European Journal of Medical Research

RESEARCH

Open Access

Predicting diabetic retinopathy based on routine laboratory tests by machine learning algorithms

Xiaohua Wan[1,2,3], Ruihuan Zhang[4], Yanan Wang[4], Wei Wei[5], Biao Song[4*], Lin Zhang[5,6,7*] and Yanwei Hu[1,2*]

Abstract

Objectives This study aimed to identify risk factors for diabetic retinopathy (DR) and develop machine learning (ML)-based predictive models using routine laboratory data in patients with type 2 diabetes mellitus (T2DM).

Methods Clinical data from 4259 T2DM inpatients at Beijing Tongren Hospital were analyzed, divided into a model construction data set ($N = 3936$) and an external validation data set ($N = 323$). Using 39 optimal variables, a prediction model was constructed using the eXtreme Gradient Boosting (XGBoost) algorithm and compared with four other algorithms: support vector machine (SVM), gradient boosting decision tree (GBDT), neural network (NN), and logistic regression (LR). The Shapley Additive exPlanation (SHAP) method was employed to interpret the XGBoost model. External validation was performed to assess model performance.

Results DR was present in 47.69% ($N = 1877$) of T2DM patients in the model construction data set. Among the models tested, the XGBoost model performed best with an AUC of 0.831, accuracy of 0.757, sensitivity of 0.754, specificity of 0.759, and F1-score of 0.752. SHAP explained feature importance for XGBoost model and identified key risk factors for DR. External validation yielded an accuracy of 0.650 for the XGBoost model.

Conclusions The XGBoost-based prediction model effectively assesses DR risk in T2DM patients using routine laboratory data, aiding clinicians in identifying high-risk individuals and guiding personalized management strategies, especially in medically underserved areas.

Keywords Type 2 diabetes mellitus, Diabetic retinopathy, Routine laboratory tests, Machine learning, XGBoost, Predictive model

*Correspondence:
Biao Song
songbiao_511@163.com
Lin Zhang
13661181680@163.com
Yanwei Hu
ywhu@mail.ccmu.edu.cn
[1] Department of Clinical Laboratory, Beijing Chao-Yang Hospital, Capital Medical University, Beijing, People's Republic of China
[2] Beijing Center for Clinical Laboratories, Beijing, People's Republic of China
[3] Department of Clinical Laboratory, Beijing Tongren Hospital, Capital Medical University, Beijing, People's Republic of China
[4] The Inner Mongolia Medical Intelligent Diagnostics Big Data Research Institute, Inner Mongolia, People's Republic of China
[5] Department of Medical Record, Beijing Tongren Hospital, Capital Medical University, Beijing, People's Republic of China
[6] Department of Endocrinology, Beijing Tongren Hospital, Capital Medical University, Beijing, People's Republic of China
[7] Beijing Diabetes Research Institute, Beijing, People's Republic of China

Introduction

Diabetic retinopathy (DR) is a predominant microvascular complication associated with diabetes and stands as the primary cause of blindness among working-aged people [1, 2]. Current global prevalence estimates indicate that 22.27% of individuals with diabetes develop DR, with 6.17% suffering from vision-threatening DR (VTDR) and 4.07% experiencing clinically significant macular edema [3]. In China, the number of patients with diabetes exceeds 140 million, and T2DM accounts for over 90%, accounting for 24% globally of patients with diabetes [2, 4, 5]. Although DR is often insidious and asymptomatic at the early stages, it might quickly progress into VTDR without awareness and intervention, and then could lead to irreversible vision impairment [2].

Early detection and treatment interventions for DR can reduce the risk of severe visual loss by approximately 90% [1]. DR screening is primarily conducted through two methods: fundus examinations and fundus photography. Fundus examinations, carried out by ophthalmologists or optometrists, demand specialized tools, such as binocular indirect ophthalmoscopes or slit lamps, along with advanced expertise. Fundus photography, performed by ophthalmologists or trained technicians, relies on expensive equipment, such as optical coherence tomography (OCT) machines and digital fundus cameras [6]. Although the American Diabetes Association (ADA) advocates for annual ophthalmic exams as the gold standard for DR screening, adherence is poor, with less than 60% of diabetes patients following recommended guidelines [7]. This shortfall is mainly due to the high costs and limited accessibility of ophthalmologic services [2–5].

Some studies have reported that annual screening rates for inner-city diabetic patients are below 25%, further hindered by a shortage of eye specialists in underserved urban areas [8, 9]. The rising patient numbers and diverse individual needs make timely eye exams challenging [10–12]. This study uniquely focuses on routine laboratory data, independent of ophthalmological and other imaging examinations and diagnoses. The aim is to primarily establish a simple, practical DR prediction model using only laboratory results and minimal basic characteristics. By enabling early detection of high-risk individuals without additional costs or complex diagnostics, the model will help prioritize retinal screenings, aiding in DR prevention and reducing the risk of blindness, especially in medically underserved areas.

Machine learning (ML) has emerged as a powerful tool in public health research, offering effective solutions to complex challenges in the field [13, 14]. Its strengths in generalizability and handling high-dimensional data make it particularly suited for analyzing intricate real-world data sets. Recently, deep learning system for DR have been widely explored for the analysis of fundus photographs [13]. Innovations like Bourouis et al.'s smartphone-integrated algorithm with microscopic lenses for retinal imaging have also been introduced [14]. However, the reliance on costly fundus cameras for retinal imaging remains a significant barrier, restricting the use of these advanced screening methods to well-resourced healthcare facilities.

In recent years, the integration of laboratory tests and ML has been extensively explored to develop DR prediction models. Chen et al. [6] investigated early stage DR detection using accessible data, such as demographics, comorbidities, and routine laboratory results. They developed and evaluated various temporal deep learning models on a large-scale, real-world data set, demonstrating superior performance over baseline random forest models in key metrics. Similarly, Homayouni et al. [15] introduced an innovative "Progressive Ablation Feature Selection method with XGBoost," which streamlined the prediction model, achieving an impressive Area Under the Curve (AUC) of 96.61%. Their findings identified creatinine(Cr) levels as the strongest predictor of DR, with neuropathy and nephropathy also playing critical roles. Ogunyemi et al. [9] constructed a deep neural network model incorporating variables, such as nephropathy, neuropathy, stroke, and insulin dependence, while Yang et al. [16] developed a nomogram that included predictors, such as diabetes duration, diabetic neuropathy, diabetic foot, diabetic kidney disease, hyperlipidemia, and hypoglycemic drug use.

Despite these advancements, most studies rely on additional variables, such as comorbidities, imaging tests, medication usage, and diabetes duration, which can be difficult to obtain. As a result, there is a critical need for a simple, practical predictive model that identifies high-risk DR individuals using only laboratory test data and easily accessible basic characteristics. Such a model would be more universal and effective, overcoming barriers to early DR diagnosis and establishing a new standard of care that enhances healthcare quality and compliance without incurring additional costs [17, 18].

In this retrospective cohort study, we analyzed routine laboratory test data from hospitalized patients with T2DM to assess the ability of ML models to predict DR risk. We also employed the Shapley Additive exPlanation (SHAP) method to identify the most effective prediction model and quantify the impact of various indicators on DR. Our goal was to develop an optimal and practical predictive model to effectively assess DR risk in T2DM patients using routine laboratory data. This model aims to assist clinicians, particularly primary care physicians in underdeveloped regions, in identifying high-risk

individuals and guiding personalized management strategies, particularly in resource-limited regions.

Methods

Research design and study patients

The clinical data of 4259 inpatients diagnosed with T2DM in the Department of Endocrinology at Beijing Tongren Hospital, Capital Medical University, from December 2013 to April 2024 were collected. In instances of multiple hospitalizations, only data from the initial hospital admission were included. Patient identifiers were not included in the data set. The clinical data of 4259 inpatients was divided into two data sets based on the patient's admission time: the model construction data set from December 2013 to April 2023 and the external validation data set from May 2023 to April 2024. The model construction data set, including 3936 patients, was mainly used for feature selection, model construction training, and internal validation; The external validation data set, comprising 323 newly admitted T2DM inpatients, was primarily used to validate the optimized models. None of these patients were included in the model construction data set, ensuring an independent assessment of the model's performance. This study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of Beijing Tongren Hospital, Capital Medical University (approval number TREC2024-KY040). Written informed consent was obtained from all the participants.

Data collection and clinical definitions

The basic characteristics, routine laboratory tests, and DR status of T2DM patients were extracted from electronic medical records. The basic characteristics included gender, age, T2DM duration, body mass index (BMI), pulse, waist-to-hip ratio (WHR), systolic blood pressure (SBP), and diastolic blood pressure (DBP). Routine laboratory tests encompassed blood routine tests (24 indicators), biochemical routine tests (30 indicators), and coagulation routine tests (6 indicators). In addition, the fundus examination of DR was collected for all patients.

BMI was calculated by dividing weight in kilograms by height in meters squared (kg/m^2) [19]. The diagnostic criteria for DR adhered to the International Clinical Diabetic Retinopathy Severity Scale [20]. Diagnosis was based on macula-centered 45° fundus photography and indirect ophthalmoscopy after pupil dilation. A double-blind assessment was conducted by two experienced ophthalmologists from Beijing Tongren Hospital, with a third ophthalmologist consulted in case of disagreement. DR was defined by the presence of any of the following lesions: microaneurysms, retinal hemorrhages, soft exudates, hard exudates, or vitreous haemorrhage [20]. Patients with DR (DR group) were considered positive samples, while those without DR (non-DR group) were considered negative samples.

Inclusion and exclusion criteria

The inclusion criteria consisted of patients with a discharge diagnosis of T2DM and patients aged 20 years or older. T2DM diagnosed according to 1999 World Health Organization (WHO) criteria and Chinese Diabetes Association criteria [21, 22]. We collected data from these patients' initial admission, including detailed basic characteristics, routine laboratory tests, and DR status.

Participants meeting the following criteria would be excluded: (1) diagnosis of type 1 diabetes; (2) presence of any other eye diseases or history of eye surgery, such as severe cataracts, glaucoma, or severe corneal opacity; (3) poor quality of fundus photographs, rendering them unsuitable for DR diagnosis; (4) acute metabolic disorders (such as diabetic ketoacidosis, hyperglycemia, or hypertonic state), or acute inflammatory diseases; and (5) presence of serious systemic diseases other than diabetes, such as severe cardiac or cerebrovascular diseases, or cancer [9, 11].

Model development and internal validation based on model construction data set

Data set division

The model construction data set included clinical data of 3936 inpatients from December 2013 to April 2023. Prior to data partitioning, preprocessing steps were undertaken to eliminate redundant information, including noise, missing values, and unknown data, ensuring data quality and accuracy. The data set was subsequently divided into training and internal validation sets in a 9:1 ratio. The training set was used for model development, while the validation set was utilized to evaluate the model's generalization ability and performance.

Feature selection

Feature selection is a critical procedure in the modelling process, as it enhances model performance, expedites computation, improves generalization, and enhances interpretability [23, 24]. In this study, Random Forest was chosen for feature selection due to its key strengths: (1) Capturing nonlinear relationships and complex interactions: as an ensemble of decision trees, it adapts to nonlinear patterns and intricate feature dependencies without strict data assumptions, handling both categorical and continuous variables; (2) Generating feature importance scores: it evaluates feature significance through average impurity reduction (e.g., Gini impurity or mean squared error) across trees, highlighting impactful features; (3) Robust performance with

high-dimensional data: it excels with high-dimensional data sets, even with noisy features, by assessing feature contributions across all trees; and (4) Permutation importance support: by shuffling feature values and measuring performance changes, it provides a robust, scale-insensitive measure of feature relevance [23, 24]. These qualities make Random Forest a highly effective and flexible tool for feature selection in complex data sets, ensuring both reliability and interpretability. The DR risk prediction model was subsequently built using the most pertinent features identified.

