Diabetes, Metabolic Syndrome and Obesity

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

Machine Learning-Based Predictive Modeling of Diabetic Nephropathy in Type 2 Diabetes Using Integrated Biomarkers: A Single-Center Retrospective Study

Ying Zhu*, Yiyi Zhang*, Miao Yang, Nie Tang, Limei Liu, Jichuan Wu, Yan Yang

Department of Endocrinology, Sichuan Provincial People's Hospital, School of Medicine, University of Electronic Science and Technology of China, Chengdu, 610072, People's Republic of China

*These authors contributed equally to this work

Correspondence: Yan Yang, Department of Endocrinology, Sichuan Provincial People's Hospital, School of Medicine, University of Electronic Science and Technology of China, Chengdu, 610072, People's Republic of China, Email yangyan_2012@126.com

Purpose: Diabetic nephropathy (DN), a major complication of diabetes mellitus, significantly impacts global health. Identifying individuals at risk of developing DN is crucial for early intervention and improving patient outcomes. This study aims to develop and validate a machine learning-based predictive model using integrated biomarkers.

Methods: A cross-sectional analysis was conducted on a baseline dataset involving 2184 participants without DN, categorized based on their development of DN over a follow-up period of 36 months: DN (n=1270) and Non-DN (n=914). Various demographic and clinical parameters were analyzed. The findings were validated using an independent dataset comprising 468 participants, with 273 developing DN and 195 remaining as Non-DN over the follow-up period. Machine learning algorithms, alongside traditional descriptive statistics and logistic regression were used for statistical analyses.

Results: Elevated levels of serum creatinine, urea, and reduced eGFR, alongside an increased prevalence of retinopathy and peripheral neuropathy, were prominently observed in those who developed DN. Validation on the independent dataset further confirmed the model's robustness and consistency. The SVM model demonstrated superior performance in the training set (AUC=0.79, F1-score =0.74) and testing set (AUC=0.83, F1-score=0.82), outperforming other models. Significant predictors of DN included serum creatinine, eGFR, presence of diabetic retinopathy, and peripheral neuropathy.

Conclusion: Integrating machine learning algorithms with clinical and biomarker data at baseline offers a promising avenue for identifying individuals at risk of developing diabetic nephropathy in type 2 diabetes patients over a 36-month period.

Keywords: diabetic nephropathy, prediction, machine learning, biomarkers, risk stratification, type 2 diabetes

Introduction

Diabetic nephropathy (DN) is one of the most severe complications of type 2 diabetes mellitus (T2DM) and represents a significant global healthcare burden.[1,2] After entering end-stage renal disease (ESRD), the 5-year survival rate of patients receiving hemodialysis was 41.4%, compared with 46.9% for patients receiving peritoneal dialysis.[3] Early identification and intervention in DN are crucial to prevent the progression of kidney damage and reduce the risk of ESRD. Traditional clinical risk factors, such as glycemic control and blood pressure management, have been valuable in assessing DN risk, but they often lack the sensitivity and specificity needed for early detection.[4,5]

In recent years, there has been a growing interest in exploring the potential of incorporating specific biomarkers into predictive models to enhance the accuracy of DN risk assessment. In particular, integrated biomarker prediction models have made substantial strides in improving risk assessment and prognosis in various medical conditions, such as

Received: 2 February 2024
Accepted: 16 April 2024
Published: 10 May 2024

cardiovascular disease,[6] cancer,[7] and renal disorders.[8] Several potential biomarkers, including albuminuria, estimated glomerular filtration rate (eGFR), neutrophil gelatinase-associated lipocalin (NGAL),[9] and urinary kidney injury molecule-1 (uKIM-1), have emerged as potential early indicators of DN.[10] Han and colleagues employed machine learning algorithms to screen and validate specific biomarkers in patients with DN exhibiting glomerular injury. They identified that Protein Kinase, cAMP-Dependent, Regulatory, Type II, Beta (PRKAR2B) and Transforming Growth Factor Beta-Induced (TGFBI) play crucial roles in the immunological microenvironment and can serve as diagnostic biomarkers.[11] Yu et al analyzed differentially expressed genes in DN patients and found that Apolipoprotein C1 (APOC1) may be a candidate intervention target for DN.[12] These biomarkers have shown promise in various studies, but their integration into a comprehensive predictive model for DN remains an area of active investigation.

Our single-center study sets out to address this gap by developing and rigorously validating a novel biomarker-integrated predictive model for DN in T2DM patients. By weaving together specific biomarkers with traditional clinical data in a robust predictive model, we hope to furnish clinicians with a powerful tool to recognize DN at its nascent stages, thereby enabling timely and personalized interventions to alleviate the burden of this formidable complication.

Methods

Study Design and Participants

This retrospective study, conducted at Sichuan Provincial People's Hospital, aimed to develop and validate a machine learning-based model for predicting DN in type 2 diabetes patients who are managed by informational system. For the development of our model, the training cohort consisted of 2184 T2DM patients, all initially free from DN. To evaluate the performance of our predictive model, we divided the data into two distinct sets: a training set comprising the aforementioned 1270 patients who developed DN and 914 who remained DN-free, and a testing set including 468 patients, with 273 developing DN and 195 remaining DN-free during the follow-up period. Inclusion criteria encompassed adult individuals (≥18 years) diagnosed with T2DM, with regular follow-up visits and comprehensive medical records. We excluded patients with Type 1 Diabetes, other primary renal diseases, or incomplete data on critical variables. This study was approved by the Ethics Committee of Sichuan Provincial People's Hospital. This study was conducted as a retrospective study and patients' informed consent was waived.

