Original Research Article

DIGITAL HEALTH

DIGITAL HEALTH
Volume 10: 1–11

DOI: 10.1177/20552076241265220
journals.sagepub.com/home/dhj

Sage

A 10-year retrospective cohort of diabetic patients in a large medical institution: Utilizing multiple machine learning models for diabetic kidney disease prediction

Guangpu Li[1,2,3,4,#], Jia Li[2,3,4,#], Fei Tian[1], Jingjing Ren[5], Zuishuang Guo[1], Shaokang Pan[1], Dongwei Liu[1], Jiayu Duan[1,2,3,4] and Zhangsuo Liu[1,2,3,4]

Abstract

Objective: As the prevalence of diabetes steadily increases, the burden of diabetic kidney disease (DKD) is also intensifying. In response, we have utilized a 10-year diabetes cohort from our medical center to train machine learning-based models for predicting DKD and interpreting relevant factors.

Methods: Employing a large dataset from 73,101 hospitalized type 2 diabetes patients at The First Affiliated Hospital of Zhengzhou University, we analyzed demographic and medication data. Machine learning models, including XGBoost, CatBoost, LightGBM, Random Forest, AdaBoost, GBDT (gradient boosting decision tree), and SGD (stochastic gradient descent), were trained on these data, focusing on interpretability by SHAP. SHAP explains the output of the models by assigning an importance value to each feature for a particular prediction, enabling a clear understanding of how individual features influence the prediction outcomes.

Results: The XGBoost model achieved an area under the curve (AUC) of 0.95 and an area under the precision-recall curve (AUPR) of 0.76, while CatBoost recorded an AUC of 0.97 and an AUPR of 0.84. These results underscore the effectiveness of these models in predicting DKD in patients with type 2 diabetes.

Conclusions: This study provides a comprehensive approach for predicting DKD in patients with type 2 diabetes, employing machine learning techniques. The findings are crucial for the early detection and intervention of DKD, offering a roadmap for future research and healthcare strategies in diabetes management. Additionally, the presence of non-diabetic kidney diseases and diabetes with complications was identified as significant factors in the development of DKD.

Keywords

Diabetic kidney disease, machine learning, diabetes mellitus

Submission date: 19 February 2024; Acceptance date: 13 June 2024

[1]Department of Integrated Traditional and Western Nephrology, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China
[2]Research Institute of Nephrology, Zhengzhou University, Zhengzhou, China
[3]Henan Province Research Center for Kidney Disease, Zhengzhou, China
[4]Key Laboratory of Precision Diagnosis and Treatment for Chronic Kidney Disease in Henan Province, Zhengzhou, China
[5]Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA

[#]Guangpu Li and Jia Li made equal contributions to this work.

Corresponding authors:
Zhangsuo Liu, Department of TCM Integrated, Department of Nephrology, The First Affiliated Hospital of Zhengzhou University, No. 1 Jianshe East Road, Zhengzhou 450052, Henan Province, China.
Email: zhangsuoliu@zzu.edu.cn

Jiayu Duan, Department of TCM Integrated, Department of Nephrology, The First Affiliated Hospital of Zhengzhou University, No. 1 Jianshe East Road, Zhengzhou 450052, Henan Province, China.
Email: jyduan@zzu.edu.cn

Introduction

According to the International Diabetes Federation (IDF) 2021 Atlas, 537 million adults aged 20–79 worldwide, approximately 1 in 10, are currently living with diabetes.[1,2] This alarming statistic is projected to rise to 784 million by 2045, with China bearing one of the heaviest burdens of diabetes worldwide.[1,3] Type 2 diabetes accounts for approximately 90% of adult diabetes and is the primary focus of our research.[1,4] Diabetic kidney disease (DKD), a common complication of diabetes, arises from damage to small blood vessels, leading to impaired kidney function or failure.[5,6] The global incidence of DKD due to type 2 diabetes has surged from approximately 1.4 million cases in 1990 to 2.4 million in 2017, indicating a 74% increase.[1,2]

Diabetes and DKD have become significant global public health concerns.[2,7,8] Despite numerous studies analyzing risk factors for diabetic kidney damage in type 2 diabetes patients,[9–11] there remains a need for research that utilizes a more comprehensive set of multidimensional clinical data to explore these associations in depth. Our study addresses this gap by incorporating a wide range of clinical variables and employing efficient machine learning techniques for a thorough analysis. Our study used a large-scale dataset encompassing 70,000 individuals, incorporating more than 60 feature variables including epidemiological characteristics, geographical information, clinical test results, and medical prescription histories.

Identifying kidney complications in patients with diabetes based on clinical data was a crucial task in our research. Early detection and intervention for kidney damage in patients with diabetes are essential for patients and healthcare professionals.[12] Compared with regression analyses, machine learning methods offer unparalleled advantages in classifying high-dimensional data.[11–13] Leveraging a retrospective cohort with 10 years of electronic medical record data from our hospital, focusing on inpatients with type 2 diabetes, we trained several interpretable machine learning models.[14,15] These models aim to predict the occurrence of DKD in type 2 diabetes and analyze the factors influencing its development. Our research seeks to provide data and model support for the establishment of a comprehensive DKD prevention and control system to enhance the accuracy of DKD diagnosis and to provide early prevention.

Data and method

Ethics approval statements

Data collection and analysis in this study were approved by the Ethics Committee of the First Affiliated Hospital of Zhengzhou University (No. 2023-KY-0810). All research processes were anonymous and retrospective, with the need for written informed consent waived.

