NEPHROLOGY - ORIGINAL PAPER

Analysis of clinical predictors of kidney diseases in type 2 diabetes patients based on machine learning

Dongna Hui[1,2] · Yiyang Sun[3] · Shixin Xu[3] · Junjie Liu[5] · Ping He[5] · Yuhui Deng[5] · Huaxiong Huang[4,5,6] · Xiaoshuang Zhou[2] · Rongshan Li[1,2]

Abstract

Background The heterogeneity of Type 2 Diabetes Mellitus (T2DM) complicated with renal diseases has not been fully understood in clinical practice. The purpose of the study was to propose potential predictive factors to identify diabetic kidney disease (DKD), nondiabetic kidney disease (NDKD), and DKD superimposed on NDKD (DKD + NDKD) in T2DM patients noninvasively and accurately.

Methods Two hundred forty-one eligible patients confirmed by renal biopsy were enrolled in this retrospective, analytical study. The features composed of clinical and biochemical data prior to renal biopsy were extracted from patients' electronic medical records. Machine learning algorithms were used to distinguish among different kidney diseases pairwise. Feature variables selected in the developed model were evaluated.

Results Logistic regression model achieved an accuracy of 0.8306 ± 0.0057 for DKD and NDKD classification. Hematocrit, diabetic retinopathy (DR), hematuria, platelet distribution width and history of hypertension were identified as important risk factors. Then SVM model allowed us to differentiate NDKD from DKD + NDKD with accuracy 0.8686 ± 0.052 where hematuria, diabetes duration, international normalized ratio (INR), D-Dimer, high-density lipoprotein cholesterol were the top risk factors. Finally, the logistic regression model indicated that DD-dimer, hematuria, INR, systolic pressure, DR were likely to be predictive factors to identify DKD with DKD + NDKD.

Conclusion Predictive factors were successfully identified among different renal diseases in type 2 diabetes patients via machine learning methods. More attention should be paid on the coagulation factors in the DKD + NDKD patients, which might indicate a hypercoagulable state and an increased risk of thrombosis.

Keywords Kidney diseases diagnosis · Type 2 diabetes mellitus · Machine learning · Noninvasive

Dongna Hui and Yiyang Sun contributed equally.

✉ Huaxiong Huang
hhuang@uic.edu.cn

✉ Xiaoshuang Zhou
zhouxiaoshuang@sxmu.edu.cn

✉ Rongshan Li
rongshanli@126.com

[1] Institute of Biomedical Sciences, Shanxi University, No. 92 Wucheng Road, Xiaodian District, Taiyuan 030006, Shanxi, China

[2] Department of Nephrology, Shanxi Provincial People's Hospital, No. 29 Shuangta Street, Yingze District, Taiyuan 030012, Shanxi, China

[3] Zu Chongzhi Center for Mathematics and Computational Sciences (CMCS), Data Science Research Center (DSRC), Duke Kunshan University, 8 Duke Ave, Kunshan, Jiangsu, China

[4] Research Center for Mathematics, Beijing Normal University, Zhuhai, China

[5] BNU-HKBU United International College, 2000 Jintong Road, Tangjiawan, Zhuhai 519087, Guangdong, China

[6] Department of Mathematics and Statistics, York University, Toronto, ON, Canada

Introduction

The growing prevalence of Diabetes Mellitus (DM) worldwide, especially type 2 DM (T2DM), greatly increases the incidence of microvascular complications in DM patients without well intervention on glucose control [1]. Diabetic kidney disease (DKD) is a typical representative of microvascular damage in T2DM and characterized with slow progression that needs several years develop from microalbuminuria to macroalbuminuria clinically [2]. Due to the heterogeneity in T2DM patients with renal lesions, there exists three histopathological patterns after renal biopsy, including pure DKD, nondiabetic kidney disease (NDKD) alone and DKD superimposed on NDKD (DKD + NDKD). How to identify various types of kidney diseases in patients with T2DM has become a primary issue in the clinical diagnosis of them.

The diagnosis and treatments of DKD are quite different from NDKD, which is not caused by the hyper-glycemia environment [3]. NDKD can be cured by one or combination of a variety of therapies (such as immunosuppressants), while the treatment of DKD passively focus on delaying progression to end-stage renal disease (ESRD). Compared to NDKD patients, patients with DKD, or with DKD superimposed on NDKD were associated with poor prognosis and had high mortality after 10 years of renal biopsy [4, 5]. Although kidney biopsy is the gold standard to identify specific kidney diseases, it is not generally applied in T2DM patients due to risk of bleeding and inflammation [6]. Therefore, it is desirable to develop noninvasive methods to assist diagnosis of these three types of kidney diseases.

Many efforts have been devoted to establishing models or identify risk factors associated with DKD and NDKD through conventional statistical analysis [7–15]. Recently, machine learning methods are also used to predict progression of DKD [16, 17]. However, there are few studies on specific indicators for clinical diagnosis of DKD superimposed on NDKD. On the one hand, patients with both DKD and NDKD form a complex disease type that comprised of both DM-related renal injury and damage to the kidney itself. On the other hand, cases of samples are limited comparing to DKD and NDKD.

