Journal of Diabetes & Metabolic Disorders (2022) 21:1433–1441
https://doi.org/10.1007/s40200-022-01076-2

RESEARCH ARTICLE

Predicting diabetic nephropathy in type 2 diabetic patients using machine learning algorithms

Seyyed Mahdi Hosseini Sarkhosh[1] · Alireza Esteghamati[2] · Mahboobeh Hemmatabadi[3] · Morteza Daraei[4]

Received: 9 February 2022 / Accepted: 19 June 2022 / Published online: 26 July 2022

Abstract

Background Global healthcare centers today are challenged by the dramatic increase in the prevalence of diabetes. Also, complications from diabetes are a major cause of deaths worldwide. One of the most frequent microvascular complications in diabetic patients is diabetic nephropathy (DN) which is the leading cause of death and end-stage renal disease (ESRD). Despite the different risk factors for DN identified in previous research, machine learning (ML) methods can help determine the importance of the predictors and prioritize them.

Objective The main focus of this investigation is on predicting the incidence of DN in type 2 diabetic mellitus (T2DM) patients using ML algorithms.

Methods Demographic information, laboratory results, and examinations on 6235 patients with T2DM covering a period of 10 years (2011–2020) were extracted from the electronic database of the Diabetes Clinic of the Imam Khomeini Hospital Complex (IKHC) in Iran. Recursive feature elimination using the cross-validation (RFECV) technique was then used with the three classification algorithms to select the important risk factors. Next, five ML algorithms were used to construct a predictive model for DN in T2DM patients. Finally, the results of the algorithms were evaluated according to the AUC criteria and the one with the best performance in terms of prediction and classification was selected.

Results The 18 DN risk factors selected by RFECV were age, diabetes duration, BMI, SBP, hypertension, retinopathy, ALT, CVD, 2HPP, uric acid, HbA1c, waist-to-hip ratio, cholesterol, LDL, HDL, FBS, triglyceride, and serum insulin. Based on a 10-fold cross-validation, the best performance among the five classification algorithms was that of the random forest with 85% AUC.

Conclusions This investigation validates the known risk factors for DN and emphasizes the importance of controlling the blood pressure, weight, cholesterol, and blood sugar of T2DM patients. In addition, as an example of the application of ML approaches in medical predictions, the findings of this study demonstrate the advantages of using these techniques.

Keywords Type 2 diabetes · Diabetic nephropathy · Machine learning · Risk factors · Imam Khomeini Hospital Complex

✉ Seyyed Mahdi Hosseini Sarkhosh
sm.hosseini@fmgarmsar.ac.ir

✉ Mahboobeh Hemmatabadi
hemmatabadi55@yahoo.com

Alireza Esteghamati
esteghamati@tums.ac.ir

Morteza Daraei
mortezadaraie@gmail.com

[1] Department of Industrial Engineering, University of Garmsar, Garmsar, Iran

[2] Department of Internal Medicine, School of Medicine, Psychosomatic Medicine Research Center, Imam Khomeini Hospital, Tehran University of Medical Sciences, Tehran, Iran

[3] Department of Internal Medicine, School of Medicine, Vali Asr Hospital, Tehran University of Medical Sciences, Tehran, Iran

[4] Department of Internal Medicine, School of Medicine, Imam Khomeini Hospital, Tehran University of Medical Sciences, Tehran, Iran

Introduction

One of the most important non-communicable and chronic diseases is type 2 diabetes mellitus (T2DM). The prevalence of diabetes increased from 5.1% of the global adult population (aged 20–79 years) in 2003 to 8.8% in 2017 and is forecast to reach 9.9% by 2045 [1]. A principal cause of disease and death is complications from diabetes which afflict many people around the world. In addition, more than 80% of diabetes treatment costs are for treating complications arising from it [2]. A major microvascular complication in diabetic patients and a principal cause of end-stage renal disease (ESRD) and death in T2DM patients is diabetic nephropathy (DN) [3–5]. According to statistics, nearly 40% of all diabetic patients are affected by DN [6].

Due to the expanding application of large data systems in hospitals and clinics [7], data analysis can be extremely useful in enhancing the accuracy of diagnoses, improving results, and reducing costs in health care systems [8]. In this regard, successful implementation of the machine learning (ML) approach can enhance the productivity of health systems and the effectiveness of services provided by physicians [9]. Despite the greater complexity of ML compared to conventional statistical analysis, its inherent advantages in knowledge discovery procedures lie in identifying latent risk factors for different diseases, the potential for generating novel hypotheses, making personalized risk profiles, and customizing clinical judgments in high-dimensional data [10]. In addition, multivariate ML models have the advantage of integration and instruction and can be used in clinical settings. By extracting knowledge from big data databases, such as electronic health records (EHR), it is possible to perform disease screening methods on a large scale, with greater accuracy, and enhanced customization that ultimately aid physicians in precise judgment and decision-making. ML models can make significant improvements in physicians' diagnosis accuracy [11] and have been used for predicting various diseases [12, 13] including diabetes [14, 15], identifying hypertension in diabetic patients [16], and classifying diabetic patients with cardiovascular disease [17].