Study population

Our study used a large, retrospective, single-center cohort from the First Affiliated Hospital of Zhengzhou University and comprised data from hospitalized patients with type 2 diabetes from January 2013 to December 2022. The dataset included information from 101,896 individuals. Individuals aged ≥18 years diagnosed with type 2 diabetes were included. The exclusion criteria included missing data, an initial hospitalization diagnosis of advanced cancer, or death (Figure 1(a)). Finally, 73,101 individuals were included in the analysis.

The provided data had been anonymized and included demographic characteristics, laboratory data, comorbidities, and medication information. In total, 67 distinct variables were considered in our analysis, encompassing height, weight, blood pressure, laboratory results including blood and urine analyses, and medication details, all uniformly recorded at admission. Laboratory data, including blood and urine analyses, were primarily obtained from the laboratory department. Urine protein quantification and urinary albumin-to-creatinine ratio (UACR) assessments are mainly performed in the renal disease laboratory. In this study, the endpoint event was defined as the clinical diagnosis of DKD.

Comorbidities and medication history

The occurrence of diabetic complications was primarily determined based on the information provided in medical records, and these data are routinely submitted to health authorities for quality assessment. We defined the incidence of complications such as coma; ketoacidosis ophthalmic, neurological, and peripheral circulatory complications; and diabetic arthropathy based on ICD-10 codes[16] (refer to the figure for the incidence rates of these complications in both groups). ICD-10 codes were used to determine the occurrence of hypertension (I10–I15), heart disease (I20–I52), and cerebrovascular diseases (I60–I69).

Information regarding medication history was sourced from prescription records and follow-up data, and the use of combination drugs was determined based on their primary components. This approach ensured a comprehensive understanding of the occurrence of comorbidities and medication usage patterns and provided valuable insights for our analysis.

Datasets and classifiers

The final dataset, consisting of 73,101 individuals, was randomly divided into 80% training and 20% testing datasets. All model training was conducted in Python 3.8, which utilizes widely adopted machine learning methods for binary classification.[12,13] These methods include gradient

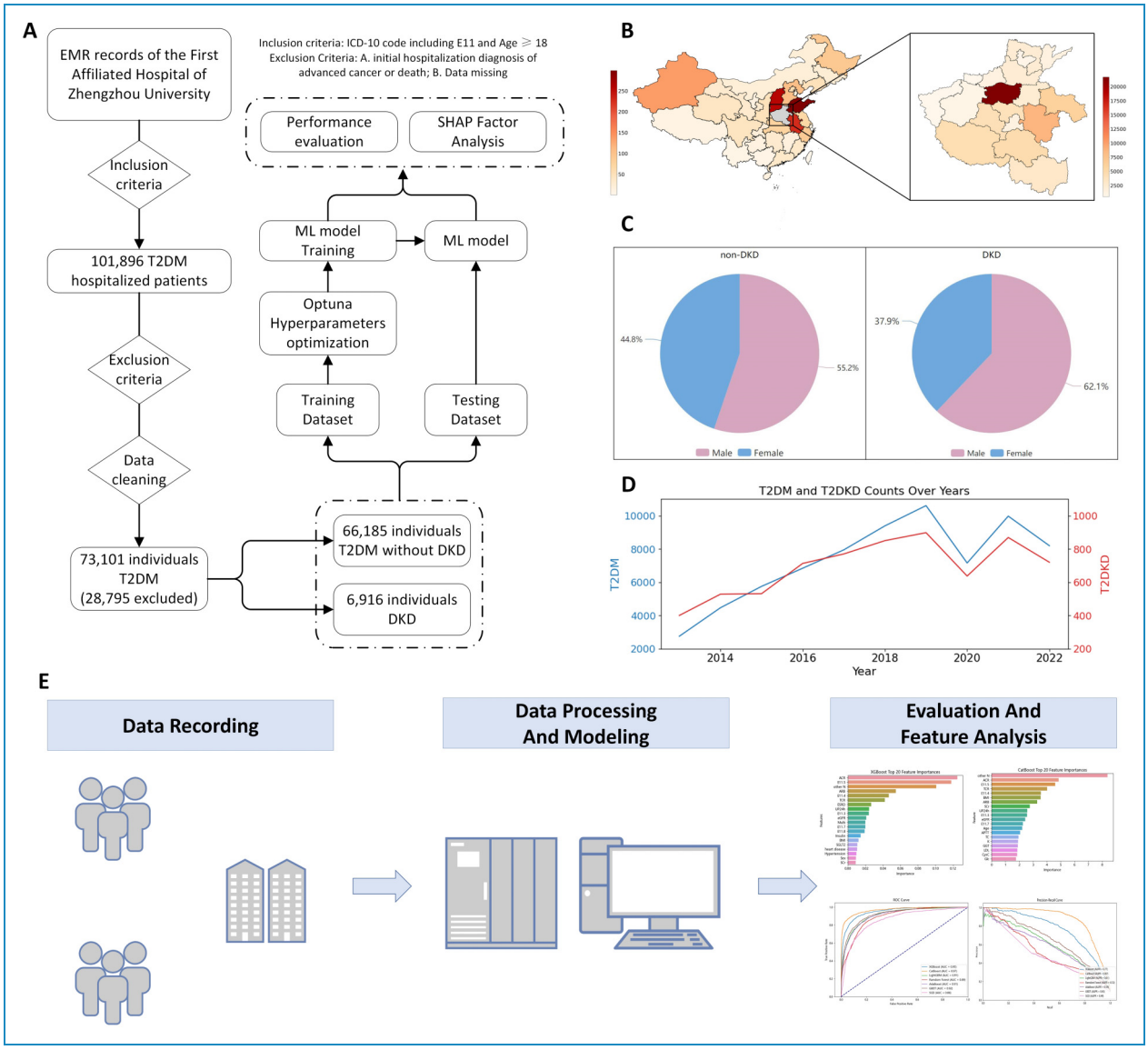

Figure 1. Geographic, demographic, and temporal insights, and study workflow in diabetes patient hospitalization analysis. (a) Data curation and data analysis process. (b) Residential geographic distribution of the population. (c) Gender distribution in the NDKD and DKD groups. (d) Number of person with diabetes newly admitted in different years. (e) Overall process of the study: diabetes patients hospitalized once or multiple times, data recorded in the hospital electronic medical record system; after data acquisition, the model is trained, evaluated, and analyzed.