In this paper, we present a retrospective study concerning T2DM with renal diseases after kidney biopsy using machining learning algorithms. The aim of this study not limited to the feature variables between DKD and NDKD, but also further analyze the risk factors for differentiating DKD + NDKD from DKD and DKD + NDKD from NDKD.
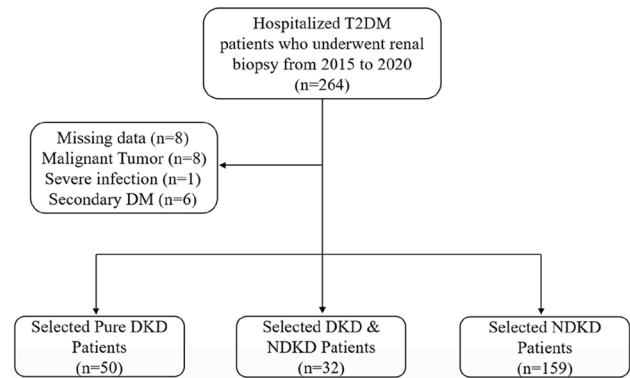

Fig. 1 Flow diagram of patient selection

Methods

Patients' selection

The study was carried out by screening T2DM hospitalized patients who underwent renal biopsy from January 2015 to December 2020 in the Department of Nephrology, Shanxi Provincial People's Hospital. Only patients with clinical diagnosis of T2DM and definite pathological diagnosis were included. Patients with incomplete data, malignancy, severe infection or secondary diabetes mellitus were excluded in this study. The flowchart of patients' selection is shown in Fig. 1. The diagnosis of T2DM was based on the criteria established by the World Health Organization in 1998 [18]. Indications for kidney biopsy were compiled based on the criteria of 2007 KDOQI guidelines [19]. The study design was approved by the Ethics Review Committee of Shanxi Provincial people's Hospital ([2021] Provincial Medical Opinions No. 354). The informed consents were agreed to waive due to its retrospective research design.

Feature selection and pre-processing

Compared with previous study that used the criteria of 2007 KDOQI guidelines [19] as the prior for factor analysis, all the biochemical parameters and clinical data of the patients were used here to identify the leading risk factors for assisting diagnose. Because the same medical items had been checked more than once during the patients' hospitalization, the data of biochemical parameters prior to renal biopsy were collected. Features with more than 40% missing from the records of patients were removed. Finally, 99 features were used for statistical and machine learning analysis as shown in Supplementary Table 1.
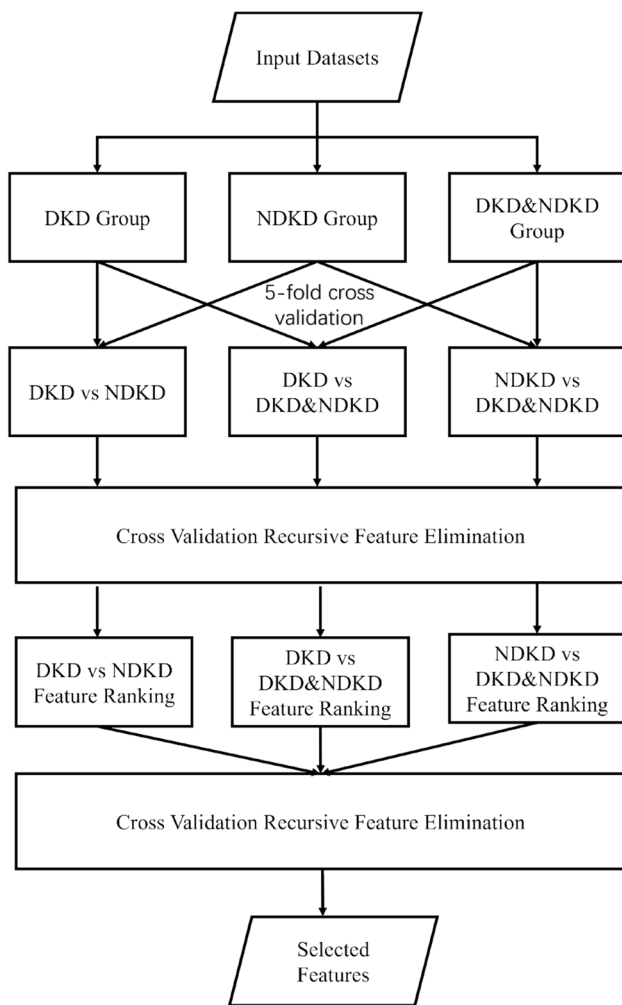
Fig. 2 The model construction workflow

Statistical analysis

Statistical analysis was performed using Python 3.7. Descriptive statistics were used to estimate those clinical and laboratory parameters. Numerical data were expressed in median (interquartile range, IQR) and categorical data presented in absolute values and percentages. To compare the distribution differences between various pairs of different groups, Mann–Whitney rank sum test was used since not all the features were normally distributed. Missing values of the selected variables were imputed using the median. In the tests of two comparison pairs: DKD and NDKD, NDKD and DKD with NDKD, features with P value less than 0.01 were considered as significant; the DKD and DKD with NDKD, features with P value less than 0.05 were considered as significant.

Machine learning analysis

Lasso Logistic Regression, Random Forest, Extreme Gradient Boosting Tree, kernel Support Vector Machine (kernel SVM) were used to identify important factors distinguishing those three groups. Figure 2 shows the procedure of model construction and it was divided into two parts: feature ranking and model evaluation. After splitting the datasets for three comparison tasks: DKD vs NDKD, DKD vs DKD + NDKD and NDKD *vs* DKD + NDKD, all feature values were scaled by Min–Max Scalar, and Recursive Feature Elimination Cross Validation (RFECV) in Scikit-Learn 0.21.1 was applied to obtain the average important rank of each parameter for 100 times. By adding parameters into the previous established models by average rank order, we were able to obtain the best performing model with the highest