A key focus of this study is in identifying important risk factors for DN to prevent its occurrence in T2DM patients. Although the ML approach has not been widely used in this field, some scholars have attempted to model the prediction of DN occurrence. In one study, for example, medical data were collected from 4321 diabetic patients, and the process took ten years in order to develop a novel system with the support vector machine classification algorithm. This model could estimate the initiation of DN at least two-to-three months earlier than the actual diagnosis and with higher predictive accuracy [18].

In another study, data from 10,251 patients using the Action to Control Cardiovascular Risk in Diabetes (ACCORD) and six types of machine learning algorithms were used to develop a DN risk prediction model. The results showed that the logistic regression and random forest model provided the most favorable prediction performance in both train and test datasets. Among the available predictors, eGFR decline was the most important determinant of DN progression [1].

In another study, researchers used the Roche/IBM algorithm based on the LR algorithm as well as a real-world data model for the initiation of chronic kidney disease in patients with diabetes and made a comparison between its performance and the results of several large randomized controlled trials (RCTs). This study showed that the results based on the proposed algorithm were consistent with or even more accurate than those employing randomized controlled trial data [19].

Although models have already been developed for DN prediction, determining the importance and priority of the identified risk factors and further calibration of prediction models, especially with the ML approach, still need further investigation and exploration. Thus, this research mainly aimed at identifying the important risk factors in predicting DN among T2DM patients using ML algorithms and patient data from the electronic database of the Diabetes Clinic of the Imam Khomeini Hospital Complex (IKHC) in Iran.

Methods

Figure 1 shows the steps taken to develop a new predictive model for identifying risk factors for DN in T2DM patients. These include collecting raw data from diabetic patients and data preprocessing, selecting and sorting the most important features (risk factors), and finally evaluating and validating the prediction models.

In Step 1, the information of patients with T2DM was obtained from the electronic database of the Diabetes Clinic of the IKHC. This dataset included demographic information, physical examinations, blood and urine laboratory results, and the history of diabetic patients referred to this medical center over a period of 10 years (2011–2020). Samples with missing values were processed and normalization performed based on each of the features.

Step 2 involved reducing the quantity of identified characteristics in order to determine the most significant ones and to enhance computational accuracy. In selecting the risk factors with the best performance and the most importance in DN prediction, recursive feature elimination using cross-validation (RFECV) with the following three classification algorithms was used: random forest (RF), extreme gradient boosting (XGBoost), and support

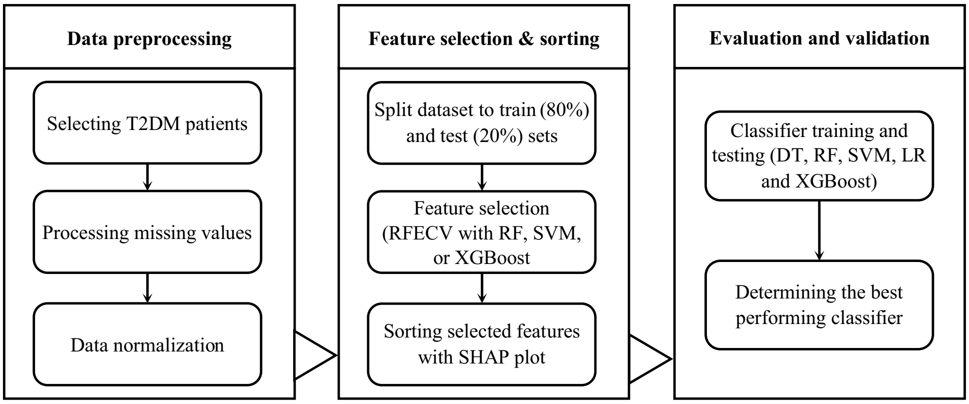
Data preprocessing
Selecting T2DM patients
Processing missing values
Data normalization
Feature selection & sorting
Split dataset to train (80%) and test (20%) sets
Feature selection (RFECV with RF, SVM, or XGBoost
Sorting selected features with SHAP plot
Evaluation and validation
Classifier training and testing (DT, RF, SVM, LR and XGBoost)
Determining the best performing classifier

Fig. 1 Steps for developing and evaluating DN prediction model

vector machine (SVM). This method allows the optimal number of features and the importance of each in different conditions to be determined. Moreover, to model and evaluate the classification algorithms, the main dataset was divided into train (80%) and test (20%) sets.