boosting-based algorithms such as XGBoost,[17] CatBoost,[18] LightGBM,[19] AdaBoost,[20] gradient boosting decision tree (GBDT), stochastic gradient descent (SGD), and other commonly used tree-based methods including decision tree and random forest. The use of these diverse classifiers ensured a comprehensive exploration of the predictive models in our analysis.

Hyperparameter optimization

Hyperparameter optimization of the models was conducted using Optuna (v3.5.0 https://optuna.org/).[21] Optuna is an open-source project that employs Bayesian optimization algorithms to search the hyperparameter space, providing support for a wide range of machine learning frameworks. This approach enhanced the efficiency of our models by systematically tuning the hyperparameters to achieve optimal performance.

Performance evaluation

To assess model performance, we calculated and compared various metrics, including the accuracy, precision, recall, area under the receiver operating characteristic curve

(AUC), and the area under the precision-recall curve (AUPR). These metrics provide comprehensive insights into the predictive capabilities of the models.[22]

Precision is defined as the ratio of correctly predicted positive observations to the total number of predicted positives and indicates the accuracy of the positive predictions made by the model, revealing the number of predicted positive instances that are true positives. Recall is the ratio of correctly predicted positive observations to all the actual positives. Recall assesses the ability of a model to capture all the actual positive instances, highlighting its sensitivity to positive cases.

In addition, we obtained the importance levels of the different factors within each model and conducted a comparative analysis to interpret the significance of these factors in the occurrence of DKD. This comprehensive evaluation provided a nuanced understanding of the predictive capabilities of the models and insights into the relative importance of individual factors in the development of DKD.

SHAP (SHapley Additive exPlanations)

Shapley values, which have desirable properties, are widely used in cooperative game theory.[23] We used the SHAP Python package(v0.44) to explain progressively more complex models, such as XGBoost and CatBoost.[24] SHAP values not only can provide a unified explanation for different models but can also explain how different factors influence predictions in various directions. We used SHAP to create bar charts of the average SHAP values and beeswarm summary plots for the groups with and without diabetic nephropathy. Additionally, we provided an example waterfall plot for one patient in each group. The waterfall plot starts with the base value (the average prediction across the dataset) and sequentially adds the effect of each feature, thus illustrating how the cumulative addition of features leads to the final prediction. This method provides a clear and intuitive way to understand the incremental impact of each feature on the model's output.

Statistical analysis and software

Statistical analyses were performed utilizing Excel (Microsoft Corporation) and Python (v3.8, Python Software Foundation). For visualization, we employed matplotlib (v3.8.2)[25] and the machine learning framework utilized was Scikit-learn (v1.4).[15] For normally distributed continuous data, descriptive statistics, such as means and standard deviations, were employed. Non-normally distributed continuous data were characterized using the median and interquartile range. Categorical data are presented as percentages. Univariate analysis was used to identify potential correlation factors. Statistical tests, including the Mann–Whitney *U* test and analysis of variance, were used to assess normally distributed continuous data. For comparisons involving categorical data and non-normally distributed continuous data, the Kruskal–Wallis *H*, Wilcoxon, and Chi-squared tests were employed. Statistical significance was set at a threshold of $P < 0.05$.

Result

Demographic and clinical characteristics

We analyzed the data of 73,101 patients hospitalized with type 2 diabetes at The First Affiliated Hospital of Zhengzhou University between January 2013 and December 2022. The dataset included demographic characteristics, laboratory data, comorbidities, and medication information, as shown in Table 1. Our study population covered all provincial-level administrative units in China, excluding patients from the Macao Special Administrative Region (Figure 1(b)), demonstrating considerable representativeness, especially in the Yellow and Huai River Basins. The average age of diabetes patients in our cohort was 59.68 ± 12.23 years, with men constituting 44.8% in the non-diabetic kidney disease (NDKD) group and 37.9% in the DKD group (Figure 1(c)). The independent sample *t*-test, Mann–Whitney *U* test, and Chi-squared test revealed significant mean differences between the DKD and NDKD groups for variables such as age, weight, body mass index (BMI), systolic blood pressure, diastolic blood pressure, and various laboratory indicators (Table 1).

In analyzing complications and medication information, it was evident that hypertension was the most prevalent non-diabetes-related complication in DKD, a finding corroborated by a macro-level data analysis.[1,26] In the hospitalized DKD patient group, 68.8% also experience hypertension, compared with 52.8% for patients with NDKD. Peripheral circulatory complications emerged as the most common diabetic complications in patients with DKD, affecting approximately 34.8%. Ophthalmic and neurological complications occurred in 29.9% and 27.5% of patients, respectively. In comparison, among patients without DKD, these numbers were notably lower at 10.2%, 7.5%, and 6.5%, respectively (Supplementary Table S1). Ophthalmic complications in diabetes mellitus (E11.3) include diabetic cataract (H28.0*) and diabetic retinopathy (H36.0*); neurological complications encompass diabetic amyotrophy (G73.0*), diabetic autonomic neuropathy (G99.0*), diabetic mononeuropathy (G59.0*), diabetic polyneuropathy (G63.2*), and diabetic autonomic neuropathy (G99.0*). These complications were categorized according to ICD-10 diagnostic codes.