Table 1 Clinical characteristic of the selected patients

Basic information	DKD	DKD + NDKD	NDKD	DKD vs NDKD	DKD vs DKD + NDKD	NDKD vs DKD + NDKD
No. of patients	50	32	159			
Age	51.74 ± 10.86	52.62 ± 10.31	51.63 ± 11.98	0.660	0.797	0.882
Gender (male)	16 (32.0%)	72 (45.3%)	12 (37.5%)	0.157	0.676	0.488
Diabetes duration (years)	9.20 ± 6.76	7.86 ± 5.97	3.20 ± 4.89	< 0.001	0.414	< 0.001
> 3	36 (72.0%)	23 (71.9%)	47 (29.6%)			
> 5	34 (68.0%)	18 (56.2%)	32 (20.1%)			
> 10	19 (38.0%)	9 (28.1%)	13 (8.2%)			
Hypertension (years)	3.00 ± 4.65	5.57 ± 8.68	4.39 ± 6.38	0.765	0.790	0.941
> 3	14 (28.0%)	15 (46.9%)	65 (40.9%)			
> 5	10 (20.0%)	12 (37.5%)	52 (32.7%)			
> 10	7 (14.0%)	9 (28.1%)	31 (19.4%)			
Systolic blood pressure	149.40 ± 24.10	140.19 ± 20.75	137.62 ± 18.73	0.004	0.115	0.482
Diastolic blood pressure	84.82 ± 13.92	83.00 ± 13.08	85.56 ± 12.38	0.420	0.725	0.337
Diabetic retinopathy	35 (70.0%)	14 (43.8%)	21 (13.2%)	< 0.001	0.008	0.087

Table 2 Rank sum results of biochemical and clinical variables

Features	DKD	DKD+NDKD	NDKD	DKD vs NDKD	DKD vs DKD+NDKD	NDKD vs DKD+NDKD
MONO# (median, IQR)	0.43 (0.18)	0.50 (0.24)	0.43 (0.26)	0.022	0.917	0.012
Hematuria (%)	18 (36%)	3 (9%)	87 (55%)	0.006	0.020	<0.001
RDW-CV	12.60 (0.9)	13 (1.425)	12.8 (1)	0.155	0.048	0.222
WBC	6.43 (2.4)	7.23 (2.35)	6.6 (2.85)	0.048	0.170	0.229
DR	35 (70%)	14 (43.8%)	21 (13.2%)	0.008	<0.001	0.087
AT-III	111.5 (13)	107 (26.25)	106 (21.5)	<0.001	0.122	0.457
HbA1c	6.55 (2.83)	6.55 (0.83)	6.55 (0.3)	0.0007	0.223	0.046
Histone	0.04 (0.04)	0.04 (0.11)	0.04 (0)	0.003	0.516	0.116
SG	1.02 (0.01)	1.02 (0.01)	1.02 (0.01)	0.004	0.621	0.069
SBP	145.00 (41.75)	139 (28.5)	136.00 (24)	0.004	0.115	0.482
BMG	2.56 (1.74)	3.58 (3.29)	2.40 (0.69)	<0.001	0.230	<0.001
BUN	7.81 (3.32)	8.03 (6.5)	5.33 (2.76)	<0.001	0.988	<0.001
C3	1.01 (0.2)	1.12 (0.16)	1.14 (0.27)	<0.001	0.789	0.006
Cys-C	1.18 (0.27)	1.51 (0.64)	1.18 (0.29)	<0.001	0.286	<0.001
D-dimer	195.0 (208.75)	279.5 (285.75)	187.0 (228.5)	0.367	0.069	0.008
HCT	0.35 (0.08)	0.36 (0.1)	0.41 (0.07)	<0.001	0.776	<0.001
Hb	119.90±21.15	120.41±26.55	137.10±20.82	<0.001	0.984	<0.001
K	3.97 (0.52)	4.10 (0.68)	3.82 (0.5)	0.014	0.428	0.004
RBC	3.99 (0.94)	4.39 (1.29)	4.66 (0.68)	<0.001	0.453	0.005
SCr	109.00 (65.46)	118.19 (77.79)	74.57 (32.02)	<0.001	0.885	<0.001
TG	1.94 (1.26)	1.66 (1.48)	2.3 (1.87)	0.012	0.406	<0.001
eGFR	89.03 (40.3)	89.03 (37.66)	89.03 (14.73)	<0.001	0.711	0.005

MONO# monocyte count; hematuria, microscopic red blood cell count in urine; *RDW-CV* variation coefficient of red blood cell distribution width; *WBC* White blood cell; *DR* diabetic retinopathy; *AT-III* antithrombin-III; *HbA1c* Glycated hemoglobin; *SG* specific gravity; *SBP* systolic blood pressure; *BMG* β2-micro-globulin; *BUN* urea nitrogen; *C3* complement 3; *Cys-C* cystatin C; *HCT* hematocrit; *Hb* hemoglobin; *K* potassium; *RBC* red blood cell count; *SCr* serum creatinine; *TG* triglycerides; *eGFR**, estimated glomerular filtration rate; DM (year), diabetes duration

*The value of eGFR is calculated through CKD-EPI formula

precision, recall and accuracy score in each pair of the three groups.

Results
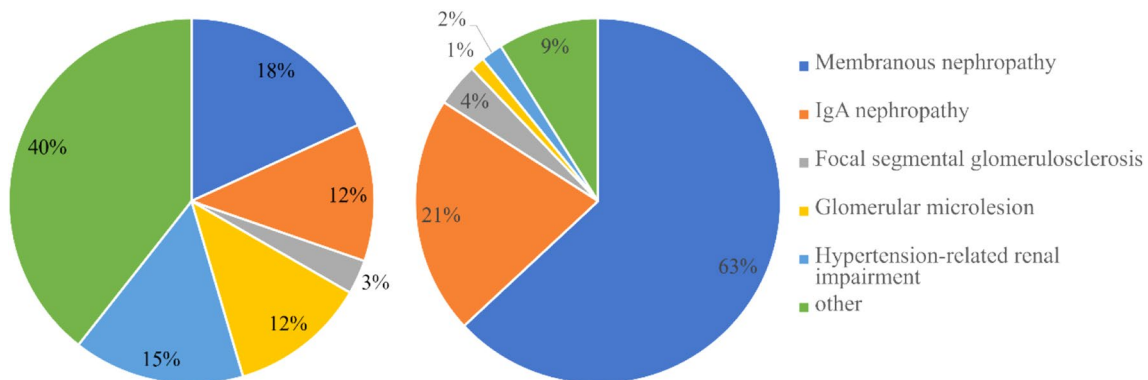
Description and rank sum test for the cohorts