Data Collection and Variables

The study utilized historical patient data extracted from electronic health records (EHR), spanning January 2018 to December 2020. Our study cohort included 2652 T2DM patients initially free from DN, diagnosed according to the American Diabetes Association (ADA) criteria:[13] HbA1c ≥ 6.5%, Fasting Blood Sugar (FBS) ≥ 126 mg/dL, or 2-Hour Postprandial Glucose (2HPP) ≥ 200 mg/dL. In our study, DN diagnosis was not solely based on eGFR levels. Instead, it relied on specific criteria, including persistent albuminuria (albumin-to-creatinine ratio ≥30 mg/g in two out of three samples over 3–6 months) and a multifactorial assessment in line with the guidelines and research findings in the field of diabetic kidney disease. Notably, the classification of participants without DN despite an eGFR <60 mL/min/1.73 m² was due to the absence of other DN markers, primarily persistent albuminuria. This approach aligns with the nuanced understanding that reduced eGFR can indicate various stages of kidney health and is not exclusively indicative of DN. In addition, we only included patients with diabetes duration of more than 10 years and excluded patients with other kidney diseases. These criteria are in accordance with the guidelines and research findings in the field of diabetic kidney disease.[14] Patients were followed up for a period of 36 months to monitor the development of DN. For patients with multiple or regular visits to the diabetes clinic, only the initial diagnosis of DN was considered. To calculate HbA1c variability, we employed the standard deviation of HbA1c levels, determined from a minimum of three readings obtained in the 12 months preceding the study baseline. This approach ensured a robust assessment of glycemic variability over time. Patient characteristics were categorized into continuous and categorical variables, with appropriate handling of missing values. Variables collected from the EHR encompassed a broad spectrum of demographic, clinical, and laboratory data. Demographic data included age and gender. Clinical parameters captured were smoking habits, alcohol consumption, exercise habits, age at diabetes diagnosis, duration of diabetes, insulin use, presence of hypertension, systolic and diastolic blood pressure, and medication use for hypertension, cardiovascular diseases, and hyperlipidemia. Additionally, we documented body mass index (BMI), presence of diabetic peripheral neuropathy, diabetic retinopathy, diabetic ketoacidosis, ischemic heart disease, vascular bypass, and peripheral vascular disease. Peripheral neuropathy was assessed using a combination of clinical examination findings and patient-reported symptoms. Specifically, we utilized the Michigan Neuropathy Screening Instrument, which includes both a patient questionnaire to capture symptoms of neuropathy and a physical assessment conducted by trained healthcare professionals to observe signs of neuropathic damage. For the diagnosis of diabetic retinopathy, all participants underwent a detailed eye examination by experienced ophthalmologists, including fundus photography and optical coherence tomography, when available. The grading of retinopathy was based on the International Clinical Diabetic Retinopathy Disease Severity Scale, which classifies the condition into several levels ranging from no apparent retinopathy to proliferative retinopathy. Key laboratory markers relevant to T2DM and DN were also included: total cholesterol, low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C), triglycerides, variation in HbA1c levels, fasting plasma glucose, serum creatinine, blood urea nitrogen (BUN), and eGFR. Variables such as the albumin excretion rate (AER) and the presence of diabetic complications were also considered.

All data were anonymized. Data quality checks ensured the accuracy and completeness of the dataset. Missing values were addressed using established imputation techniques, contributing to the robustness of the dataset for subsequent predictive modeling.

Feature Selection and Elimination for Predictive Modeling of DN

In our study, the feature selection and elimination process was a critical phase in developing a predictive model for DN in patients with T2DM. The process began by tentatively identifying potential predictors of DN based on multivariable logistic regression analysis results. Following this initial statistical analysis, we employed Random Forest (RF) and Support Vector Machine (SVM) algorithms to further refine our feature set. The RF algorithm was utilized for its capacity to handle high-dimensional data and assess the importance of each feature through a bootstrapping method involving multiple decision trees. Features consistently contributing to accurate predictions across various trees were deemed important. The SVM, known for its effectiveness in classification problems, was applied to evaluate how each feature contributed to correctly classifying patients into DN risk categories by constructing hyperplanes in a multi-dimensional space. Features that contributed to a clearer margin of separation in the SVM model were prioritized. The combined use of RF and SVM allowed for a comprehensive evaluation of each potential predictor's role in DN risk stratification, ensuring the final model was not only statistically valid but also practically relevant in a clinical setting.

The performance of the feature set selected through RF and SVM was rigorously evaluated using cross-validation techniques. We partitioned our dataset into training (2, 184) and testing (468) sets to validate the model's performance. The evaluation focused on key metrics such as accuracy, precision, F1 score, and the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve. This dual approach in feature selection and validation ensured that our predictive model for DN was robust, accurate, and clinically applicable.