In Step 3, after extracting the important features from the original sample, a new sample of the dataset was selected. The new sample was then divided into train datasets and test datasets. Due to the characteristics of the dataset and the labeled outcomes, the decision tree (DT), SVM, LR (logistic regression), RF, and XGBoost supervised ML algorithms were used to construct the prediction model. DT achieves good results in classification work and has good interpretability and generalizability. SVM has a strong theoretical basis and can ensure that the extremum solution is the optimal global solution rather than the local minimum, meaning that it has the ability to generalize well to unknown samples. LR is also a simple and efficient algorithm that provides the regression coefficients and confidence intervals. RF and XGBoost are two algorithms from the ensemble learning family. RF is a simple, understandable, and highly efficient way of extracting features and has shown powerful performance in many real-world prediction problems. XGBoost is also a relatively new algorithm with high efficiency, flexibility and portability, and has achieved excellent results in many classification tasks [20, 21]. Subsequently, a set of train data was entered into the model to obtain the results. Next, the results of the prediction model were compared according to AUC criteria, and finally, the prediction model with the best performance was selected.

The Python package (version 3.7.2) was used to perform all the required analyses with the following libraries: pandas and NumPy for data preprocessing; scikit-learn and XGBoost for prediction model building and evaluation; and Matplotlib, seaborn, and SHAP for visualizing and plotting the results.

Data preprocessing

Criteria for data inclusion and exclusion

To determine patients with T2DM, the results of all four diagnostic tests available in the electronic database of the IKHC Diabetes Clinic, namely fasting blood sugar (FBS), glycohemoglobin (HbA1c), 2-hours Postprandial Glucose Test (2HPP), and glucose tolerance test (GTT) were obtained. The recommended criteria of the American Diabetes Association (ADA) [22] were used to diagnose T2DM patients. Individuals with HbA1c that was higher than or equal to 6.5% or FBS higher than or equal to 126 mg/dl or 2HPP higher than or equal to 200 mg/dl were considered T2DM patients. The presence or absence of nephropathy was defined as a binary variable in the prediction model. According to ADA diagnostic criteria for the definition of DN, T2DM patients with urinary microalbumin higher than or equal to 30 mg or estimated glomerular filtration rate (eGFR) less than or equal to 60 mg/dl were considered as patients with DN and labeled 1. For regular patients of the Diabetes Clinic, only information on when they were first identified as having DN was included.

The dataset of diabetes patients in this study included only non-pregnant adults aged at least 20. With this limitation, it was possible to focus on predicting DN in T2DM patients and exclude other types of diabetes from the analysis, including gestational diabetes, that only occurring in pregnant women, and type 1 diabetes that generally occurs among adolescents and children.

Processing missing values

Due to the characteristics of the T2DM patients' datasets, the features and samples with numerous missing values were excluded because they did not provide enough valuable information and knowledge. In this study, all samples with more than 50% missing values were considered noise data and removed from the dataset. In addition, features with more than 50% of null column values were also omitted.

To fill in the other missing values the features were divided into continuous and discrete parts before being processed in the prediction model. If the null value was of the continuous type, the lost value was filled based on the feature's average value in other samples. If the null value was discrete, the maximum frequency or mode of the feature in other samples was used to fill the missing value.

Based on previous research on DN risk factors [1, 4, 23–25] and available features in the electronic database of the Diabetes Clinic, the details of T2DM patient including demographic information, examinations, blood and urine laboratory results, and history were extracted. After removing some features as a result of the high number of missing values, the following risk factors remained for further analysis: gender, age, waist-to-hip ratio, body mass index (BMI), diastolic blood pressure (DBP), systolic blood pressure (SBP), HDL, LDL, total cholesterol, triglyceride, HbA1c, FBS, alkaline phosphatase (ALP), aspartate aminotransferase (AST), alanine amine (ALT), serum insulin, serum vitamin D, potassium, sodium, uric acid, history of hypertension, taking biguanides to control diabetes, taking sulfonylureas to control diabetes, insulin use, taking lipid-lowering drugs, history of retinopathy, history of diabetic foot ulcers, history of neuropathy, family history of diabetes, history of cardiovascular disease (CVD), family history of heart disease, family history of hypertension, fatty liver disease, smoking, and duration of diabetes.

Data normalization

Normalization was performed to equalize and eliminate different scales in the features. Different scales can ignore some features and affect the results of the prediction model. Once normalized, all the main properties of the features were scaled and lost their dimensions for comprehensive comparison and evaluation. In this study, data normalization was performed using the standardization method and z-score calculation.

Selecting important features

The RFECV technique was used in this study to select the most important features. This technique, together with a classification algorithm, can identify the most important features and improve prediction performance. The RFE algorithm begins the search with a complete set of features and an evaluation criterion like the AUC of the classifier. In the last iteration of this algorithm, the most irrelevant features are eventually removed and the most relevant features for sorting placed at the top. According to the sorting table of features generated by the evaluation criterion, RFE produces different subsets of features. RFE can be combined with other classifiers [26]. In this study, the RFECV method was used with three classification algorithms i.e., RF, SVM, and XGBoost, and 10-fold cross-validation.

