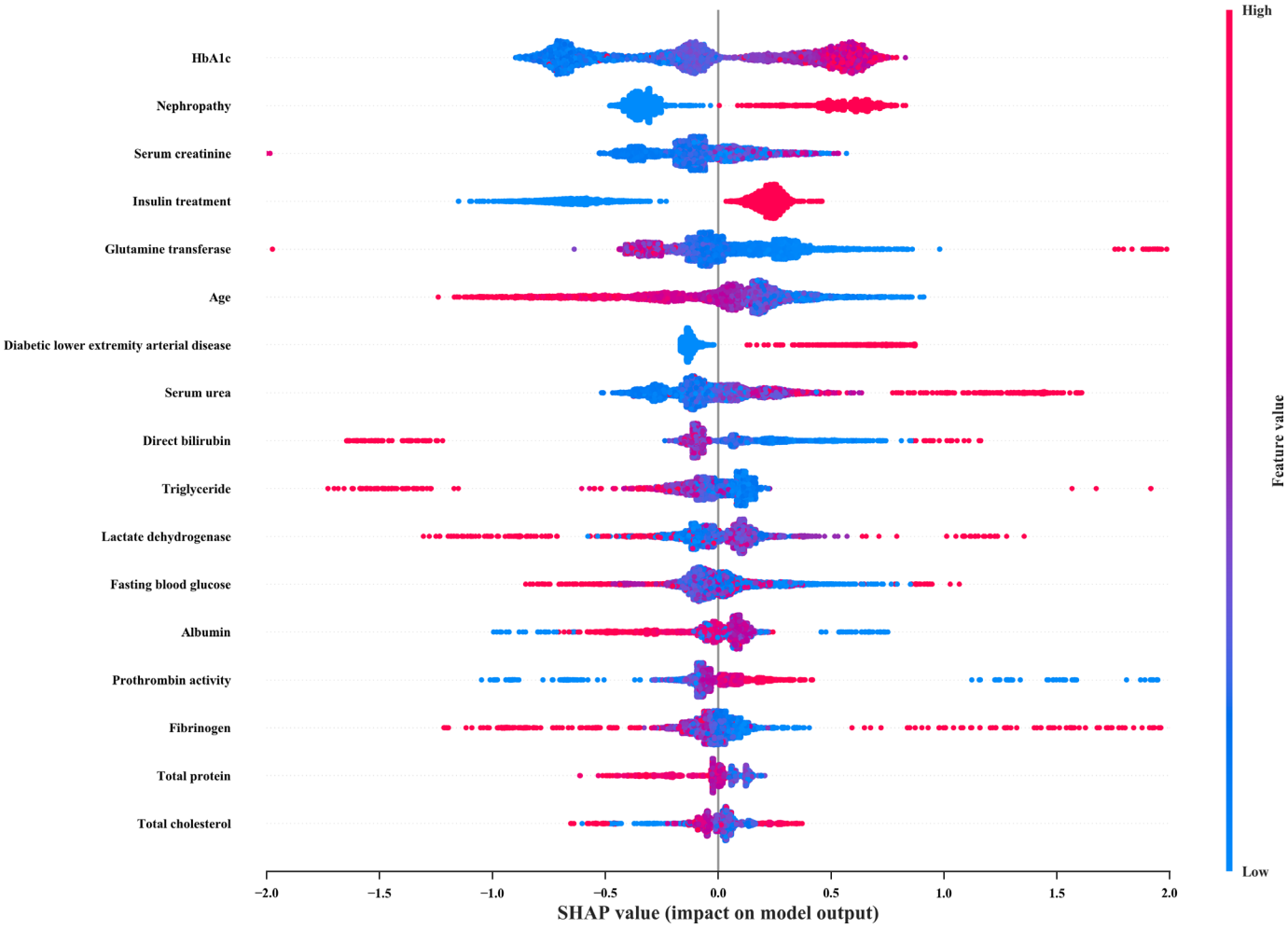

Figure 4 SHAP summary plot of the XGBoost model. The higher the SHAP value of a feature, the higher the risk of DR. The contribution of each feature of each patient to the model corresponds to a dot. The dots are coloured according to the values of features. Red represents a higher feature value, and blue represents a lower feature value. The higher the SHAP value of a feature, the more likely DR occurrence. DR, diabetic retinopathy; SHAP, Shapley Additive exPlanation.