Statistical Analysis and Machine Learning Models

Descriptive statistics were first employed to summarize the demographic, clinical, and laboratory data. Continuous variables were expressed as means ± standard deviations (SD), while categorical variables were summarized as frequencies and percentages. To compare the characteristics between the DN and non-DN groups, independent t-tests (for continuous variables) and Chi-square tests (for categorical variables) were conducted. A P-value of less than 0.05 was considered statistically significant. In our multivariate logistic regression analysis to identify DN predictors, we recognized the potential for multicollinearity among variables, notably between hypertension, systolic blood pressure, use of anti-hypertensive medication, eGFR, and BUN. To address this: We performed a detailed Variance Inflation Factor (VIF) analysis to quantify multicollinearity among these variables. This analysis helped us understand the extent to which multicollinearity might affect our regression coefficients and the overall model validity. Based on the VIF results, adjustments were made by selecting representative variables or combining correlated variables where clinically and statistically appropriate, ensuring no single variable unduly influenced the model due to high correlation with others. The final model included variables adjusted for multicollinearity, ensuring a more accurate and reliable analysis. This refined approach allowed us to maintain the clinical relevance of our model without compromising statistical integrity. Variables that showed statistical significance in univariate analyses were included in the multivariate model. The results were presented as odds ratios (ORs) with 95% confidence intervals (CIs), providing insights into the strength and direction of the associations.

Prior to modeling, data underwent preprocessing including imputation of missing values using the median for continuous variables and mode for categorical variables. Features were standardized to have a mean of zero and a standard deviation of one. Feature engineering included the creation of interaction terms deemed clinically relevant.

For predictive modeling, we utilized several machine learning algorithms, including Random Forest (RF), Support Vector Machine (SVM), Gradient Boosting, and Logistic Regression. The primary objective was to develop a model that accurately predicts the onset of DN in T2DM patients.

Random Forest (RF): An ensemble learning method that constructs multiple decision trees and merges their predictions. RF is known for its high accuracy and ability to handle large datasets with numerous variables.

Support Vector Machine (SVM): A supervised learning model that identifies the best hyperplane to separate different classes in the data. SVM is effective for classification tasks with complex datasets.

Gradient Boosting: An approach that builds an additive model in a forward stage-wise fashion, allowing optimization of arbitrary differentiable loss functions.

Logistic Regression: A statistical model that estimates the probability of a binary outcome based on one or more predictor variables. The performance of these models was evaluated using cross-validation techniques, specifically a tenfold cross-validation approach. Model performance metrics included accuracy, precision, recall, F1 score, and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). The model exhibiting the best performance metrics was selected as the primary predictive model for DN. Hyperparameter tuning was conducted for each model to optimize performance. We used grid search and random search methods to find the most effective parameter combinations, ensuring that the models were robust and generalized well to new data. The final models were tested on an independent dataset to validate their predictive accuracy. The results from these machine learning models were compared with traditional statistical methods to assess their relative effectiveness in predicting DN in T2DM patients.

Results

Demographic Characteristics of the Patients

The detailed demographic characteristics of the two patient groups were presented in Table 1. The average age in the DN group was slightly higher at 51.50 years compared to 51.00 years in the non-DN group, with this difference being statistically significant (P=0.033). Similar proportions of gender, as well as comparable habits regarding smoking, drinking, and exercising, were observed in both groups. A significantly longer duration of diabetes and a lower frequency of insulin usage were noted in the DN group compared to the non-DN group (P<0.001). Hypertension was found to be more prevalent in the DN group (P=0.013). Significant differences between the groups were observed in terms of systolic blood pressure (P<0.001), diabetic peripheral neuropathy (P=0.003), and diabetic retinopathy (P<0.001). Furthermore, notable differences were detected in levels of HDL-C, HbA1c variability, eGFR, and BUN (P<0.001).

Risk Factors for Diabetic Nephropathy

In the multivariate logistic regression analysis of baseline data with significant differences, the duration of diabetes, insulin usage, hypertension, systolic blood pressure, changes in HbA1c variability, diabetic peripheral neuropathy, diabetic retinopathy, eGFR, BUN, and triglycerides were identified as potential significant risk factors for DN patients. Age, treatment with antihypertensive drugs, and body mass index were not found to be significantly associated with DN (Table 2).

