Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Prediction and analysis of risk factors for diabetic retinopathy based on machine learning and interpretable models

Xu Wang, Weijie Wang, Huiling Ren *, Xiaoying Li, Yili Wen

Institute of Medical Information/Medical Library, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China

ARTICLE INFO

Keywords:
Diabetic retinopathy
Machine learning
CatBoost
SHAP
Prediction model

ABSTRACT

Objective: Diabetic retinopathy is one of the major complications of diabetes. In this study, a diabetic retinopathy risk prediction model integrating machine learning models and SHAP was established to increase the accuracy of risk prediction for diabetic retinopathy, explain the rationality of the findings from model prediction and improve the reliability of prediction results.
Methods: Data were preprocessed for missing values and outliers, features selected through information gain, a diabetic retinopathy risk prediction model established using the CatBoost and the outputs of the mode interpreted using the SHAP model.
Results: One thousand early warning data of diabetes complications derived from diabetes complication early warning dataset from the National Clinical Medical Sciences Data Center were used in this study. The CatBoost-based model for diabetic retinopathy prediction performed the best in the comparative model test. ALB_CR, HbA_{1c}, UPR_24, NEPHROPATHY and SCR were positively correlated with diabetic retinopathy, while CP, HB, ALB, DBILI and CRP were negatively correlated with diabetic retinopathy. The relationships between HEIGHT, WEIGHT and ESR characteristics and diabetic retinopathy were not significant.
Conclusion: The risk factors for diabetic retinopathy include poor renal function, elevated blood glucose level, liver disease, hematonosis and dysarteriotony, among others. Diabetic retinopathy can be prevented by monitoring and effectively controlling relevant indices. In this study, the influence relationships between the features were also analyzed to further explore the potential factors of diabetic retinopathy, which can provide new methods and new ideas for the early prevention and clinical diagnosis of subsequent diabetic retinopathy.

1. Introduction

According to the most recent data from the International Diabetes Federation, 537 million individuals (aged 20–79 years old) suffered from diabetes by 2021 worldwide. It is predicted that the number will increase to approximately 643 million by 2030. China has the largest number of diabetic patients in the world, and there has been over 141 million diabetic patients in China (aged 20–79 years old) by 2021 [1]. A significant complication of diabetes mellitus is diabetic retinopathy (DR), which affects approximately one third of diabetic patients to some extent and results in visual impairment in 10 % of the patients with DR. Diabetic retinopathy is irreversible. Hyperglycemia can induce the activation of retinal glial cells, the loss of neuronal apoptosis and retinal atrophy, resulting

* Corresponding author. Institute of Medical Information/Medical Library, Chinese Academy of Medical Sciences & Peking Union Medical College, No.69, Dongdan North Street, Beijing, China.
E-mail address: ren.huiling@imicams.ac.cn (H. Ren).

https://doi.org/10.1016/j.heliyon.2024.e29497
Received 24 January 2024; Received in revised form 9 April 2024; Accepted 9 April 2024
Available online 10 April 2024

in retinal damage in patients [2]. Early screening and diagnosis can not only prevent the damage associated with retinopathy, but also improve the level of comprehensive diabetes care [3].

Due to the large sample size, many influential factors, information loss or redundancy of diabetes patients and other imbalanced samples, a simple statistical analysis will lead to underfitting or overfitting of disease prediction results and loss of accuracy [3,4]. Thus machine learning has been widely used in the risk prediction for diabetes mellitus and its complications [5–7]. There are still major obstacles in the application of machine learning-based models in clinical disease prediction. However, most of the current machine learning models used for disease prediction are black box models, resulting in the lack of intuitive and credible interpretation of most models, and there are still major obstacles in the application of machine learning models to clinical disease prediction. In the previous studies, researchers devoted to improving the predictive performance of the model, but paid less attention to the analysis of the model, and then ignored disease causes and the association between disease risks, so the research was limited to the technical surface.

Therefore, the main objective of this study was to improve the accuracy and interpretability of prediction results for diabetic retinopathy through machine learning models. In addition, the article also reveals the main and potential risk factors and associations between factors of diabetic retinopathy, which can provide references for machine learning-assisted screening of diabetic complications, risk assessment, early diagnosis, and clinical intervention. With this goal, we conducted a lot of experiments and data analysis to build a diabetic retinopathy risk prediction model based on CatBoost, which has better performance than the machine learning model. SHapley Additive exPlanations (SHAP) was applied to interpret and analyze the research results to overcome the shortcomings of the black box model. In the second part, we expounded the application of machine learning in diabetes prediction and the basic principle and application of SHAP interpretation model. In the third part, we introduced the dataset and research methods used in the study. In the fourth part and the fifth part, we showed all the experimental results, explained and discussed risk factors for diabetic retinopathy in the context of past research and clinical findings.

2. Related studies

2.1. Machine learning-based diabetes risk prediction

Within the healthcare area, machine learning is frequently used. Clinical professionals may receive further support in making diagnostic decisions in addition to key features from medical data being identified. This study examined the advances in machine learning for diabetes prediction. Several researchers employed various machine learning algorithms to establish prediction models and obtain certain outcomes. Ensemble learning and non-ensemble learning are currently the two modes of machine learning available for prediction of diabetes risks.

2.1.1. Non-ensemble learning

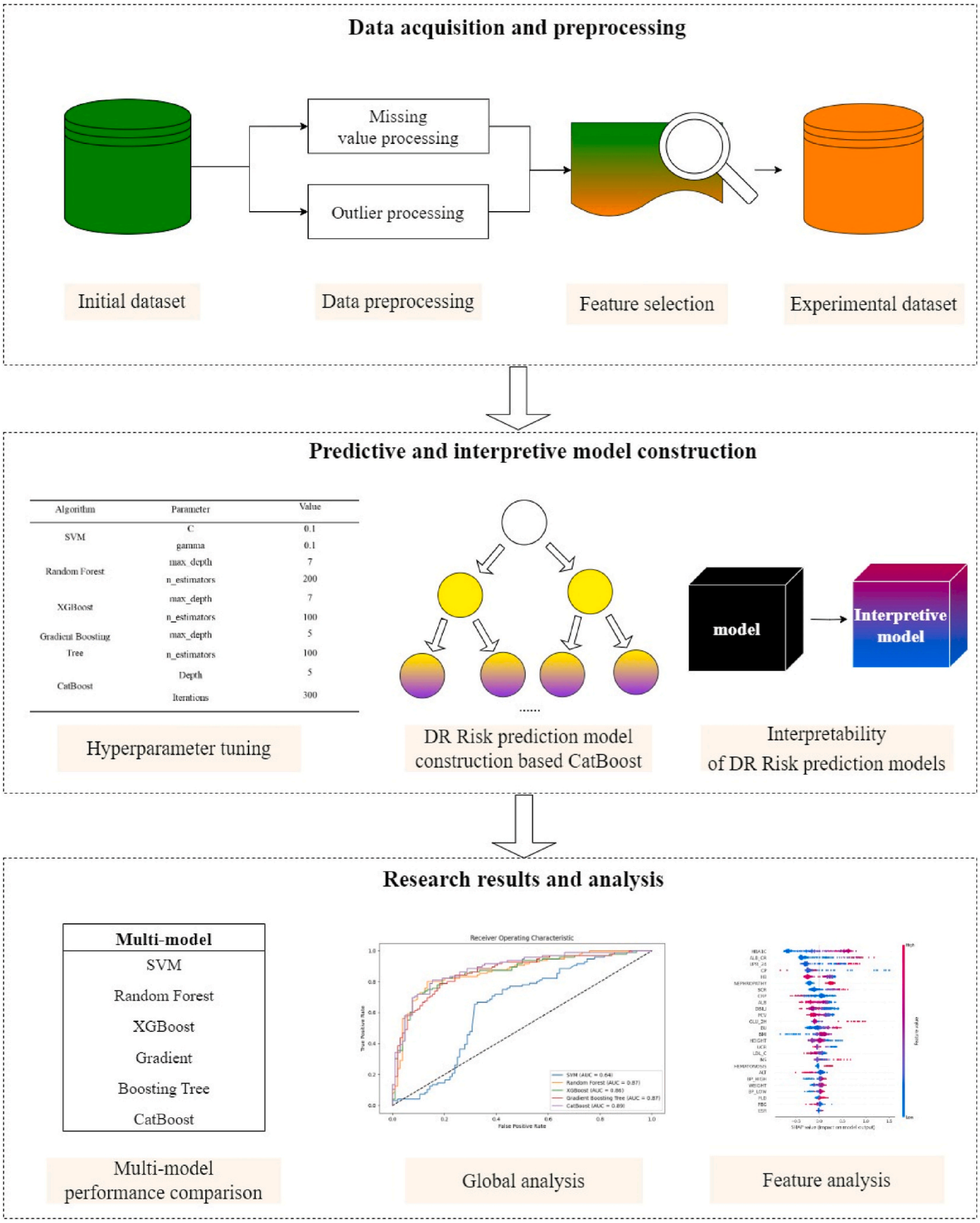
When it comes to non-ensemble learning, researchers typically employ the artificial neural network (ANN) and support vector machine (SVM) to predict illness.

One supervised learning algorithm is SVM. For the purpose of data categorization, it divides the current dataset by the decision hyperplane and determines which side of the new dataset is on the hyperplane. SVM is extensively applied in a variety of domains, including economic decision-making and disease diagnosis. In predicting the presence of diabetes, Jegan conducted follow-up investigations after utilizing the support vector machine as a classifier for diabetes diagnosis [8]. Tan et al. developed an insulin evaluation model using the support vector machine technique to visually assess patients' insulin and provide guidance for the prevention and management of diabetes [9]. Additionally, Zheng et al. estimated the risk of gestational diabetes using the support vector machine to serve as a guide for routine prenatal examinations in the early stages of pregnancy [10].

ANN is a research hotspot following the rise of artificial intelligence, for which the machine learning-based model is established by simulating human neuron network. Li et al. employed the artificial neural network technique to build a diabetic complication risk prediction model, with which the accuracy of prediction of different complications ranged from 64.71 % to 82.35 % [7]. Hou et al. developed a model for predicting diabetes risk factors based on the decision tree, logistic regression, and neural network models. They found that the neural network model was superior to the decision tree and logistic regression models [6]. Ashiquzzaman et al. established a diabetes prediction model based on the deep neural network, and its accuracy was up to 88.41 % [5]. With the continuous development of deep learning, ANN has ushered in a new research boom. Multilayer perceptron (MLP), recurrent neural network (RNN) and convolutional neural network (CNN) are commonly used neural networks nowadays. Hu et al. built a MLP to predict diabetes, and the findings from their study demonstrated the high accuracy and strong generalization capacity of the MLP in the early detection of diabetes [11]. Zhu et al. predicted the hospitalization of patients with respiratory system problems and diabetes mellitus using a short-term memory recurrent neural network. They confirmed that the prediction results were correct, and this might help choose when to and when not to utilize medical resources during polluted weather [12]. R. Yasashvini et al. classified diabetic retinopathy using the CNN and hybrid deep convolutional neural network. They also established an improved hybrid CNN model with DenseNet fusion to allow a more effective automated diagnosis of diabetic retinopathy [13].

2.1.2. Ensemble learning

Learning based on a single model is referred to as non-ensemble learning. Its operation is quick and easy, but ensemble learning usually outperforms it in terms of the prediction accuracy and performance. The goal of ensemble learning is to improve computing performance by integrating many models according to the predetermined criteria. Prediction using a single model is no longer adequate due to the technical advancement and maturity. In order to improve the model performance and the prediction accuracy,



Algorithm	Parameter	Value
SVM	C	0.1
	gamma	0.1
Random Forest	max_depth	7
	n_estimators	200
XGBoost	max_depth	7
	n_estimators	100
Gradient Boosting Tree	max_depth	5
	n_estimators	100
CatBoost	Depth	5
	Iterations	300

Fig. 1. Risk prediction for diabetic retinopathy based on a combination of CatBoost and SHAP.

there is now a push to enhance deep learning and introduce multi-model fusion. Bagging, stacking, and boosting are examples of popular ensemble learning techniques employed in the field of disease prediction.