Machine learning algorithms

ML algorithms were chosen based on the following criteria. First, the algorithms were required to demonstrate evaluation capabilities across mixed data types, encompassing numerical and categorical variables. Second, selected algorithms were expected to possess broad applicability and a track record of successful utilization in relevant domains [18, 25].

Following the outlined criteria, we selected five machine learning algorithms for this study: (1) eXtreme Gradient Boosting (XGBoost): an efficient gradient boosting decision tree algorithm known for its high accuracy and fast training. It utilizes parallel computing and regularization, making it ideal for classification and regression tasks [26]; (2) Support vector machine (SVM): a method that finds the optimal hyperplane for classification or regression. It performs well with high-dimensional data and handles both linear and non-linear problems effectively [27]; (3) Gradient boosting decision tree (GBDT): an iterative algorithm that builds decision trees sequentially, with each tree correcting errors from the previous one. It is versatile for regression and classification tasks [28]; (4) Neural network (NN): a model inspired by biological neurons, which learns complex patterns and features in data through multiple layers of nonlinear transformations. It excels with high-dimensional data [29]; and (5) Logistic regression (LR): a straightforward and interpretable linear model for binary or multi-class classification. It uses the Sigmoid function to map input features to probabilities [30].

Hyperparameter tuning

The study explored various aspects of ML algorithms, focusing on enhancing their performance through the intricate process of hyperparameter optimization. This step is crucial for fine-tuning key model parameters to achieve optimal predictive accuracy. We employed the GridSearchCV method to determine the optimal key hyperparameters for the five ML models. GridSearchCV works by generating a grid of hyperparameter values, testing every possible combination, evaluating model performance, and selecting the optimal set of hyperparameters, thereby facilitating effective model optimization [31, 32].

Model evaluation and internal validation

The models underwent testing and internal validation through tenfold cross-validation [18]. The evaluation of the developed models' performance was based on sensitivity, specificity, accuracy, AUC of the receiver operating characteristic (ROC), and F1-score on the test data set.

Performance metrics were computed from the confusion matrix, which includes the following four measures: (1) true positive (TP): instances where the model correctly predicts DR when the actual class is DR; (2) false positive (FP): instances where the model incorrectly predicts DR when the actual class is non-DR; (3) false negative (FN): cases where the model incorrectly classifies an instance as non-DR when the actual condition is DR; and (4) true negative (TN): cases where the model correctly predicts non-DR when the actual class is non-DR [18, 23].

Sensitivity, commonly referred to as the true positive rate (TPR) or recall(R), quantifies a model's ability to correctly identify positive instances: sensitivity = TP/(TP + FN). Minimizing false negatives is key to achieving higher sensitivity [33].

Specificity, or true negative rate (TNR), measures correct identification of negative instances: specificity = TN/(TN + FP). To achieve higher specificity, the number of false positives (FPs) should be minimized [33].

Accuracy is a key metric used to evaluate a model's performance in ML. It measures the proportion of correct predictions relative to the total number of predictions, offering an overall assessment of the model's correctness. Accuracy is calculated as: accuracy = number of correct predictions/total number of predictions [33].

The AUC evaluates classifier performance, with values from 0 to 1. Higher AUC indicates better performance, plotting the trade-off between TPR and False Positive Rate (FPR = 1—Specificity) across thresholds [34].

The F1-score is a vital metric for assessing ML models, particularly in imbalanced data sets. It balances precision and recall into one measure, providing a comprehensive accuracy assessment. Precision is the ratio of correctly identified positives to all predicted positives: precision = TP/(TP + FP). F1-score was calculated as

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

As the harmonic mean of precision and recall, the F1-score ranges from 0 to 1, where 1 reflects optimal balance and performance, and 0 indicates poor

performance. Intermediate scores suggest a precision–recall imbalance [33].

Model interpretation

To improve the interpretability of the model's decision-making process and prediction outcomes, the SHAP method was employed. SHAP is a widely adopted interpretability framework in ML that quantifies the contribution of individual features to model predictions using SHAP values. Rooted in Shapley values from cooperative game theory, SHAP values quantify the impact of individual features on model predictions. It offers both global explanations (for overall model behavior) and local explanations (for individual predictions), making it a versatile tool for model interpretation, feature selection, and debugging. Despite its computational demands, SHAP's strong theoretical foundation and intuitive interpretability render it an essential resource for understanding and explaining ML models [35, 36].

Model external validation based on external validation data set

To further evaluate the model's generalization ability, external validation is used to assess its performance on unknown data. External validation refers to evaluating the predictive performance of a model on different relevant data sets, which helps to obtain an accurate, objective, and unbiased estimate of the model's performance. The external validation data set comprised clinical data from 323 newly admitted inpatients collected between May 2023 and April 2024, and none of the patients were included in the model construction data set. The optimal model was selected from five candidates and further evaluated on this unseen data to assess its performance and generalizability.

Statistical analysis

Development and validation of the models were conducted using Python 3.8.3 (library, sci-kit-learn). Description analyses were conducted using SPSS 27.0 (IBM, Chicago, IL, USA). Continuous variables were expressed as means with standard deviation (SD) or medians with interquartile ranges (IQRs), while categorical variables were reported as numbers. The basic characteristics and routine laboratory tests were compared between DR and non-DR groups using a t test or Mann–Whitney U test for continuous variables and chi-square test for categorical variables, respectively. All two-tailed tests with a $P < 0.05$ were considered statistically significant.

Results

Characteristics of the enrolled T2DM patients

A flowchart of the study design is shown in Fig. 1. A total of 4259 patients' clinical data were collected from December 2013 to April 2024. All selected patients aged 20 years or older, with 57.62% (2454/4259) males and 42.38% (1805/4259) females. Among the 4259 cases, 47.26% (2013/4259) patients were positive cases clinically diagnosed with DR, while 52.74% (2246/4259) negative cases did not suffer from DR. The clinical data of 4259 inpatients was divided into two data sets, including 3936 cases in the model construction data set and 323 cases in the external validation data set. We divided the model construction data set into training and internal validation sets in a random ratio of 9:1.

We conducted feature selection methods to select the optimal feature subset from the training set, and then train it using five ML algorithms: XGBoost, SVM, GBDT, NN, and LR. By comparing the performance results of five ML models, XGBoost model showed the best performance. We also used SHAP method to interpret and analyze the impact of features on the prediction results of XGBoost model. Further evaluation of the model's generalization ability was conducted on an external validation set. For the 323 cases in the external validation data set, 136 patients were clinically diagnosed with DR, while 187 patients did not suffer from DR (Fig. 1).

In the model construction data set, 3936 cases included 2059 patients without DR (non-DR group) and 1877 patients with DR (DR group), encompassing a total of 76 variables. Subsequently, due to data missing greater than 30%, seven variables were removed, resulting in 69 variables remaining (Tables 1–4).

The basic characteristics of the 3936 cases in the model construction data set are shown in Table 1. The average age of the 3936 T2DM patients was 57.12 ± 11.86 years, with 2259 (57.39%) males and 1677 (42.61%) females. Among them, 1877 (47.69%) patients were diagnosed with DR (DR group), with 1053 (56.10%) males and 824 (43.90%) females. Patients of DR group exhibited higher levels of age, T2DM duration, Pulse, SBP, and DBP, compared to those of non-DR group. Conversely, patients of DR group had lower levels of BMI. There were no statistically significant differences in the levels of WHR.

The blood routine indicators of the model construction data set are shown in Table 2. Patients of DR group exhibited higher levels of neutrophil percentage (NEUT%), neutrophil count (NEUT#), neutrophil to lymphocyte ratio (NLR), mean platelet volume (MPV), platelet volume distribution width (PDW) and erythrocyte sedimentation rate (ESR), compared to those of non-DR group. Conversely, patients of DR group had lower levels of lymphocyte percentage (LYMPH%), lymphocyte count

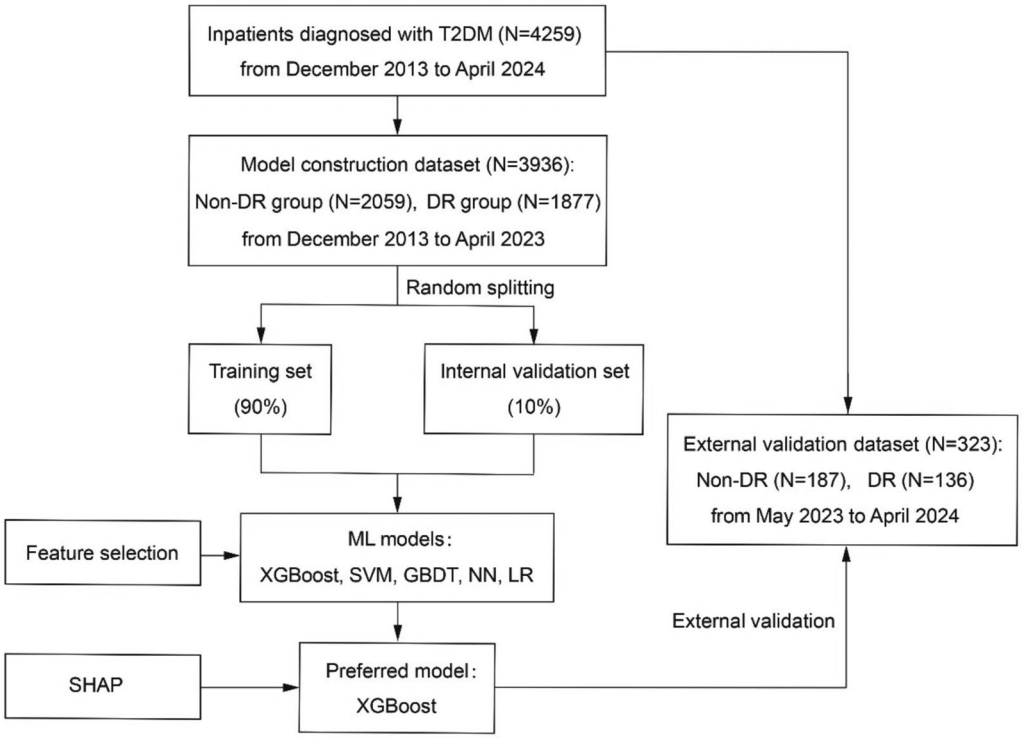

Fig. 1 Flowchart of study design. The non-DR group were defined as patients without DR, and DR group were patients with DR. T2DM, type 2 diabetes mellitus; DR, diabetic retinopathy; ML, machine learning; XGBoost, eXtreme Gradient Boosting; SVM, support vector machine; GBDT, gradient boosting decision tree; NN, neural network; LR, logistic regression; SHAP, Shapley Additive exPlanation

Table 1 Basic characteristics of the included T2DM patients in the model construction data set

Variables	Description	Total (*N* = 3936)	Non-DR group (*N* = 2059)	DR group (*N* = 1877)	*P*
Gender(Male/female)	Gender	2259/1677	1206/853	1053/824	0.117
Age(year-old)	Age of the patient	57.12 ± 11.86	56.44 ± 12.22	57.88 ± 11.39	< 0.001
T2DM duration (years)	Duration of T2DM	10.00 (5.00–17.00)	8.00(3.00–14.00)	14.00(10.00–20.00)	< 0.001
BMI (kg/m^2)	Body mass index	25.84 ± 3.70	25.98 ± 3.77	25.70 ± 3.62	0.018
Pulse (/min)	Pulse rate	74.51 ± 10.04	73.69 ± 9.81	75.40 ± 10.21	< 0.001
WHR	Waist to hip ratio	1.05 ± 3.03	1.07 ± 3.51	1.02 ± 2.38	0.568
SBP (mmHg)	Systolic blood pressure	131.14 ± 16.96	127.78 ± 15.33	134.82 ± 17.88	< 0.001
DBP (mmHg)	Diastolic blood pressure	77.48 ± 10.59	76.41 ± 10.30	78.33 ± 10.84	< 0.001

The continuous variables were expressed as mean ± standard deviation (SD) or the medians with interquartile ranges (IQRs) after the normality distribution test

The categorical variables were expressed as number

T2DM, type 2 diabetes mellitus; DR, diabetic retinopathy

(LYMPH#), red blood cell count (RBC), hemoglobin (HGB), hematocrit (HCT), mean corpuscular volume (MCV) and mean corpuscular hemoglobin (MCH).