Table 1 Demographic and Clinical Characteristics of Participants at Baseline

Feature	Non-DN (n=914)	DN (n=1270)	P-value
Age (years)	51.0 (35.0, 64.0)	51.5 (36.0, 67.0)	0.033*
Sex (female)	502 (54.9%)	525 (41.34)	0.670
Smoking Habit	442 (48.4%)	647 (50.9%)	0.251
Alcohol Habit	448 (49.0%)	642 (50.6%)	0.506
Exercise Habit	302 (33.0%)	442 (34.8%)	0.522
Duration of Diabetes (years)	21.0 (11.0, 30.0)	29.0 (20.0, 39.0)	<0.001*
Insulin use	413 (45.2%)	131 (10.3%)	<0.001*
Hypertension (%)	423 (46.3%)	656 (51.7%)	0.015*
Systolic Blood Pressure (SBP, mmHg)	115.9 ± 41.7	133.3 ± 21.6	<0.001*
Diastolic Blood Pressure (DBP, mmHg)	80.1 ± 8.7	78.1 ± 9.7	0.931
Medicine (%)			
Hypertension Medicine	404 (44.2%)	625 (49.2%)	0.013*
Cardiovascular Medicine	464 (50.8%)	612 (48.2%)	0.252
Hyperlipidemia Medicine	474 (51.9%)	660 (52.0%)	0.995
Total Cholesterol (mmol/L)	5.33 (4.42, 6.15)	5.31 (4.4125, 6.25)	0.813
LDL_C (mmol/L)	3.075 (2.48, 3.67)	3.09 (2.44, 3.6875)	0.624
HDL_C (mg/dL)	49.0 (34.0, 65.0)	39.97 (25.29, 55.45)	<0.001*
Variation HbA1c (%)	10.0 (5.0, 16.0)	13.92 (8.40, 18.39)	<0.001*
Variation in Fasting Plasma Glucose (mmol/L)	10.06 ± 1.70	10.08 ± 1.55	0.931
Diabetes complications (%)			
Diabetic Peripheral Neuropathy (%)	366 (40.0%)	615 (48.43%)	0.003*
Diabetic Retinopathy, n(%)	476 (52.08%)	870 (68.50%)	<0.001*
Diabetic Ketoacidosis	0.0 (0.0, 1.0)	1.0 (0.0, 1.0)	0.195
Estimated Glomerular Filtration Rate (eGFR, mL/min/1.73 m²)	63.00 (36.00, 90.00)	34.17 (15.00, 61.65)	<0.001*
HbA1c variability	1.04 (0.57, 1.52)	1.00 (0.48, 1.51)	0.311
Serum Creatinine (mg/dL)	1.015 (0.766, 1.251)	1.373 (1.119, 1.263)	<0.001*
Blood Urea Nitrogen (BUN, mg/dL)	20.71 (15.49, 25.23)	21.29 (16.34, 26.05)	0.001*
Body Mass Index (BMI, kg/m²)	23.84 (20.95, 26.78)	24.31 (21.29, 27.37)	0.002*
Triglycerides (mg/dL)	150.25 ± 58.56	180.79 ± 58.85	<0.001*
Ischemic Heart Disease (%)	459 (50.2%)	618 (48.7%)	0.500
Peripheral Vascular Disease (%)	469 (51.3%)	654 (51.5%)	0.967

Notes: Data are presented as mean ± standard deviation, median (25th percentile, 75th percentile), or number (percentage). P-values marked with an asterisk (*) indicate statistical significance (P < 0.05).

Abbreviations: LDL-C, Low-Density Lipoprotein Cholesterol; HDL-C, High-Density Lipoprotein Cholesterol; HbA1c, Hemoglobin A1c; eGFR, Estimated Glomerular Filtration Rate; Scr, Serum Creatinine; BUN, Blood Urea Nitrogen; BMI, Body Mass Index.

Comparison of Machine Learning Models

After establishing machine learning models using significant risk factors, these models were validated on both training and testing sets. In the training set, the SVM demonstrated superior performance (AUC=0.80, F1-score=0.74). This performance surpassed other models including RF, Extra Trees, and Gradient Boosting, indicating that SVM was more effective in accurately identifying patients at risk of diabetic nephropathy based on the training data (Table 3, Figure 1). When tested on the validation dataset, the proposed SVM model again exhibited the highest accuracy (AUC=0.83, F1-score=0.82) (Table 4, Figure 2). The consistency of performance across the training and validation datasets underscored the robustness and reliability of the SVM in predicting DN in patients with T2DM.

Predictive Features of the Model

Furthermore, this study identified the most significant features used in the predictive models. The results revealed that serum creatinine and estimated eGFR were among the most crucial predictors across all models. The presence of diabetic retinopathy also emerged as an important predictive factor. Other key features included HbA1c variability, reflecting

Table 2 Multivariate Logistic Regression Analysis of Risk Factors for DN

Risk Factor	Odds Ratio (OR)	95% Confidence Interval (CI)	P-value
Age (years)	1.02	(1.00, 1.04)	0.051
Duration of Diabetes (years)	1.05	(1.03, 1.07)	<0.001*
Insulin use	1.50	(1.20, 1.85)	<0.001*
Hypertension (%)	1.30	(1.10, 1.55)	0.002*
SBP (mmHg)	1.01	(1.00, 1.02)	0.025*
Hypertension Medicine	0.90	(0.75, 1.10)	0.300
HDL_C (mmol/L)	0.95	(0.91, 0.99)	0.020*
Variation HbA1c (%)	1.20	(1.10, 1.30)	<0.001*
Diabetic Peripheral Neuropathy (%)	2.00	(1.60, 2.50)	<0.001*
Diabetic Retinopathy (%)	2.80	(2.20, 3.60)	<0.001*
eGFR (mL/min/1.73 m²)	0.97	(0.95, 0.99)	0.005*
BUN (mg/dL)	1.03	(1.01, 1.05)	0.010*
Serum Creatinine (mg/dL)	1.15	(1.05, 1.25)	0.003*
BMI (kg/m²)	1.02	(0.99, 1.05)	0.120
Triglycerides (mg/dL)	1.01	(1.00, 1.02)	0.030*

Notes: Data are presented as the odds ratio (OR) with the 95% confidence interval (CI) for each risk factor. P-values marked with an asterisk (*) indicate statistical significance (P < 0.05).