As previous studies have showed, the complexity of DR lies in the multifactorial mechanisms that affect both the development of diabetes and DR, such as the duration of diabetes, level of blood glucose, HbA1c and hypertension and so on.[18 19] Clinically, although blood glucose is an absolutely important factor in the occurrence and progression of DR, it is evidently not the only determinant.[17 20] Assessing the risk of DR should combine the control of blood glucose and systemic factors. Machine learning algorithms has gained widespread attention regarding applications in the analysis of electronic health record data including DR.[10 12 20 21] Oh *et al* demonstrated that the LASSO (Least Absolute Shrinkage and Selection Operator) model had a higher AUC (81%) than traditional indicators (AUC of fasting blood glucose 54%; AUC of glycated haemoglobin 69%) when diagnosing DR.[11] However, they mainly compared different models without specific explanation of variables retained in the model. By comparing several machine learning algorithms, Tsao *et al* identified the use of insulin and duration of diabetes as features to identify the high-risk patients for DR.[12] In their study, the limitations lay in the non-DR and DR samples of only 106 patients and there were only 10 clinical indicators were included.

Owing to the numerous factors affecting the occurrence of DR, a large sample size is needed to systematically study the risk factors of DR and develop prediction models. When the sample size and dimension of the data set are large, the XGBoost algorithm has advantages over the logistic regression algorithm. Because logistic regression is a linear model, the high correlation between independent variables will distort the weight parameter estimation of the model. The XGBoost algorithm is an ensemble algorithm based on decision trees, and it is a non-parametric estimation. The correlation of independent variables has no significant impact on the model.

Although the performance of our XGBoost prediction model is not as good as that obtained via the artificial intelligence (AI) fundus recognition system reported in the past,[22] their objectives are different. The AI fundus recognition system is based on the acquired fundus images of the patient and aims to replace the ophthalmologist in accurately diagnosing fundus diseases; moreover, its requirements for equipment are very strict. The

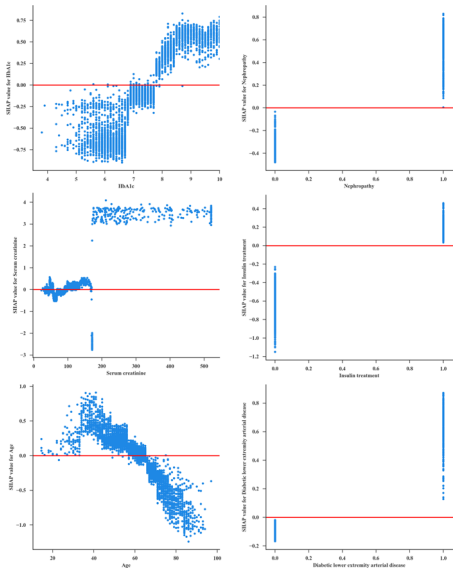

Figure 5 SHAP dependence plot of the XGBoost model. The SHAP value of each feature exceeded zero, indicating an increased risk of DR. HbA1c, nephropathy, serum creatinine, insulin treatment and diabetic lower extremity arterial disease were risk factors of DR. Age was a protective factor of DR. DR, diabetic retinopathy; SHAP, Shapley Additive exPlanation.

purpose of our prediction model was to assist doctors in the health management of diabetic patients and to increase the fundus screening rate of patients as much as possible; this is the requirement before a fundus examination. Similar research has also been reported before, and the comparison of the different types of models is presented in table 3.