Performance comparison

We employed a diverse array of machine learning methods utilizing 67 variables as input features to predict DKD.

Table 1. Demographic and clinical characteristics.

Factors	Total	NDKD	DKD	*P*-value
No.	73,101	66,185	6916	
Demographic characteristics				
Age (years)	59.68 (12.23)	59.93 (12.17)	57.31 (12.62)	<0.001
Female sex (%)	44.1 (32,259)	44.8 (29,640)	37.9 (2619)	<0.001
Height (m)	1.67 (0.10)	1.67 (0.10)	1.67 (0.09)	0.476
Weight (kg)	62.21 (10.6)	61.81 (10.32)	66 (12.33)	<0.001
BMI (kg/m^2)	22.16 (2.81)	22.02 (2.70)	23.53 (3.47)	<0.001
SBP (mmHg)	134.27 (18.96)	133.64 (18.65)	140.28 (20.83)	<0.001
DBP (mmHg)	80.71 (11.4)	80.45 (11.31)	83.25 (11.92)	<0.001
Laboratory data				
Hemoglobin (g/L)	125.2 (21.36)	125.87 (20.84)	118.84 (24.98)	<0.001
Leukocyte (×109/L)	7.17 (3.27)	7.18 (3.31)	7.08 (2.82)	0.298
BUN (mmol/L)	6.45 (4.36)	6.17 (3.95)	9.09 (6.61)	<0.001
Scr (μmol/L)	65 (54,81)	64 (53,78)	82 (62,150)	<0.001
Uric acid (μmol/L)	278.93(109.44)	274.58 (108.2)	320.59 (112.44)	<0.001
Serum protein (g/L)	65.37 (7.78)	65.62 (7.57)	62.97 (9.23)	<0.001
Serum albumin (g/L)	39.61 (6.2)	39.81 (5.98)	37.75 (7.72)	<0.001
eGFR (mL/min/1.73 m^2)	97.58 (77.06,106.05)	98.47 (80.76,106.22)	74.22 (24.50,103.26)	<0.001
TG (mmol/L)	1.82 (1.57)	1.8 (1.54)	2.05 (1.81)	<0.001
TC (mmol/L)	4.2 (1.28)	4.17 (1.25)	4.53 (1.50)	<0.001
HDL (mmol/L)	1.06 (0.29)	1.06 (0.28)	1.07 (0.31)	0.141
LDL (mmol/L)	2.51 (0.87)	2.49 (0.84)	2.68 (1.10)	<0.001
Cystatin C (mg/L)	1.05 (0.59)	1.01 (0.50)	1.37 (1.08)	<0.001
Blood glucose (mmol/L)	8.62 (4.15)	8.57 (4.07)	9.08 (4.82)	<0.001
HbA1c (%)	8.38 (2.68)	8.35 (2.67)	8.58 (2.74)	<0.001
AST (IU/L)	27.04 (98.39)	30.52 (679.88)	21.39 (22.69)	<0.001
ALT (IU/L)	27.83 (67.30)	28.35 (68.89)	22.94 (49.35)	<0.001
γ-GT (IU/L)	45.56 (91.87)	46.12 (92.35)	40.13 (86.98)	<0.001

(continued)

Table 1. Continued.

Factors	Total	NDKD	DKD	*P*-value
ALP (IU/L)	83.85 (60.53)	84.05 (61.79)	81.97 (46.67)	0.446
PT (s)	10.53 (2.49)	10.56 (2.54)	10.25 (1.94)	<0.001
APTT (s)	32.05 (8.51)	32.02 (8.27)	32.35 (10.47)	<0.001
TT (s)	15.2 (11.21)	15.18 (11.31)	15.42 (10.13)	<0.001
CRP (mg/L)	13.93 (40.69)	14.12 (41.1)	12.09 (36.5)	<0.001
ESR (mm/h)	22.62 (21.12)	21.79 (19.89)	30.57 (29.31)	<0.001
Urine protein				<0.001
0	78.7 (57,497)	82.5 (54,605)	41.8 (2892)	
1	10.8 (7924)	10.0 (6641)	18.6 (1283)	
2	4.8 (3518)	3.7 (2416)	15.9 (1102)	
3	5.7 (4162)	3.8 (2523)	23.7 (1639)	
24h-UP (g)	0.56 (1.69)	0.4 (1.38)	2.13 (3.06)	<0.001
UACR (mg/mmol)	43.17 (122.23)	30.66 (101.19)	162.92 (209.95)	<0.001
Potassium (mmol/L)	4.57 (0.61)	4.54 (0.59)	4.8 (0.75)	<0.001
Sodium (mmol/L)	141 (4.37)	141.03 (4.38)	140.79 (4.34)	<0.001

Data are expressed as percentage (*n*), mean (standard deviation), or median (interquartile range), as appropriate.
BMI: body mass index; SBP: systolic pressure; DBP: diastolic pressure; BUN: blood urea nitrogen; eGFR: estimated glomerular filtration rate; TG: triglycerides; TC: total cholesterol; Scr: serum creatinine; HbA1c: glycated hemoglobin A1c; AST: aspartate aminotransferase; ALT: alanine aminotransferase; γ-GT: γ-glutamyl transpeptidase; ALP: alkaline phosphatase; PT: prothrombin time; APTT: activated partial thromboplastin time; TT: thrombin time; CRP: C-reactive protein; ESR: erythrocyte sedimentation rate; 24h-UP: 24-h urine protein quantification; UACR: urine albumin-to-creatinine ratio.