Fig. 3 The pathological types of DKD+NDKD (left) and NDKD (right)

The clinical baseline characteristics are presented in Table 1. The mean age of the patients was 52.57 ± 10.65 (ranging from 24 to 73 years). The ratio of female to male was 1.38:1. The mean of eGFR was 87.6 mL/min/1.73 m^2 in total. After feature selection and pre-processing, the remaining 99 features were divided into two categories: 10 clinical factors and 89 biochemical factors such as blood test, urine test, renal function test and liver function test results. Based on the rank sum results from Table 2, factors with significant differences within each pair were listed. For DKD and DKD + NDKD patients, the factors were hematuria, RDW-CV, WBC, DR ($P < 0.05$). In the DKD and NDKD pair, hematuria, DR, AT-III, HbA1c, Histone, SG, SBP, BMG, BUN, C3, Cys-C, HCT, Hb, RBC, SCr, eGFR and the duration of DM had significant differences ($P < 0.01$). For DKD + NDKD and NDKD, hematuria, BUN, Cys-C, D-DIMER, HCT, Hb, K, RBC, SCr, eGFR and duration of DM showed differences ($P < 0.01$).

Pathological types of renal biopsy

In the retrospective cohort study, 241 patients were analyzed, and the distributions were the following: 50 (20.7%) with DKD, 159 (66.0%) with NDKD and 32 (13.3%) with DKD superimposed on NDKD. The pathological types of NDKD and DKD + NDKD groups after renal biopsy are shown in Fig. 3. The common renal lesions in the DKD + NDKD group were membranous nephropathy and IgA nephropathy, accounting for 6 (18%) and 4 (12%), respectively. Other main types were 1 (3%) focal segmental glomerulosclerosis, 4 (12%) glomerular microlesion and 5 (15%) hypertension-related renal impairments. The remaining types in DKD + NDKD patients were 3 of cute interstitial nephritis, 2 of tubular interstitial injury and 4 of membranous-related renal diseases, 1 of immune renal disease and 1 of hepatitis B virus associated glomerulonephritis. For NDKD patients, the most common kidney injury were also membranous nephropathy and IgA nephropathy,

Table 3 Machine learning comparison

Pair of patients	Include factors (order by its importance)	Precision score	Recall score	Accuracy score	Model
DKD (0) vs NDKD (1)	HCT, DR, Hypertension (year), PDW, hematuria	0.8626 ± 0.0062	0.9170 ± 0.0043	0.8306 ± 0.0057	Logistic regression (penalty: $l1$, $C = 1e - 1$)
DKD (0) vs DKD + NDKD (1)	D-Dimer, hematuria, INR, SBP, DR	0.6822 ± 0.0322	0.4950 ± 0.0271	0.7168 ± 0.0641	Logistic regression (penalty: $l1$, $C = 1$)
DKD + NDKD (0) vs NDKD (1)	hematuria, DM (year), INR, HDL-C, D-dimer, BMG, DR, TP, TG, MPO, MCV, LEU	0.9736 ± 0.022	0.8642 ± 0.035	0.8686 ± 0.052	Kernel SVM (kernel: linear, penalty: $l2$, $C = 5e - 3$)

PDW platelet distribution width; *INR* international normalized ratio; *HDL-C* high-density lipoprotein cholesterol; *TP* total protein; *TG* triglyceride; *MPO* myeloperoxidase; *MCV* mean corpuscular volume; *LEU* urine leukocyte