In this study, we performed a baseline analysis of 60 variables and then adopted 17 variables via the RFE method with RF as the basic classifier. Feature selection helps to obtain a more reliable weighted ranking of XGBoost in risk factors analysis. XGBoost is a highly flexible non-parametric model that integrates many other machine learning models (decision trees). A few significant advantages of this algorithm are that it supports multi-threaded calculations, is less time-consuming, and has high model accuracy and good robustness.[23] Compared with LR, SVM and RF, the XGBoost model achieved the highest AUC value (0.90) on the internal validation set, this indicates that the XGBoost algorithm is more reliable when analysing high-dimensional data. In addition, the XGBoost model does not only have good performance but also it allows for strong interpretability. The SHAP value is a good for rendering the output of the XGBoost model clinically interpretable. The most critical risk factors of DR and the cut-off of risk factors is found by SHAP method, in order to provide more targeted suggestions for the treatment and management of type 2 diabetes inpatients.

In this study, the XGBoost model showed that HbA1c was the most important risk factor of DR, and insulin treatment also ranked high in the result. An HbA1c value above 8% and the need for insulin treatment increased the risk of DR. Insulin treatment suggests that the glycaemia levels of patients have not been able to return to normal levels through exercise, diet or oral hypoglycaemic agents. The level of hyperglycaemic, as measured using HbA1c determination at a baseline examination, was found to be a strong and independent predictor of the incidence of any retinopathy, and progression of proliferative retinopathy.[24] Variation in FPG (fasting plasma glucose) levels was found to be a risk factor for microvascular complications.[7] The UK Prospective Diabetes Study and the Kumamoto Study have shown that intensive glycaemic control has a significant negative correlation with the rate of microvascular complications in people with type 2 diabetes.[25 26] Many studies have examined the optimal cut-off values of HbA1c to predict the presence of retinopathy, and the results were different.[27–29] Meanwhile, DR as a specific complication of diabetes has been historically accepted as the best criterion to compare glycaemic measures.[30]

Nephropathy and serum creatinine ranked second and third in the list of influencing factors. Suffered from nephropathy, or a serum creatinine value greater than 100 µmol/L increased the risk of DR. This result is consistent with previous studies that indicated patients with chronic kidney disease (CKD) experienced a higher incidence of DR compared with patients without CKD.[31–33] Both retina and kidney are terminal perfusion organs supplied by microvasculature, which are sensitive to fluctuations in blood flow.[34] DR and CKD may progress in parallel. Previous studies indicated a bidirectional relationship between CKD and DR supporting the same pathology because of the shared risk factors such as

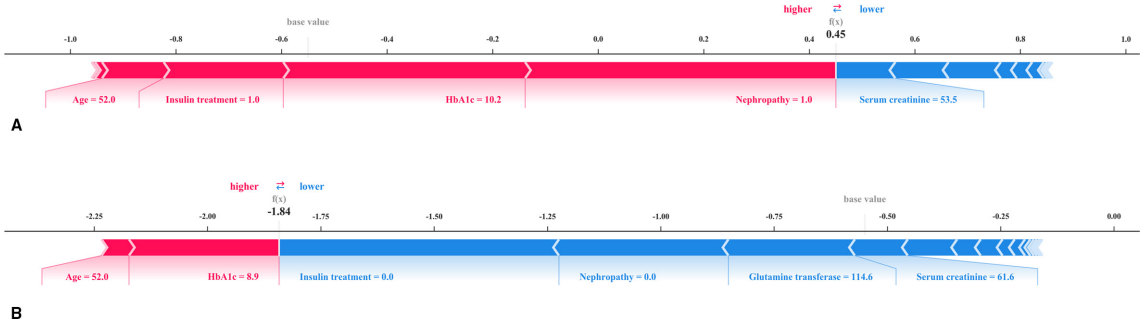

Figure 6 Shapley Additive exPlanation force plot for diabetic retinopathy (DR) patient and non-DR patient.