One of the most common bagging techniques is random tree (RF), based on which many decision trees are established to learn and predict data samples. In prior studies, Su et al. analyzed the risk factors of type 2 diabetes mellitus based on least absolute shrinkage and selection operator (LASSO) regression and the random forest algorithm [14]. Stacking is an integrated model that stacks multiple single models. It is usually composed of two layers, of which the first layer is the base model, and the second layer is a new model output after the training of the first-layer model. Shen et al. built a combination model based on SVM, RF, and logistic regression using the stacking approach. According to investigation, the combination model operated better in prediction than the single model, and could be used to diagnose and predict diabetic retinopathy [15]. With the purpose of estimating the incidence of type 2 diabetes, Singh et al. developed an improved ensemble learning system called “NSGA–II–Stacking” based on stacking, which surpassed conventional

Table 1
Continuous variable description.

variable name	variable meaning	sample mean	variable name	variable meaning	sample mean
AGE	age	57.36	PLT	platelet	215.28
HEIGHT	height	167.07	ESR	erythrocyte sedimentation rate	25.45
WEIGHT	weight	73.79	TBILI	total bilirubin	10.34
BP_HIGH	systolic pressure	139.05	DBILI	direct bilirubin	3.08
BP_LOW	diastolic pressure	80.83	TP	total protein	65.23
HEART_RATE	heart rate	80.96	ALB	serum albumin	39.52
BMI	BMI	26.33	LDH_L	lactic dehydrogenase	173.72
GLU	Fasting blood glucose	8.58	ALT	glutamic-pyruvic transaminase	25.43
GLU_2H	Blood sugar 2 h after a meal	14.89	AST	glutamic oxalacetic transaminase	21.32
HBA1C	glycated hemoglobin	7.78	GGT	transglutaminase	43.34
GSP	glycated serum protein	225.70	ALP	alkaline phosphatase	74.00
TG	triglyceride	2.07	LP_A	lipoprotein	35.03
TC	total cholesterol	4.67	PL	phospholipid	2.52
HDL_C	high density lipoprotein cholesterol	1.07	PT	prothrombin time	13.07
LDL_C	low density lipoprotein cholesterin	2.90	PTA	prothrombin activity	99.70
FBG	fibrinogen	7.84	APTT	partial activation of prothrombin time	36.51
UPR_24	24 h urinary trace protein	1.39	FIBRIN	fibrous protein	10.54
BUN	blood urea nitrogen	0	ALB_CR	rapid trace urine protein/creatinine assay	155.16
BU	blood urea	7.20	LPS	Serum lipase	160.91
SCR	serum creatinine	106.76	CA199	Tumor marker CA199	21.80
UCR	urine creatinine	5.73	CRP	C-reactive protein	1.18
SUA	serum uric acid	332.31	M1_M2	macrophage	36
HB	hemoglobin	132.81	TH2	T helper cell TH2	0.45
CP	fasting c-peptide	2.33	IBILI	indirect bilirubin	7.25
INS	fasting insulin	16.44	GLO	globulin	25.70
PCV	packed cell volume	0.39			

Table 2
Discrete variable description.

variable name	variable meaning	value	variable name	variable meaning	value
label	type of disease	0- diabetes; 1- diabetic retinopathy	MI	myocardial infarction	0-No; 1-Yes
SEX	gender	0- male; 1- female	CHF	Cardiac insufficiency and heart failure	0-No; 1-Yes
NATION	nation	0- Han nationality; 1- other nationalities	ARRHYTHMIAS	arrhythmology	0-No; 1-Yes
MARITAL_STATUS	marital status	0- single; 1- married	RESPIRATORY_SYSTEM_DISEASE	respiratory disease	0-No; 1-Yes
HYPERTENTION	hypertension	0-No; 1-Yes	LEADDP	lower extremity arterial disease	0-No; 1-Yes
HYPERLIPIDEMIA	hyperlipemia	0-No; 1-Yes	HEMATONOSIS	hematonosis	0-No; 1-Yes
A_S	atherosclerosis	0-No; 1-Yes	RHEUMATIC_IMMUNITY	rheumatic immune disease	0-No; 1-Yes
CEREBRAL_APOPLEXTY	cerebral apoplexy	0-No; 1-Yes	PREGNANT	pregnancy	0-No; 1-Yes
CAROTID_ARTERY_STENOSIS	carotid artery stenosis	0-No; 1-Yes	ENDOCRINE_DISEASE	endocrine diseases	0-No; 1-Yes
FLD	fatty liver	0-No; 1-Yes	MEN	endocrine adenoma	0-No; 1-Yes
CIRRHOSIS	liver cirrhosis	0-No; 1-Yes	PCOS	polycystic ovarian syndrome	0-No; 1-Yes
CLD	chronic liver diseases	0-No; 1-Yes	DIGESTIVE_CARCINOMA	digestive carcinoma	0-No; 1-Yes
PANCREATIC_DISEASE	exocrine diseases of the pancreas	0-No; 1-Yes	UROLOGIC_NEOPLASMS	urologic neoplasms	0-No; 1-Yes
BILIARY_TRACT_DISEASE	biliary tract disease	0-No; 1-Yes	GYNECOLGICAL_TUMOR	gynecological oncology	0-No; 1-Yes
NEPHROPATHY	nephropathy	0-No; 1-Yes	BREAST_TUMOR	breast tumor	0-No; 1-Yes
RENAL_FALIURE	renal failure	0-No; 1-Yes	LUNG_TUMOR	lung tumor	0-No; 1-Yes
NERVOUS_SYSTEM_DISEASE	nervous system disease	0-No; 1-Yes	INTRACRANIAL_TUMOR	intracranial tumours	0-No; 1-Yes
CHD	coronary heart disease	0-No; 1-Yes	OTHER_TUMOR	other tumor	0-No; 1-Yes

Table 3
Optimal optimum parameters.

Algorithm	Parameter	Value
SVM	C	0.1
	gamma	0.1
RF	max_depth	7
	n_estimators	200
XGBoost	max_depth	7
	n_estimators	100
GBT	max_depth	5
	n_estimators	100
CatBoost	Depth	5
	Iterations	300

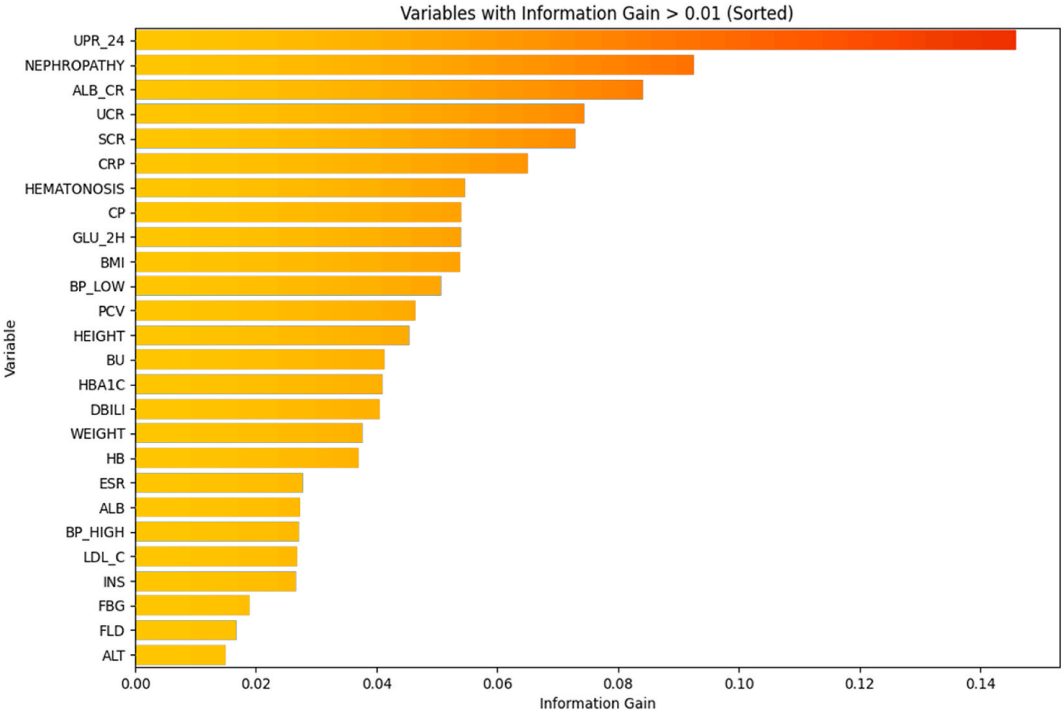


Fig. 2. Feature correlation ranking.

Table 4
Comparison of the performance results among models.

Algorithm	Accuracy	Precision	Recall	F_1 value
SVM	0.7850	0.8046	0.7292	0.7650
RF	0.8100	0.8352	0.7917	0.7935
XGBoost	0.8000	0.8045	0.7292	0.7989
GBT	0.7700	0.7692	0.7292	0.7553
CatBoost	0.8250	0.8211	0.8125	0.8168

machine learning techniques [16]. Boosting, as the most widely used ensemble learning algorithm, works by generating multiple weak learners and combining them to form a new model. Dong et al. employed seven machine learning methods to forecast performance, among which the gradient boosting tree (GBT) performed the best, and SHAP was then used to interpret the outputs from such model [17]. Liu et al. proposed a diabetes risk prediction and characteristic analysis model based on eXtreme gradient boosting (XGBoost), and its accuracy was up to 96.85 % after parameter tuning [18]. Miao et al. used the CatBoost algorithm to predict diabetes, and obtained better performance results [19]. Rufo et al. developed a diabetes diagnostic model based on the light gradient boosting machine (LightGBM). They compared it to six existing machine algorithms and found that the newly developed model did the best, with an accuracy of 98.1 % [20].

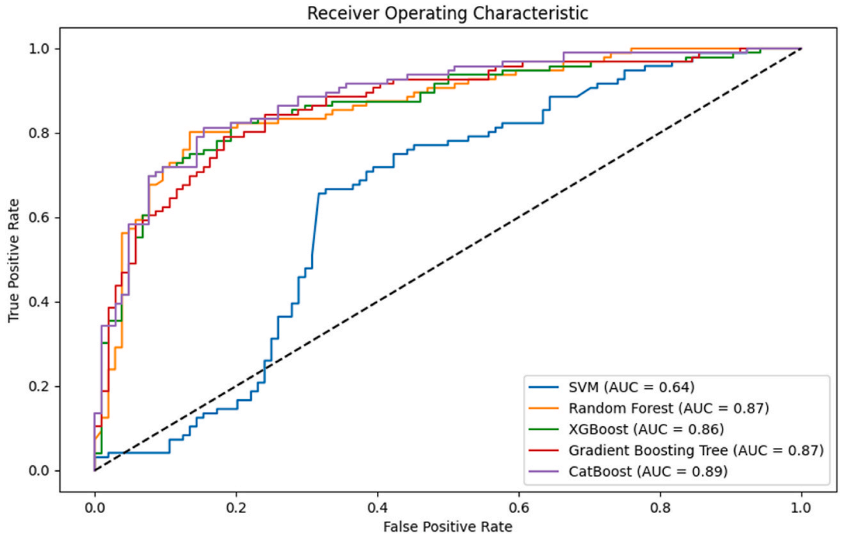


Fig. 3. Comparison of ROCs.

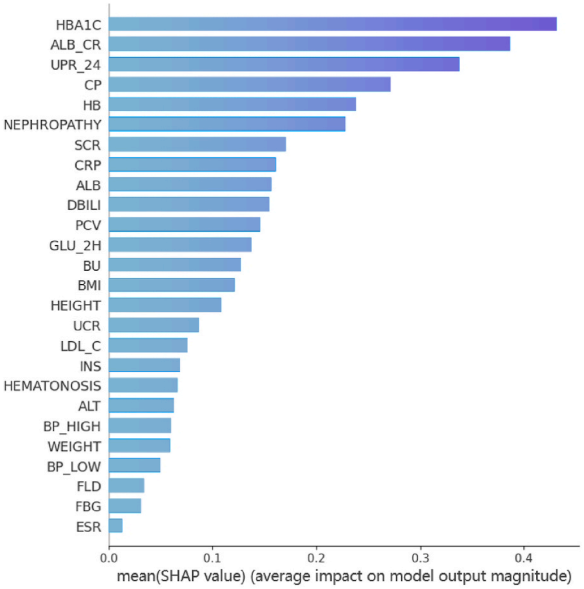


Fig. 4. Feature importance ranking.

2.2. SHAP-based explanatory algorithm

Lundberg and Lee first proposed the SHAP algorithm in 2017 [21]. The concept of SHAP was derived from the game theory and mostly pertained to the Shapley value computation technique. For SHAP, the subset of the combination of features is mainly selected, and the marginal contribution is calculated. The process is to first calculate the marginal contribution of a feature in a certain model, then calculate the different marginal contributions of the feature in all feature sequences, and finally obtain the SHAP value, i.e., the mean value of the overall marginal contribution. SHAP can be used to interpret the prediction using a machine learning-based model, namely to determine the reason behind and extent to which the model influences the prediction of the sample. The SHAP-based explanatory algorithm is widely used with machine learning algorithms to make up for the weak interpretation for the black box model. It has been widely applied in the fields of user behavior analysis, financial risk prediction, intelligent medical diagnosis, etc.