Abbreviations: SBP, Systolic Blood Pressure; HDL-C, High-Density Lipoprotein Cholesterol; HbA1c, Hemoglobin A1c; eGFR, Estimated Glomerular Filtration Rate; BUN, Blood Urea Nitrogen; Scr, Serum Creatinine; BMI, Body Mass Index.

Table 3 Overall Performance Evaluation for Applied Predictive Models Using Training Dataset

Algorithms	Accuracy	AUC	Precision	Recall	F1-Score
Random Forest	0.78	0.75	0.72	0.68	0.70
Extra Trees	0.76	0.73	0.70	0.69	0.69
Gradient Boosting	0.80	0.77	0.74	0.71	0.72
Support Vector Machine	0.81	0.79	0.75	0.73	0.74

average blood glucose levels, and urea, indicating blood nitrogen levels. These features were ranked according to their importance in the predictive models, with serum creatinine and estimated eGFR consistently being the most significant predictors (Table 5).

Discussion

Our study successfully developed a machine learning-based model for predicting DN using integrated biomarkers. Among four different algorithms, we found that the SVM exhibited superior performance in predicting DN. SVM, a supervised learning method for data classification, operates on the principles of minimizing empirical risk and maximizing the margin between classes, ensuring high classification accuracy even with a limited number of samples.[15] Currently, SVM has been applied in the auxiliary diagnosis of various diseases. In a recent study, compared to other machine learning models, the SVM model demonstrated effective performance in classifying Parkinson's disease patients, with an accuracy of up to 92.3% achieved through Bayesian optimization.[16] Almansour and colleagues suggested that Artificial Neural Networks outperformed SVM in the early diagnosis of chronic kidney disease using machine learning.[17] Another study indicated that among Gradient Boosting Machines, SVM, Logistic Regression, and RF algorithms, the RF algorithm showed the best performance in predicting the progression to End-Stage Renal Disease, exhibiting the highest AUC and ACC.[18] Although different from our findings, our study aimed to provide a novel method for early diagnosis and risk stratification for individuals at risk of DN within 36 months. This was similar to the findings

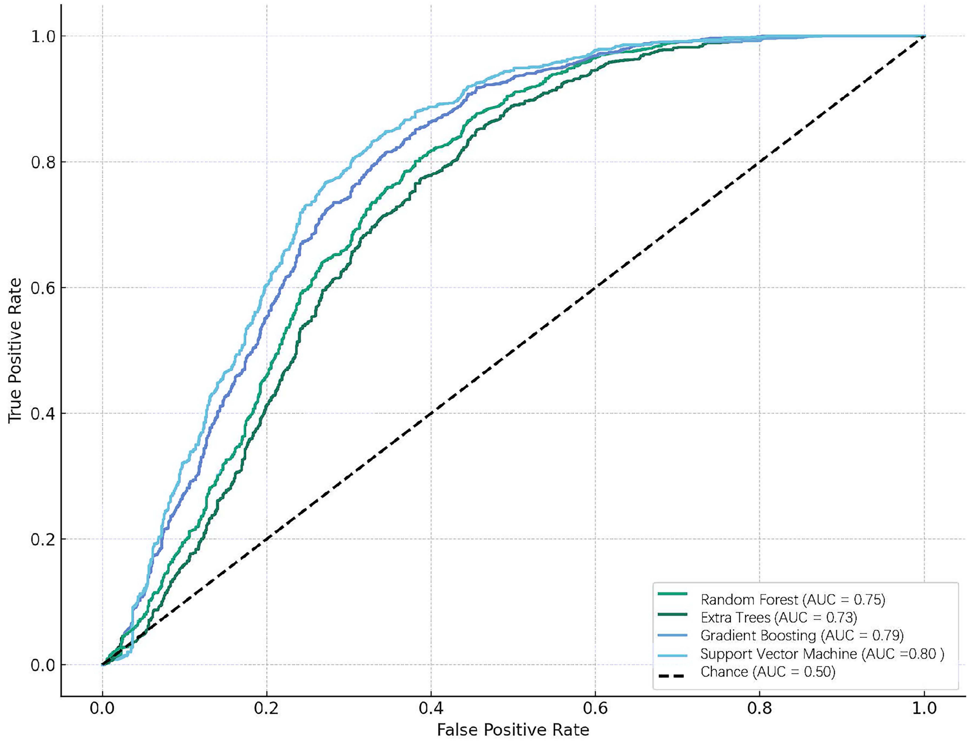

Figure 1 ROC curve for prediction of DN over 36 months in the train dataset.

of Leung and colleagues, who demonstrated that SVM showed the best predictive accuracy in their study on the association between inflammatory and lipid metabolism-related gene polymorphisms and DN, compared to other algorithms.[19]