Table 3 Comparison with other previous DR prediction or diagnosis model

Author	Gulshan *et al*[22]	Liao *et al*[47]	Mendoza-Herrera *et al*[48]	Tsao *et al*[12]	The present prediction model
Published time	2016	2018	2017	2018	/
Number of samples	9963 (EyePACS-1 data set) 1748 (Messidor-2 data set)	1055	1000	536	32 452
Algorithm (best result)	Deep convolutional neural network	Logistic regression	Probit model	Support vector machines	XGBoost
Sensitivity (validation)	0.975(EyePACS-1 data set) 0.961(Messidor-2 data set)	NA	NA	0.933	0.70
Specificity (validation)	0.934 (EyePACS-1 data set) 0.939 (Messidor-2 data set)	NA	NA	0.724	0.90
Accuracy (validation)	NA	NA	NA	0.795	0.90
ROC-AUC (internal validation)	0.991 (EyePACS-1 data set) 0.990 (Messidor-2 data set)	0.744	0.778	0.839	0.90

chronic hyperglycaemic.[35–37] Several studies have shown that diabetic microvascular complication of DR and diabetic nephropathy (DN) are multifactorial diseases involving multiple pathways, oxidative stress, aldose reductase pathway, activation of PKC and complement activation.[38 39] If damage caused by the inflammatory process occurs in the kidneys, it causes DN. If it occurs in retina, it causes DR. The biomarkers of kidney function, such as serum creatinine and serum urea, may reflect the function of retina. It also suggests that the diagnosis of DR should prompt a recommendation to identify if the deterioration of the kidney function of patients is caused by DN, whereas the kidney lesion of diabetes patients without DR is more likely to be due to non-DN such as IgA nephropathy and membranous nephropathy.[40–42]

Age was identified as a protective factor of DR, and it ranked fifth in the influencing list of XGBoost results. An age over 65 was associated with a decreased risk of DR. This is confirmed by a few previous studies. A review showed that an age <45 years was related to severe fibrovascular proliferation (p=0.005)[19]; furthermore, Klein *et al*[43] found that DR patients less than 30 years of age showed a higher risk of PDR. By contrast, in a prospective cohort study, the frequency of DR increased with age; however, the difference was not statistically significant.[44] The relationship between age and the occurrence of DR is influenced by many factors, such as work status, social activities, life schedule and diet. Hence, more causal studies are required to explore and understand this controversial question.

In addition, suffered from DLEAD were associated with an increased risk of DR. LEAD is a diabetic macrovascular complication associated with consistent disability in both clinical symptoms and functions, which can result in diabetic foot.[45] Our results are consistent with previous studies in demonstrating the relationship between LEAD and DR. A previous multicentre observational study reported that, compared with patients without DR, patients with DR were more likely to undergo lower limb amputation (p<0.001).[46] Nwanyanwu *et al*[8] aimed to identify the factors associated with the progression of DR. After adjustments for confounders, those with non-healing ulcers had a 54% (OR=1.54, 95% CI 1.15 to 2.07) increased chance of developing proliferative DR. In our study, complications such as nephropathy and DLEAD were included in the prediction model, which could increase the awareness regarding the existence and importance of comorbidities and serve as a reminder to patients to focus on the prevention and treatment of different complications from DM.

The significance of this study is that it is a real-world risk assessment study, based on 32 452 samples, which was performed by comparing four machine learning algorithms. The best prediction model, the XGBoost model, has a better generalisability benefit from its algorithm. Moreover, using the advantages of a machine learning algorithm, the analysis can include different types of indicators, including blood glucose, kidney function, liver function, coagulation function, and therefore, it can be used to comprehensively analyse the influencing factors. In addition, the SHAP method is a reliable method to enable the output of the XGBoost model to be clinically interpretable. Doctors can propose reasonable referral suggestions and individualised DR health management recommendations to diabetes patients.

There are, however, several limitations of this study. This is a single-centre study with only an internal validation. Furthermore, the deficiency of an important indicator—the duration of diabetes—is due to the limitation of natural language processing capabilities to extract an item from the medical record. More effort will be made in multi-centre prospective study depending on more opportunities for multi-centre cooperation and improvements to data mining capabilities in future work.