For diabetes risk prediction, SHAP is usually based on a machine learning model to further interpret the findings. Wang et al. developed a diabetes prediction model based on LightGBM and presented the SHAP-based algorithm so as to experimentally compare it with other machine learning models based on SVM, RF, decision tree, XGBoost. The model developed outperformed those models by a large margin on the basis of the findings of the five performance metrics that were examined [22]. Prendin et al. found that the

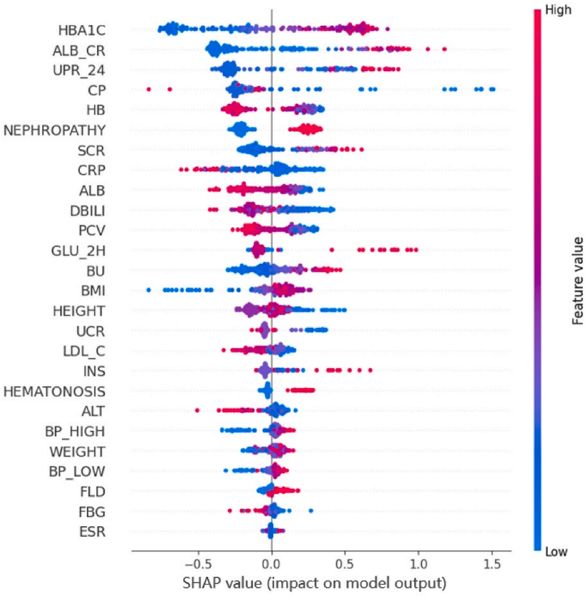


Fig. 5. Summary chart of features.

Table 5
Feature classification description.

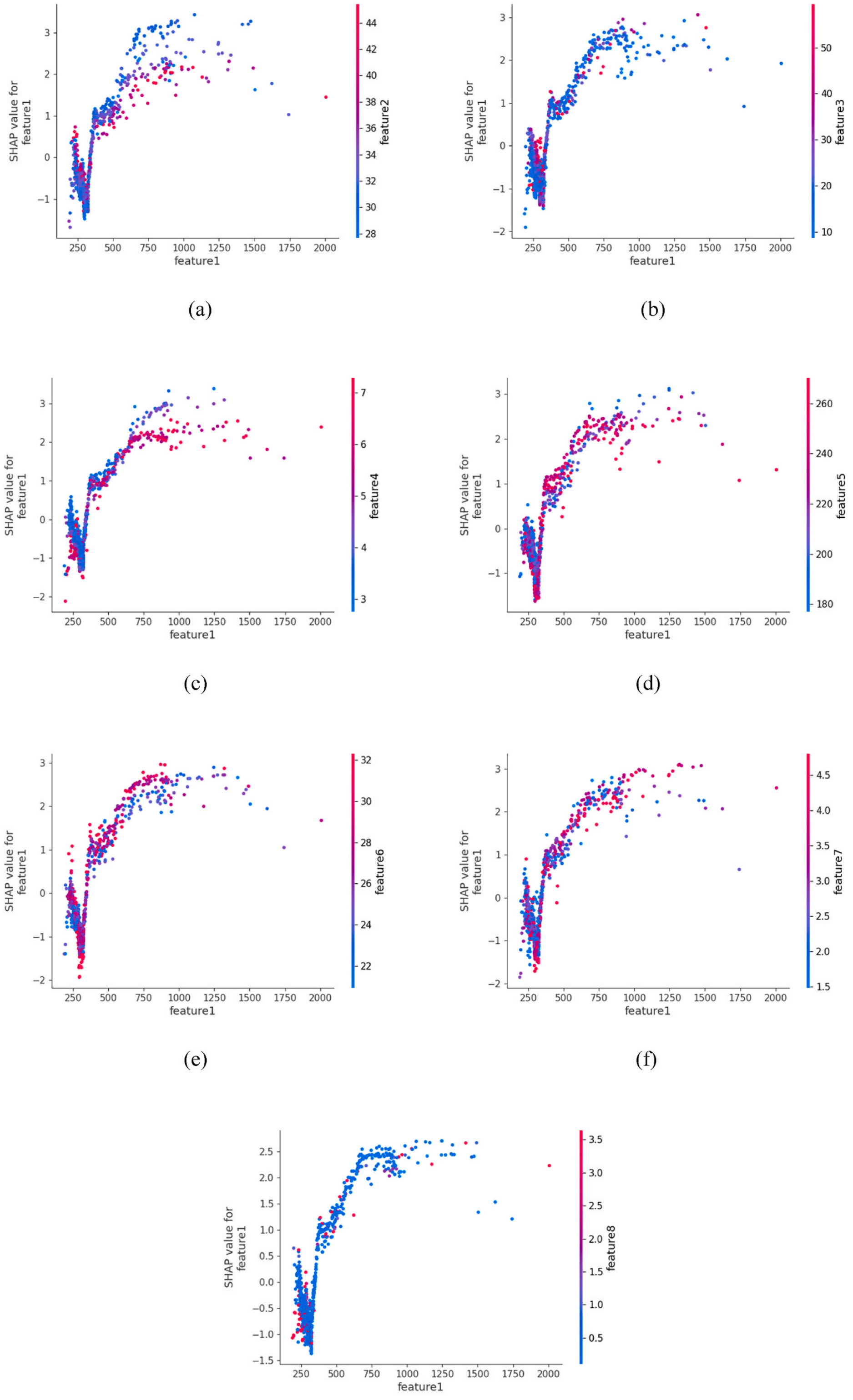
No.	Classification	Feature name	Feature description
1	Renal function indices	ALB_CR	Rapid trace urine protein/creatinine assay
		UPR_24	24 h urinary microprotein
		NEPHROPATHY	Nephropathy
		SCR	Serum creatinine
		HB	Hemoglobin
		ALB	Serum albumin (35–50 g/L)
		DBILI	Direct bilirubin (0-8.6 μmol/L)
		BU	Blood urea nitrogen
		UCR	Urine creatinine
2	Glycemic indices	HBA1C	Glycosylated hemoglobin
		CP	Fasting C-peptide
		GLU_2H	Blood glucose 2 h after a meal
		INS	Fasting insulin
3	Liver function indices	ALT	Alanine aminotransferase
		FLD	Fatty liver
4	Blood indices	HEMATONOSIS	Hematonosis
		FBG	Fibrinogen
		PCV	Hematocrit
5	Blood pressure	BP_HIGH	Systolic blood pressure
		BP_LOW	Diastolic blood pressure
6	Other indices	BMI	Body mass index
7		LDL_C	Low-density lipoprotein cholesterol (0–3.4 mmol/L)
8		CRP	C-reactive protein

prediction accuracy of neural networks with long- and short-term memory was comparable. They employed SHAP to examine the predictions of blood glucose under diabetes condition using two samples in order to further clarify the prediction effects [23]. Jin et al. established a diabetic retinopathy detection model based on an interpretable machine learning algorithm in an effort to further enhance the interpretability of the machine learning-based model, i.e., the black box model. They applied SHAP to provide individual, feature, and global explanations for lesions [24].

3. Materials and methods

3.1. Overview of methods

Our main work is to build a risk prediction model for diabetic retinopathy and conduct interpretative analysis. The implementation



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(caption on next page)

Fig. 6. Correlations between renal function indices and other measures. (a) Relationship between feature 1 and feature 2 based on SHAP measurement. (b) Relationship between feature 1 and feature 3 based on SHAP measurement. (c) Relationship between feature 1 and feature 4 based on SHAP measurement. (d) Relationship between feature 1 and feature 5 based on SHAP measurement. (e) Relationship between feature 1 and feature 6 based on SHAP measurement. (f) Relationship between feature 1 and feature 7 based on SHAP measurement. (g) Relationship between feature 1 and feature 8 based on SHAP measurement.

process can be summarized as follows:

- Data were preprocessed to enhance their quality, which improve the performance of the diagnostic process. We use quartile method to supplement missing values for missing values and Z-scores method to identify outliers.
- We tested and processed the normality of the dataset, and took the information gain method to select the features.
- We established a CatBoost-based risk prediction model for diabetic retinopathy, and employed the grid search approach to tune its parameters for better performance. Moreover, it was compared with SVM, RF, XGBoost and GBT in terms of the prediction performance.
- We also analyzed the direct and possible causes, as well as the risk factors of diabetic retinopathy and their interactions from the point of view of global features and inter-feature correlations, using SHAP to conduct an explanatory analysis of the expected results.

The research technology route is shown in Fig. 1.

3.2. Dataset description

The data in this study, which included 500 cases of diabetes and 500 cases of diabetes complicated by retinopathy, were derived from early warning data collection of diabetic complications from the National Clinical Medical Sciences Data Center [25]. It also includes 87 variables, of which 36 are discrete and 51 are continuous. The variable descriptions are shown in Table 1 and Table 2.

3.3. Data preprocessing

3.3.1. Missing value processing

During data collection, missing data are common, which have a significant impact on the final results of prediction. Filling, removal, and direct use are typically the options available for the management of missing values. Too many missing values will affect the distribution characteristics of the data. To avoid affecting the results, this study will delete the features with missing values greater than or equal to 90 %. For features with less than 90 % missing values, we decided to use the quartile method to fill in missing values due to the different types of variables of dataset. The quartile method fills in the missing values through the quantile, which can retain the distribution characteristics. Meanwhile, the quartile is a statistic with strong robustness, and the influence of outliers on the filling results can be reduced after filling which makes the data more stable. The operation method is as follows: Determine the lower quartile (Q1) and upper quartile (Q3) of the features, and calculate the quartile distance (IQR) of the features, i.e., Q3-Q1. Depending on the size of the quartile distance, we can create a random integer between Q1 and Q3 to fill in the missing value [26].

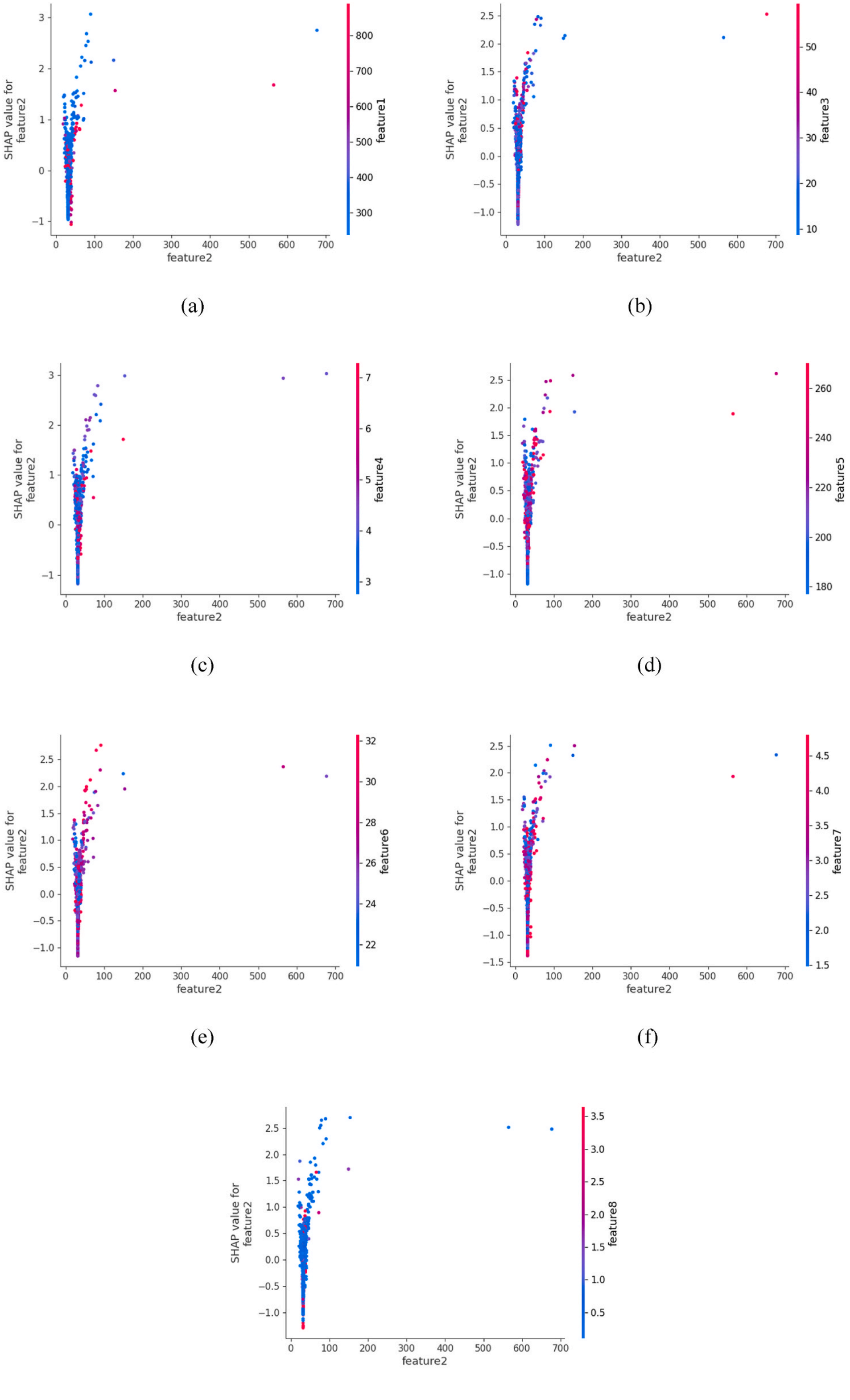
3.3.2. Outlier processing

The term “outlier” describes a value that deviates from the majority of observed values for a sample. Outlier identification, screening, and processing are three different approaches for outlier handling. Some outliers during data preparation may include important information, and whether or not to eliminate them are determined as the case may be. We use the Z-score method to identify and process outliers, which is a common method for outlier processing. The operation method is to calculate the distance between each data and the mean value and compare it with the 3 times standard deviation of the data. Data exceeding 3 times standard deviation will be identified as outliers [27]. The number of outliers in the dataset of this study is small. Through manual observation, some outliers may be abnormal indicators caused by other diseases of the patients, which can be retained. The remaining outliers deviated from the facts were deleted and filled with quartiles.