Our results highlighted that elevated serum creatinine, urea, a reduction in eGFR, and the presence of retinopathy and peripheral neuropathy were identified as key biomarkers associated with DN. Serum creatinine, commonly used to assess kidney function, was observed to reflect a decline in renal filtration capacity when levels were elevated.[20] In the context of diabetic nephropathy, an increase in creatinine levels was often indicative of glomerular damage and reduced function.[21] This correlation was confirmed in our study, where elevated serum creatinine levels were significantly associated with an increased risk of DN. Urea, as another indicator of renal function, was found to suggest weakened renal excretory function when elevated. The progression of DN was marked by changes in urea levels, potentially reflecting a decline in the kidney's ability to process metabolic waste, linked to an exacerbation of renal pathology.[22] The eGFR served as a crucial metric for assessing overall kidney function.[23] However, a study conducted in Japan revealed no significant difference in subsequent eGFR changes between individuals who received routine renal disease monitoring and those who did not,[24] which contrasted with our findings. In our research, a reduction in eGFR was closely associated with the development of DN, aligning with previous studies,[25] and indicating that a decline in eGFR was a significant

Table 4 Performance Evaluation of Predictive Models on the Validation Dataset

Algorithms	Accuracy	AUC	Precision	Recall	F1-Score
Logistic Regression	82.0	0.78	80.0	75.0	0.77
Random Forest	80.0	0.77	78.0	74.0	0.76
Gradient Boosting	84.0	0.81	83.0	79.0	0.81
Support Vector Machine	85.0	0.83	82.0	81.0	0.82

Note: AUC, area under the receiver operating characteristic curve.

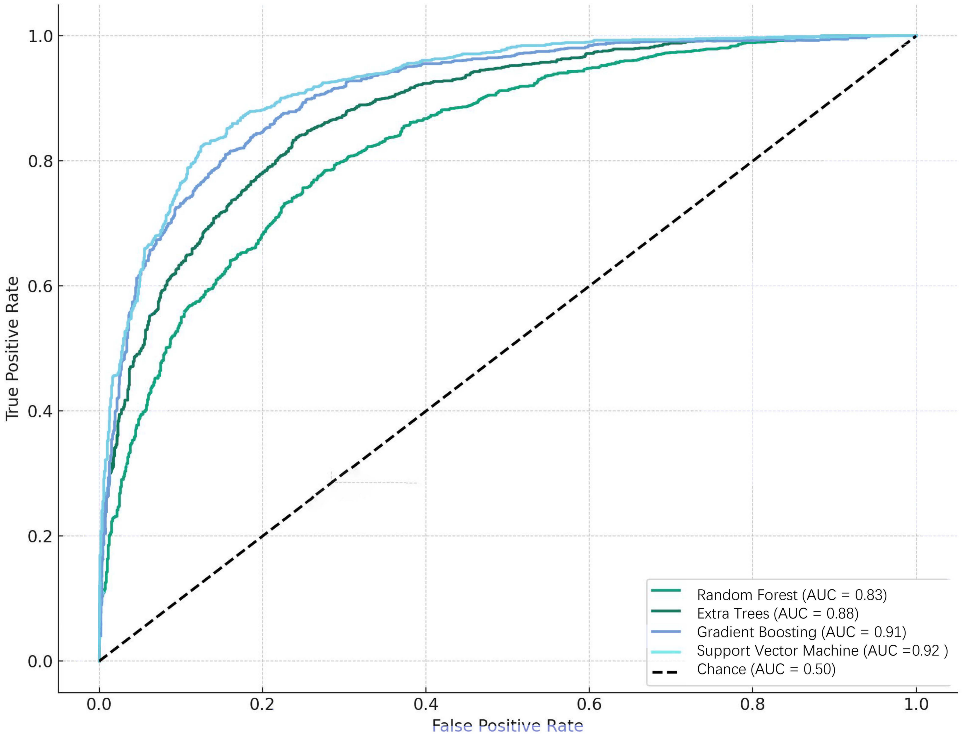

Figure 2 ROC curve for prediction of DN over 36 months in the validation dataset.

predictor of DN progression. The effectiveness of this marker, however, might vary depending on the stage of DN in different patients. Retinopathy and peripheral neuropathy, as prevalent complications of diabetes, were not only indicative of the disease's impact on the eyes and nervous system but were also found to be potentially related to the development of renal lesions. This association was consistent with the findings of Saini et al.[26] In our study, the presence of these complications was linked to an increased risk of DN, possibly due to the cumulative effect of microvascular damage caused by diabetes across various organs.

In summary, the presence of elevated serum creatinine, urea, reduced eGFR, retinopathy, and peripheral neuropathy not only holds significant clinical relevance for the diagnosis and monitoring of DN, but also demonstrated their

Table 5 Predominant Predictors for Diabetic Nephropathy Across Multiple Machine Learning Models

Predictive Feature	Logistic Regression	Random Forest	Gradient Boosting	Support Vector Machine	Frequency	Description
Serum Creatinine (mg/dL)	2	1	3	2	4	Indicator of renal function; elevated levels correlate with DN risk
Estimated GFR (mL/min/1.73 m²)	1	3	1	5	3	Filtration efficiency of the kidneys; decreased in DN
Presence of Diabetic Retinopathy	3	2	2	1	4	Ocular manifestation of diabetes; associated with nephropathy
Glycated Hemoglobin (HbA1c, %)	5	6	4	3	3	Reflects average glucose levels; higher values indicate poor glycemic control

(*Continued*)

Table 5 (Continued).