CONCLUSION

Compared with LR, SVM and RF, the XGBoost model achieved the highest AUC value (0.90) on the internal validation set, this indicates that the XGBoost algorithm is more reliable when analysing high-dimensional data. The SHAP method is a reliable method to make the output

of the XGBoost model clinically interpretable. HbA1c, nephropathy, serum creatinine, insulin treatment and DLEAD were associated with an increased risk of DR, and age was associated with a decreased risk of DR.

Author affiliations
[1]Medical School of Chinese PLA, Beijing, China
[2]Department of Ophthalmology, Chinese PLA General Hospital, Beijing, China
[3]Medical Big Data Research Center, Medical Innovation Research Division of Chinese PLA General Hospital, Beijing, China
[4]Department of Endocrinology, Chinese PLA General Hospital, Beijing, China
[5]Information Management Department, Chinese PLA General Hospital, Beijing, China
[6]Information Center, Logistics Support Department, Central Military Commission, Beijing, China

Acknowledgements The authors thank the computer office of PLA general hospital for its assistance in extracting clinical data.

Contributors All authors made a substantial contribution to this study. WL interpreted the data, designed the data analysis scheme, and drafted the manuscript. YS interpreted the data, designed the data analysis scheme, analysed the data and drafted the manuscript. YZ and MZ conceived and designed the study. YZ acts as ganrantor. KC interpreted the data. JY, ZZ, SQ and MY acquired and interpreted the data. All the authors reviewed the manuscript for important intellectual content and approved the final version of the manuscript submitted.

Funding This work was funded by Chinese PLA general hospital medical big data programme (2017MBD-020).

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval This study was conducted at the Chinese People's Liberation Army (PLA) General Hospital, and was approved by the institutional clinical research ethics committee (No. S2019-326-02), adhering to the tenets of the Declaration of Helsinki.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available. The data used to support the findings of this study have not been made available because the dataset was built on the hospital's local area network. In military hospitals, computers connected to the local area network cannot exchange information with computers connected to the Internet.

ORCID iDs
Wanyue Li http://orcid.org/0000-0002-9419-6256
Yanan Song http://orcid.org/0000-0002-4422-6349