Evaluation and validation

In this study, the ML-based classification algorithms used were the decision tree (DT), logistic regression (LR), support vector machine (SVM), extreme gradient boosting (XGBoost), and random forest (RF). The receiver operating characteristic curve (ROC) was used to evaluate the generalizability of the classification algorithm. The area under ROC (AUC) is the area under the receiver operating characteristic curve. AUC is a comprehensive index that shows the continuous variables of the sensitivity and accuracy of the classification algorithm. An AUC value above 70% indicates an appropriate performance of the classification algorithm while a value below 50% shows that it is unable to differentiate the true outcomes from the false ones.

Findings

Demographics, clinical, and laboratory information

The demographic characteristics and the laboratory and clinical results of 6235 patients with T2DM were obtained from the electronic database of the Diabetes Clinic of IKHC (Table 1). Of the diabetic patients referred to the clinic, 2210 (35%) were diagnosed as DN. These patients were mainly middle-aged or elderly patients (mean age 58 years) with an approximate diabetes history of 10 years. More than half (56%) were women, and most (68%) had a family background of diabetes. Also, the mean HbA1c, FBS, and 2HPP of the patients was 7.73%, 160 mg/dl, and 220.85 mg/dl, respectively.

Sorting important features

As mentioned above, the RFECV method was used to determine the importance of each feature, rank it, and select a subset of the best-performing ones. This study employed the three classification algorithms of RF, SVM, and XGBoost with 10-fold cross-validation in the RFECV method. In addition, a gradual method was used to select the features in deciding the quantity of characteristics with the optimum performance. Figure 2 shows the change in AUC based on the number of characteristics with the optimum performance in each of the three classification algorithms. As can be seen, the RF algorithm performs better in predicting DN than the other two algorithms. The 18 features selected produced

Table 1 Demographics, clinical, and laboratory characteristics of patients

Feature	Mean (SD) or frequency (%)
Age, year	57.60 (11.29)
Gender, Female	3481 (56%)
BMI, kg/m^2	30.43 (6.12)
Waist to Hip ratio	0.94 (0.15)
Diabetes duration, year	10.36 (7.93)
Hypertension	2680 (43%)
SBP, mmHg	130.14 (27.55)
DBP, mmHg	78.87 (9.04)
Biguanides drug intake	2656 (43%)
Sulfonylureas drug intake	4943 (79%)
Insulin use	394 (6%)
Lipid-lowering drug intake	3811 (61%)
Retinopathy	643 (12%)
Neuropathy	803 (22%)
Diabetic foot ulcer	82 (2%)
Fatty liver	1689 (47%)
History of CVD	1634 (26%)
Family background of Diabetes	4181 (68%)
Family background of Hypertension	2070 (40%)
Family background of CVD	1511 (29%)
Smoking	206 (5%)
Cholesterol	
HDL, mg/dL	45.18 (11.72)
LDL, mg/dL	101.64 (38.97)
Triglyceride, mg/dL	175.62 (109.29)
Total, mg/dL	179.90 (45.26)
HbA1c, %	7.73 (1.64)
FBS, mg/dL	160.35 (57.49)
2HPP, mg/dL	220.85 (88.42)
ALP, IU/L	158.29 (77.21)
AST, IU/L	22.65 (13.06)
ALT, IU/L	27.72 (18.04)
Potassium, mmol/L	4.32 (1.83)
Sodium, mmol/L	140.56 (4.38)
Uric acid, mg/dL	5.17 (2.23)
Vitamin D, ng/mL	24.25 (15.20)
Serum insulin, mcU/mL	10.49 (6.53)

an AUC value of 83%, and adding additional features did not result in any significant increase in this criterion. Then, the SHapley Additive exPlanations (SHAP) method was employed to show the importance of the key DN predictive factors [27], as illustrated in Fig. 3.

SHAP is an interpretability method for analyzing the importance of features on the basis of their impacts on the model output. The significance of adding each characteristic to the model was calculated in the entire potential feature sequences. For SHAP, positive and negative values reflect positive and negative effects on the model output, respectively. The 18 most significant characteristics selected were age, duration of diabetes, BMI, SBP, history of hypertension, history of retinopathy, ALT, CVD, 2HPP, uric acid, HbA1c, waist-to-hip ratio, total cholesterol, LDL, HDL, FBS, triglyceride, and serum insulin. Figure 3 shows that older patients, and those with longer duration of diabetes, higher BMI, higher SBP, history of hypertension, and history of retinopathy are more likely to develop DN.

The SHAP force plot in Fig. 4 illustrates the individual predictions for the RF classification algorithm. Features that result in lower and higher SHAP values are displayed in blue and red, respectively. The size of each feature is associated with its impact on the classification model's output. This figure shows an individual sample correctly classified as a DN patient. In this sample, age, duration of diabetes, and HbA1c have a greater effect on the model's output in reducing the risk of DN. In contrast, LDL, BMI, and the waist-to-hip ratio of the sample have a greater effect on the model's output in increasing the risk of DN.