The models applied include XGBoost,[17] CatBoost,[18] LightGBM,[19] AdaBoost, GBDT, SGD, and Random Forest.[15] The receiver operating characteristic and precision-recall curves are illustrated in Figure 2(a) and (b), respectively. The model performance metrics, including the AUC, AUPR, accuracy, precision, and recall, are presented in Table 2.

Among the various models, CatBoost and XGBoost demonstrated particularly promising outcomes. In the test set, CatBoost achieved an AUC of 0.97 and an AUPR of 0.84, while XGBoost exhibited an AUC of 0.95 and an AUPR of 0.76. These models showcase the ability to accurately predict the occurrence of DKD. A comprehensive evaluation of the model performance metrics provided insights into the effectiveness of different machine learning approaches in the context of DKD prediction.

Factor analysis

The machine learning models we selected were interpretable,[14,15] and the top 20 important features of XGBoost and CatBoost are depicted in Supplementary Figure S3. Among them, UACR, diabetes with peripheral circulatory complications, and other kidney diseases emerged as the top three factors in terms of importance in both models. Supplementary Figure S3 illustrates the frequency with which each factor entered the top 20 feature importance levels across all seven models. Diabetes with peripheral circulatory complications and BMI stood out as the most frequent features, consistently appearing in the top 20 features of every model (Figure 2(c)). This emphasis on interpretability enhances the transparency of our machine learning approach, shedding light on the key factors influencing the prediction of DKD.