Figure (a) DKD vs NDKD

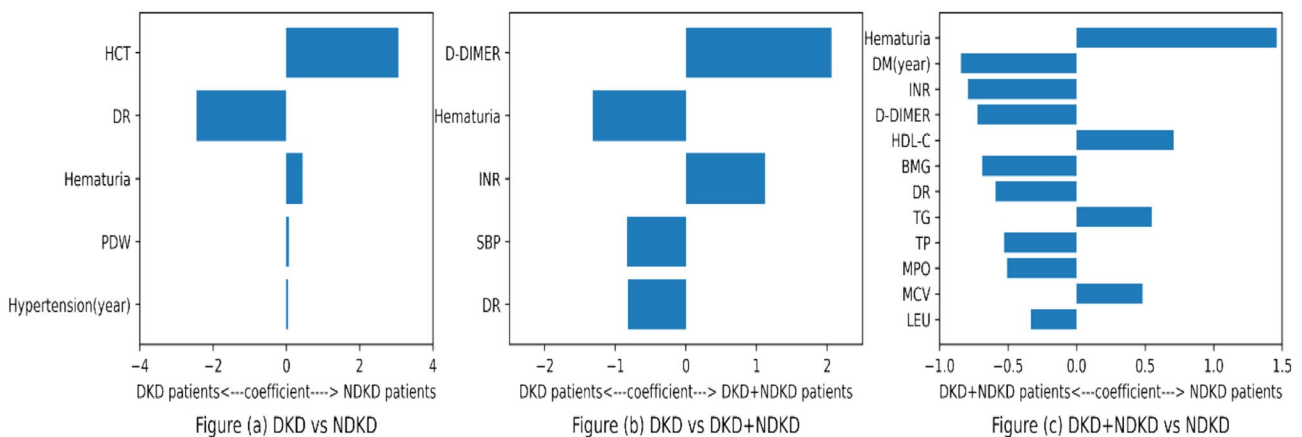Figure (b) DKD vs DKD+NDKD

Figure (c) DKD+NDKD vs NDKD

Fig. 4 Features importance in different pairs of kidney diseases with diabetes patients. The x-axis represents the ranking contribution for different groups of patients. The direction of each contribution represented the positive effect of this feature to this group. The larger absolute value of this feature had, the more important it was in classifying this pair of patients
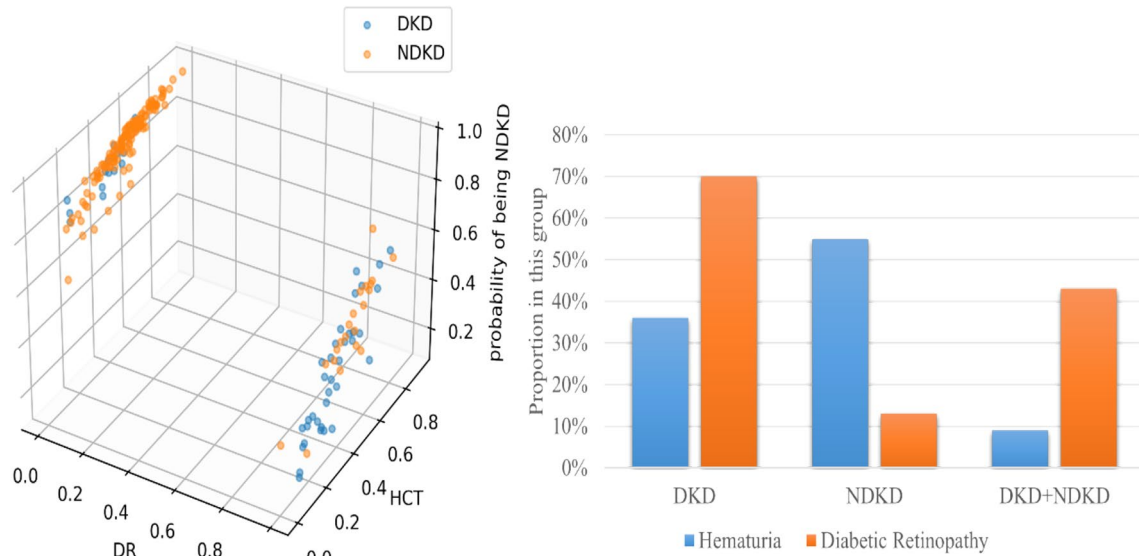
Fig. 5 The scatterplot for diabetic retinopathy (DR) and HCT performance in the model between DKD and NDKD (left); the histogram for DR and hematuria in different groups (right)

accounting for 99 (63%) and 33 (21%), respectively. 6 (4%) patients suffered from focal segmental glomerulosclerosis, 2 (1%) suffered from glomerular microlesion and 3 (2%) suffered from hypertension-related renal impairments in the NDKD group. Other types for NDKD patients included 3 mild hyperplastic glomerulonephritis, 1 membranoproliferative nephritis, 2 ANCA-associated nephritis, 3 chronic interstitial nephritis, 1 nodular diabetic glomerulosclerosis, 1 crescentic glomerulonephritis, 1 obesity-associated nephritis and 1 allergic purpura nephritis.

Classification models and predictors

The models to differentiate various groups are summarized in Table 3 and the risk factors ranked are shown in Fig. 4. The best performing model for the DKD and NDKD pair was Lasso Logistic Regression with precision of 0.8626 ± 0.0062, recall of 0.9170 ± 0.0043. Five feature variables were selected, and their relative importance was HCT > DR > hematuria > PDW > history of hypertension. Being DR and having lower hematuria were more likely to be DKD patients, while higher HCT, higher PDW, and years of hypertension had higher possibility of NDKD, as shown in Fig. 4a. For the DKD and DKD + NDKD pair, the best performing model was Lasso Logistic Regression using 5 features (D-DIMER, hematuria, INR, SBP and DR) with precision of 0.6822 ± 0.0322, recall of 0.4950 ± 0.0271. Figure 4b shows that higher D-DIMER and INR were critical indicators for the DKD + NDKD group. For NDKD and DKD + NDKD patients, the best performing model were Linear SVM using 12 features, the precision score
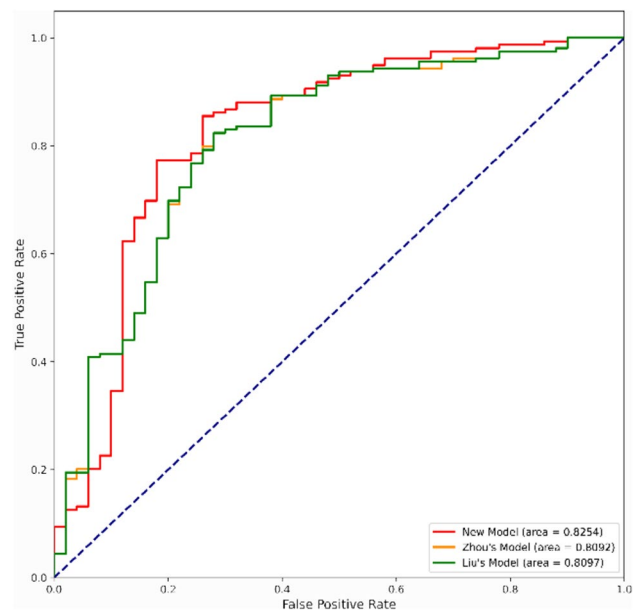

Fig. 6 The ROC curve comparison between our new model of DKD vs NDKD and the existing models

Table 4 DKD vs NDKD model comparison

	AUC	IDI	IDI (*P* value)	NRI
Our model	0.8254	0.000	–	0.000
Zhou's model	0.8092	0.050	< 0.05	0.012
Liu's model	0.8097	0.045	< 0.05	0.008

was 0.9736 ± 0.0220 and recall score was 0.8642 ± 0.0350. Figure 4c shows that hematuria was an important indicator of NDKD; longer duration of diabetes, higher INR, higher D-DIMER, higher BMG, presence of DR was more prone to be DKD + NDKD. The proportion of hematuria and DR is shown in Fig. 5 right panel. For the DKD group, most patients accompanied with DR, while the proportion of DR in NDKD was the lowest. Finally, most NDKD patients suffered from hematuria, while the amount of hematuria in DKD + NDKD patients was the least. All models developed by different machine learning algorithms are shown in Supplementary Table 2.