Evaluation of classification algorithms

In this study, 18 features that had the best performance rating in DN prediction were obtained. Subsequent analyses were based on the performance evaluations of the prediction model with the same optimal number of features. After several experiments using the GridSearchCV method and changing the parameters, the best combination of parameters was determined for each classification algorithm. Figure 5 shows the performance of each classification algorithm on the basis of the ROC curve for the training dataset. As shown, the best results, with a prediction performance of 84.6% in AUC, were obtained for RF while the worst performing classification algorithm was DT.

Discussion

Assessing the effects of risk factors in predicting the likelihood of DN using the ML approach can be highly beneficial for early prevention and detection among T2DM patients. According to the findings of the proposed model in this study, age is the most important feature in DN, followed by the duration of diabetes, BMI, SBP, hypertension, history of retinopathy, ALT, CVD, 2HPP, uric acid, HbA1c, waist-to-hip ratios, cholesterol, LDL, HDL, FBS, triglycerides, and serum insulin.

As in previous research, the findings of this study affirm the role of older age as a DN risk factor. The association between a higher risk for DN and older age in T2DM patients has been reported in various investigations [28–30]. These studies mainly examined the changes in

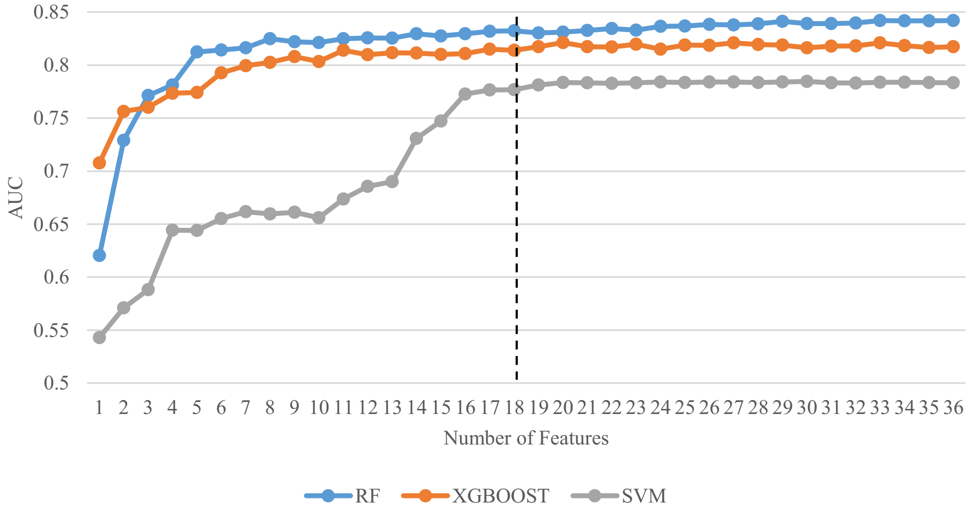

Fig. 2 Change of AUC based on the number of features

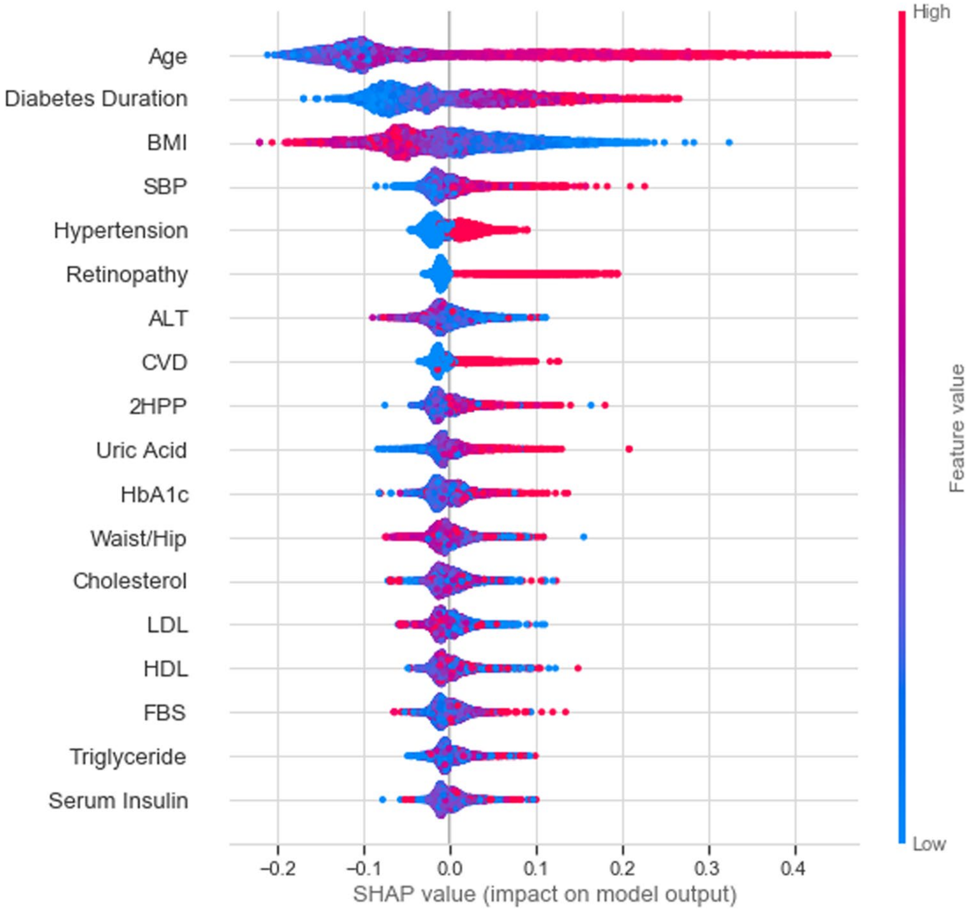

Fig. 3 SHAP summary plots for important DN risk factors

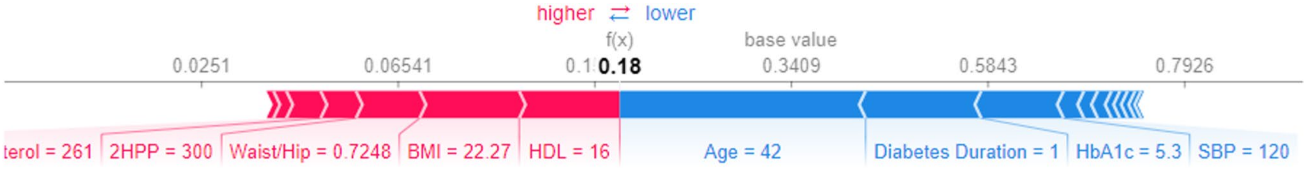

Fig. 4 SHAP force plot for a sample DN patient

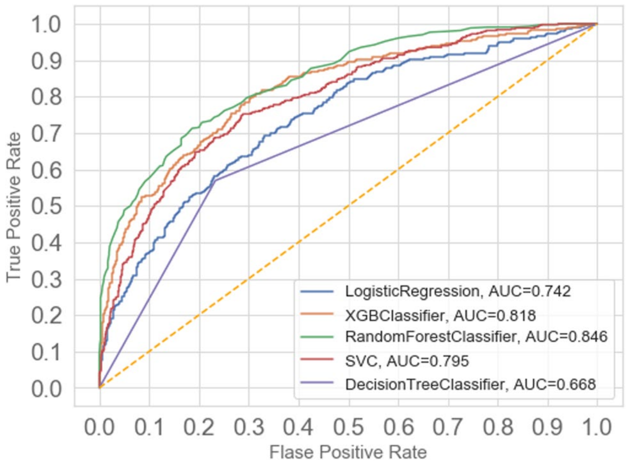

Fig. 5 ROC curve to evaluate the prediction models

eGFR. This result is consistent with the gradual decrease in eGFR that occurs in the normal population after about age 40 [31].

Hyperglycemia may result in microvascular complications, such as nephropathy, neuropathy, and retinopathy, and macrovascular complications, such as peripheral, cerebrovascular, and cardiovascular disorders by causing multiple cellular damages in the blood vessels of a person with T2DM. The longer the hyperglycemia duration, the greater the damage to arteries. Thus, this study, along with many other investigations [32, 33], confirms the impact of the duration of diabetes in DN development and in other important complications of the disease.