3.4. Feature selection

The feature selection procedure that is frequently applied in the decision tree indicates the extent to which information complexity or uncertainty is reduced under a given set of conditions. The degree of reduction in information uncertainty increases with the amount of information gain. The selection of discrete and continuous features may be done using information gain, which is computed without assuming the value of features. The calculation method is as follows:

First, the information entropy of the target variable in the dataset is calculated to represent the degree of uncertainty [28]. For the target variable X, we assume that all possible values are x_1, x_2, …, x_k, the corresponding probability of occurrence is $p(x_1)$, $p(x_2)$, …, $p(x_k)$. The calculation formula of the information entropy $H(X)$ of the target variable X is shown in (1).


(a)

(b)

(c)

(d)

(e)

(f)

(g)

(caption on next page)

Fig. 7. Correlations between blood glucose and other indices. (a) Relationship between feature 2 and feature 1 based on SHAP measurement. (b) Relationship between feature 2 and feature 3 based on SHAP measurement. (c) Relationship between feature 2 and feature 4 based on SHAP measurement. (d) Relationship between feature 2 and feature 5 based on SHAP measurement. (e) Relationship between feature 2 and feature 6 based on SHAP measurement. (f) Relationship between feature 2 and feature 7 based on SHAP measurement. (g) Relationship between feature 2 and feature 8 based on SHAP measurement.

$$H(X) = -\sum_{i=1}^{k} p(x_i) log_2\, p(x_i) \tag{1}$$

Second, the characteristic variable Y is assumed to have all conceivable values of y_1, y_2, ..., y_n. The associated probability of occurrence is $p(y_1)$, $p(y_2)$, ..., $p(y_n)$. The formula for calculating the conditional entropy $H(X|Y)$ is shown in (2).

$$H(X|Y) = \sum_{j=1}^{n} p(y_j) \left(-\sum_{i=1}^{k} p(x_i|y_j) log_2\, p\left(x_i|y_j\right) \right) \tag{2}$$

Finally, the degree to which the feature might reduce the uncertainty of the target variable is described by subtracting the conditional entropy from the information entropy to determine the information gain of the feature to the target variable [29]. The formula of information gain is shown in (3).

$$Information\ Gain = H(X) - H(X|Y) \tag{3}$$

In this study, we selected variables with values greater than 0.01 to reconstruct the experimental dataset during model training and verification, and graded the computed information gain results from high to low.

3.5. Model construction

3.5.1. DR risk prediction model

Yandex enhanced the symmetric tree mechanism method CatBoost based on the Gradient Boosting Decision Tree (GBDT) architecture. This method is to train several decision trees iteratively so as to build an ensemble model [30]. CatBoost automatically processes class features, reduces the probability of overfitting and automatically modifies parameters during training to optimize the model performance. The CatBoost model is rarely used in the studies that are currently available to predict the risk of diabetes complicated by retinopathy. In this study, a risk prediction model for diabetes complicated by retinopathy was constructed using CatBoost.

First, the categorical features in the dataset are processed and converted into numerical features. Compared with GBDT, CatBoost introduces the prior value, which is the pre-estimation of the target variable. Using the prior value as the initial predicted value, the model can not only find the approximate solution faster to accelerate the training process of the model, but also prevent the overfitting phenomenon caused by too few samples. In CatBoost, the prior value is obtained by calculating the average of the target variables in the dataset. In the dataset $\mathrm{T} = \{(X_1, Y_1)\ ,\ (X_2, Y_2)\ ,\ \ldots\ ,\ (X_n, Y_n)\}, X_1 = (x_{i1}, x_{i2}, \ldots, x_{im})$ is the m-dimensional variable, Y_i is the marker value, q is the prior value, $\alpha(\alpha > 0)$ is the weight coefficient of the prior value. Assuming that $\mu = (\mu_1, \mu_1, \ldots \mu_1)$ is a random ordering sequence, the categorical features can be converted to numerical values according to formula (4).

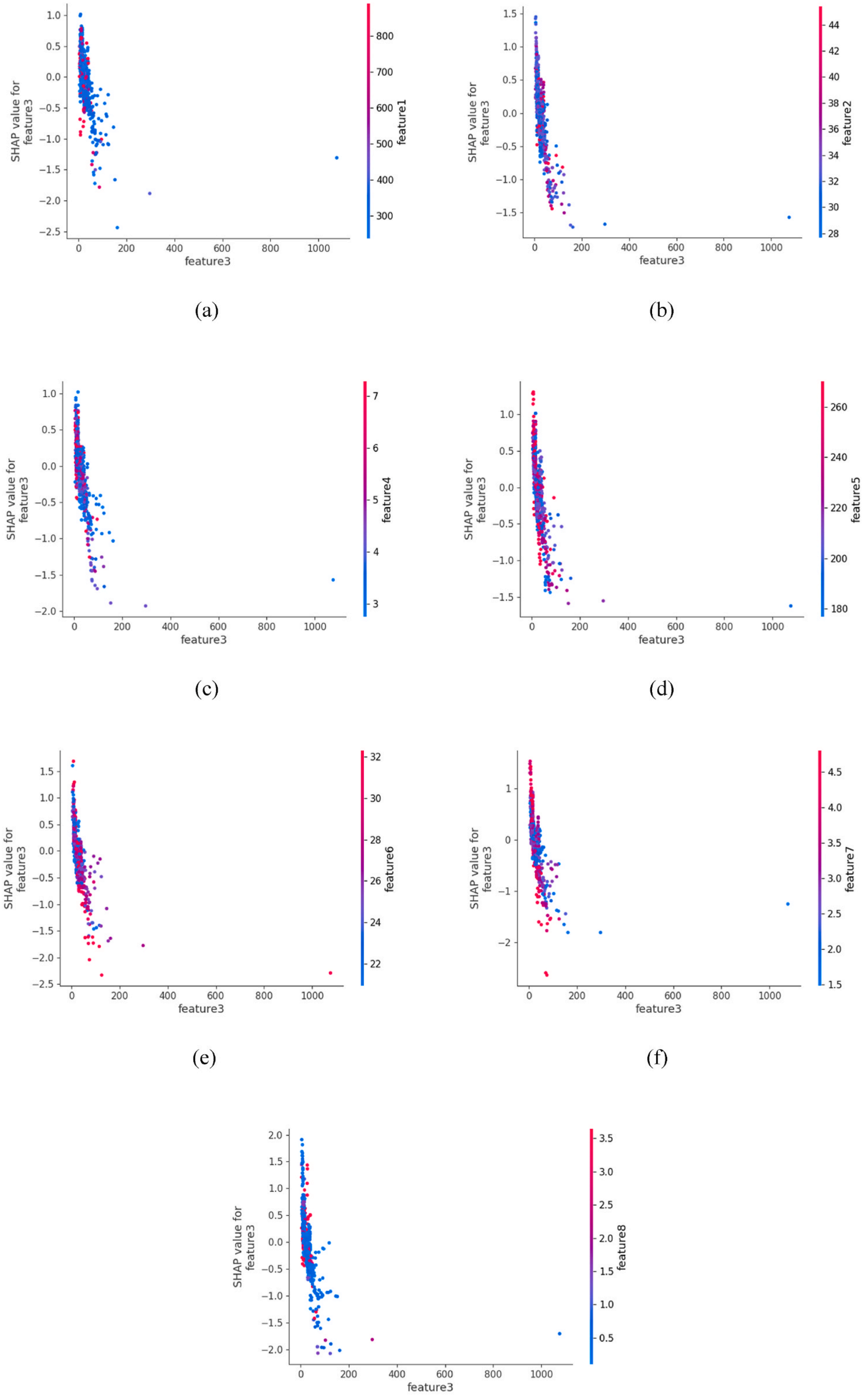
$$X_{\mu q,k} = \frac{\sum_{j=1}^{q-1} \left[X_{\mu j,k} = X_{\mu q,k}\right] * Y_{\mu j} + \alpha * q}{\sum_{j=1}^{q-1} \left[X_{\mu j,k} = X_{\mu q,k}\right] + \alpha} \tag{4}$$

Then, multiple rounds of iteration train the decision tree. Each tree is trained on the residual of the previous round of trees, and the greedy algorithm selects the best split point to minimize the loss function. CatBoost can stop processes early by monitoring loss function values on validation sets to prevent overfitting and improve model generalization. Finally, the predicted value is obtained by combining the predicted results of all decision trees [31,32].

3.5.2. Interpretability of the DR risk prediction model

As the black box mechanism of CatBoost may lead to issues with poor interpretation, we employ the SHAP for attribution analysis to enhance the interpretability and reliability of model prediction findings [33]. SHAP determines the contribution degree of each feature quantitatively, and can further explore the relationship between features, which greatly improves the interpretability of machine learning model.

We assume that F is the set of all features, S is any subset of F, x_s is the feature of the set S, $|F|$ is the total number of all features, and $|S|$ is the size of the set. In order to calculate the contribution degree of a single feature, a model $f_{S\cup\{i\}}$ containing feature i and a model f_S without feature i are trained respectively, and the prediction results between them are compared, namely, $f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_s)$, which means that the shap values of feature i are calculated only with a subset of other features except feature i. Since the prediction results are affected by other features, the difference between all possible subsets $S \subseteq \mathrm{F} \setminus \{\mathrm{i}\}$ is also calculated [34]. The calculation formula of shap values Φ_i of feature i is shown in (5).



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(caption on next page)

Fig. 8. Correlations between liver function indices and other indices. (a) Relationship between feature 3 and feature 1 based on SHAP measurement. (b) Relationship between feature 3 and feature 2 based on SHAP measurement. (c) Relationship between feature 3 and feature 4 based on SHAP measurement. (d) Relationship between feature 3 and feature 5 based on SHAP measurement. (e) Relationship between feature 3 and feature 6 based on SHAP measurement. (f) Relationship between feature 3 and feature 7 based on SHAP measurement. (g) Relationship between feature 3 and feature 8 based on SHAP measurement.

$$\Phi_i = \sum_{S\subseteq F\setminus\{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!}\left[f_{S\cup\{i\}}\left(x_{S\cup\{i\}}\right)-f_S(x_s)\right] \tag{5}$$

3.6. Model evaluation index

The following metrics were selected for this study: F_1 value, area under the receiver operating characteristic curve (ROC), accuracy, precision and recall. The area under the ROC curve was a metric to assess the wind prediction model's performance for diabetic retinopathy.

Accuracy is defined as the percentage of the total number of correct predictions. The formula for its calculation is given in (6).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \tag{6}$$

The term "accuracy rate" describes how well the model predicts true positive class samples, or more specifically, how many true positive class samples the model can accurately predict. The formula for its calculation is given in (7).

$$Precision = \frac{TP}{TP+FP} \times 100\% \tag{7}$$

Recall rate is the percentage of all actual positive class samples that the model correctly predicts as positive class samples. The recall rate quantifies the model's capacity to recognize positive class samples, or the number of actual positive class samples that the model can properly locate. The formula for its calculation is given in (8).

$$Recall = \frac{TP}{TP+FN} \times 100\% \tag{8}$$

The weighted harmonic mean of accuracy rate and recall rate, with 1 denoting the highest effect and 0 the worst, is known as the F_1 value. This value can more accurately reflect the model's accuracy and recall rates and allow a more thorough and efficient evaluation of the model. The formula for its calculation is given in (9).

$$F_1 = \frac{2\times P\times R}{P+R} \tag{9}$$

Area Under Curve (AUC), which is typically represented via the ROC curve, illustrates the model's prediction effect under various classification thresholds and can provide a more thorough indication of the model's predictive capacity. Typically, the AUC value falls between 0.5 and 1. The greater the value, or the area under the curve is, the more effective the model is.