REFERENCES

1 Congdon NG, Friedman DS, Lietman T. Important causes of visual impairment in the world today. *JAMA* 2003;290:2057–60.
2 Yang Q-H, Zhang Y, Zhang X-M, *et al*. Prevalence of diabetic retinopathy, proliferative diabetic retinopathy and non-proliferative diabetic retinopathy in Asian T2DM patients: a systematic review and meta-analysis. *Int J Ophthalmol* 2019;12:302–11.
3 Menini S, Iacobini C, Vitale M, *et al*. The inflammasome in chronic complications of diabetes and related metabolic disorders. *Cells* 2020;9. doi:10.3390/cells9081812. [Epub ahead of print: 30 Jul 2020].
4 Chan TC, Wilkinson Berka JL, Deliyanti D, *et al*. The role of reactive oxygen species in the pathogenesis and treatment of retinal diseases. *Exp Eye Res* 2020;201:108255.
5 Roy S, Kim D. Retinal capillary basement membrane thickening: role in the pathogenesis of diabetic retinopathy. *Prog Retin Eye Res* 2021;82:100903.
6 Ding Y, Zhao J, Liu G, *et al*. Total bilirubin predicts severe progression of diabetic retinopathy and the possible causal mechanism. *J Diabetes Res* 2020;2020:7219852.
7 Takao T, Ide T, Yanagisawa H, *et al*. The effect of fasting plasma glucose variability on the risk of retinopathy in type 2 diabetic patients: retrospective long-term follow-up. *Diabetes Res Clin Pract* 2010;89:296–302.
8 Harris Nwanyanwu K, Talwar N, Gardner TW, *et al*. Predicting development of proliferative diabetic retinopathy. *Diabetes Care* 2013;36:1562–8.
9 Waldron L, Pintilie M, Tsao M-S, *et al*. Optimized application of penalized regression methods to diverse genomic data. *Bioinformatics* 2011;27:3399–406.
10 Kavakiotis I, Tsave O, Salifoglou A, *et al*. Machine learning and data mining methods in diabetes research. *Comput Struct Biotechnol J* 2017;15:104–16.
11 Oh E, Yoo TK, Park E-C. Diabetic retinopathy risk prediction for fundus examination using sparse learning: a cross-sectional study. *BMC Med Inform Decis Mak* 2013;13:106.
12 Tsao H-Y, Chan P-Y, Su EC-Y. Predicting diabetic retinopathy and identifying interpretable biomedical features using machine learning algorithms. *BMC Bioinformatics* 2018;19:283.
13 Expert Committee on the Diagnosis and Classification of Diabetes Mellitus. Report of the expert Committee on the diagnosis and classification of diabetes mellitus. *Diabetes Care* 2003;26 Suppl 1:S5–20.
14 Wilkinson CP, Ferris FL, Klein RE, *et al*. Proposed International clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology* 2003;110:1677–82.
15 Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Proceedings of the 22nd acm sigkdd International Conference on knowledge discovery and data mining. *ACM* 2016:785–94.
16 Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Sys-tems* 2017:4765–74.
17 Song P, Yu J, Chan KY, *et al*. Prevalence, risk factors and burden of diabetic retinopathy in China: a systematic review and meta-analysis. *J Glob Health* 2018;8:010803.
18 Cui J, Ren J-P, Chen D-N, *et al*. Prevalence and associated factors of diabetic retinopathy in Beijing, China: a cross-sectional study. *BMJ Open* 2017;7:e015473.
19 Wu Y-B, Wang C-G, Xu L-X, *et al*. Analysis of risk factors for progressive fibrovascular proliferation in proliferative diabetic retinopathy. *Int Ophthalmol* 2020;40:2495–502.
20 Park Y-M, Ko S-H, Lee J-M, *et al*. Glycaemic and haemoglobin A1c thresholds for detecting diabetic retinopathy: the fifth Korea National health and nutrition examination survey (2011). *Diabetes Res Clin Pract* 2014;104:435–42.
21 Park C, Took CC, Seong J-K. Machine learning in biomedical engineering. *Biomed Eng Lett* 2018;8:1–3.
22 Gulshan V, Peng L, Coram M, *et al*. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus Photographs. *JAMA* 2016;316:2402–10.
23 Ruan Y, Bellot A, Moysova Z, *et al*. Predicting the risk of inpatient hypoglycemia with machine learning using electronic health records. *Diabetes Care* 2020;43:1504–11.
24 Klein R, Klein BE, Moss SE, *et al*. Glycosylated hemoglobin predicts the incidence and progression of diabetic retinopathy. *JAMA* 1988;260:2864–71.
25 Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33). UK prospective diabetes study (UKPDS) group. *Lancet* 1998;352:837–53.
26 Effect of intensive blood-glucose control with metformin on complications in overweight patients with type 2 diabetes (UKPDS 34). UK prospective diabetes study (UKPDS) group. *Lancet* 1998;352:854–65.
27 Colagiuri S, Lee CMY, Wong TY, *et al*. Glycemic thresholds for diabetes-specific retinopathy: implications for diagnostic criteria for diabetes. *Diabetes Care* 2011;34:145–50.
28 Sabanayagam C, Liew G, Tai ES, *et al*. Relationship between glycated haemoglobin and microvascular complications: is there a natural cut-off point for the diagnosis of diabetes? *Diabetologia* 2009;52:1279–89.
29 Nakagami T, Takahashi K, Suto C, *et al*. Diabetes diagnostic thresholds of the glycated hemoglobin A1c and fasting plasma glucose levels considering the 5-year incidence of retinopathy. *Diabetes Res Clin Pract* 2017;124:20–9.