According to the results of this study, people with retinopathy or CVD are more likely the first to have DN. Many previous works [34, 35] have also highlighted that most patients with DN initially had retinopathy.

The findings of the predictive model in this study demonstrate that a higher BMI is a significant risk factor for DN. This association between high BMI and the likelihood of developing DN has been also reported in previous research. This problem was associated with obesity-related focal glomerular sclerosis regardless of the complication [36–38]. For example, an association between high BMI and decreased eGFR was reported in a survey of 105 patients with DN and type 2 diabetes [39].

This result, along with some other studies, shows that patients with T2DM and high blood pressure have a higher risk of developing DN. Hypertension is considered another independent risk factor for nephropathy [40, 41]. The relationship between low blood pressure and a lower risk of progression from moderate albuminuria to its severe form or end-stage renal disease has been previously established [42]. In addition, low blood pressure in T2DM patients has been linked to moderate albuminuria to normoalbuminuria [43].

HbA1c indicates moderate levels of blood sugar within the past two to three months, and long-term hyperglycemia leads to non-enzymatic glycosylation of proteins, which in turn leads to systemic vascular damage and increased development of microangiopathy [44]. Based on previous research [25] on T2DM, the findings of this study confirm that careful control of blood sugar can lead to a significant decrease in the risk of diabetic complications, including DN.

The importance of cholesterol control in diabetic patients with DN was verified in this study. Although numerous investigations have not declared HDL as a risk factor for DN, these results appear to be consistent with previous reports that categorize dyslipidemia among the potential risk factors for DN [45, 46]. Our model reports LDL and HDL as important factors for predicting DN.