The biochemical indicators of the model construction data set are shown in Table 3. Patients of DR group exhibited higher levels of glycated hemoglobin (HbA1c), potassium (K), fasting blood glucose (FBG), blood urea nitrogen (BUN), Cr, creatine kinase-MB (CKMB), lactate dehydrogenase (LDH), lipoprotein(a) (LP(a)), total cholesterol (TC), high-density lipoprotein cholesterol (HDL-C) and low-density lipoprotein cholesterol (LDL-C), compared to those of non-DR group. Conversely, patients of DR group had lower levels of calcium (CA), total protein (TP), albumin (ALB), total bilirubin (TBIL),

Table 2 Blood routine examination indicators of the included T2DM patients in the model construction data set

Indicators	Description	Total (*N* = 3936)	Non-DR group (*N* = 2059)	DR group (*N* = 1877)	*P*
WBC (10^9/L)	White blood cell count	6.44 ± 1.66	6.42 ± 1.59	6.46 ± 1.72	0.393
LYMPH% (%)	Lymphocyte percentage	33.42 ± 7.91	34.31 ± 7.63	32.45 ± 8.10	< 0.001
MONO% (%)	Monocyte percentage	7.69 ± 1.95	7.70 ± 1.94	7.68 ± 1.97	0.781
NEUT% (%)	Neutrophil percentage	55.70 ± 8.26	54.78 ± 7.93	56.71 ± 8.51	< 0.001
EOS% (%)	Eosinophils percentage	2.30(1.50–3.44)	2.30(1.54–3.50)	2.24(1.44–3.40)	0.263
BASO% (%)	Bashophils percentage	0.40(0.24–0.64)	0.40(0.24–0.64)	0.40(0.24–0.60)	0.677
LYMPH# (10^9/L)	Lymphocyte count	2.12 ± 0.67	2.18 ± 0.68	2.06 ± 0.66	< 0.001
MONO# (10^9/L)	Monocyte count	0.49 ± 0.16	0.49 ± 0.15	0.49 ± 0.16	0.719
NEUT# (10^9/L)	Neutrophil count	3.62 ± 1.23	3.54 ± 1.14	3.71 ± 1.32	< 0.001
NLR	Neutrophil to lymphocyte ratio	1.66 (1.30–2.16)	1.61(1.26–2.03)	1.75(1.36–2.29)	< 0.001
EOS# (10^9/L)	Eosinophils count	0.14(0.09–0.22)	0.14(0.09–0.22)	0.14(0.09–0.22)	0.464
BASO# (10^9/L)	Bashophils count	0.03 (0.01–0.04)	0.03(0.01–0.04)	0.03(0.01–0.04)	0.826
RBC (10^{12}/L)	Red blood cell count	4.52 ± 0.52	4.59 ± 0.49	4.45 ± 0.53	< 0.001
HGB (g/L)	Hemoglobin	135.37 ± 16.00	137.98 ± 15.13	132.52 ± 16.43	< 0.001
HCT	Hematocrit	0.40 ± 0.04	0.41 ± 0.04	0.39 ± 0.04	< 0.001
MCV (fL)	Mean corpuscular volume	89.55 ± 4.53	89.94 ± 4.51	89.13 ± 4.52	< 0.001
MCH (pg)	Mean corpuscular hemoglobin	30.02 ± 1.83	30.14 ± 1.84	29.90 ± 1.81	< 0.001
MCHC (g/L)	Mean corpuscular hemoglobin concentration	335.23 ± 11.61	335.07 ± 11.77	335.41 ± 11.44	0.359
RDW (%)	Red cell distribution width	12.67 ± 0.88	12.67 ± 0.90	12.68 ± 0.85	0.337
PLT (10^9/L)	Absolute platelet count	216.75 ± 58.48	217.19 ± 57.54	216.27 ± 59.51	0.624
MPV (fL)	Mean platelet volume	10.79 ± 0.96	10.74 ± 0.94	10.84 ± 0.98	0.001
PCT	Plateletcrit	0.23 ± 0.26	0.23 ± 0.06	0.23 ± 0.06	0.597
PDW (fL)	Platelet volume distribution width	12.97 ± 2.13	12.89 ± 2.07	13.06 ± 2.19	0.011
ESR (mm/h)	Erythrocyte sedimentation rate	14.00 (7.00–24.00)	12.00(7.00–19.00)	16.00(9.00–28.00)	< 0.001

The continuous variables were expressed as mean ± standard deviation (SD) or the medians with interquartile ranges (IQRs) after the normality distribution test

T2DM, type 2 diabetes mellitus; DR, diabetic retinopathy

direct bilirubin (DBIL), indirect bilirubin (IBIL), total bile acids (TBA), alanine aminotransferase (ALT), aspartate aminotransferases (AST), gamma-glutamine transferase (GGT) and hypersensitivity C-reactive protein (hsCRP).

The coagulation indicators of the model construction data set are shown in Table 4. Patients of DR group exhibited higher levels of prothrombin time activity percentage (PT%), fibrinogen (FIB) and D-dimer (DD), compared to those of non-DR group. Conversely, patients of DR group had lower levels of prothrombin time (PT) and international normalized ratio (INR). There were no statistically significant differences in the levels of activated partial thromboplastin time (APTT) between DR group and non-DR group.

Feature selection

Initially, based on the statistical analysis results of 68 variables in the non-DR and DR groups, 55 variables were selected. Subsequently, the random forest method was employed to rank the importance of the 55 variables. The results indicated that a significant portion of the T2DM duration variable influenced the performance of the subsequent models. Consequently, the T2DM duration variable was excluded, and ultimately, 54 variables were retained for analyzing the data of non-DR and DR groups. All 24 blood routine indicators were included in the selected 54 variables due to their ease of accessibility and broad clinical utility. In addition, the retained variables encompassed 6 basic characteristics, 20 biochemical indicators, and 4 coagulation indicators. The data distribution characteristics and correlation of 54 variables in T2DM patients of non-DR and DR groups were shown through violin plots, butterfly chart, and heatmap in Fig. 2.

The violin plots were utilized to visualize the distribution of data and compare the differences between the non-DR and DR groups. Analysis of the violin plots reveals slight differences in SBP, HbA1c, TBA, ALT, AST, and ESR between the two groups, while the variations in other parameters are not significant (Fig. 2A).

The butterfly chart was used to illustrate the normalized median values of various indicators for two groups,

Table 3 Biochemical indicators of the included T2DM patients in the model construction data set

Indicators	Description	Total (*N* = 3936)	Non-DR group (*N* = 2059)	DR group (*N* = 1877)	*P*
HbA1c (%)	Glycated hemoglobin	8.98 ± 1.97	8.74 ± 2.00	9.24 ± 1.90	< 0.001
K (mmol/L)	Potassium	3.95 ± 0.37	3.90 ± 0.34	4.00 ± 0.39	< 0.001
NA (mmol/L)	Sodium	139.68 ± 2.36	139.75 ± 2.29	139.60 ± 2.44	0.047
CL (mmol/L)	Chlorine	104.67 ± 2.87	104.61 ± 2.74	104.73 ± 3.01	0.166
CA (mmol/L)	Calcium	2.27 ± 0.10	2.27 ± 0.99	2.26 ± 0.10	0.001
PHOS (mmol/L)	Phosphorus	1.27 ± 0.18	1.28 ± 0.18	1.27 ± 0.18	0.436
FBG(mmol/L)	Fasting blood glucose	7.96 ± 2.74	7.66 ± 2.52	8.30 ± 2.92	< 0.001
BUN (mmol/L)	Blood urea nitrogen	5.36 ± 1.96	4.98 ± 1.58	5.79 ± 2.23	< 0.001
Cr (μmol/L)	Creatinine	70.87 ± 25.71	68.12 ± 17.82	73.90 ± 31.95	< 0.001
URIC (μmol/L)	Uric acid	333.89 ± 86.48	332.41 ± 86.25	335.51 ± 86.72	0.260
TP (g/L)	Total protein	64.58 ± 5.50	65.01 ± 5.25	64.11 ± 5.73	< 0.001
ALB (g/L)	Albumin	38.66 ± 3.84	39.43 ± 3.47	37.82 ± 4.05	< 0.001
TBIL (μmol/L)	Total bilirubin	13.99 ± 5.45	14.78 ± 5.56	13.13 ± 5.18	< 0.001
DBIL (μmol/L)	Direct bilirubin	2.34 ± 1.10	2.51 ± 1.10	2.16 ± 1.06	< 0.001
IBIL (μmol/L)	Indirect bilirubin	11.65 ± 4.64	12.28 ± 4.77	10.97 ± 4.40	< 0.001
TBA (μmol/L)	Total bile acids	3.30 (2.20–5.10)	3.40(2.30–5.20)	2.10(3.30–5.00)	0.003
ALT (U/L)	Alanine aminotransferase	19.00(14.00–27.00)	21.00(15.00–30.00)	18.00(13.00–24.00)	< 0.001
AST (U/L)	Aspartate aminotransferases	20.00 (16.00–26.00)	21.00(17.00–28.00)	19.00(16.00–25.00)	< 0.001
ALP (U/L)	Alkaline phosphatase transferase	70.64 ± 22.18	70.54 ± 21.65	70.70 ± 22.75	0.815
GGT (U/L)	Gamma-glutamyltransferase	24.00 (17.00–35.00)	25.00(18.00–37.00)	22.00(17.00–32.00)	< 0.001
CK (U/L)	Creatine kinase	74.00(54.00–105.00)	73.00(54.00–103.00)	75.00(54.00–108.00)	0.253
CKMB (U/L)	Creatine kinase-MB	11.79 ± 9.11	11.46 ± 6.58	12.15 ± 11.24	0.019
LDH (U/L)	Lactate dehydrogenase	155.19 ± 36.11	151.89 ± 34.09	158.80 ± 37.88	< 0.001
hsCRP (mg/L)	Hypersensitivity C-reactive protein	1.14 (0.50–2.81)	1.22(0.54–2.96)	1.04(0.45–2.70)	0.003
LP(a) (mg/dL)	Lipoprotein(a)	14.63 (6.71–29.29)	13.10(6.33–26.60)	16.30(7.27–32.55)	< 0.001
TG (mmol/L)	Triglyceride	1.54 (1.10–2.23)	1.51(1.10–2.16)	1.56(1.10–2.34)	0.243
TC (mmol/L)	Total cholesterol	4.51 ± 1.14	4.44 ± 1.06	4.59 ± 1.21	< 0.001
HDL-C (mmol/L)	High-density lipoprotein cholesterol	1.05 ± 0.31	1.03 ± 0.30	1.06 ± 0.31	0.007
LDL-C (mmol/L)	Low-density lipoprotein cholesterol	2.75 ± 0.94	2.71 ± 0.87	2.80 ± 1.01	0.002
CO_2CP (mmol/L)	Carbon dioxide binding capacity	25.96 ± 2.22	25.91 ± 2.12	26.01 ± 2.32	0.189

The continuous variables were expressed as mean ± standard deviation (SD) or the medians with interquartile ranges (IQRs) after the normality distribution test; T2DM, type 2 diabetes mellitus; DR, diabetic retinopathy