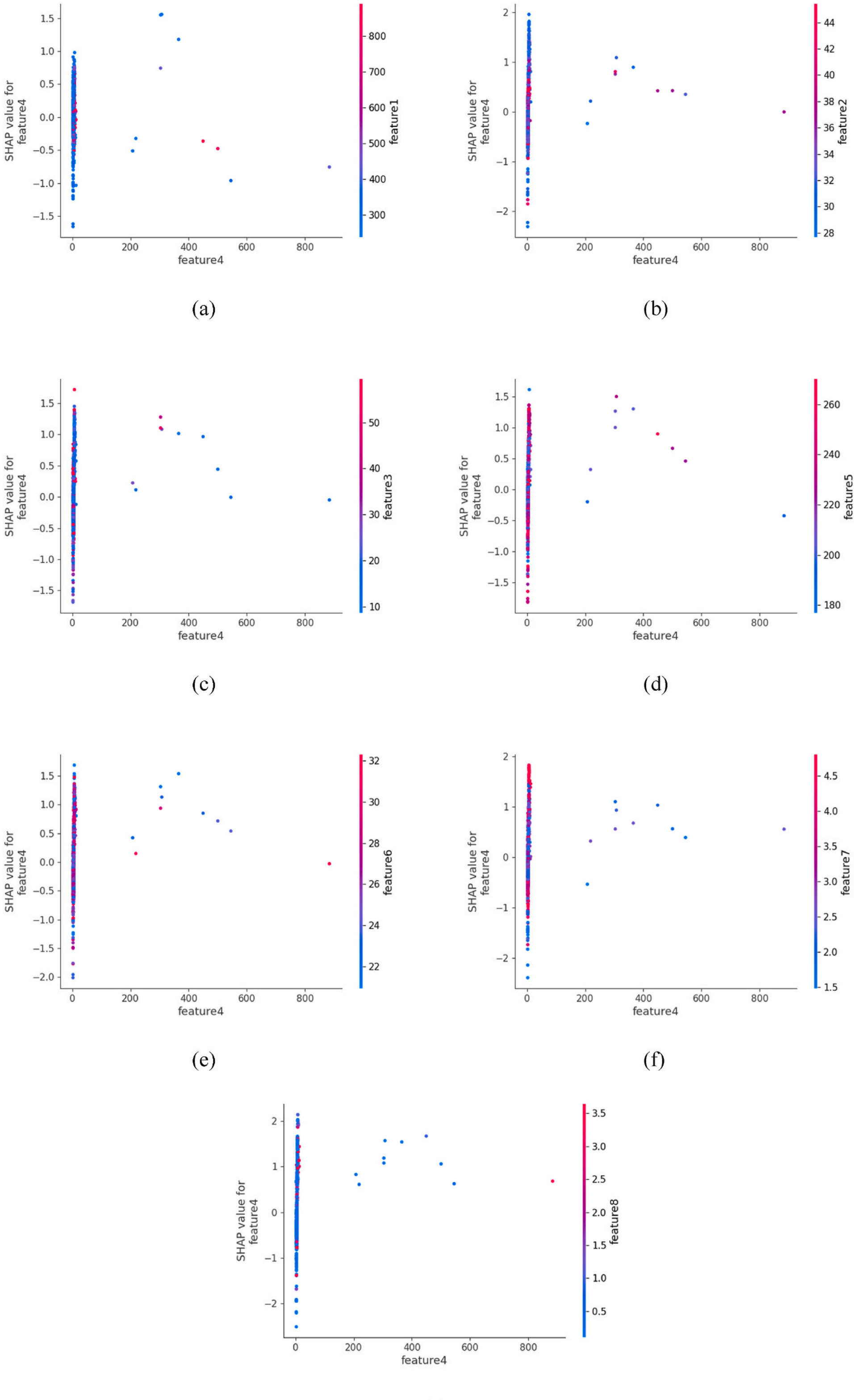
4. Results and analysis

4.1. Dataset partitioning

Using a 4:1 ratio, 1000 diabetic experimental data were randomly jumbled and split into training and test sets. There were 800 pieces of data in the training set to help determine the best model's parameters, and 200 pieces of data in the test set to help with model evaluation and output.

4.2. Hyperparameter tuning

The amount of parameters requiring adjustment is growing with the complexity of the algorithm. Consequently, all possible parameter combinations are matched in this study by employing the grid search algorithm, and the ideal parameters are filtered by applying 50 % cross validation once all combinations have been explored. First, the parameter grid search space is constructed, which contains various hyperparameter candidates for every model. GridSearchCV is a hyperparameter tuning tool for the scikit-learn library in Python that performs cross-validation in a given parameter grid to determine the best combination of parameters. Next, each model and the associated parameter search space are used to generate and store the GridSearchCV object. Fit is then called for model training and hyperparameter search by iterating the GridSearchCV object of each model, and the optimal hyperparameter combination is output for each model. Finally, hyperparameter search is used to find the ideal settings. The optimal parameters in this study are shown in Table 3.



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(caption on next page)

Fig. 9. Correlations between blood indices and other indices. (a) Relationship between feature 4 and feature 1 based on SHAP measurement. (b) Relationship between feature 4 and feature 2 based on SHAP measurement. (c) Relationship between feature 4 and feature 3 based on SHAP measurement. (d) Relationship between feature 4 and feature 5 based on SHAP measurement. (e) Relationship between feature 4 and feature 6 based on SHAP measurement. (f) Relationship between feature 4 and feature 7 based on SHAP measurement. (g) Relationship between feature 4 and feature 8 based on SHAP measurement.

4.3. Results of feature selection

Upon computation, 26 features with information gain greater than 0.01 were selected as shown in Fig. 2. UPR_24 (24-h urinary microprotein) exhibited the highest information gain and the highest correlation with the target variable, which thus had the greatest impact on the model. All the values of NEPHROPATHY (Nephropathy), ALB_CR (Rapid trace urine protein/creatinine assay), UCR (Urine creatinine), SCR (Serum creatinine), CRP (C-reactive protein), HEMATONOSIS (Hematonosis), CP (Fasting C-peptide), GLU_2H (Blood glucose 2 h after a meal), BMI (Body mass index) and BP_LOW (Diastolic blood pressure) were higher than 0.05, which showed a strong correlation with the target variable and thus some influence on the model. PCV (Hematocrit), HEIGHT (Height), BU (Blood urea nitrogen), HbA$_{1c}$ (Glycosylated hemoglobin), DBILI(Direct bilirubin), WEIGHT(Weight), HB (Hemoglobin), ESR (Blood sedimentation), ALB (Serum albumin), BP_HIGH (Systolic blood pressure), LDL_C (Low-density lipoprotein cholesterol), INS (Fasting insulin), FBG (Fibrinogen), FLD (Fatty liver), and ALT (Alanine aminotransferase) all exhibited information gains less than 0.05, which had little effect on the model and little association with the target variable.

4.4. Comparative analysis of model performance

The findings were achieved by applying the ten-fold cross-validation approach. The performance of the CatBoost-based model was compared with that of SVM, RF, XGBoost and GBT, common models for disease prediction in previous studies. The results are shown in Table 4.

As shown in Table 4, CatBoost exhibited the best overall performance, with the accuracy value of 0.8250, the precision value of 0.8211, the recall value of 0.8125 and the F1 value of 0.8168. CatBoost showed better accuracy, recall, and F1 values than other models, and its prediction values were second only to random forest. In a word, the CatBoost model outperformed the other four models in terms of the prediction efficiency. CatBoost was more accurate and more adept at handling sorting issues in classification tasks than SVM, RF, XGBoost and GBT. Furthermore, CatBoost was more resilient and tolerant to issues like noise, allowing it to produce more consistent and trustworthy outcomes.

To facilitate a clearer comparison of each model's predictive ability, ROC curves were used to illustrate the AUC values of five different algorithms.

According to Fig. 3, the AUC values of SVM, RF, XGBoost, GBT and CatBoost were 64 %, 87 %, 86 %, 87 % and 89 %, respectively. CatBoost had the highest AUC value, namely the largest area under the ROC curve. Among the five algorithms, it exhibited the highest prediction accuracy. SVM had an AUC value of 64 % and poor predictive performance, RF, XGBoost, and GBT all had AUC values of more than 80 % and exhibited superior predictive performance.

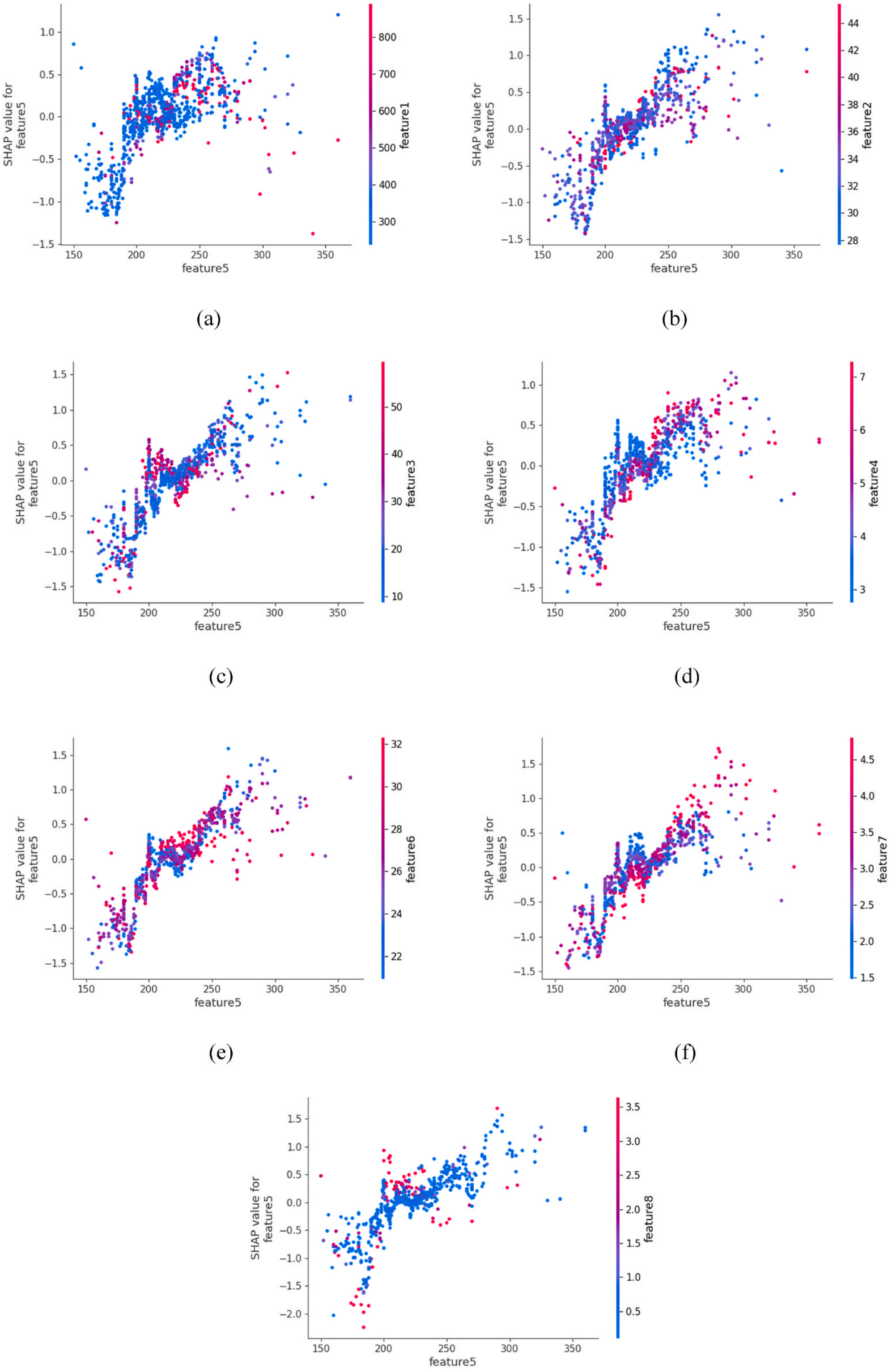
4.5. Analysis of risk factors of DR based on SHAP

4.5.1. Global analysis

SHAP was applied to interpret and analyze features based on the CatBoost-based model analysis. Fig. 4 illustrates how important features were. The SHAP values of HbA$_{1c,}$ ALB_CR, and UPR_24 were much higher than those of the other features, so they had the greatest impact on diabetic retinopathy. An analysis of the summary chart of features was conducted to more clearly illustrate each feature's impact on the target variable. The results are presented in Fig. 5.

Features are arranged in an ascending order by significance on the left. According to a horizontal analysis, all of the experimental data samples are represented by the points that are behind each feature. In this case, red denotes the sample with higher values, and blue the samples with the lower values. The color symbolizes the size of the feature value. For example, a higher value of HbA$_{1c}$ corresponds to a higher chance of diabetic retinopathy, and a lower value corresponds to a lower probability of diabetic retinopathy. Furthermore, these two values are positively associated. Overall, there was a positive correlation found between diabetic retinopathy and HbA$_{1c}$, ALB_CR, UPR_24, NEPHROPATHY, SCR, GLU_2H, BMI, BU, INS, HEMATONOSIS, BP_HIGH, BP_LOW, and FLD, and there was a negative correlation found between diabetic retinopathy and CP, HB, ALB, DBILI, CRP, PCV, LDL_C, ALT, UCR and FBG.