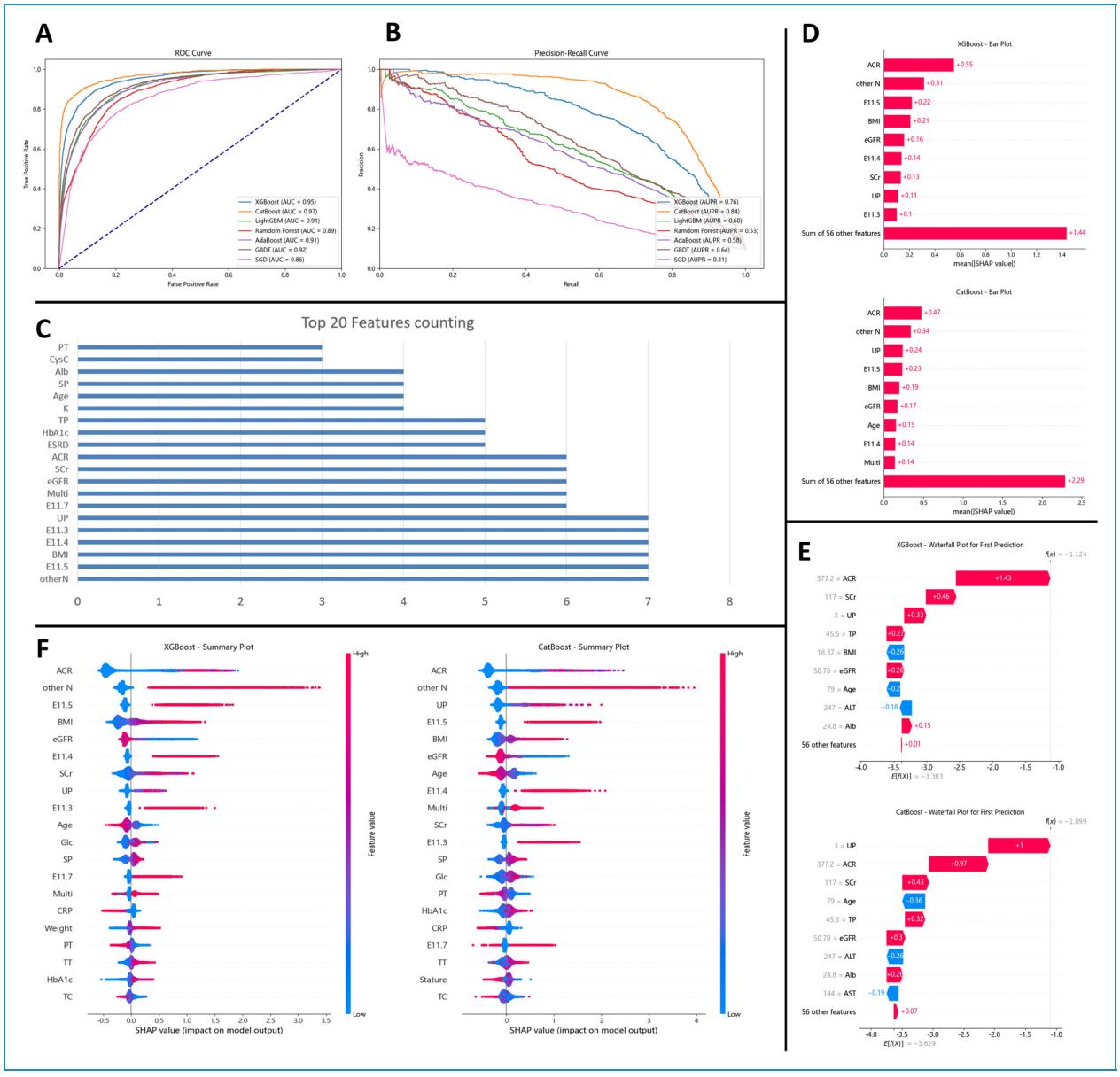

Figure 2. Performance evaluation and SHAP for model interpretation. (a) Receiver operating characteristic (ROC) curve of the testing dataset; (b) precision-recall curve of the testing dataset; (c) frequency count of the top 20 important features for each variable across different models. (d) SHapley Additive exPlanations (SHAP) bar plots of XGBoost and CatBoost: Display the mean absolute SHAP values of each factor, with UACR and non-diabetic kidney disease (NDKD) consistently ranking in the top 2 for both models. (e) Summary plots for both models. (f) Waterfall plot drawn from a sample dataset of the models: it reveals UACR, serum creatinine, among others, as the primary risk factors leading the model to predict DKD (indicated in red for positive SHAP values, signifying these factors as risk elements for the occurrence of DKD, with blue indicating the opposite). BMI: body mass index; UP: urine protein; UACR or ACR: urine albumin-to-creatinine ratio; TCR: urine total protein-to-creatinine ratio; UP24h: 24-h urine protein quantification; SCr: serum creatinine; WBC: leukocyte; UA: urine acid; CysC: cystatin C; eGFR: estimated glomerular filtration rate; AST: aspartate aminotransferase; ALT: alanine aminotransferase; TP: serum total protein; Alb: serum albumin; K: serum potassium; ARB: angiotensin receptor blocker; E11.3: with ophthalmic complications; E11.4: with neurological complications; E11.5: with peripheral circulatory complications; E11.7: with multiple diabetic complications; Multi: multiple other complications.

In addition, we employed SHAP to conduct a factor analysis and an interpretation of two models to demonstrate performance: XGBoost and CatBoost. The bar plots delineate the absolute SHAP values for each factor within the models. Notably, renal diseases not attributed to diabetes and the UACR were among the top three factors influencing the onset of DKD, underscoring the pivotal role of proteinuria as an early marker for DKD identification (Figure 2(d)).

Table 2. Performance of models.

	Accuracy%	Precision	Recall	AUC	AUPR
XGBoost	94.39	0.84	0.50	0.95	0.76
CatBoost	95.93	0.92	0.62	0.96	0.84
LightGBM	90.55	0.72	0.43	0.91	0.60
RF	92.44	0.94	0.42	0.88	0.53
AdaBoost	92.44	0.66	0.40	0.90	0.58
GBDT	93.18	0.75	0.41	0.91	0.64
SGD	90.62	0.60	0.21	0.75	0.31

AUC: area under the receiver operating characteristic curve; AUPR: area under the precision-recall curve.

The summary plot rendered for each sample delineated the Shapley values of the individual features, highlighting the most influential factors and their respective impacts on the dataset (Figure 2(e)). The risk of developing DKD is elevated in patients with renal diseases unrelated to diabetes. Furthermore, consistent with previous analyses, BMI was identified as one of the top five factors across both models, indicating that a higher BMI correlates with an increased risk of DKD. Additionally, we randomly selected a diabetic patient's data to construct a Waterfall Plot using SHAP, where blue denotes a negative contribution of the factor and red signifies a positive contribution, illustrating the role of each indicator within the model (Figure 2(f)).

Discussion

In this retrospective study, we used a dataset comprising 67 variables, including demographic characteristics, laboratory data, comorbidities, and medication information, to train and validate multiple machine learning models. This study aimed to predict the onset of DKD in patients with diabetes. Among the models, CatBoost and XGBoost demonstrated superior performance on our dataset, including both the training and validation sets.

In the feature importance analysis of various models, the presence of NDKD emerged as a significant factor for the development of DKD. Patients with other kidney diseases prior to the onset of diabetic renal damage or diabetes are more likely to experience a vicious cycle of diabetic kidney injury. This is considered a dynamic process. Clinically, DKD, NDKD, and a combination of both are often considered static diagnoses. However, NDKD in patients with diabetes may evolve into a combined state of DKD and NDKD owing to the progression of diabetes. Diabetes with peripheral circulatory complications also ranked highly as a risk factor in many studies[27–29] specifically diabetic gangrene, diabetic peripheral angiopathy, and diabetic ulcerative DKDs, as microvascular complications, are more highly correlated with diabetic retinopathy than any other diabetes-related complications. This association has been consistently reported in several studies.[9,10,30] In our model, diabetes with ophthalmic complications ranked after peripheral circulatory complications and neurological complications.[31–33] Compared to the insidious onset of renal damage, peripheral circulatory complications, which are more dangerous among diabetic complications, tend to present more symptomatic changes due to the severe hyperglycemic state[34] and heightened oxidative stress and inflammatory responses.[35] They share common risk factors for DKD. Urine protein remains a critical early marker for the identification of DKD. In several models predicting end-stage renal disease in DKD, the urinary albumin level can be used to assess the severity of DKD and is a key predictor of its progression.[36–40]

Consequently, patients with other NDKDs may exhibit urinary albumin abnormalities. Even with a confirmed diagnosis of NDKD, vigilance for DKD is of paramount importance. BMI was one of the most frequently top-ranked indicators across the models, indicating an intrinsic link between BMI and DKD. This includes obesity-related renal damage and the insulin-resistant state associated with BMI,[41] complicating glycemic control and maintenance, potentially affecting renal function through various mechanisms, including increased renal blood flow and pressure, chronic inflammation, and perpetuating a vicious cycle that accelerates the onset of DKD.

Additionally, we used regular expressions and large language models for text analysis to extract data from past EMRs to obtain the duration of diabetes and the age of onset of diabetes. In our dataset, only 43,849 patients had these data, and the data were not entirely accurate. However, we conducted similar analyses on these data: we divided the 43,849 records into 80% training set and 20% testing set. For the duration of diabetes, we categorized the data into three groups: 0–5 years as 0, 6–10 years as 1, and more than 10 years as 2. As shown in Supplementary Figure S5, both the duration of diabetes (T2DMTRANK) and the age of onset of diabetes (T2DMAGE) were among the top three important variables in the CatBoost, LGB, and AdaBoost models, highlighting their significance. However, data missingness is unavoidable in many large retrospective datasets and real-world data. We processed a large amount of medical text information, which was challenging and might not be entirely accurate. Nonetheless, based on this information, we achieved good results in our newly set dataset: the AUCs for XGBoost, CatBoost, LightGBM, RF, AdaBoost, and GBDT were all above 0.9 in both the training and testing

datasets (Supplementary Figures S6 and S7). Adding the duration of diabetes and the age of onset of diabetes to the models did not significantly change the final evaluation results, suggesting that other indicators might also relate to these variables. Nevertheless, the duration of diabetes is undeniably an important and objective factor in the occurrence of DKD.