Table 4 Coagulation indicators of the included T2DM patients in the model construction data set

Indicators	Description	Total (*N* = 3936)	Non-DR group (*N* = 2059)	DR group (*N* = 1877)	*P*
PT (s)	Prothrombin time	11.36 ± 0.95	11.41 ± 1.02	11.31 ± 0.86	0.001
PT% (%)	Prothrombin time activity percentage	106.89 ± 13.59	106.21 ± 13.46	107.63 ± 13.68	0.001
INR	International normalized ratio	0.96 ± 0.08	0.97 ± 0.09	0.96 ± 0.08	0.001
APTT (s)	Activated partial thromboplastin time	25.66 ± 3.01	25.65 ± 2.89	25.67 ± 3.14	0.831
FIB (g/L)	Fibrinogen	3.02 ± 0.79	2.87 ± 0.68	3.18 ± 0.87	< 0.001
DD (mg/L)	D-dimer	0.20(0.13–0.34)	0.18(0.12–0.29)	0.23(0.14–0.37)	< 0.001

The continuous variables were expressed as mean ± standard deviation (SD) or the medians with interquartile ranges (IQRs) after the normality distribution test

T2DM, type 2 diabetes mellitus; DR, diabetic retinopathy

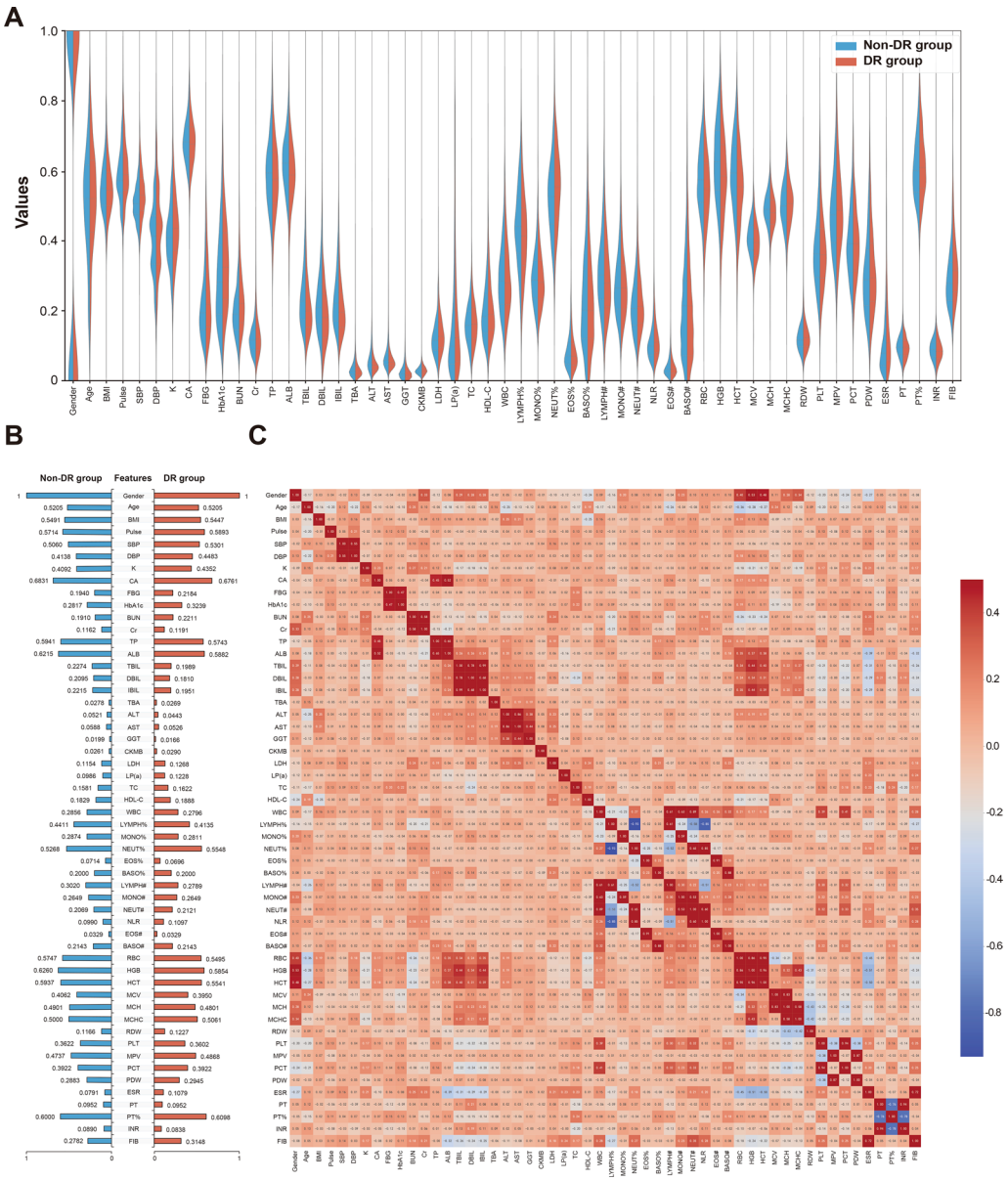

Fig. 2 Data distribution characteristics and correlation of variables in T2DM patients of non-DR and DR groups from the model construction data set. **A** Violin plots were employed to visualize the data distribution of the non-DR (blue area) and DR (yellow area) groups. The width of the plot indicates data density, with broader sections signifying higher density. The shape of the plot reveals the data distribution, offering insights into its spread and concentration. **B** Butterfly chart was used to display the normalized median values of different indicators for the non-DR (blue area) and DR (yellow area) groups. **C** Correlation between the variables was analyzed by a heatmap. Abbreviations as in Tables 1–4

with values ranging between 0 and 1 (Fig. 2B). The levels of Pulse, SBP, DBP, HbA1c, K, FBG, BUN, LDH, LP(a), TC, NEUT%, MPV, PDW, ESR, and FIB in patients of the DR group are significantly higher than those in the non-DR group. Conversely, levels of CA, TP, ALB, ALT, AST, TBIL, DBIL, IBIL, LYMPH%, LYMPH#, RBC, HGB, HCT, and MCV in patients of the DR group were significantly lower than those in the non-DR group.

The correlation between variables was analyzed using a heatmap (Fig. 2C), with different block colors representing correlation coefficients to determine the magnitude of correlation between variables. It is important to note that correlation coefficients can only measure linear correlations between variables. From the heatmap, a linear relationship can be observed between LYMPH% and NLR, LYMPH% and NEUT%, PT and PT%, and PT% and

INR, while the linear relationship between the remaining variables is relatively insignificant.

To identify the key features for predicting DR, the Random Forest algorithm was utilized for feature selection and contribution calculation. It ranked the initial 54 variables and selected the most relevant ones, followed by the removal of low-contribution variables based on this ranking. The feature selection accuracy curve (Fig. 3) demonstrated that the model achieved its highest accuracy when the number of variables was 39. The optimal 39 variables which chosen to establish the prediction model included 5 basic characteristics (SBP, age, BMI, pulse and gender), 16 blood routine indicators (HCT, MCV, PLT, NEUT%, LYMPH%, EOS%, PDW, MCHC, MONO%, MCH, RDW, RBC, WBC, NEUT#, MONO# and BASO%), 16 biochemical indicators (BUN, HbA1c, ALT, FBG, ALB, K, TBIL, LDH, Cr, TBA, LP(a), TC, DBIL, HDL-C, GGT and TP), and 2 coagulation indicators (FIB and PT%) (Fig. 4). Based on the feature importance analysis, SBP, FIB, BUN, HbA1c, and ALT was highly ranked. In addition, FBG, ALB, K, HCT, and TBIL were identified as potential strong predictors for DR.

Hyperparameters utilizing the GridSearchCV method

The hyperparameters for each of the five ML models were determined through fivefold cross-validation using GridSearchCV method. The optimal configurations are as follows: (1) XGBoost: n_estimators = 80, max_depth = 16, learning_rate = 0.05, min_child_weight = 1. (2) SVM: C = 10, gamma = 0.001, kernel = 'rbf'. (3) GBDT: learning_rate = 0.1, n_estimators = 100, max_depth = 4. (4) NN: num_layers = 4, units = 128, epochs = 80. (5) LR: C = 1, penalty = 'l2'. These hyperparameter combinations were identified as the best-performing during the grid search process, significantly improving the predictive accuracy of the models. The parameter grid and the optimal outcomes from GridSearchCV for all models are summarized in Table 5.

Model performance based on the internal validation set

As shown in Table 6 and Fig. 5, the model's performance on the internal validation set was evaluated primarily through predicted evaluation indicators and the ROC curve. Table 6 presents the discriminative performance of the algorithms using tenfold cross-validation with a 9:1 ratio for training and internal validation. Figure 5 illustrates the performance of five different models. Among these, the XGBoost algorithm exhibited the highest performance with an AUC of 0.831. The AUCs for the other models, SVM, GBDT, NN, and LR were 0.800, 0.811, 0.726, and 0.803, respectively.

Assessment of DR influencing predictors by SHAP

SHAP is a powerful method for interpreting the output of ML model, regardless of its complexity. It calculates the mean prediction (base value) and quantifies

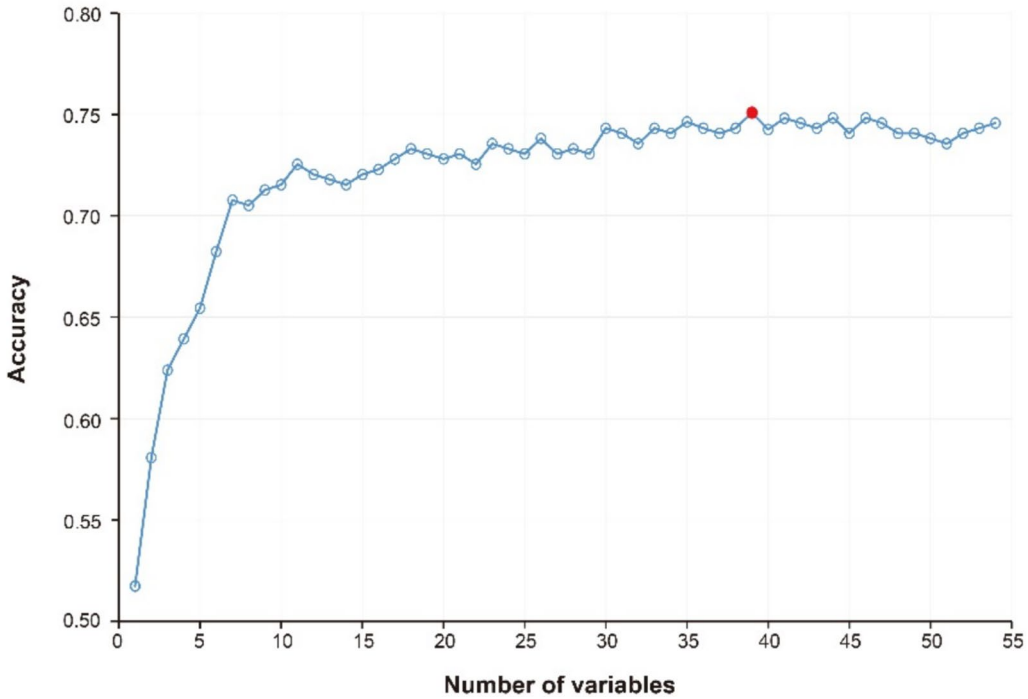

Fig. 3 Feature selection accuracy curve. The accuracy got the highest value when the number of variables was 39 (represented as a red solid point)