Besides, it was unclear how specifically HEIGHT, WEIGHT, ESR, and diabetic retinopathy related to one another. The distribution of the red and blue sample points of HEIGHT above the mean value of SHAP, as illustrated in Fig. 5, suggests that the unique influence of patient height on diabetic retinopathy could not be explained. Because distinct forms of diabetes are not separated in the experimental dataset, the distribution of sample points for WEIGHT is relatively concentrated. Due to inadequate insulin function, those with type 1 diabetes are typically skinny, while those with type 2 diabetes are typically fat due to insulin resistance. As a result, a patient's weight cannot be used to directly assess diabetic retinopathy. The ESR distribution of the figure is unclear, and further data should be included for a more in-depth analysis.



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(caption on next page)

Fig. 10. Correlations between blood pressure levels and other indices. (a) Relationship between feature 5 and feature 1 based on SHAP measurement. (b) Relationship between feature 5 and feature 2 based on SHAP measurement. (c) Relationship between feature 5 and feature 3 based on SHAP measurement. (d) Relationship between feature 5 and feature 4 based on SHAP measurement. (e) Relationship between feature 5 and feature 6 based on SHAP measurement. (f) Relationship between feature 5 and feature 7 based on SHAP measurement. (g) Relationship between feature 5 and feature 8 based on SHAP measurement.

4.5.2. *Analysis of features*

The feature indicators with similar characteristics were categorized to serve as a point of reference for the relationship between the features described subsequently, enabling a more thorough analysis of the potential causes of diabetic retinopathy. The classification results are provided in Table 5.

With the feature fusion approach, the relationship between the characteristics of calculation for various types of characteristic analysis of the interaction is handled under the classification index by using the presented shapes model.

(1) Renal function indices

Fig. 6 (a) to (g) show the relationship between feature 1 and feature 2 to 8. The right vertical axis shows the distribution of other features impacted by the feature, the horizontal axis shows the sample value of the feature, the color indicates the size of the feature value, with red denoting a sample with a higher value and blue denoting a sample with a lower value. For instance, the vertical axes on the left in Fig. 6 (a) represent the SHAP value of feature 1, the distribution of feature 2 following feature 1 change, and the horizontal axis represents the sample value of feature 1.

In [−1, 1], the SHAP value is focused when feature 1 is set to [250, 500]. Features 2's distribution is not evident. A curve-like distribution of distribution points can be created by setting feature1's value to [500, 2000], and [1,3] is where the concentration of SHAP values occurs. Blue distribution is depicted in Fig. 6(b) and (g). Low values are found for features and feature 8, no matter how feature 1 changes. Feature 1 greatly influences feature 4. The number of feature 4 items rises in proportion to the value of feature 1. Fig. 6 (d) and 6 (e) are mainly distributed in red. Features 5 and 6 have a greater value than feature 1, no matter how feature 1 changes. The color distribution is uneven in Fig. 6 (a) and Fig. 6 (f), feature 1 has no obvious influence on feature 2 and feature 7.

(2) Glycemic indices

As shown in Fig. 7 (a) to (g), the overall image distribution is in a straight line. As the value of feature 2 increases, the image curves upward. Fig. 7(a)–. 7 (b), and Fig. 7 (g) are displayed in blue. Regardless of the value of feature 2, feature 1, feature 3, and feature 8 are low in value. The colors in Fig. 7 (c), 7 (d), 7 (e), and 7 (f) are not evenly distributed. It is not possible to directly determine the impact of feature 2 on feature 4, feature 5, feature 6, and feature 7.

(3) Liver function indices

As shown in Fig. 8 (a) to (g), the overall image distribution shows a decreasing curve. As the value of feature 3 increases, its SHAP value decreases. Fig. 8 (a) and Fig. 8 (g) are displayed in blue. No matter how feature 3 changes, the values of feature 1 and feature 8 are low. As the colors in Fig. 8 (b), 8 (c), 8 (d), 8 (e), and 8 (f) are not evenly distributed, it is impossible to directly determine the impact of feature 3 on feature 2, feature 4, feature 5, feature 6, and feature 7.

(4) Blood indices

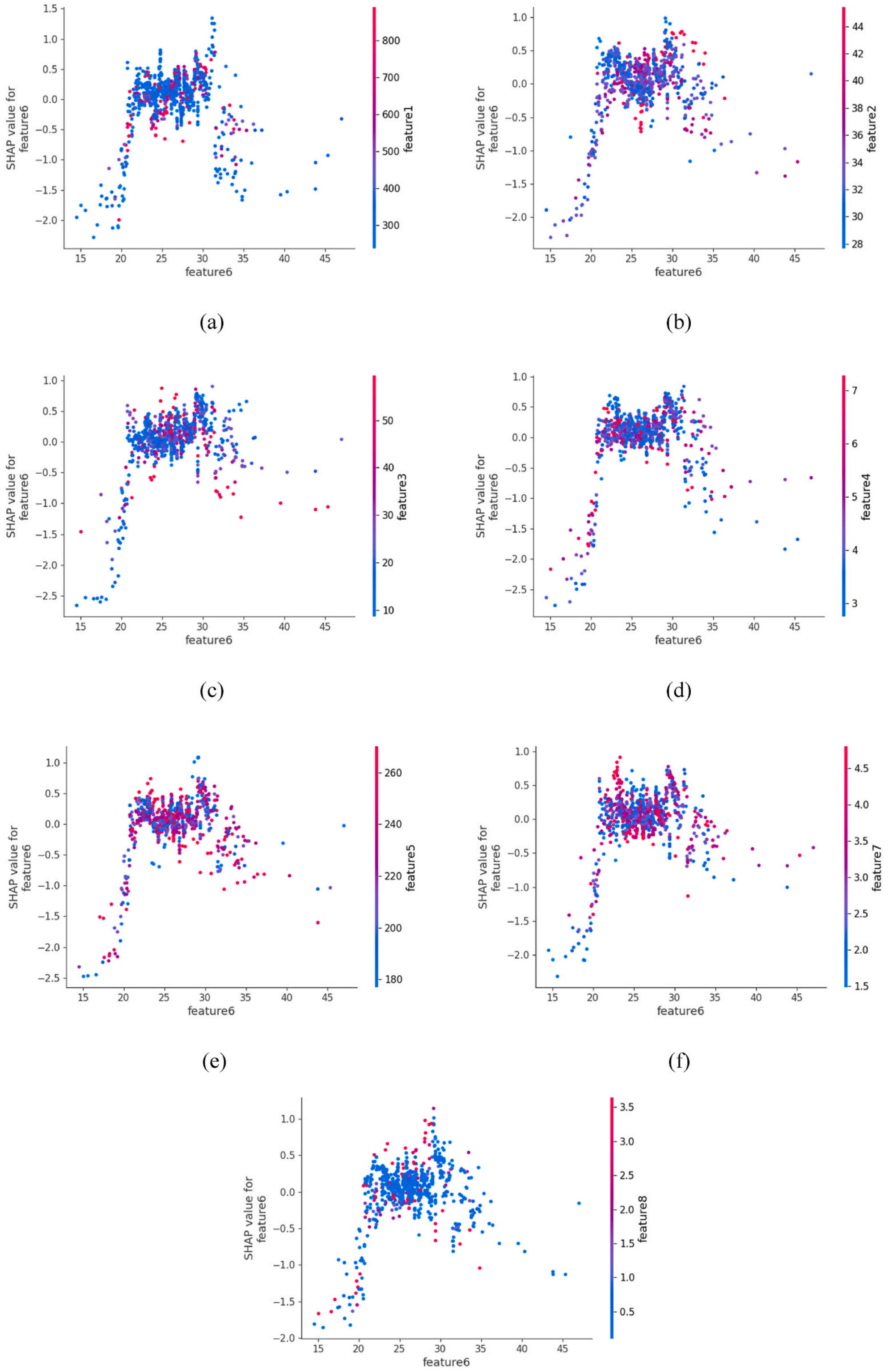
As shown in Fig. 9 (a) to (g), the overall image distribution is a straight line, and a small number of samples are distributed outside the line. Fig. 9 (a) and Fig. 9 (g) are displayed in blue. No matter how feature 2 changes, the values of feature 1 and feature 8 are low. Fig. 9 (d) is displayed in red. No matter how feature 2 changes, feature 5 has a high value. The colors of Fig. 9 (b), 9 (c), 9 (e), and 9 (f) are not evenly distributed. It is not possible to directly determine the impact of feature 4 on feature 2, feature 3, feature 6, and feature 7.

(5) Blood pressure

As shown in Fig. 10 (a) to (g), the image is roughly distributed in an increasing straight line, and a large number of sample points are distributed outside the line. Fig. 10 (g) is displayed in blue. No matter how feature 5 changes, the value of feature 8 is low. As the colors of Fig. 10 (a), 10 (b), 10 (c), 10 (d), 10 (e), and 10 (f) are not evenly distributed, it is impossible to directly determine the impact of feature 5 on feature 1, feature 2, feature 3, feature 4, feature 6, and feature 7.

(6) Other indices

As shown in Fig. 11 (a) to (g), the images are distributed in the interval of feature 6 value [20,35], corresponding to SHAP value [−0.5, 1]. In addition, a few samples are distributed. Fig. 11 (a) and Fig. 11 (g) are displayed in blue. No matter how feature 6 changes,



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(caption on next page)

Fig. 11. Correlations between BMI and other indices. (a) Relationship between feature 6 and feature 1 based on SHAP measurement. (b) Relationship between feature 6 and feature 2 based on SHAP measurement. (c) Relationship between feature 6 and feature 3 based on SHAP measurement. (d) Relationship between feature 6 and feature 4 based on SHAP measurement. (e) Relationship between feature 6 and feature 5 based on SHAP measurement. (f) Relationship between feature 6 and feature 7 based on SHAP measurement. (g) Relationship between feature 6 and feature 8 based on SHAP measurement.

feature 1 and feature 8 are low in value. The colors of Fig. 11 (b), 11 (c), 11 (d), 11 (e), and 11 (f) are not evenly distributed. It is not possible to directly determine the impact of feature 6 on feature 2, feature 3, feature 4, feature 5, and feature 7.

As shown in Fig. 12 (a) to (g), the image trend is distributed in an irregular pattern. It is concentrated in the range of feature 7 value [2,6], corresponding to the SHAP value in the range [−1, 1]. A few samples are also distributed. Fig. 12 (a) and Fig. 12 (g) are all in blue. No matter how feature 7 is changed, the values of feature 1 and feature 8 are low. As the colors in Fig. 12 (b), 12 (c), 12 (d), 12 (e), and 12 (f) are not evenly distributed, it is impossible to directly determine the impact of feature 7 on feature 2, feature 3, feature 4, feature 5, and feature 6.

As shown in Fig. 13 (a) to (g), the overall image distribution is in a straight line. As the value of feature 8 increases, the image curves downward. Fig. 13 (a) is displayed in blue. No matter how the value of feature8 changes, feature 1 is low. As the colors of Fig. 13 (b), 13 (c), 13 (d), 13 (e), and 13 (f) are not evenly distributed, it is impossible to directly determine the impact of feature 8 on feature 2, feature 3, feature 4, feature 5, feature 6, and feature 7.

5. Discussion

The SHAP employed in this study improved the consistency of the experimental findings and provided a clear explanation of the outcomes expected from the model. The search of clinical study data was conducted to establish more connections with practice, and a discussion on the risk factors of diabetic retinopathy was made.