Predictive Feature	Logistic Regression	Random Forest	Gradient Boosting	Support Vector Machine	Frequency	Description
Urea (mg/dL)	4	5	6	4	3	Blood nitrogen levels; increased with renal impairment
Blood Pressure (SBP/DBP, mmHg)	7	4	5	6	3	Hypertension is a known risk factor for DN
Duration of Diabetes (years)	6	7	7	7	3	Longer duration often correlates with an increased risk of complications
Peripheral Neuropathy Presence	8	8	8	8	3	Neuropathic complications of diabetes; indicative of systemic disease progression
Triglycerides (mmol/L)	6	14	4	10	4	Dyslipidemia is linked with diabetic kidney disease

Notes: The numerical values represent the rank of importance attributed to each feature by the corresponding predictive model, with a lower number denoting higher importance. The "Frequency" column denotes the count of models that identified the feature as a significant predictor for diabetic nephropathy (DN). Features were deemed significant if they appeared as key predictors in at least three out of the four models. Statistical significance was set at a p-value of <0.05.
Abbreviations: GFR, Glomerular Filtration Rate; SBP, Systolic Blood Pressure; DBP, Diastolic Blood Pressure; HbA1c, Hemoglobin A1c.

importance as predictive factors in our machine learning model. These findings underscore the importance of monitoring and assessing these biomarkers in the management of DN, while also providing new directions for future research.

Moreover, this study has its limitations. Firstly, as a retrospective study, inherent limitations include potential selection and information biases. Our sample was drawn from a single medical center, which may limit the generalizability and extrapolation of our findings. The development of diabetic nephropathy may vary significantly across different regions and populations, hence, our results might not be entirely applicable to a broader demographic. Secondly, the study included a limited comparison of algorithms, necessitating further comparisons with a wider range of machine learning methods. Lastly, our research primarily focused on biomarkers and clinical parameters, without delving into other factors such as lifestyle, dietary habits, and socio-economic factors that might influence the progression of diabetic nephropathy. The absence of these factors could limit the comprehensiveness and predictive capability of our model.

Conclusion

In conclusion, these results demonstrate the efficacy of machine learning models, particularly the SVM classifier, in predicting DN in patients with T2DM. The identification of key predictive features further enhances our understanding of the disease mechanisms and offers potential pathways for early intervention and targeted treatment in high-risk populations. In future research, we aim to validate our model in a broader and more diverse population to enhance its generalizability and practical utility.

Data Sharing Statement

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Ethics Statement

This study was approved by the Ethics Committee of Sichuan Provincial People's Hospital. According to the guidance of the Ethics Committee of Sichuan Provincial People's Hospital, this study, as a retrospective study, did not involve direct intervention or potential risks to patients, and informed consent requirements for patients could be waived. This study strictly adhered to the principles of the Declaration of Helsinki, and all patient information had been anonymized prior to collection to prevent any possible identification. Only authorized researchers had access to the data, and access was limited to research purposes. All data storage and transmission processes are encrypted to protect data security.

Author Contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Funding

This study was supported by the Science and Technology Support Program of Sichuan Province (No. 2016FZ0090), Beijing Medical Award Foundation-Ruichuang Chronic Disease Management Fund (No. YXJL-2023-0314-0025).

Disclosure

The authors declare that they have no conflicts of interest in this work.