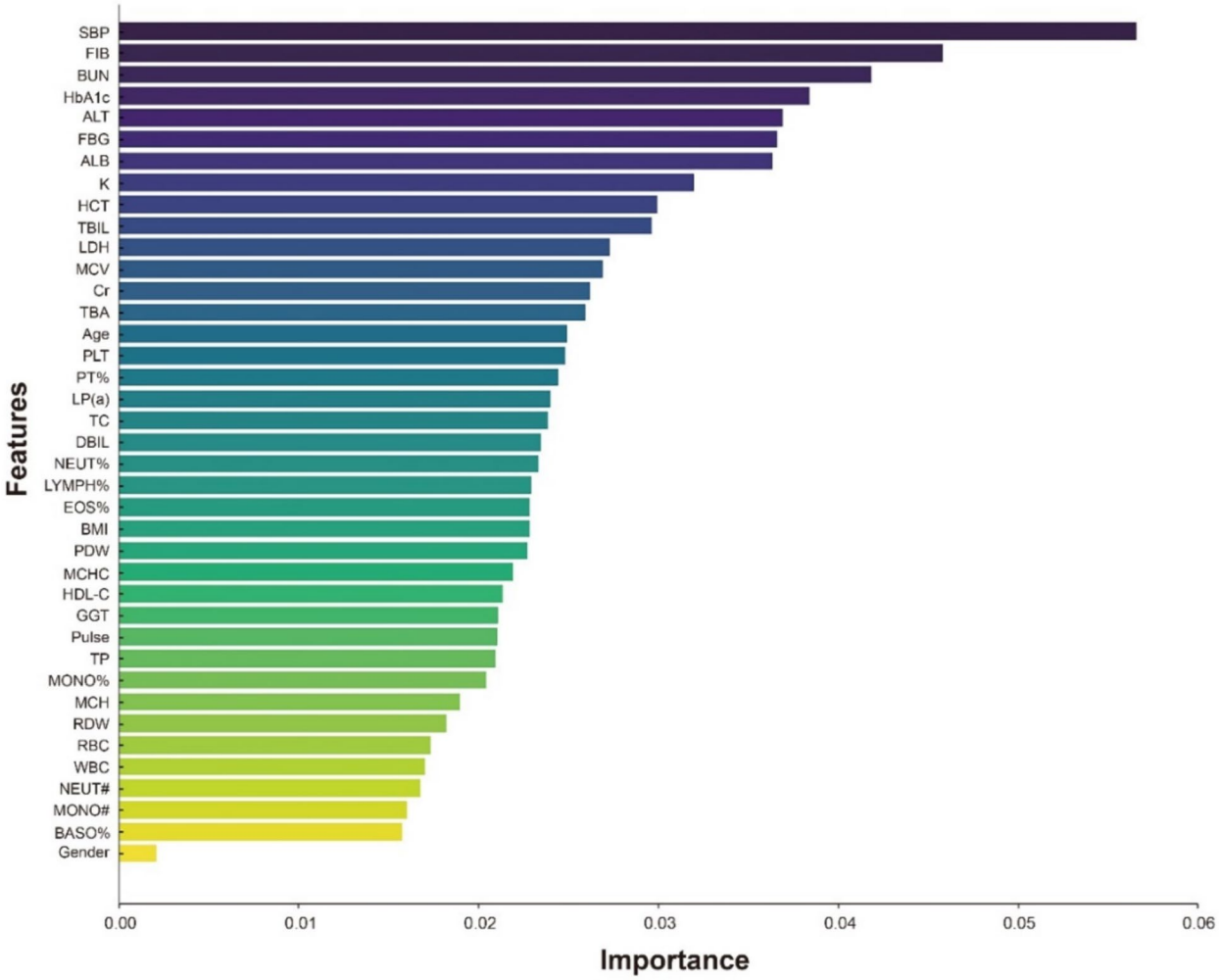

Fig. 4 Feature importance ranking of diabetes retinopathy risk based on Random Forest algorithm in T2DM patients in the model construction data set. The longer the bar chart, the greater the impact of the variable on the prediction results, and the more valuable it is for decision-making reference. Abbreviations as in Tables 1–4

each feature's contribution to deviations from this base, providing both global and local explanations [23, 35, 36]. For the XGBoost model, global feature importance is depicted in Fig. 6. In Fig. 6A, the *y*-axis lists features, such as "SBP," "BUN," and "HbA1c," with longer bars indicating greater influence. The *x*-axis shows the mean absolute SHAP value, representing the average impact of each feature on the model's output.

The inset PieDonut chart categorizes features (outer ring) and details individual variable contributions (inner ring) (Fig. 6A). Within the "Basic characteristics" category, SBP demonstrates the highest feature contribution. For "Biochemical indicators" category, BUN, HbA1c, ALB, K, and ALT show significant contributions. MCV is the most influential feature in the "Blood routine indicators" category, while FIB leads in the "Coagulation indicators" category. The outer pie chart reveals that "Biochemical indicators" category account for 52.01% of the total contribution, followed by "Blood routine indicators" category at 26.63%. The "Basic characteristics" category contributes 15.76%, while "Coagulation indicators" have the smallest contribution at 5.60%. The inner circle further breaks down the contributions of specific features within each category, providing an intuitive understanding of their importance to the model's predictions.

As shown in Fig. 6B, the SHAP summary plot reveals that higher SHAP values correlate with a greater likelihood of DR. Red dots represent higher feature values, and blue dots denote lower values, helping identify key features and improve model interpretability. Among the 39 predictive factors, the top 10 predictive factors were SBP, BUN, HbA1c, ALB, K, ALT, MCV, FIB, FBG, and LDH (Fig. 6).

Table 5 Parameter grid and GridSearchCV best outcome for the five machine learning models

Classifier	Hyperparameters	Possible values	GridSearchCV outcome
XGBoost	n_estimators	80, 100, 150, 200	80
	max_depth	10, 12, 14, 16	16
	learning_rate	0.05, 0.1, 0.2	0.05
	min_child_weight	1, 3, 5	1
SVM	C	0.01, 0.1, 1, 10, 100	10
	gamma	0.001, 0.01, 0.1, 1, 10, 100	0.001
	kernel	rbf, linear, poly, sigmoid	rbf
GBDT	learning_rate	0.08, 0.09, 0.1, 0.11, 0.12	0.1
	n_estimators	60, 80, 100, 120	100
	max_depth	2, 3, 4, 5, 6	4
NN	num_layers	2, 3, 4, 5	4
	units	32, 64, 128, 256	128
	epochs	60, 80, 100, 120	80
LR	C	0.001, 0.01, 0.1, 1, 10, 100	1
	penalty	l1, l2	l2

XGBoost, eXtreme Gradient Boosting; SVM, Support Vector Machine; GBDT, Gradient Boosting Decision Tree; NN, Neural Network; LR, Logistic Regression; n_estimators, number of estimators; max_depth, maximum depth; min_child_weight, minimum child weight; learning_rate, learning rate; num_layers, number of layers

Table 6 Predictive performance of machine learning models for diagnosing diabetic retinopathy in T2DM patients in the model construction data set

Models	AUC	Accuracy	Sensitivity	Specificity	F1-Score
XGBoost	0.831	0.757	0.754	0.759	0.752
SVM	0.800	0.716	0.721	0.711	0.713
GBDT	0.811	0.751	0.737	0.765	0.743
NN	0.726	0.678	0.670	0.685	0.670
LR	0.803	0.719	0.737	0.701	0.719

AUC, area under the receiver operating characteristic (ROC) curve

Abbreviations as in Tables 1 and 5

For local explanations, the SHAP Force Plot illustrates how individual features influence predictions for specific instances. Features pushing predictions higher (red) or lower (blue) relative to the base value are highlighted. The value of f(x) represents the predicted log-odds of the outcome for that instance [36]. Two examples of the local explanation of the predictions using SHAP values are shown in Fig. 7. In Fig. 7A, features such as SBP, Pulse, ALT, FBG, DBIL, and FIB collectively reduce the prediction, while in Fig. 7B, features such as HCT, TC, TBIL, DBIL, BUN, SBP, and FIB collectively increase it.

In addition, the SHAP dependence plot of the top-10 important features demonstrates the impact of individual features on the XGBoost model's output (Fig. 8). The y-axis represents the SHAP values for the variables, while the x-axis shows the standardized variable values. The blue dots represent the eigenvalues and the SHAP values corresponding to each observation. The red line represents the SHAP values equal to zero. A SHAP value greater than zero suggests an increased risk of developing DR. As shown in Fig. 8, SHAP values exhibit a clear upward trend with increases in SBP, BUN, FIB, FBG, and LDH, while ALB, ALT, and MCV show a downward trend. HbA1c and K display an initial rise followed by a decline in SHAP values.

External validation of XGBoost model

In this study, the external validation data set included 323 medical records of T2DM patients in Beijing Tongren Hospital from May 2023 to April 2024, and each patient had only one medical record. The selected data dimension was consistent with the internal validation set dimension. This data set included 136 clinically diagnosed DR patient records and 187 non-DR patient records. After preprocessing the external validation data set, the XGBoost model was used to test the data set. Using the XGBoost model, 86 DR patients were correctly predicted as DR in the 136 clinically diagnosed DR patients. Similarly, for 187 non-DR patients, 118 non-DR patients were accurately predicted as non-DR. In the external validation data set, the AUC, accuracy, sensitivity, specificity, and F1-Score were 0.709, 0.650, 0.669, 0.636, and 0.652, respectively (Table 7).

Discussion

In this study, five ML models were developed to evaluate the risk of DR in patients with T2DM based on routine laboratory tests. The XGBoost model demonstrated superior performance compared to the others. The variables selected for this study are commonly available, as they are typically derived from routine blood tests. We recommend utilizing this model to generate predictions and to guide specific DR examinations.

In this study, 47.26% (2013/4259) of T2DM patients were diagnosed with DR, a rate significantly higher than previously reported in the general Chinese population [37, 38]. This discrepancy likely arises, because the cases were drawn from a leading comprehensive hospital in China specializing in ophthalmic disease diagnosis and treatment. Many diabetes patients are admitted to the endocrine department to stabilize blood sugar levels before undergoing ophthalmic examinations and treatments [11]. In China, there are 42 million DR patients, and the high severity and prevalence of DR underscore

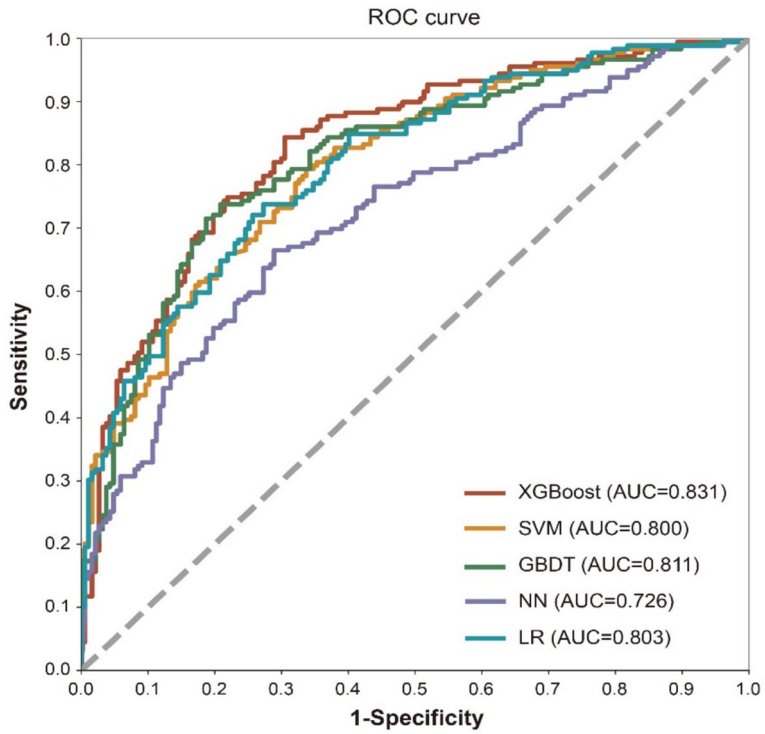

Fig. 5 Receiver operating characteristic (ROC) curves of five algorithms for detecting diabetic retinopathy based on 39 important variables in T2DM patients from the model construction data set. Abbreviations as in Tables 1, 5, and 6

the urgent need for early screening. According to the National Committee for the Prevention of Blindness, there were only 44,800 ophthalmologists in China in 2018, a number insufficient to meet the growing demand for DR care [37, 38]. Another key factor contributing to the low screening rate is the lack of awareness among T2DM patients about retinopathy, as early DR symptoms are often subtle. However, DR-related blindness can occur suddenly, leaving many asymptomatic patients undiagnosed. This highlights the critical need for enhanced screening programs and public awareness initiatives to address this significant healthcare challenge.