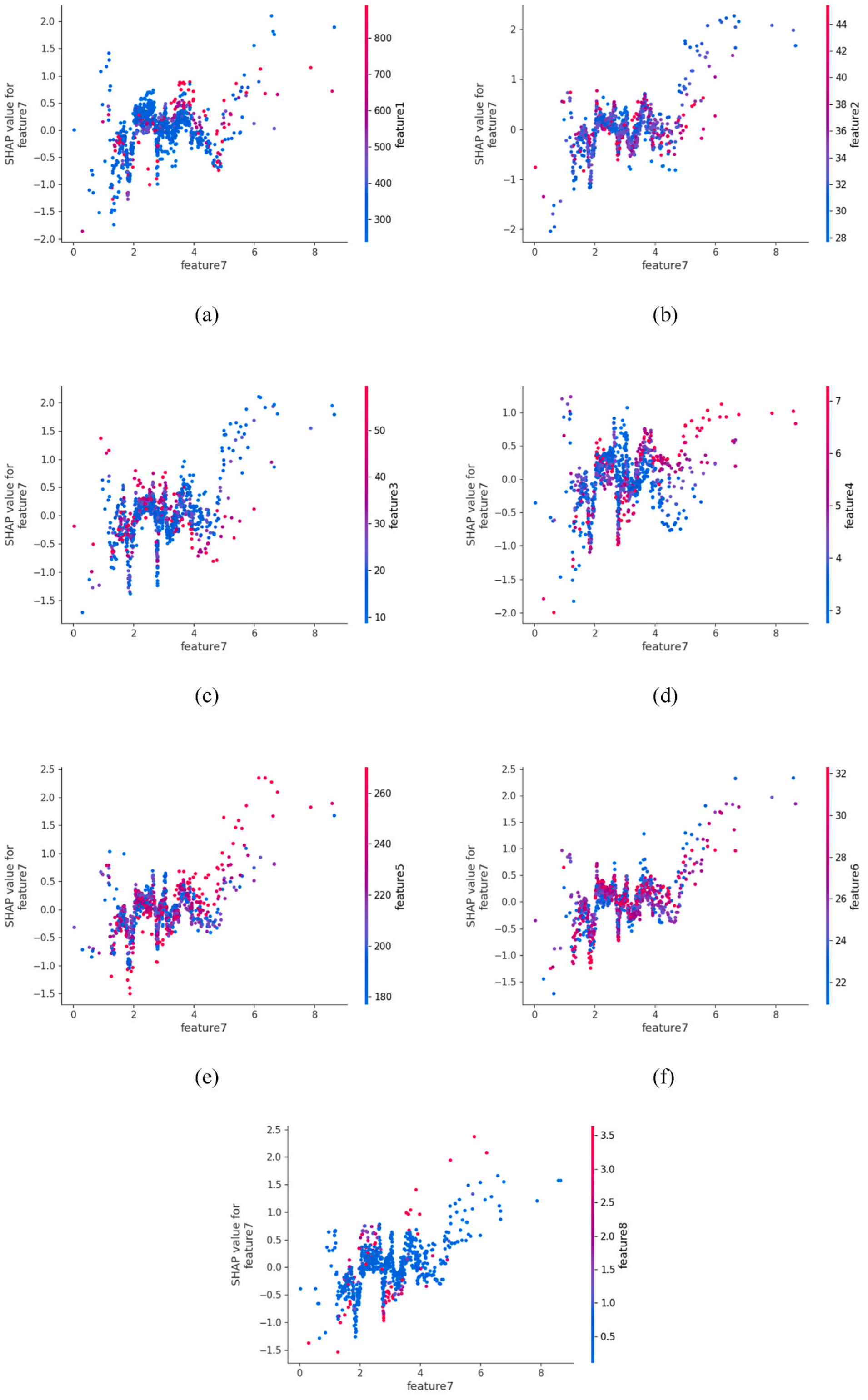
Firstly, the risk of diabetic retinopathy was greatly increased for abnormal renal function. According to the above data, there were different levels of influence that ALB_CR, UPR_24, HB, NEPHROPATHY, SCR, ALB, DBILI, BU, UCR, and other renal function associated indices had on diabetic retinopathy, and the risk was greatly elevated in cases of chronic renal disease or abnormal kidney function indications. It has been demonstrated that diabetic retinopathy and diabetic nephropathy share a similar pathophysiology and are frequently concomitant [35], consistent with the outcomes of this study. Regarding the etiology and progression of diabetic retinopathy and nephropathy, the scientific community remains divided. Diabetic retinopathy was found to be a predictor of the onset of diabetic nephropathy in an analysis by Zou Guming et al. using clinical cases confirmed by renal biopsy [36]. However, Mastropasqua L et al. noted that diabetic nephropathy might cause diabetic retinopathy and cause problems in patients' eyes [37]. More research and discussion are still required to determine the precise association and order of the two diseases.

Based on the characteristic analysis, the liver function indices, CRP, and renal function indices all had an impact on one another. Patients with diabetic nephropathy are at risk for hypoproteinemia as a result of renal injury, and persistent hypoproteinemia can either induce or aggravate the symptoms of anemia [38]. In addition, fat cells can produce CRP and other inflammatory mediators as insulin resistance occurs, which results in aberrant blood markers. Consequently, CRP may be a sign of diabetes, and complex kidney diseases could be the source of aberrant blood markers in diabetic individuals. There are few studies on the relationship between diabetic nephropathy and liver disease. Some scholars believe that kidney damage in diabetic patients may cause complications of liver disease [39], and the specific mechanism of action remains to be verified.

Secondly, diabetic retinopathy can result from elevated blood glucose level. Diabetes is mostly brought on by abnormal blood glucose levels associated with insulin resistance and secretion. The findings from this study demonstrate that high blood glucose levels are also likely to result in diabetic retinopathy, that insulin secretion and resistance are linked to CP and INS, and that diabetic retinopathy and HbA1c and GLU_2H are positively correlated. The blood glucose level is increased by both high INS and low CP. Currently, it is believed that a persistent increase in the blood glucose level will either induce or worsen retinopathy [40]. In addition, Chen Jiaxian et al. analyzed the cases of diabetic retinopathy of different degrees by the logistic regression method, and summarized the influencing factors that induced the lesions, indicating that HbA1c was a risk factor for diabetic retinopathy [41]. Cheng Zhuo verified that there was a significant correlation between the blood glucose level at 2 h after meals and diabetic retinopathy [42]. Wang Jing et al. confirmed CP as an important protective factor against diabetic retinopathy by regression analysis [43]. The aforementioned research perspectives can bolster the findings of this study. Furthermore, any alteration in the blood glucose level will also result in a change in CRP. In his research, Fukuhara, M. noted that blood glucose regulation could effectively lower CRP, and thus lower the inflammatory response in the body [44].

Thirdly, diabetic retinopathy is linked to liver health. Fig. 5 illustrates the positive correlation between FLD and diabetic retinopathy and the negative correlation between ALT and FBG. Low ALT and FBG may result in aberrant liver function in individuals, which will impact glucose tolerance and human metabolism, potentially precipitating or exacerbating hepatogenic diabetes [45–47]. This increases the chance of developing diabetic retinopathy. Renal function and CRP are also associated with liver function indices, and liver illness may occur as elevated CRP, an inflammatory marker. There appears to be a link between liver disease and kidney disease in diabetes patients, as stated in the first section. However, further research is required to determine the precise nature of this association and how it relates to clinical practice.

Fourthly, diabetic retinopathy may be affected by hematonosis. Diabetic individuals with blood illnesses are more prone to develop diabetic retinopathy, and hematonosis, FBG, PCV, and other blood indices have some bearing on the condition. Low PCV suggests



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(caption on next page)

Fig. 12. Correlations between LDL_C and other indices. (a) Relationship between feature 7 and feature 1 based on SHAP measurement. (b) Relationship between feature 7 and feature 2 based on SHAP measurement. (c) Relationship between feature 7 and feature 3 based on SHAP measurement. (d) Relationship between feature 7 and feature 4 based on SHAP measurement. (e) Relationship between feature 7 and feature 5 based on SHAP measurement. (f) Relationship between feature 7 and feature 6 based on SHAP measurement. (g) Relationship between feature 7 and feature 8 based on SHAP measurement.

anemia or renal illness, low FBG points to poor blood clotting capacity, and low PCV and FBG together raise the risk of diabetic retinopathy. The direct relationship between blood health and diabetic retinopathy has not yet been thoroughly examined in many researches. Blood pressure, CRP, and renal function are all connected with blood indices based on the characteristic analysis. As mentioned previously, nephropathy is the source of abnormal blood indices in diabetic patients, and abnormal blood indices themselves may be signs of diabetic nephropathy. Blood disorders may present with CRP, an inflammatory marker, and blood pressure stability may be influenced by the blood quality. Although the exact mechanism of action of blood disorders on diabetic retinopathy cannot currently be confirmed, this study suggests that aberrant blood indices of diabetes may contribute to or exacerbate other diabetic problems, hence raising the chance of diabetic retinopathy.

Fifthly, the blood pressure level influences diabetic retinopathy. A rise in either of the two blood pressure indices — diastolic or systolic blood pressure — will lead to diabetic retinopathy. As of right now, research indicates that maintenance of a stable systolic blood pressure and management of blood pressure can greatly lower the incidence of diabetic retinopathy [48–50]. A characteristic study revealed that elevated blood pressure was associated with CRP.

Last, other features have impact on diabetic retinopathy. An increase in BMI also has an effect on diabetic retinopathy, which has been confirmed in existing studies [50]. Consistent with the findings from this study, it was also noted in pertinent research that increased expression of LDL_C was one of the reasons causing diabetic retinopathy [51]. Upon additional examination in conjunction with Fig. 5, it was shown that LDL_C had a negative correlation with diabetic retinopathy overall; however, some samples with high feature values were dispersed among zones with elevated SHAP values. The short experimental dataset used in this study may be connected to this occurrence, thus more data will be needed to confirm this feature. One measure of inflammation is CRP. When combined with the earlier data, CRP is linked to all disease risks. Patients with diabetes actually exhibit a more pronounced autoinflammatory response. The further findings from the study indicate that elevated CRP levels are associated with an increased risk of diabetic retinopathy and possible triggers for aberrant CRP levels. Abnormal CRP may potentially have an association with lipid metabolism. The study conducted by Zhang revealed a direct correlation between diabetic retinopathy and aberrant lipid metabolism. Specifically, faulty lipid metabolism increases the risk of retinopathy by inducing the production of CRP [51].

While this study has made some contributions to the construction and interpretation of the diabetic retinopathy risk prediction model, it has certain limitations that have to be acknowledged. On the one hand, the dataset used is small and the sample size is limited. It is difficult to determine the precise and definitive effect of several features, such as LDL_C and ESR, on diabetic retinopathy. These aspects should be further investigated by expanding the dataset in the future. On the other hand, we use SHAP to mine and visualize the correlation between features, indicating that there are a certain number of relationships between features, but it cannot verify the causal relationship between features, which needs further discussion.

6. Conclusion

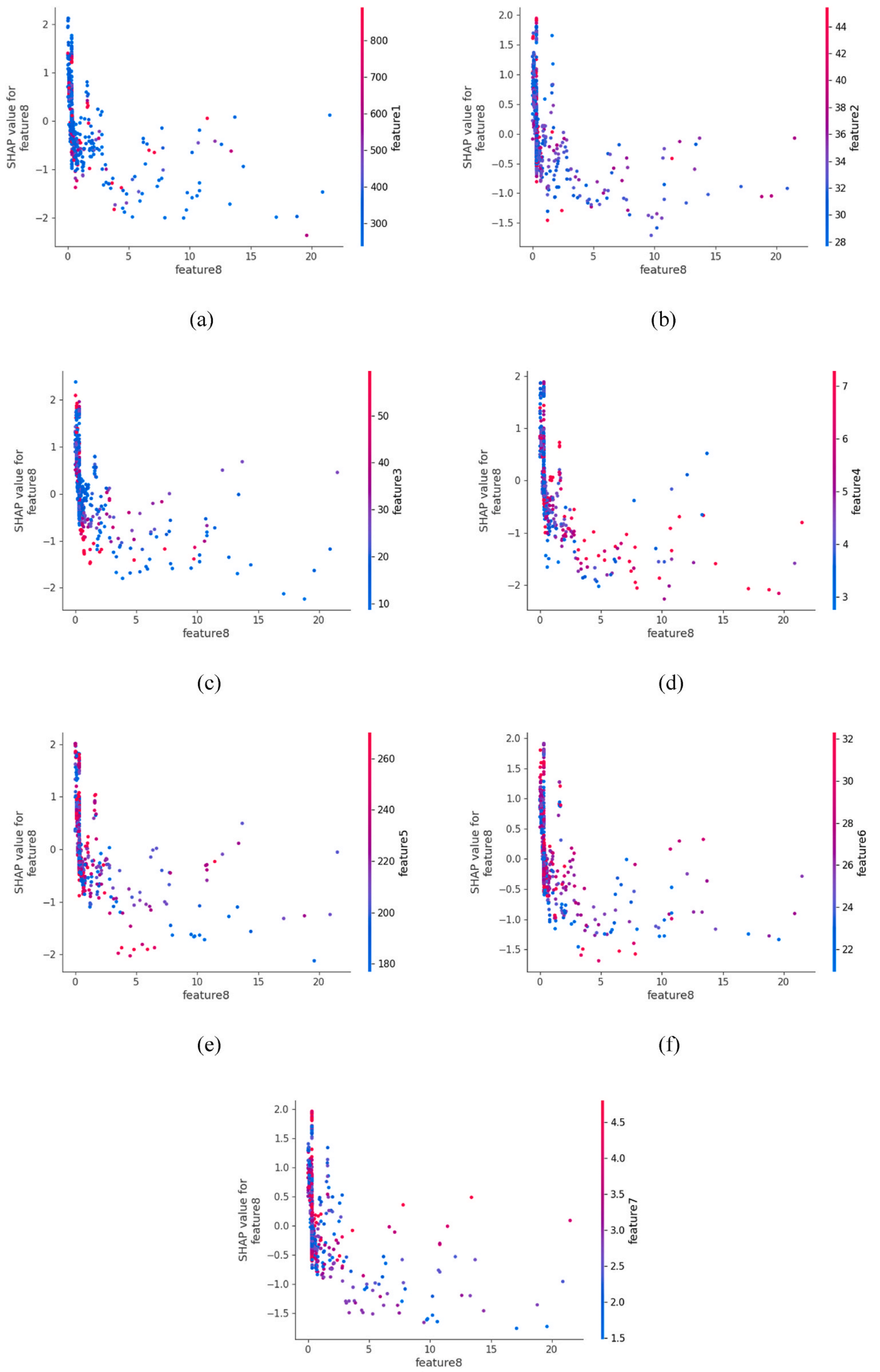
The diabetes complication prediction dataset from the National Clinical Medical Sciences Data Center served as the subject of this study. A CatBoost-based diabetic retinopathy risk prediction model was created by extracting and screening the typical indications of diabetic retinopathy, manipulating and improving the model parameters, and using the SHAP to investigate the risk factors of diabetic retinopathy. Study results show that, compared with machine learning-based models such as SVM, RF, XGBoost and GBT, the CatBoost-based prediction model for diabetic retinopathy showed the best performance. The explanation results from SHAP showed that ALB_CR, HbA_{1c}, UPR_24, NEPHROPATHY and SCR were positively correlated with diabetic retinopathy, while CP, HB, ALB, DBILI and CRP were negatively correlated with diabetic retinopathy. From the above discussion, it can be concluded that kidney disease, abnormal blood glucose, liver disease, blood pressure, and other characteristic factors can cause or aggravate diabetic retinopathy. Additionally, these factors can interact with one another to indirectly cause or exacerbate diabetic retinopathy. Thus, to preserve eye health, diabetic individuals should monitor their kidney, liver, and blood conditions, and follow a balanced diet.