Our medical center, located in an urban area, has observed a notable trend in our diabetic patient population; farmers constitute the largest occupational group, accounting for approximately 38.38% (28,053 individuals). This finding highlights the severe burden of DKD in rural areas. According to data from the IDF, by 2045, rural China is projected to have the highest number of individuals with diabetes in the world.[1] Consequently, our efforts to prevent and treat diabetes and its complications should extend to rural areas. Our center is situated in the Huang-Huai region, which encompasses the Yellow and Huai River basins. This area has been densely populated since the pre-Qin era and has been a long-standing agricultural civilization. To date, it remains a core agricultural zone comparable to the Corn Belt in the central United States and the Chornozem (Black Earth) region of Ukraine. With a high proportion of rural population, this region presents a crucial direction for future research and applications in rural.

Our ongoing efforts to enhance patient care and medical research include the development of a sophisticated and convenient follow-up and predictive modeling system that leverages the capabilities of mobile networks and applications. This innovative approach involves deploying predictive models to send alerts about abnormal values and increased risks of DKD directly to patients, physicians, and researchers. Furthermore, these alerts will be shared with the patients' family doctors, integrating our efforts with tiered healthcare systems. Currently, we are updating our follow-up data system and integrating it into a mobile application. This initiative aims to make the follow-up system accessible not only to researchers and physicians, but also to patients, offering a user-friendly interface for seamless data exchange. This dual approach empowers both healthcare professionals and patients to proactively engage in the early prevention of DKD and other related complications.

This study had several limitations. Although our cohort was large and the data were derived from four different hospital branches in various administrative districts of Zhengzhou City, this remains a single-center study. Fortunately, our data comprehensively recorded patients' residential and work locations. The distribution of our data broadly represents inpatients with diabetes from both rural and urban areas of the Yellow and Huai River basins. As our research progresses and with the development of the aforementioned data center and updates to the follow-up system, we plan to collaborate with hospitals in surrounding cities for multicenter studies. This expansion will serve to validate and enhance our predictive models and enlarge the cohort of hospitalized patients with diabetes.

Conclusion

Diabetic nephropathy often has a concealed onset, making it particularly important for people with diabetes to identify those among them who are suffering from DKD. This study provides a comprehensive approach for predicting DKD in patients with type 2 diabetes, employing a large dataset and machine learning techniques. This approach allowed us to identify factors involved in the development of DKD. We believe that our study makes a significant contribution to the literature because we show that UACR, peripheral circulatory complications, other kidney diseases, and BMI were the most important factors predicting DKD.

Acknowledgements: The authors are grateful for the organizational support of the Clinical Big Data Center of Henan Province.

Contributorship: ZL, JD and GL designed the study. GL and JL drafted the article. GL, FT and SP developed the project code. GL, JR, DL, ZG and JD extracted and cleaned the data. Every author participated in data interpretation, provided critical revisions to the manuscript, and gave their approval for the final version. ZL and GL were responsible for the overall coordination and completion of the work.

Declaration of conflicting interests: The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical approval: Data collection and analysis were approved by the Ethics Committee of the First Affiliated Hospital of Zhengzhou University (No. 2023-KY-0810). All processes were anonymous and retrospective.

Funding: The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was supported by grants from the National Natural Science Young Scientists Foundation of China (No. 82103916), the General Program of the National Natural Science Foundation of China (No. 81970633), the National Nature Science Foundation of China (Joint project No. U21A20348), and Outstanding Young Scientists Fund of Henan Health Commission of the People's Republic of China (Grant No. YXKC2022054).

Guarantor: GL and ZL.

ORCID iD: Guangpu Li https://orcid.org/0000-0003-1370-9425

Supplemental material: Supplemental material for this article is available online.