Model comparison for DKD vs NDKD

We compared our model for differentiating DKD from NDKD with the models in [7] (including DM (year), SBP, HbA1c, hematuria and DR) and in [8] (including years of DM, SBP, hematuria, DR and Hb). To evaluate the performance of each model, receiver operating characteristic (ROC) curve of each model in fivefold cross validation was presented, and the results are shown in Fig. 6. The higher value of area under curve (AUC) indicates the better performance for the model. The results showed that our new models performed better than the previously used models (0.8254). Furthermore, the reclassification improvement (NRI) and integrated discrimination improvement (IDI) were also used for model comparison. Table 4 suggests that our models outperform previous models with higher NRI and IDI score. The IDI index satisfied the z distribution, and it was proved to be significant ($P < 0.05$) with mean 0 based on z test.

Discussion

An accurate noninvasive method to differentiate three types of kidney diseases in T2DM patients can improve the ability of clinical diagnosis and identify patients who did not undergo renal biopsy. It will also reduce the rate of misclassification and misdiagnosis of that population. In this retrospective study, we applied machine learning algorithm to analyze T2DM patients with renal damage and identify distinct features among them. Our models performed better than previously developed ones with high precision values of pairs of groups DKD *vs* NDKD (91.3%) and NDKD vs DKD + NDKD (91.03%). Although the performance of our model for distinguishing DKD from DKD + NDKD was not ideal, the features such as D-Dimer and INR found in the study could serve as meaningful indicators.

Lasso Linear Regression was used for classifying DKD and NDKD and further compared with models from Chen's team [7, 8]. The comparison was based on the similar

enrolled criteria and conducted in a Chinese people, which would increase the reliability and applicability of the results. The reason for the better performance of our model for differentiating DKD and NDKD could be closely related to the choice of weights of multivariate factors by machine learning algorithm. Subsequently, we successfully established a model to identify NDKD with DKD + NDKD using Kernel SVM after analyzing more than 18,000 medical data. This model adequately reflected the advantage of SVM in dealing with small samples and nonlinear feature interactions. Finally, we attempted to construct a model to predict DKD + NDKD and DKD, the precision value of which was 68.22%. Compared with these results, the performance of DKD and DKD + NDKD model was not as accurate, the reason for which might be explained by the smaller sample size of the DKD + NDKD group. We plan to collect more cases of DKD + NDKD to improve the model's performance in a follow-up study.

HCT, DR and hematuria were three most important characteristic variables to differentiate DKD from NDKD, and the others were PDW and history of hypertension. Our results showed that only DR was a positive risk factor in predicting DKD. The predictive value of DR alone had been reported in two meta-analysis. The results showed that DR was a strong predictor in diagnosing or screening for DKD with a pooled sensitivity of 0.65/0.67 and specificity of 0.75/0.78, respectively. Meanwhile, they emphasized the high diagnosis value of proliferative DR in predicting DKD [20, 21]. A recent review also indicated that patients with DR would increase the risk of DKD by 31% [22]. All the evidences indicated that DR was a key element applied in non-invasive clinical diagnosis. In addition, we have shown that HCT, hematuria, PDW and history of hypertension are implicated as risk factors for developing NDKD. The indicator of hematuria and history of hypertension are consistent to the findings in previous studies [7, 9, 10, 15, 23]. Hct is considered as a determinant factor of oxygen supply and blood viscosity and can provide physicians useful information about red blood cell volume [24]. In DKD patients, long-term hyperglycemia leads to renal interstitial hypoxia, aggravates inflammatory response, prevents erythropoietin (EPO) production, resulting in anemia [25]. In contrast, NDKD patients are less closely associated with anemia. PDW reflects platelet activation and is related to microthrombi formation. Patients with DKD exhibit hemostatic abnormalities, which can aggravate renal ischemia and hypoxia under the hypercoagulable state [26]. Our study confirmed that Hct and PDW were valuable predictors in differentiating DKD and NDKD.

We reported that hematuria, duration of diabetes, INR, D-Dimer, HDL-C, BMG, DR, TG, TP, MPO, MCV and LEU as predictors for identifying NDKD and DKD + NDKD. Among these variables, hematuria, HDL-C, TG and MCV

were identified as risk factors for NDKD. These findings indicated that microscopic hematuria and higher level of plasma lipids could be two representative clinical manifestations in NDKD patients in contrast to the DKD + NDKD. So, controlling the modifiable factor of lipid as early as possible might benefit patients with NDKD. The risk factors for the DKD + NDKD mainly reflected in hyperglycemia conditions and coagulation function, such as duration of diabetes and D-dimer. Previous studies suggested that there existed an increased D-dimer level and hypercoagulable state in DKD patients [27]. Hypercoagulation had a tendency to increase thrombosis through upregulation of coagulation factors [28]. So, we inferred that the higher level of D-dimer in DKD + NDKD patients might exhibit thrombotic tendencies under a double damage to the kidney. Therefore, with regard to the DKD + NDKD, it might be necessary to raise the patients' awareness of blood glucose control and improve the ability of health management in this population. Moreover, there was high prevalence of NDKD and the DKD + NDKD (79.25%) in patients underwent kidney biopsy, which again reminded us of the importance of identifying different patterns of kidney diseases in T2DM.

Differentiating DKD + NDKD from DKD was a challenging problem in clinical diagnosis. Current studies were limited to describe the incidence and types of pathology in the DKD + NDKD patients [5, 7, 10, 12, 13, 29]. Based on the successful construction of the two prediction models, we conducted a tentative study on this group and found that D-dimer and INR were two important parameters for predicting DKD + NDKD when compared with DKD. In the present study, the level of D-dimer in the DKD + NDKD was twice more than those of the other groups (shown in Table 2). Previous studies observed higher D-dimer level in DKD patients, which indicated that renal injury was associated with disorder of the coagulation system and alerted the progression of the disease via coagulation-protease-dependent signal [27, 30]. Our findings revealed that it would be beneficial for monitoring coagulation function early in the DKD + NDKD patients.