Ethical approval statement

Review and/or approval by an ethics committee was not needed for this study because the database used in this study is from the public data platform.

Data availability statement

The name of the dataset of this study was to diabetes complications early warning Data set (CSTR: A0006.11 A0005.202006.001018). Dataset from the National Clinical Medical Science Data Center stored at the platform of China National Population Health Data Center and was open to the public. We have submitted an application for data use to the provider through the platform and obtained the approval. The dataset is linked to: https://www.ncmi.cn/phda/dataDetails.do?id=CSTR:A0006.11.A0005.



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(caption on next page)

Fig. 13. Correlations between CRP and other indices. (a) Relationship between feature 8 and feature 1 based on SHAP measurement. (b) Relationship between feature 8 and feature 2 based on SHAP measurement. (c) Relationship between feature 8 and feature 3 based on SHAP measurement. (d) Relationship between feature 8 and feature 4 based on SHAP measurement. (e) Relationship between feature 8 and feature 5 based on SHAP measurement. (f) Relationship between feature 8 and feature 6 based on SHAP measurement. (g) Relationship between feature 8 and feature 7 based on SHAP measurement.

202006.001018-V1.0.

Fundings

This work was supported by the Peking Union Medical College “Subject Evaluation Service based on benchmarking institution analysis” project(Education 0036).

CRediT authorship contribution statement

Xu Wang: Writing – original draft, Software, Methodology, Formal analysis. **Weijie Wang:** Writing – review & editing, Resources, Data curation. **Huiling Ren:** Writing – review & editing. **Xiaoying Li:** Writing – review & editing. **Yili Wen:** Visualization.

Declaration of competing interest

No conflict of interest exits in the submission of this manuscript.

Acknowledgments

We are grateful to the National Clinical Medical Science Data Center and China National Population Health Data Center for providing us with the “Diabetes Complication Early Warning Dataset”. These data resources are reliable and truthful, for which this study was carried out smoothly.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e29497.

References

[1] International Diabetes Federation, Facts & Figures, 2023, 2023-11-2.

[2] T. Derong, Y. Yuwen, S. Rui, L. Dandan, L. Rong, Analysis of pathological regularity and related risk factors of retinal nerve injury in patients with diabetes mellitus, Recent Advances in Ophthalmology 43 (12) (2023) 964–969.

[3] G.D. Calderon, O.H. Juarez, G.E. Hernandez, S.M. Punzo, Z.D. De la Cruz, Oxidative stress and diabetic retinopathy: development and treatment, Eye (Lond). 31 (8) (2017) 1122–1130.

[4] Chinese clinical Guidelines for the prevention and treatment of type 2 diabetes in the elderly (2022 edition), Chinese Journal of Diabetes 1 (30) (2022) 2–51.

[5] A. Ashiquzzaman, A.K. Tushar, M.R. Islam, et al., Reduction of Overfitting in Diabetes Prediction Using Deep Learning Neural Network, Springer, Singapore, 2018.

[6] Y. H, Y. Z, L. Z, S. W, Q. G, Application of decision tree model in risk prediction of type 2 diabetes mellitus, Chin. J. Health Statistics 6 (33) (2016) 976–978.

[7] G. L, L. J, Establishing a model for predicting diabetes complications based on the LVQ neural work, Chinese Journal of natural medicine 4 (2006) 254–258.

[8] V.A. Kumari, R. Chitra, Classification of diabetes disease using support vector machine, International Journal of Engineering Research and Applications 3 (2) (2013) 1797–1801.

[9] J. T, Z. C, B. Y, Q. T, To explore the human insulin level evaluation model based on SVM algorithm, Information technology and informatization (8) (2022) 33–37+42.

[10] L. Z, S. N, Establishment of a prediction model of gestational diabetes mellitus based on support vector machine, Anhui Journal Of Preventive Medicine 6 (25) (2019) 465–468.

[11] Q. H, J. H, Multi-layer perceptron diabetes prediction model combined with batch normalization, Computer Systems & Applications 5 (29) (2020) 182–188.

[12] Q. Z, M. Z, Y. H, et al., Research on prediction ot dally admissions ot respiratory diseases with comoroid ciabetes in Bering based on ong short-term memory network.recurrent neural, Journal of Zhejiang University(Medical Sciences) 1 (51) (2022) 1–9.

[13] R. Yasashvini, V.R.M. Sarobin, R. Panjanathan, G.S. Jasmine, J.L. Anbarasi, Diabetic retinopathy classification using CNN and hybrid deep convolutional neural networks, Symmetry (Basel). 14 (9) (2022).

[14] Y. S, S. L, W. X, et al., A risk factor analysis for type 2 diabetes mellitus based on LASSC regression and random forest algorithm, Journal of Environmental Hygiene 7 (13) (2023) 485–495.

[15] S.D.L. Siyuan, Risk disclosure and key factors analysis of diabetic retinopathy, Chinese Journal of Medical Physics 6 (39) (2022) 783–787.

[16] N. Singh, P. Singh, Stacking-based multi-objective evolutionary ensemble framework for prediction of diabetes mellitus - ScienceDirect, Biocybern. Biomed. Eng. 40 (1) (2020) 1–22.

[17] D. Z, W. Q, K. Y, Prediction of 3-year risk of diabetic kidney disease using machine learning based on electronic medical records, J. Transl. Med. 20 (1) (2022) 143.

[18] L.Y.M.E.A. Qiaohong, Classification prediction and application of diabetes based on XGBoost model, Mod. Instrum. 4 (29) (2023) 1–6.

[19] F. M, Y. L, C. G, M. W, D. L, Diabetes prediction method based on CatBoost algorithm, Computer Systems & Applications 9 (28) (2019) 215–218.

[20] D.D. Rufo, T.G. Debelee, A. Ibenthal, W.G. Negera, Diagnosis of diabetes mellitus using gradient boosting machine (LightGBM), Diagnostics 11 (9) (2021).

[21] S. Lundberg, S. Lee, A Unified Approach to Interpreting Model Predictions, 2017.

[22] X. W, B. L, M. L, R. S, Combination of LightGBM and SHAP for diabetes prediction and feature analysis, Journal of Chinese Mini-Micro Computer Systems 9 (43) (2022) 1877–1885.
[23] P. Francesco, P. Jacopo, C. Giacomo, et al., The importance of interpreting machine learning models for blood glucose prediction in diabetes: an analysis using SHAP, Sci. Rep. 1 (13) (2023) 16865.
[24] D. J, C. G, F. P, et al., From SHAP to probability—an lnterpretable machine learning framework for targeted lipidomics study on diabetic retinopathy, Chin. J. Health Statistics 4 (40) (2023) 511–515.
[25] Chinese People's Liberation Army General Hospital. Shared Cup Edition _ Diabetes Complication warning dataset, National Population Health Sciences Data Center Data Warehousing PHDA, 2020.
[26] J. Junping, Fundamental Statistics, 2010.
[27] M. Guojun, D. Lijuan, Principles and Algorithms of Data Mining, 2016.
[28] C. Shannon, W. Weaver, The mathematical theory of communication, Phil. Rev. 60 (1949) 144.
[29] W. Zhijin, Information Retrieval and Processing, 2015.
[30] L. Prokhorenkova, G. Gusev, A. Vorobev, A.V. Dorogush, A. Gulin, CatBoost: Unbiased Boosting with Categorical Features, 2017.
[31] A.V. Dorogush, V. Ershov, A. Gulin, CatBoost: Gradient Boosting with Categorical Features Support, 2018.
[32] L. Prokhorenkova, G. Gusev, A. Vorobev, A.V. Dorogush, A. Gulin, CatBoost: Unbiased Boosting with Categorical Features, 2017.
[33] J X, Breast cancer prediction and feature analysis model based on CatBoost and SHAP, Comput. Mod. 10 (2023) 32–38.
[34] S. Lundberg, S.I. Lee, A Unified Approach to Interpreting Model Predictions, 2017.
[35] J.W.Y. Yau, S.L. Rogers, R. Kawasaki, E.L. Lamoureux, J. Vioque, Global prevalence and major risk factors of diabetic retinopathy, Diabetes Care 35 (3) (2012) 556–564.
[36] G. Z, N. Z, H. G, et al., Correlation between type 2 diabetic nephropathy and retinopathy, Journal of Clinical Nephrology 4 (15) (2015) 208–211.
[37] L. Mastropasqua, A. Verrotti, L. Lobefalo, F. Chiarelli, G. Verdesca, Visual field defects in diabetic children without retinopathy : relation between visual function and microalbuminuria, Acta Ophthalmol. Scand. 2 (73) (1995) 125–128.
[38] M. T, R. M, D. Z, et al., Analysis of anemia status and its correlation with serum biochemical indexes in diabetic nephropathy patients, Modern Medicine Journal of China 6 (25) (2023) 36–38.
[39] X.B.B. Ruikui, Research progress on the correlation of non-alcoholic fatty liver disease with type 2 diabetes and diabetic nephropathy, Med. Recapitulate 1 (21) (2015) 107–108.
[40] S.D. Solomon, M.F. Goldberg, ETDRS grading of diabetic retinopathy: still the gold standard? Ophthalmic Res.: Journal for Research in Experimental and Clinical Ophthalmology 62 (4) (2019) 190–195.
[41] C.Y.W.E.A. Jiaxian, An analysis of influencing factors of diabetic retinopathy among type 2 diabetic patients: a nested case-controlstudy, Chin. J. Dis. Control Prev. 26 (3) (2022) 269–273.
[42] C. Zhuo, Clinical analysis of diabetic retinopathy and blood glucose control, China Health Standard Management 9 (11) (2018) 34–36.
[43] W.J.D. Jing, Corelation between fasting C-peptide and retinopathy in elderly patients with type 2 diabetes mellitus and its influencing factors, Hebei Medical Journal 45 (15) (2023) 2324–2326.
[44] M. Fukuhara, K. Matsumura, M. Wakisaka, et al., Hyperglycemia promotes microinflammation as evaluated by C-reactive protein in the very elderly, Intern. Med. 46 (5) (2007) 207–212.
[45] R. Cano, J.L. Perez, L.A. Davila, et al., Role of endocrine-disrupting chemicals in the pathogenesis of non-alcoholic fatty liver disease: a comprehensive review, Current drug targets-The International journal for timely in-depth reviews on drug targets 22 (9) (2021).
[46] Q.W.Z.E.A. Jialin, Influence of hepatitis B virus infection and liver function status on the occurrence of gestational diabetes mellitus, Med. J. Chin. Peoples Lib. Army 48 (2) (2023) 218–223.
[47] S. Yanqiu, Effect of liraclutide on clucose and lipid metabolism and liver function in patients with type 2 diabetes complicated with metabolic fatty liver disease 7 (21) (2023) 90–93.
[48] X. B, J. S, F. L, et al., Factors affecting diabetic retinopathy among Chinese adults: a meta-analysis, Prev. Med. 7 (35) (2023) 595–601.
[49] V. Foo, Cheung G. JoannebQuah, N. Chun, T. Kyi, HbA1c, systolic blood pressure variability and diabetic retinopathy in Asian type 2 diabetics, J. Diabetes 9 (2) (2016) 200–207.
[50] L.W.H.E.A. Jin, Development and validation of a predictive risk model for vision-threatening diabetic retinopathy in patients with type 2 diabetes, J. Sun Yat-sen Univ. (Soc. Sci. Ed.) 6 (44) (2023) 1–9.
[51] Z. Qiuyan, Changes of serum lipids in elderly diabetic retinopathy patients and their relationship with severity of disease, Clin. Res. 1 (30) (2022) 183–186.