References

1. Thipsawat S. Early detection of diabetic nephropathy in patient with type 2 diabetes mellitus: a review of the literature. *Diab Vasc Dis Res*. 2021;18 (6):14791641211058856. doi:10.1177/14791641211058856
2. Gupta S, Dominguez M, Golestaneh L. Diabetic Kidney Disease: an Update. *Med Clin North Am*. 2023;107(4):689–705. doi:10.1016/j.mcna.2023.03.004
3. Gupta R, Woo K, Yi JA. Epidemiology of end-stage kidney disease. *Semin Vasc Surg*. 2021;34(1):71–78. doi:10.1053/j.semvascsurg.2021.02.010
4. Cheneke W, Suleman S, Yemane T, et al. Assessment of glycemic control using glycated hemoglobin among diabetic patients in Jimma University specialized hospital, Ethiopia. *BMC Res Notes*. 2016;9:96. doi:10.1186/s13104-016-1921-x
5. Hadjkacem F, Triki F, Frikha H, et al. Masked arterial hypertension in patients with type 2 diabetes mellitus: prevalence, associated factors and cardiovascular impact. *Ann Cardiol Angeiol*. 2022;71(3):136–140. doi:10.1016/j.ancard.2021.10.018
6. Vernooij LM, van Klei WA, Moons KG, et al. The comparative and added prognostic value of biomarkers to the Revised Cardiac Risk Index for preoperative prediction of major adverse cardiac events and all-cause mortality in patients who undergo noncardiac surgery. *Cochrane Database Syst Rev*. 2021;12(12):CD013139. doi:10.1002/14651858.CD013139.pub2
7. Fang Z, Hang D, Wang K, et al. Risk prediction models for colorectal cancer: evaluating the discrimination due to added biomarkers. *Int, J, Cancer*. 2021;149(5):1021–1030. doi:10.1002/ijc.33621
8. Gao J, Ye F, Han F, et al. A radiogenomics biomarker based on immunological heterogeneity for non-invasive prognosis of renal clear cell carcinoma. *Front Immunol*. 2022;13:956679. doi:10.3389/fimmu.2022.956679
9. Motawi TK, et al. Potential serum biomarkers for early detection of diabetic nephropathy. *Diabet Res Clin Pract*. 2018;136:150–158. doi:10.1016/j.diabres.2017.12.007
10. Penno G, Solini A, Bonora E, et al. Clinical significance of nonalbuminuric renal impairment in type 2 diabetes. *J Hypertens*. 2011;29 (9):1802–1809. doi:10.1097/HJH.0b013e3283495cd6
11. Han H, Chen Y, Yang H, et al. Identification and Verification of Diagnostic Biomarkers for Glomerular Injury in Diabetic Nephropathy Based on Machine Learning Algorithms. *Front Endocrinol*. 2022;13:876960. doi:10.3389/fendo.2022.876960
12. Yu K, Li S, Wang C, et al. APOC1 as a novel diagnostic biomarker for DN based on machine learning algorithms and experiment. *Front Endocrinol*. 2023;14:1102634. doi:10.3389/fendo.2023.1102634
13. American Diabetes A. Classification and Diagnosis of Diabetes: standards of Medical Care in Diabetes-2021. *Diabetes Care*. 2021;44(Suppl 1):S15–S33. doi:10.2337/dc21-S002
14. Ikizler TA, Burrowes JD, Byham-Gray LD, et al. KDOQI Clinical Practice Guideline for Nutrition in CKD: 2020 Update. *Am J Kidney Dis*. 2020;76(3 Suppl 1):S1–S107. doi:10.1053/j.ajkd.2020.05.006
15. Valkenborg D, Rousseau A-J, Geubbelmans M, et al. Support vector machines. *Am J Orthod Dentofacial Orthop*. 2023;164(5):754–757. doi:10.1016/j.ajodo.2023.08.003
16. Elshewey AM, Shams MY, El-Rashidy N, et al. Bayesian Optimization with Support Vector Machine Model for Parkinson Disease Classification. *Sensors*. 2023;23(4):2085. doi:10.3390/s23042085
17. Almansour NA, Syed HF, Khayat NR, et al. Neural network and support vector machine for the prediction of chronic kidney disease: a comparative study. *Comput Biol Med*. 2019;109:101–111. doi:10.1016/j.compbiomed.2019.04.017
18. Zou Y, Zhao L, Zhang J, et al. Development and internal validation of machine learning algorithms for end-stage renal disease risk prediction model of people with type 2 diabetes mellitus and diabetic kidney disease. *Ren Fail*. 2022;44(1):562–570. doi:10.1080/0886022x.2022.2056053
19. Leung RK, Wang Y, Ma RC, et al. Using a multi-staged strategy based on machine learning and mathematical modeling to predict genotype-phenotype risk patterns in diabetic kidney disease: a prospective case-control cohort analysis. *BMC Nephrol*. 2013;14(1):162. doi:10.1186/1471-2369-14-162
20. Perrone RD, Madias NE, Levey AS. Serum creatinine as an index of renal function: new insights into old concepts. *Clin Chem*. 1992;38 (10):1933–1953. doi:10.1093/clinchem/38.10.1933
21. Nikolov D, Stoyanova VK, Vladimirova-Kitova L, et al. Analysis and evaluation of correlation between DNA polymorphism in the genes MTHFR, PAI-1 and serum creatinine, creatinine clearance and albumin/creatinine ratio in morning urine of patients with type 2 diabetes mellitus and diabetic nephropathy. *Folia Med (Plovdiv)*. 2022;64(6):896–904. doi:10.3897/folmed.64.e67912
22. Cao P, Huang B, Hong M, et al. Association of amino acids related to urea cycle with risk of diabetic nephropathy in two independent cross-sectional studies of Chinese adults. *Front Endocrinol*. 2022;13:983747. doi:10.3389/fendo.2022.983747

23. Tonneijck L, Muskiet MHA, Smits MM, et al. Glomerular Hyperfiltration in Diabetes: mechanisms, Clinical Significance, and Treatment. *J Am Soc Nephrol*. 2017;28(4):1023–1039. doi:10.1681/ASN.2016060666
24. Ono S, Ono Y, Koide D, et al. Association Between Routine Nephropathy Monitoring and Subsequent Change in Estimated Glomerular Filtration Rate in Patients With Diabetes Mellitus: a Japanese Non-Elderly Cohort Study. *J Epidemiol*. 2020;30(8):326–331. doi:10.2188/jea.JE20180255
25. Rudberg S, Persson B, Dahlquist G. Increased glomerular filtration rate as a predictor of diabetic nephropathy--an 8-year prospective study. *Kidney Int*. 1992;41(4):822–828. doi:10.1038/ki.1992.126
26. Saini DC, Kochar A, Poonia R. Clinical correlation of diabetic retinopathy with nephropathy and neuropathy. *Indian J Ophthalmol*. 2021;69 (11):3364–3368. doi:10.4103/ijo.IJO_1237_21