Based on the comprehensive analysis of the selected features among three groups of patients, we found that hematuria as a common variable included in all three models. T2DM patients were more likely to develop NDKD when hematuria level was high, which was used as an indication for renal biopsy [19, 31]. In the present study, hematuria was indeed an important feature of NDKD, and its relative importance gradually reduced from the first place to the third in three models. The prevalence of hematuria in NDKD (55%) was higher than in the other groups, which is consistent with a previous study [4]. However, there were opposite views that hematuria might be a clinical manifestation of DKD [32]. The discordance might

ascribe to the difference in the selection of subjects and definition of hematuria. Recently, a meta-analysis showed that the predictive value of hematuria in NDKD was low with the pooled sensitivity of 0.42 and specificity of 0.72, and further revealed that urinary dysmorphic erythrocytes might be more effective in predicting NDKD [31].

DR was the second common risk factor in each group. Presence or absence of DR as predictors for DKD or NDKD has been widely studied. On the other hand, it was pointed out that the severity of DR might not be correlated with the presence of DKD [33]. This indicated that DR alone might be insufficient as a predictor of DKD, and it should be combined with other indicators to improve predictive accuracy in T2DM patients with kidney damage. Interestingly, our findings also showed that the importance of DR as a risk factor varied significantly among the three groups of patients. The importance of DR ranked second in distinguishing DKD and NDKD, seventh in NDKD and DKD + NDKD, and fifth in DKD and DKD + NDKD. This suggested that DR was more likely to be related to DM-associated kidney diseases.

Another important indicator for separating DKD from NDKD patients was anemia. It was reported that anemia occurred earlier and more severe in DKD patients due to reduction of erythropoietin production and other pathophysiology mechanisms [34]. Currently, one study from histopathological perspective revealed that severe interstitial fibrosis and tubular atrophy (IF/TA) independent of global glomerulosclerosis caused damage to erythropoietin production, resulting in earlier concurrency of anemia in DKD patients [35]. Therefore, it was worthwhile to further elucidate the mechanism of early anemia and renal diseases progression in T2DM patients.

The study collected the entirety of patients' electronic medical record at the time of kidney biopsy and carried out a comprehensive analysis of the diagnosis of different kidney diseases. Secondly, machine learning techniques applied to large volume and multiple patterns of clinical data could better identify meaningful parameters without using explicit instructions and lead to stable and accurate predictive models. Finally, data with missing variables to some extent could still be useful in research without affecting final results. Nevertheless, there existed several limitations to this study. Firstly, it was a retrospective study in a single center and the recruited patients were limited in one provincial of China, which would decrease the extensive applicability of the results. Secondly, the machine learning techniques may lead to some findings of unknown or unmeaningful factors, which need future study to illustrate their significance. Moreover, the goal to find and establish a differential diagnosis of DKD and the DKD + NDKD has not been realized and needed to be followed up in a future study.

In conclusion, we successfully developed models for distinguishing NDKD from DKD and DKD + NDKD from NDKD in patients with T2DM using machine learning methods. Several meaningful risk factors were identified among these three groups, which will benefit patients with contraindications for renal biopsy. Furthermore, our results also suggested that more attention should be paid to coagulation function in DKD + NDKD patients.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11255-022-03322-1.

Declarations

Conflict of interest The authors have no conflicts of interest to declare.