Routine laboratory tests, when integrated with ML techniques, can enhance the prediction of new outcomes and aid in identifying novel diagnostic markers. Researchers have developed various ML-based prediction models to assist physicians in assessing DR risk [10, 12, 18, 36]. Tsao et al. [10] identified high-risk DR patients and pinpointed ten predictive features, such as family history of diabetes, exercise habits, and insulin treatment. Yang et al. [12] analyzed data from 1,418 diabetic patients among 8,952 rural residents to develop and validate their model, identifying 10 critical features (HbA1c, TG, Cr, SBP, BMI, age, diabetes duration, educational level, hypertension duration, and income level). Zhao et al. [18] analyzed data from 7,943 T2DM patients, creating nomograms to predict DR. They identified key laboratory tests, such as GGT, AST, and HbA1c, alongside significant clinical features, including insulin use, diuretics, statins, hypertension, smoking status, and drinking status. Similarly, Li et al. [36] constructed a DR risk prediction model using the XGBoost algorithm, incorporating 17 indicators, including age, nephropathy, insulin treatment, diabetic lower extremity arterial disease (DLEAD), and 13 routine laboratory tests. These studies highlight the significance of laboratory data and clinical diagnostic information in modeling DR risk. However, the practical application of such models may be limited due to the restricted availability of many clinical variables. In contrast, this study primarily utilizes routine laboratory test results, which are easier to obtain, objective, and more feasible for clinical implementation.

In this study, over 15% of patients reported a T2DM duration of less than 2 years, likely reflecting delayed diagnosis following the onset of DR symptoms, especially in underdeveloped regions of China. Although T2DM duration is a known risk factor for DR [10, 11], its reference value is limited and disproportionately overshadowed other critical indicators during random forest feature selection in this study. In addition, the T2DM duration may not be accurate, as many T2DM patients remain undiagnosed for years. To enhance model

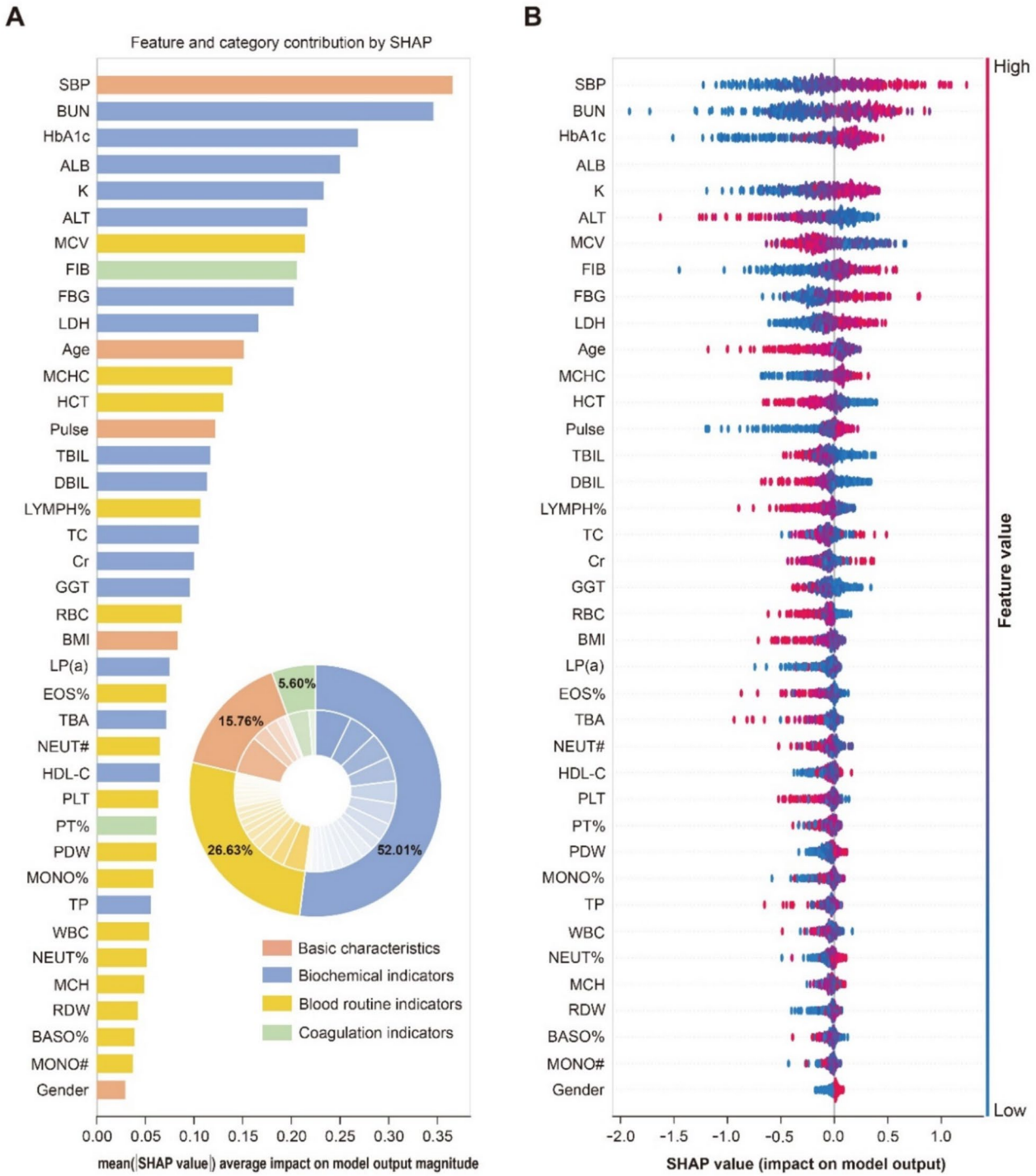

Fig. 6 SHAP explained global feature importance for XGBoost model. **A** Bar chart of the mean absolute SHAP value for each predictor. The inset PieDonut contains categorized features (out ring) and single variable contributions (inner ring). **B** SHAP summary plot. The dot's color represents the magnitude of the feature value, with red denoting higher values and blue indicating lower values. Its horizontal position corresponds to the SHAP value, reflecting the direction and strength of the feature's influence on the model's output. SHAP, Shapley Additive exPlanation; Abbreviations as in Tables 1–5

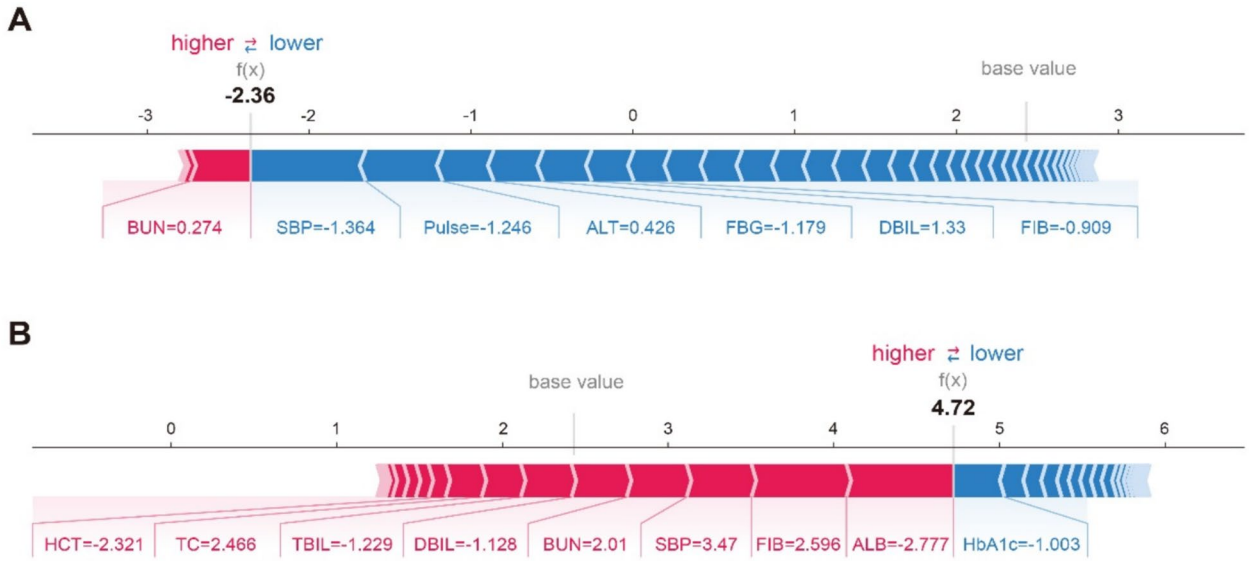

Fig. 7 Two examples of the local explanation of the predictions using the Shapley Additive exPlanation (SHAP) values. **A** Predicted T2DM patient without diabetic retinopathy. **B** Predicted T2DM with diabetic retinopathy. Factors that push the predicted score higher compared to the base value (mean prediction) are colored red, and those pushing lower the prediction are shown in blue. SHAP, Shapley Additive exPlanation; Abbreviations as in Tables 1–5

accuracy and practicality, T2DM duration was excluded from the final model, ensuring a balanced representation of multifactorial DR risks. Despite this exclusion, the model demonstrated robust predictive performance, strong discriminative ability, and reliable calibration in both test and external validation sets, confirming its effectiveness even without T2DM duration data.

The XGBoost model outperformed other algorithms in this study, achieving the highest AUC value of 0.831. This success is attributed to its ensemble learning approach, efficient handling of large-scale data, adaptability to data distribution and noise, strong regularization capabilities, and interpretability through feature importance evaluation [36, 39]. By combining multiple decision trees, processing high-dimensional features, and preventing overfitting, XGBoost proved particularly effective for analyzing high-dimensional DR data in other studies [18, 23]. For instance, Zhao et al. [18] evaluated DR data from 7,943 participants using 31 predictors and demonstrated that XGBoost outperformed RF, LR, SVM, and K-Nearest Neighbors (KNN), achieving an AUC of 0.803, with accuracy, sensitivity, and specificity rates of 88.9%, 74.0%, and 81.1%, respectively. Similarly, Islam et al. [23] analyzed data from 6,374 Chinese participants and confirmed that XGBoost surpassed other models, achieving an AUC of 0.850, further validating its superior predictive performance for DR. These attributes make XGBoost an ideal algorithm for early DR screening and related applications.

The SHAP value effectively renders the output of the XGBoost model clinically interpretable [36]. In this study, the proposed XGBoost-based model, incorporating SHAP values, identified SBP as the most significant risk factor for DR. This correlation between SBP and DR development is well-documented in existing literature [39–41]. Karoli et al. [39] demonstrated that hypertension and elevated SBP are commonly observed in patients with DR. Individuals with elevated SBP face an increased risk of developing DR. In hypertensive patients, abnormal retinal autoregulation impairs their ability to buffer changes in blood pressure exacerbated by hyperglycemia in DM patients, further compromising the regulation of retinal perfusion [40]. A comprehensive population-based cross-sectional survey in China indicated that high blood pressure accelerates the progression of DR, while lower blood pressure decelerates it [41].