30 International Expert Committee. International expert Committee report on the role of the A1c assay in the diagnosis of diabetes. *Diabetes Care* 2009;32:1327–34.
31 Edwards MS, Wilson DB, Craven TE, *et al*. Associations between retinal microvascular abnormalities and declining renal function in the elderly population: the cardiovascular health study. *Am J Kidney Dis* 2005;46:214–24.
32 Pedro R-A, Ramon S-A, Marc B-B, *et al*. Prevalence and relationship between diabetic retinopathy and nephropathy, and its risk factors in the north-east of Spain, a population-based study. *Ophthalmic Epidemiol* 2010;17:251–65.
33 Wong TY, Coresh J, Klein R, *et al*. Retinal microvascular abnormalities and renal dysfunction: the Atherosclerosis risk in Communities study. *J Am Soc Nephrol* 2004;15:2469–76.
34 Lee WJ, Sobrin L, Kang MH, *et al*. Ischemic diabetic retinopathy as a possible prognostic factor for chronic kidney disease progression. *Eye* 2014;28:1119–25.
35 Chen Y-H, Chen H-S, Tarng D-C. More impact of microalbuminuria on retinopathy than moderately reduced GFR among type 2 diabetic patients. *Diabetes Care* 2012;35:803–8.
36 Mottl AK, Kwon KS, Garg S, *et al*. The association of retinopathy and low GFR in type 2 diabetes. *Diabetes Res Clin Pract* 2012;98:487–93.
37 Kramer CK, Retnakaran R. Concordance of retinopathy and nephropathy over time in Type 1 diabetes: an analysis of data from the Diabetes Control and Complications Trial. *Diabet Med* 2013;30:1333–41.
38 Sheetz MJ, King GL. Molecular understanding of hyperglycemia's adverse effects for diabetic complications. *JAMA* 2002;288:2579–88.
39 Keir LS, Firth R, Aponik L, *et al*. VEGF regulates local inhibitory complement proteins in the eye and kidney. *J Clin Invest* 2017;127:199–214.
40 Teng J, Dwyer KM, Hill P, *et al*. Spectrum of renal disease in diabetes. *Nephrology* 2014;19:528–36.
41 Tan J, Zwi LJ, Collins JF, *et al*. Presentation, pathology and prognosis of renal disease in type 2 diabetes. *BMJ Open Diabetes Res Care* 2017;5:e000412.
42 Zhang J, Wang Y, Li L, *et al*. Diabetic retinopathy may predict the renal outcomes of patients with diabetic nephropathy. *Ren Fail* 2018;40:243–51.
43 Klein R, Klein BE, Moss SE, *et al*. The Wisconsin epidemiologic study of diabetic retinopathy. III. prevalence and risk of diabetic retinopathy when age at diagnosis is 30 or more years. *Arch Ophthalmol* 1984;102:527–32.
44 Anwar SB, Asif N, Naqvi SAH, *et al*. Evaluation of multiple risk factors involved in the development of diabetic retinopathy. *Pak J Med Sci* 2019;35:156–60.
45 Buso G, Aboyans V, Mazzolai L. Lower extremity artery disease in patients with type 2 diabetes. *Eur J Prev Cardiol* 2019;26:114–24.
46 Leveziel N, Ragot S, Gand E, *et al*. Association between diabetic macular edema and cardiovascular events in type 2 diabetes patients: a multicenter observational study. *Medicine* 2015;94:e1220.
47 Liao W-L, Lin J-M, Chen W-L, *et al*. Multilocus genetic risk score for diabetic retinopathy in the Han Chinese population of Taiwan. *Sci Rep* 2018;8:14535.
48 Mendoza-Herrera K, Quezada AD, Pedroza-Tobías A, *et al*. A diabetic retinopathy screening tool for low-income adults in Mexico. *Prev Chronic Dis* 2017;14:E95.