References

1. Federation ID. *IDF Diabetes Atlas*. 10th ed. Brussels, Belgium: International Diabetes Federation, 2021.
2. Tuttle KR, Jones CR, Daratha KB, et al. Incidence of chronic kidney disease among adults with diabetes, 2015–2020. *N Engl J Med* 2022; 387: 1430–1431.
3. Hu FB. Globalization of diabetes: the role of diet, lifestyle, and genes. *Diabetes Care* 2011; 34: 1249–1257.
4. de Boer IH, Rue TC, Hall YN, et al. Temporal trends in the prevalence of diabetic kidney disease in the United States. *JAMA* 2011; 305: 2532–2539.
5. Kumar KS, Bhowmik D, Deb L, et al. Role of pharmacist management and novel therapies of diabetic nephropathic patients. *Pharma Innov* 2012; 1: 54.
6. Prabhakar PK. Pathophysiology of secondary complications of diabetes mellitus. *Pathophysiology* 2016; 9: 32–36.
7. Duan J, Duan G, Wang C, et al. Prevalence and risk factors of chronic kidney disease and diabetic kidney disease in a central Chinese urban population: a cross-sectional survey. *BMC Nephrol* 2020; 21: 115.
8. Duan J, Wang C, Liu D, et al. Prevalence and risk factors of chronic kidney disease and diabetic kidney disease in Chinese rural residents: a cross-sectional survey. *Sci Rep* 2019; 9: 10408.
9. He F, Xia X, Wu X, et al. Diabetic retinopathy in predicting diabetic nephropathy in patients with type 2 diabetes and renal disease: a meta-analysis. *Diabetologia* 2013; 56: 457–466.
10. He R, Shen J, Zhao J, et al. High cystatin C levels predict severe retinopathy in type 2 diabetes patients. *Eur J Epidemiol* 2013; 28: 775–778.
11. Slieker RC, van der Heijden AAWA, Siddiqui MK, et al. Performance of prediction models for nephropathy in people with type 2 diabetes: systematic review and external validation study. *Br Med J* 2021; 374: n2134.
12. Loftus TJ, Shickel B, Ozrazgat-Baslanti T, et al. Artificial intelligence-enabled decision support in nephrology. *Nat Rev Nephrol* 2022; 18: 452–465.
13. Ellahham S. Artificial intelligence: the future for diabetes care. *Am J Med* 2020; 133: 895–900.
14. Sidak D, Schwarzerová J, Weckwerth W, et al. Interpretable machine learning methods for predictions in systems biology from omics data. *Front Mol Biosci* 2022; 9: 926623.
15. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Technol.* 2011; 12: 2825–2830.
16. Organization WH. *International statistical classification of diseases and related health problems*. 10th revision, 5th ed. Geneva, Switzerland: World Health Organization, 2016.
17. Chen T and Guestrin C. Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on knowledge discovery and data mining; 2016, pp.785–794.
18. Hancock JT and Khoshgoftaar TM. Catboost for big data: an interdisciplinary review. *J Big Data* 2020; 7: 1–45.
19. Ke G, Meng Q, Finley T, et al. Lightgbm: a highly efficient gradient boosting decision tree. *Adv Neural Inf Process Syst* 2017; 30: 3146–3154.
20. Ying C, Qi-Guang M, Jia-Chen L, et al. Advance and prospects of AdaBoost algorithm. *Acta Autom Sin* 2013; 39: 745–758.
21. Akiba T, Sano S, Yanase T, et al. Optuna: a next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining; 2019, pp.2623–2631.
22. Japkowicz N and Shah M. *Evaluating Learning Algorithms: A Classification Perspective*. United Kingdom: Cambridge University Press, 2011.
23. Lundberg SM and Lee S-I. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst* 2017; 30: 4768–4777.
24. Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. *Nature Mach Intellig* 2020; 2: 56–67.
25. Hunter JD. Matplotlib: a 2D graphics environment. *Comput Sci Eng* 2007; 9: 90–95.
26. Jaeger BC, Chen L, Foti K, et al. Hypertension statistics for US adults: an open-source web application for analysis and visualization of national health and nutrition examination survey data. *Hypertension* 2023; 80: 1311–1320.
27. Okada H, Tanaka M, Yasuda T, et al. Peripheral perfusion, measured by perfusion index, is a novel indicator for renal events in patients with type 2 diabetes mellitus. *Sci Rep* 2020; 10: 6054.
28. Association AD. Microvascular complications and foot care: standards of medical care in diabetes. *Diabetes Care* 2021; 44: S151–S167.
29. Borderie G, Foussard N, Larroumet A, et al. Albuminuric diabetic kidney disease predicts foot ulcers in type 2 diabetes. *J Diabetes Complications* 2023; 37: 108403.
30. Zhang K, Liu X, Xu J, et al. Deep-learning models for the detection and incidence prediction of chronic kidney disease and type 2 diabetes from retinal fundus images. *Nat Biomed Eng* 2021; 5: 533–545.
31. Liabeuf S, Pepin M, Franssen CF, et al. Chronic kidney disease and neurological disorders: are uraemic toxins the missing piece of the puzzle? *Nephrol Dial Transplant* 2022; 37: ii33–ii44.
32. Group DPPR. Long-term effects of lifestyle intervention or metformin on diabetes development and microvascular complications over 15-year follow-up: the diabetes prevention program outcomes study. *Lancet Diabetes Endocrinol* 2015; 3: 866–875.
33. Duan J, Liu D, Zhao Z, et al. Short-term duration of diabetic retinopathy as a predictor for development of diabetic kidney disease. *J Transl Internal Med* 2023; 11: 449–458.
34. Caturano A, Galiero R, Pafundi PC, et al. Does a strict glycemic control during acute coronary syndrome play a cardioprotective effect? Pathophysiology and clinical evidence. *Diabetes Res Clin Pract* 2021; 178, 108959.
35. Mohandes S, Doke T, Hu H, et al. Molecular pathways that drive diabetic kidney disease. *J Clin Invest* 2023; 133: e165654.

36. Curtis S and Komenda P. Screening for chronic kidney disease: moving toward more sustainable health care. *Curr Opin Nephrol Hypertens* 2020; 29: 333–338.
37. Jia W, Gao X, Pang C, et al. Prevalence and risk factors of albuminuria and chronic kidney disease in Chinese population with type 2 diabetes and impaired glucose regulation: Shanghai diabetic complications study (SHDCS). *Nephrol Dial Transplant* 2009; 24: 3724–3731.
38. Jiang G, Luk AOY, Tam CHT, et al. Progression of diabetic kidney disease and trajectory of kidney function decline in Chinese patients with type 2 diabetes. *Kidney Int* 2019; 95: 178–187.
39. Joshi VD. Quality of life in end stage renal disease patients. *World J Nephrol* 2014; 3, 08.
40. Song J, Jin C, Shan Z, et al. Prevalence and risk factors of hyperuricemia and gout: a cross-sectional survey from 31 provinces in mainland China. *J Transl Internal Med* 2022; 10: 134–145.
41. Group DPPR. 10-Year follow-up of diabetes incidence and weight loss in the diabetes prevention program outcomes study. *Lancet* 2009; 374: 1677–1686.