References

1. Anders HJ, Huber TB, Isermann B, Schiffer M (2018) CKD in diabetes: diabetic kidney disease versus nondiabetic kidney disease. Nat Rev Nephrol 14(6):361–377
2. Thomas MC, Brownlee M, Susztak K, Sharma K, Jandeleit-Dahm KA, Zoungas S et al (2015) Diabetic kidney disease. Nat Rev Dis Primers 30(1):1–20
3. Santoro D, Torreggiani M, Pellicanò V, Cernaro V, Messina RM, Longhitano E et al (2021) Kidney biopsy in type 2 diabetic patients: Critical reflections on present indications and diagnostic alternatives. Int J Mol Sci 22(11):5425
4. Bermejo S, González E, López-Revuelta K, Ibernon M, López D, Martín-Gómez A et al (2020) Risk factors for non-diabetic renal disease in diabetic patients. Clin kidney J 13(3):380–388
5. Soleymanian T, Hamid G, Arefi M, Najafi I, Ganji MR, Amini M et al (2015) Non-diabetic renal disease with or without diabetic nephropathy in type 2 diabetes: clinical predictors and outcome. Ren Fail 37(4):572–575
6. Bermejo S, Pascual J, Soler MJ (2018) The current role of renal biopsy in diabetic patients. Minerva Med 109(2):116–125
7. Zhou JH, Chen XM, Xie YS, Li JJ, Yamanaka N, Tong X (2008) A differential diagnostic model of diabetic nephropathy and non-diabetic renal disease. Nephrol Dial Transplant 23(6):1940–1945
8. Liu MY, Chen XM, Sun XF, Zhou JH, Zhang XG, Zhu HY et al (2014) Validation of a differential diagnostic model of diabetic nephropathy and non-diabetic renal diseases and the establishment of a new diagnostic model. J Diabetes 6(6):519–526
9. Dong ZY, Wang YD, Qiu Q, Zhang XG, Zhang L, Wu J et al (2016) Clinical predictors differentiating non-diabetic renal diseases from diabetic nephropathy in a large population of type 2 diabetes patients. Diabetes Res Clin Pract 121:112–118
10. Kritmetapak K, Anutrakulchai S, Pongchaiyakul C, Puapairoj A (2018) Clinical and pathological characteristics of non-diabetic renal disease in type 2 diabetes patients. Clin kidney J 11(3):342–347
11. Yang ZH, Feng LH, Huang Y, Xia N (2019) A differential diagnosis model for diabetic nephropathy and non-diabetic renal disease in patients with type 2 diabetes complicated with chronic kidney disease. Diabetes Metab Syndr Obes 12:1963–1972
12. Wang JL, Han QQ, Zhao LJ, Zhang JL, Wang YT, Wu YC et al (2019) Identification of clinical predictors of diabetic nephropathy and non-diabetic renal disease in Chinese patients with type 2 diabetes, with reference to disease course and outcome. Acta Diabetol 56(8):939–946
13. Li L, Yang Y, Zhu XJ, Xiong XF, Zeng LF, Xiong S et al (2020) Design and validation of a scoring model for differential diagnosis of diabetic nephropathy and nondiabetic renal diseases in type 2 diabetic patients. J Diabetes 12(3):237–246
14. Fontana F, Perrone R, Giaroni F, Alfano G, Giovanella S, Ligabue G et al (2021) Clinical predictors of nondiabetic kidney disease in patients with diabetes: a single-centre study. Int J Nephrol. https://doi.org/10.1155/2021/9999621
15. Fiorentino M, Bolignano D, Tesar V, Pisano A, Biesen WV, Tripepi G et al (2017) Renal biopsy in patients with diabetes: a pooled meta-analysis of 48 studies. Nephrol Dial Transplant 32(1):97–110
16. Makino M, Yoshimoto R, Ono M, Itoko T, Katsuki T, Koseki A et al (2019) Artificial intelligence predicts the progression of diabetic kidney disease using big data machine learning. Sci Rep 9(1):11862
17. Chan LL, Nadkarni GN, Fleming F, McCullough JR, Connolly P, Mosoyan G et al (2021) Derivation and validation of a machine learning risk score using biomarker and electronic patient data to predict progression of diabetic kidney disease. Diabetologia 64(7):1504–1515
18. Alberti KG, Zimmet PZ (1998) Definition, diagnosis and classification of diabetes mellitus and its complications: report of a WHO consultation. Part 1: diagnosis and classification of diabetes mellitus provisional report of a WHO consultation. Diabet Med 15(7):539–553
19. KDOQI (2007) KDOQI Clinical practice Guidelines and clinical practice recommendations for diabetes and chronic kidney disease. Am J Kidney Dis 49(2 Suppl 2):S12-154
20. Liang S, Zhang XG, Cai GY, Zhu HY, Zhou JH, Wu J et al (2013) Identifying parameters to distinguish non-diabetic renal diseases from diabetic nephropathy in patients with type 2 diabetes mellitus: a meta-analysis. PLoS ONE 8(5):e64184
21. He F, Xia X, Wu XF, Yu XQ, Huang FX (2013) Diabetic retinopathy in predicting diabetic nephropathy in patients with type 2 diabetes and renal disease: a meta-analysis. Diabetologia 56(3):466–475
22. Jiang WH, Wang JY, Shen XF, Lu WL, Wang Y, Li W et al (2020) Establishment and validation of a risk prediction model for early diabetic kidney disease based on a systematic review and meta-analysis of 20 cohorts. Diabetes Care 43(4):925–933
23. Liu SJ, Guo QY, Han HB, Cui PH, Liu X, Miao LN et al (2016) Clinicopathological characteristics of non-diabetic renal disease in patients with type 2 diabetes mellitus in a north-eastern Chinese medical centre: a retrospective analysis of 273 cases. Int Urol Nephrol 48(10):1691–1698

24. Brun JF, Varler-Marie E, Richou M, Raynaud de Mauverger E (2018) Seeking the optimal hematocrit: may hemorheological modelling provide a solution? Clin Hemorheol Microcirc 64(4):493–501

25. Singh DK, Winocour P, Farrington K (2009) Erythropoietic stress and anemia in diabetes mellitus. Nat Rev Endocrinol 5(4):204–210

26. Ferroni P, Basili S, Falco A, Davì G (2004) Platelet activation in type 2 diabetes mellitus. J Thromb Haemost 2(8):1282–1291

27. Pan LL, Ye YL, Wo MY, Bao DN, Zhu FJ, Cheng ML et al (2018) Clinical significance of hemostatic parameters in the prediction for type 2 diabetes mellitus and diabetic nephropathy. Dis Markers. https://doi.org/10.1155/2018/5214376

28. Hess K (2015) The vulnerable blood. Coagulation and clot structure in diabetes mellitus. Hamostaseplogie 35(1):25–33

29. Hsieh JT, Chang FP, Yang AH, Tarng DC, Yang CY (2020) Timing of kidney biopsy in type 2 diabetic patients: a stepwise approach. BMC Nephrol 21(1):131

30. Madhusudhan T, Kerlin BA, Isermann B (2016) The emerging role of coagulation proteases in kidney disease. Nat Rev Nephrol 12(2):94–109

31. Jiang SM, Wang YN, Zhang Z, Dai PL, Yang Y, Li W (2018) Accuracy of hematuria for predicting non-diabetic renal disease in patients with diabetes and kidney disease: a systematic review and meta-analysis. Diabetes Res Clin Pract 143:288–300

32. Li Y, Su XX, Ye Q, Guo XD, Xu B, Guan TJ, Chen A (2021) The predictive value of diabetic retinopathy on subsequent diabetic nephropathy in patients with type 2 diabetes: a systematic review and meta-analysis of prospective studies. Renal fail 43(1):231–240

33. Heine GH, Sester U, Girndt M, Köhler H (2004) Acanthocytes in the urine: useful tool to differentiate diabetic nephropathy from glomerulonephritis. Diabetes Care 27(1):190–194

34. Pappa M, Dounousi E, Duni A, Katopodis K (2015) Less known pathophysiological mechanisms of anemia in patients with diabetic nephropathy. Int Urol Nephrol 47(8):1365–1372

35. Ito K, Yokota S, Watanabe M, Inoue Y, Takahashi K, Himuro N et al (2021) Anemia in diabetic patients reflects severe tubulointerstitial injury and aids in clinically predicting a diagnosis of diabetic nephropathy. Intern Med 60(9):1349–1357

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.