In this study, the XGBoost model with SHAP values confirmed that HbA1c is a critical parameter for DR. Poor diabetes management leading to elevated HbA1c levels significantly increases the likelihood of developing DR [42, 43]. Oxidative stress and inflammation caused by high glucose levels exacerbate damage to the vascular endothelium and tissues, collectively advancing the pathogenesis and progression of DR [44, 45]. Furthermore, elevated HbA1c levels mediate damage to vascular endothelial cells and promote leukocyte adhesion, which enhances thrombosis risk. Consequently, patients with high HbA1c levels are at a considerably increased

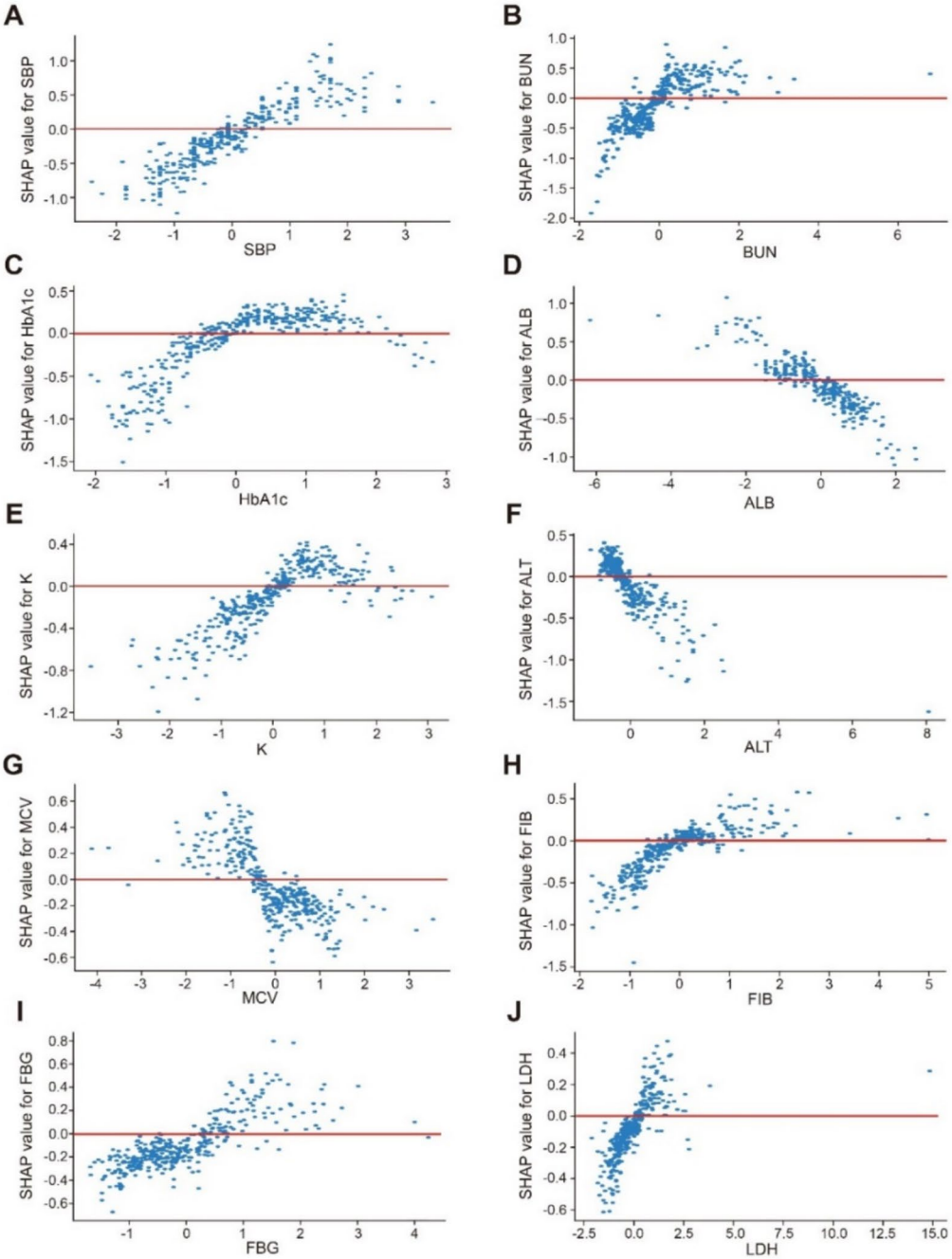

Fig. 8 SHAP dependence plots of top 10 important features in XGBoost model. **A** SHAP dependence plots of SBP; **B** SHAP dependence plots of BUN; **C** SHAP dependence plots of HbA1c; **D** SHAP dependence plots of ALB; **E** SHAP dependence plots of K; **F** SHAP dependence plots of ALT; **G** SHAP dependence plots of MCV; **H** SHAP dependence plots of FIB; **I** SHAP dependence plots of FBG; **J** SHAP dependence plots of LDH. The blue dots represent the eigenvalues and the SHAP values corresponding to each observation. The red line represents the SHAP values equal to zero. SHAP, Shapley additive explanation; Abbreviations as in Tables 1–5

risk for DR, and there is a positive correlation between the severity of DR and HbA1c levels [41]. However, clinical observations indicate that some patients with long-term controlled HbA1c still develop DR, suggesting that HbA1c is not the sole major risk factor [46].

In this study of data set of 3,936 records, a 9:1 split ratio was adopted to allow the model to fully utilize the data for training, learn rich features, and enhance its generalization capability. Although the test set is relatively small (394 records, accounting for 10%), its distribution

Table 7 Diabetic retinopathy predictive performance of XGBoost in the model construction and external validation data sets

Data sets	AUC	Accuracy	Sensitivity	Specificity	F1-Score
Model construction data set	0.831	0.757	0.754	0.759	0.752
External validation data set	0.709	0.650	0.669	0.636	0.652

XGBoost, eXtreme Gradient Boosting; AUC, area under the receiver operating characteristic (ROC) curve

is consistent with the overall data set, making it sufficient to evaluate the model's performance. The external validation data set, similar in size to the test set, aids in verifying the model's robustness. In ML, a model's performance on the internal test set is typically better than on the external validation set. The external validation set comes from a different source and may differ in feature distribution, noise levels, or other factors, challenging the model's generalization ability. The potential improvements are as follows: expand the size and diversity of the external validation set to better align with real-world data distributions and employ domain adaptation techniques to reduce the distribution gap between the training set and the external validation set.

The application of ML in medical diagnostics holds immense potential but also presents significant ethical challenges, particularly concerning the fairness of model predictions. These challenges primarily include: (1) data bias and fairness; (2) transparency and interpretability; (3) privacy and data security; (4) informed consent and autonomy; (5) resource allocation and accessibility; and (6) continuous monitoring and updates. While the prospects for ML in medical diagnostics are vast, it is crucial to address these ethical challenges carefully, ensuring that the technology is fair, transparent, secure, and respectful of patients' rights and autonomy [2, 23, 36].

The discussion of the aforementioned content highlights several significant advantages of this study. First, all variables used in this study were derived from easily accessible non-ocular examinations and questionnaires. The selected variables were primarily based on the most frequently conducted routine laboratory tests. Furthermore, only four basic characteristics (gender, age, BMI, and SBP) are required, making these data readily obtainable. The ease of data accessibility enables the integration of this model into electronic health record systems, allowing for the real-time identification of high-risk DR patients and providing stronger practical applicability. This model is particularly suitable for primary hospitals and diabetic clinics that lack expensive laboratory facilities and specialist ophthalmic equipment, proving especially beneficial in medically underserved areas. Second, our predictive model could encourage both doctors and patients to emphasize primary and secondary prevention of DR and increase the rate of fundus screening among high-risk individuals. Finally, utilizing SHAP values for interpretation guarantees that the features included in our ML model are both homogeneous in origin and readily accessible. This establishes a robust foundation for developing predictive models for clinical use.

Nevertheless, there were some limitations to this study. First, there is a potential for selection bias, as our analysis was based on hospital data, whereas population-based data might more accurately reflect the realities of a DR screening program. Second, this was a cross-sectional study, and we only evaluated the static state of indicators 4259 inpatients with T2DM. Conducting a longitudinal follow-up to observe the dynamic changes in these indicators could improve the accuracy of the nomogram. Third, due to the relatively small size of the external validation data set, there may be issues of overfitting. Future work will focus on multi-center or longitudinal research. This study did not compare models incorporating additional clinical features, such as imaging data. In future research, we plan to include additional clinical features, such as imaging and clinical diagnostics, to individualize different models for various populations, aiming to improve the identification rate of high-risk DR individuals. Finally, although the stability of our prediction model has been confirmed within this population, it has yet to be validated in other regions or countries.

Overall, the results observed in this study are promising. The main objective of our prediction model is to aid physicians in managing the health of T2DM patients and to optimize the fundus screening rate among them.

Conclusion

In this study, utilizing routine laboratory tests, we developed and evaluated a ML-based model for predicting DR risk in patients with T2DM. In addition, we demonstrated the effectiveness of XGBoost models in enabling clinicians to accurately identify individuals at high risk of DR and develop personalized management strategies, ultimately aiming to reduce the progression and incidence of DR, especially in medically underserved areas.

Abbreviations

T2DM	Type 2 diabetes mellitus
ML	Machine learning
XGBoost	EXtreme Gradient Boosting Tree
SVM	Support Vector Machine

GBDT	Gradient Boosting Decision Tree
NN	Neural Network
LR	Logistic Regression
AUC	Area under the receiver operating characteristic (ROC) curve
SHAP	Shapley Additive exPlanation

Acknowledgements
Not applicable.

Author contributions
X. W.and R.Z. searched the literature, collected data, analyzed data, drawed figures, wrote the manuscript, reviewed and edited the manuscript. Y. W. and W.W. collected data, analyzed data, reviewed and edited the manuscript.B.S. designed the study, analyze data, analyzed data, reviewed and edited the manuscript. L.Z. and Y. H. designed the study, analyze data, analyzed data, reviewed and edited the manuscript. All authors have accepted responsibility for the entire content of this manuscript and approved its submission.

Funding
This work was supported by a Grant from Capital's Funds for Health Improvement and Reseach (2024–3-1181). This study was also supported by funding from the National Natural Science Foundation of China (82372304).

Data availability
The data supporting the results of this study are available upon reasonable request from the corresponding author.

Declarations

Ethics approval and consent to participate
This study was approved by the ethics committee of Beijing Tongren Hospital, Capital Medical University (No. TREC2024-KY040).

Human ethics and consent to participate
Every human participant agreed to participate in the study and signed an informed consent form.

Competing interests
The authors declare no competing interests.

Received: 10 November 2024 Accepted: 9 March 2025
Published online: 18 March 2025

References

1. Vision Loss Expert Group of the Global Burden of Disease Study; GBD. Blindness and Vision Impairment Collaborators (2024) Global estimates on the number of people blind or visually impaired by diabetic retinopathy: a meta-analysis from 2000 to 2020. Eye (Lond). 2019;38(11):2047–57. https://doi.org/10.1038/s41433-024-03101-5.
2. Hou X, Wang L, Zhu D, Guo L, Weng J, Zhang M, et al. Prevalence of diabetic retinopathy and vision-threatening diabetic retinopathy in adults with diabetes in China. Nat Commun. 2023;14(1):4296. https://doi.org/10.1038/s41467-023-39864-w.
3. Teo ZL, Tham YC, Yu M, Chee ML, Rim TH, Cheung N, et al. Global prevalence of diabetic retinopathy and projection of burden through 2045: systematic review and meta-analysis. Ophthalmology. 2021;128(11):1580–91. https://doi.org/10.1016/j.ophtha.2021.04.027.
4. Chinese Diabetes Society. Guideline for the prevention and treatment of type 2 diabetes mellitus in China (2020 edition). Chin J Diabetes Mellit. 2021;13(4):315–409. https://doi.org/10.3760/cma.j.cn115791-20210221-00095.
5. Fundus Disease Group of Ophthalmological Society of Chinese Medical Association, Fundus Disease Group of Ophthalmologist Branch of Chinese Medical Doctor Association. Evidence-based guidelines for diagnosis and treatment of diabetic retinopathy in China (2022). Chin J Ocul Fundus Dis. 2023;39(2):99–124. https://doi.org/10.3760/cma.j.cn511434-20230110-00018.
6. Chen SH, Wang ZK, Yao B, Liu TM. Prediction of diabetic retinopathy using longitudinal electronic health records. In: Chen SH, editor. 2022 IEEE 18th international conference on automation science and engineering (CASE), Mexico. Mexico: IEEE; 2022. p. 949–54.
7. Solomon SD, Chew E, Duh EJ, Sobrin L, Sun JK, VanderBeek BL, et al. Diabetic retinopathy: a position statement by the American Diabetes Association. Diabetes Care. 2017;40(3):412–8. https://doi.org/10.2337/dc16-2641.
8. Wylie-Rosett J, Basch C, Walker EA, Zybert P, Shamoon H, Engel S, et al. Ophthalmic referral rates for patients with diabetes in primary-care clinics located in disadvantaged urban communities. J Diabetes Complic. 1995;9(1):49–54. https://doi.org/10.1016/1056-8727(94)00005-9.
9. Ogunyemi OI, Gandhi M, Lee M, Teklehaimanot S, Daskivich LP, Hindman D, et al. Detecting diabetic retinopathy through machine learning on electronic health record data from an urban, safety net healthcare system. JAMA Open. 2021;4:ooab066. https://doi.org/10.1093/jamiaopen/ooab066.
10. Tsao HY, Chan PY, Su EC. Predicting diabetic retinopathy and identifying interpretable biomedical features using machine learning algorithms. BMC Bioinform. 2018;19(Suppl 9):283. https://doi.org/10.1186/s12859-018-2277-0.
11. Pan H, Sun J, Luo X, Ai H, Zeng J, Shi R, et al. A risk prediction model for type 2 diabetes mellitus complicated with retinopathy based on machine learning and its application in health management. Front Med (Lausanne). 2023;10:1136653. https://doi.org/10.3389/fmed.2023.1136653.
12. Yang C, Liu Q, Guo H, Zhang M, Zhang L, Zhang G, et al. Usefulness of machine learning for identification of referable diabetic retinopathy in a large-scale population-based study. Front Med (Lausanne). 2021;8:773881. https://doi.org/10.3389/fmed.2021.773881.
13. Ting DSW, Cheung CY, Lim G, Tan GSW, Quang ND, Gan A, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. JAMA. 2017;318(22):2211–23. https://doi.org/10.1001/jama.2017.18152.
14. Bourouis A, Feham M, Hossain MA, Zhang L. An intelligent mobile based decision support system for retinal disease diagnosis. Decis Support Syst. 2014;59:341–50. https://doi.org/10.1016/j.dss.2014.01.005.
15. Homayouni A, Liu T, Thieu T. Diabetic retinopathy prediction using progressive ablation feature selection: a comprehensive classifier evaluation. Smart Health. 2022;26: 100343. https://doi.org/10.1016/j.smhl.2022.100343.
16. Yang H, Xia M, Liu Z, Xing Y, Zhao W, Li Y, et al. Nomogram for prediction of diabetic retinopathy in patients with type 2 diabetes mellitus: a retrospective study. J Diabetes Complic. 2022;36(11): 108313. https://doi.org/10.1016/j.jdiacomp.2022.108313.
17. Cardozo G, Pintarelli GB, Andreis GR, Lopes ACW, Marques JLB. Use of machine learning and routine laboratory tests for diabetes mellitus screening. Biomed Res Int. 2022;2022:8114049. https://doi.org/10.1155/2022/8114049.
18. Zhao Y, Li X, Li S, Dong M, Yu H, Zhang M, et al. Using machine learning techniques to develop risk prediction models for the risk of incident diabetic retinopathy among patients with type 2 diabetes mellitus: a cohort study. Front Endocrinol (Lausanne). 2022;13: 876559. https://doi.org/10.3389/fendo.2022.876559.
19. WHO Expert Consultation. Appropriate body-mass index for Asian populations and its implications for policy and intervention strategies. Lancet. 2004;363(9403):157–63. https://doi.org/10.1016/S0140-6736(03)15268-3.
20. Wilkinson CP, Ferris FL 3rd, Klein RE, Lee PP, Agardh CD, Davis M, et al. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. Ophthalmology. 2003;110(9):1677–82. https://doi.org/10.1016/S0161-6420(03)00475-5.
21. Alberti KG, Zimmet PZ. Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: diagnosis and classification of diabetes mellitus provisional report of a WHO consultation. Diabet Med. 1998;15(7):539–53. https://doi.org/10.1002/(SICI)1096-9136(199807)15:7%3c539::AID-DIA668%3e3.0.CO;2-S.
22. Jia W, Weng J, Zhu D, Ji L, Lu J, Zhou Z, et al. Standards of medical care for type 2 diabetes in China 2019. Diabetes Metab Res Rev. 2019;35(6):e3158. https://doi.org/10.1002/dmrr.3158.

23. Islam MM, Rahman MJ, Rabby MS, Alam MJ, Pollob SMAI, Ahmed NAMF, et al. Predicting the risk of diabetic retinopathy using explainable machine learning algorithms. Diabetes Metab Syndr. 2023;17(12): 102919. https://doi.org/10.1016/j.dsx.2023.102919.
24. Shi S, Gao L, Zhang J, Zhang B, Xiao J, Xu W, et al. The automatic detection of diabetic kidney disease from retinal vascular parameters combined with clinical variables using artificial intelligence in type-2 diabetes patients. BMC Med Inform Decis Mak. 2023;23(1):241. https://doi.org/10.1186/s12911-023-02343-9.
25. Padoan A, Plebani M. Artificial intelligence: is it the right time for clinical laboratories? Clin Chem Lab Med. 2022;60(12):1859–61. https://doi.org/10.1515/cclm-2022-1015.
26. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Chen T, editor. KDD '16: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. New York: Association for Computing Machinery; 2016. p. 785–94. https://doi.org/10.1145/2939672.2939785.
27. Hao PY, Chiang JH, Chen YD. Possibilistic classification by support vector networks. Neural Netw. 2022;149:40–56. https://doi.org/10.1016/j.neunet.2022.02.007.
28. Seto H, Oyama A, Kitora S, Toki H, Yamamoto R, Kotoku J, et al. Gradient boosting decision tree becomes more reliable than logistic regression in predicting probability for diabetes with big data. Sci Rep. 2022;12(1):15889. https://doi.org/10.1038/s41598-022-20149-z.
29. Shaik NB, Pedapati SR, Taqvi SA, Othman AR, Dzubir FA. A feed-forward back propagation neural network approach to predict the life condition of crude oil pipeline. Processes. 2020;8(6):661. https://doi.org/10.3390/pr8060661.
30. Sainani KL. Logistic regression. PM R. 2014;6(12):1157–62. https://doi.org/10.1016/j.pmrj.2014.10.006.
31. Kartini D, Nugrahadi DT, Farmadi A. Hyperparameter tuning using GridsearchCV on the comparison of the activation function of the ELM method to the classification of pneumonia in toddlers. In: Kartini D, editor. 2021 4th International conference of computer and informatics engineering (IC2IE). Depok: IEEE; 2021. p. 390–5. https://doi.org/10.1109/ic2ie53219.2021.9649207.
32. Gill KS, Gupta R. Chronic kidney disease detection using GridSearchCV cross validation method. In: Gill KS, editor. 2023 international conference on recent advances in electrical, electronics and digital healthcare technologies (REEDCON). New Delhi: IEEE; 2023. p. 318–22. https://doi.org/10.1109/REEDCON57544.2023.10151392.
33. Singamsetty S, Ghanta S, Biswas S, Pradhan A. Enhancing machine learning-based forecasting of chronic renal disease with explainable AI. PeerJ Comput Sci. 2024;10: e2291. https://doi.org/10.7717/peerj-cs.2291.
34. Ogunyemi O, Kermah D. Machine learning approaches for detecting diabetic retinopathy from clinical and public health records. AMIA Annu Symp Proc. 2015;2015:983–90.
35. Zong GW, Wang WY, Zheng J, Zhang W, Luo WM, Fang ZZ, et al. A metabolism-based interpretable machine learning prediction model for diabetic retinopathy risk: a cross-sectional study in Chinese patients with type 2 diabetes. J Diabetes Res. 2023;2023:3990035. https://doi.org/10.1155/2023/3990035.
36. Li W, Song Y, Chen K, Ying J, Zheng Z, Qiao S, et al. Predictive model and risk analysis for diabetic retinopathy using machine learning: a retrospective cohort study in China. BMJ Open. 2021;11(11): e050989. https://doi.org/10.1136/bmjopen-2021-050989.
37. Nation Committee for the Prevention of Blindness. Vision 2020: Eye health in China, Chinese english comparison. Beijing: People's Medical Publishing House; 2023.
38. Mayinuer Y, Wang NL. Vision 2020: the progress of blindness prevention and eye health in China. Zhonghua Yi Xue Za Zhi. 2020;100(48):3831–4. https://doi.org/10.3760/cma.j.cn112137-20200825-02468.
39. Karoli R, Fatima J, Shukla V, Garg P, Ali A. Predictors of diabetic retinopathy in patients with type 2 diabetes who have normoalbuminuria. Ann Med Health Sci Res. 2013;3(4):536–40. https://doi.org/10.4103/2141-9248.122087.
40. Curtis TM, Gardiner TA, Stitt AW. Microvascular lesions of diabetic retinopathy: clues towards understanding pathogenesis? Eye (Lond). 2009;23(7):1496–508. https://doi.org/10.1038/eye.2009.108.
41. Song P, Yu J, Chan KY, Theodoratou E, Rudan I. Prevalence, risk factors and burden of diabetic retinopathy in China: a systematic review and meta-analysis. J Glob Health. 2018;8(1): 010803. https://doi.org/10.7189/jogh.08.010803.
42. Yau JW, Rogers SL, Kawasaki R, Lamoureux EL, Kowalski JW, Bek T, et al. Global prevalence and major risk factors of diabetic retinopathy. Diabetes Care. 2012;35(3):556–64. https://doi.org/10.2337/dc11-1909.
43. Raman R, Ganesan S, Pal SS, Kulothungan V. Sharma T (2014) Prevalence and risk factors for diabetic retinopathy in rural India. Sankara Nethralaya Diabetic Retinopathy Epidemiology and Molecular Genetic Study III (SN-DREAMS III), report no 2. BMJ Open Diabetes Res Care. 2014;2(1): e000005. https://doi.org/10.1136/bmjdrc-2013-000005.
44. American Diabetes Association Professional Practice Committee. 2. Classification and diagnosis of diabetes: standards of medical care in diabetes-2022. Diabetes Care. 2022;45(Suppl 1):S17–38. https://doi.org/10.2337/dc22-S002.
45. Liu Y, Zhou Z, Wang Z, Yang H, Zhang F, Wang Q. Construction and clinical validation of a nomogram-based predictive model for diabetic retinopathy in type 2 diabetes. Am J Transl Res. 2023;15(10):6083–94.
46. Kowall B, Rathmann W. HbA1c for diagnosis of type 2 diabetes. Is there an optimal cut point to assess high risk of diabetes complications, and how well does the 6.5% cutoff perform? Diabetes Metab Syndr Obes. 2013;6:477–91. https://doi.org/10.2147/DMSO.S39093.

Publisher's Note