The findings of this study confirm the known risk factors for DN and emphasize the importance of controlling blood pressure, hyperlipidemia, weight, postprandial blood sugar, and fasting blood sugar in patients with T2DM. In addition, our study findings, as an example of the application of ML approaches in medical predictions, indicate the advantages of using these techniques.

However, the classification and analysis of the models presented in this study have some limitations. The Diabetes Clinic of the IKHC lacks new DN biomarkers. The dataset only includes commonly evaluated clinical factors, including HbA1c, SCr, etc. Testing for alternative predictive biomarkers, including vascular endothelial growth factors, tumor necrosis factor-α, or transforming growth factor-β, which may be more closely linked to the pathogenesis of DN mechanisms, was not possible [47]. In addition, some known risk factors in predicting DN, such as physical activity [4], diet [24], etc., were not included in our dataset. As a result, some important factors were not considered in the proposed prediction model. In this analysis, only the internal validation of the model was examined. Hence, these results apply to populations that meet the inclusion criteria of the present study. In order to generalize the results, more research is needed on medical datasets in other communities.

Another limitation of this study is the large amount of missing data in some features. In order to ensure that accurate paraclinical and clinical records are kept in the electronic database, common guideline can be developed to establish the same treatment procedures and follow-up courses between different physicians. Also, electronic profiles based on information in the national ID card should be developed for each patient to facilitate the accurate storage and retrieval of health records. Finally, it is recommended that patients be provided with necessary training to keep their medical records with them during their appointments at the Diabetes Clinic.

Conclusion

This study underlined the application and advantages of ML techniques in identifying the risk factors for DN in T2DM patients referred to the Diabetes Clinic of the IKHC. Of the prediction models tested the RF classification algorithm showed the best performance. This study indicated 18 important predictive factors for DN. This suggests that governments should mandate broader DN screening and increase health care for patients with T2DM, especially through interventions and monitoring of important risk factors. At the same time, diabetics should also take their own precautions, especially those with a long history of T2DM or high blood pressure in regard to the occurrence of DN.

Using predictive models to review and explore the knowledge from electronic medical records can provide useful clinical tools for identifying important risk factors in various diseases. Future studies could focus on optimizing response variables and determining different levels of DN. In addition, recognizing the importance of risk factors using ML-based methods will facilitate the development of applications to evaluate the risk of DN occurrence to screen T2DM patients.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Rodriguez-Romero V, Bergstrom RF, Decker BS, Lahu G, Vakilynejad M, Bies RR. Prediction of nephropathy in type 2 diabetes: an analysis of the ACCORD trial applying machine learning techniques. Clin Transl Sci. Wiley Online Library; 2019;12:519–528.
2. Zhao N, Gao JD, Zhang HL. Research progress of glycosylated haemoglobin as a diagnostic criterion for diabetes. Chin Clin Res. 2015;1:127–8.
3. Ahmad J. Management of diabetic nephropathy: recent progress and future perspective. Diabetes Metab Syndr Clin Res Rev. Elsevier; 2015;9:343–358.
4. Radcliffe NJ, Seah J, Clarke M, MacIsaac RJ, Jerums G, Ekinci EI. Clinical predictive factors in diabetic kidney disease progression. J Diabetes Investig. Wiley Online Library; 2017;8:6–18.
5. Narres M, Claessen H, Droste S, Kvitkina T, Koch M, Kuss O, et al. The incidence of end-stage renal disease in the diabetic (compared to the non-diabetic) population: a systematic review. PLoS One. Public Library of Science San Francisco, CA USA; 2016;11:e0147329.
6. Rao V, Rao LBV, Tan SH, Candasamy M, Bhattamisra SK. Diabetic nephropathy: an update on pathogenesis and drug development. Diabetes Metab Syndr Clin Res Rev. Elsevier; 2019;13:754–762.
7. Gans D, Kralewski J, Hammons T, Dowd B. Medical groups' adoption of electronic health records and information systems. Health Aff. Project HOPE-The People-to-People Health Foundation, Inc.; 2005;24:1323–33.
8. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. Heal Inf Sci Syst. Springer; 2014;2:3.
9. Magoulas GD, Prentza A. Machine learning in medical applications. Adv course Artif Intell. Springer; 1999. p. 300–307.
10. Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine learning and data mining methods in diabetes research. Comput Struct Biotechnol J. Elsevier; 2017;15:104–116.
11. Dinh A, Miertschin S, Young A, Mohanty SD. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. BMC Med Inform Decis Mak. Springer; 2019;19:211.
12. Alexopoulos E, Dounias GD, Vemmos K. Medical diagnosis of stroke using inductive machine learning. Mach Learn Appl Mach Learn Med Appl. 1999:20–3.
13. Kourou K, Exarchos TP, Exarchos KP, Karamouzis M V, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J. Elsevier; 2015;13:8–17.
14. Yu W, Liu T, Valdez R, Gwinn M, Khoury MJ. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. BMC Med Inform Decis Mak. Springer; 2010;10:16.
15. Semerdjian J, Frank S. An ensemble classifier for predicting the onset of type II diabetes. arXiv Prepr arXiv170807480. 2017;
16. Teimouri M, Ebrahimi E, Alavinia SM. Comparison of various machine learning methods in diagnosis of hypertension in diabetics with/without consideration of costs. Iran J Epidemiol. 2016;11:46–54.
17. Parthiban G, Srivatsa SK. Applying machine learning methods in diagnosing heart disease for diabetic patients. Int J Appl Inf Syst. Citeseer; 2012;3:25–30.
18. Cho BH, Yu H, Kim K-W, Kim TH, Kim IY, Kim SI. Application of irregular and unbalanced data to predict diabetic nephropathy using visualization and feature selection methods. Artif Intell Med. Elsevier; 2008;42:37–53.
19. Ravizza S, Huschto T, Adamov A, Böhm L, Büsser A, Flöther FF, et al. Predicting the early risk of chronic kidney disease in patients with diabetes using real-world data. Nat Med. Nature Publishing Group; 2019;25:57–59.
20. Chang W, Liu Y, Xiao Y, Yuan X, Xu X, Zhang S, et al. A machine-learning-based prediction method for hypertension outcomes based on medical data. Diagnostics. Multidisciplinary Digital Publishing Institute; 2019;9:178.
21. Choudhury A, Gupta D. A survey on medical diagnosis of diabetes using machine learning techniques. Recent Dev Mach Learn data Anal. Springer; 2019. p. 67–78.
22. American Diabetes Association. Diagnosis and classification of diabetes mellitus. Diabetes Care. Am Diabetes Assoc; 2014;37:S81–S90.
23. Tziomalos K, Athyros VG. Diabetic nephropathy: new risk factors and improvements in diagnosis. Rev Diabet Stud RDS. Society for Biomedical Diabetes Research; 2015;12:110.
24. Alicic RZ, Rooney MT, Tuttle KR. Diabetic kidney disease: challenges, progress, and possibilities. Clin J Am Soc Nephrol. Am Soc Nephrol; 2017;12:2032–2045.
25. Lou J, Jing L, Yang H, Qin F, Long W, Shi R. Risk factors for diabetic nephropathy complications in community patients with type 2 diabetes mellitus in Shanghai: logistic regression and classification tree model analysis. Int J Health Plann Manage. Wiley Online Library; 2019;34:1013–1024.
26. Granitto PM, Furlanello C, Biasioli F, Gasperi F. Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. Chemom Intell Lab Syst. Elsevier; 2006;83:83–90.

27. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. Nature Publishing Group; 2020;2:56–67.
28. Rossing K, Christensen PK, Hovind P, Tarnow L, Rossing P, Parving H-H. Progression of nephropathy in type 2 diabetic patients. Kidney Int. Elsevier; 2004;66:1596–1605.
29. Zoppini G, Targher G, Chonchol M, Ortalda V, Negri C, Stoico V, et al. Predictors of estimated GFR decline in patients with type 2 diabetes and preserved kidney function. Clin J Am Soc Nephrol Am Soc Nephrol. 2012;7:401–8.
30. Elley CR, Robinson T, Moyes SA, Kenealy T, Collins J, Robinson E, et al. Derivation and validation of a renal risk score for people with type 2 diabetes. Diabetes Care. Am Diabetes Assoc; 2013;36:3113–3120.
31. Coresh J, Astor BC, Greene T, Eknoyan G, Levey AS. Prevalence of chronic kidney disease and decreased kidney function in the adult US population: third National Health and nutrition examination survey. Am J kidney Dis. Elsevier; 2003;41:1–12.
32. Zhou JJ, Coleman R, Holman RR, Reaven P. Long-term glucose variability and risk of nephropathy complication in UKPDS, ACCORD and VADT trials. Diabetologia. Springer; 2020;63:2482–2485.
33. Larroumet A, Molina O, Foussard N, Monlun M, Blanco L, Mohammedi K, et al. Early worsening of diabetic nephropathy in type 2 diabetes after rapid improvement in chronic severe hyperglycemia. Diabetes Care 2021; 44: e55–e56. Diabetes Care. Am Diabetes Assoc; 2021;44:e110–e111.
34. Butt A, Mustafa N, Fawwad A, Askari S, Haque MS, Tahir B, et al. Relationship between diabetic retinopathy and diabetic nephropathy; a longitudinal follow-up study from a tertiary care unit of Karachi, Pakistan. Diabetes Metab Syndr Clin Res Rev. Elsevier; 2020;14:1659–1663.
35. Wu H-Q, Wei X, Yao J-Y, Qi J-Y, Xie H-M, Sang A-M, et al. Association between retinopathy, nephropathy, and periodontitis in type 2 diabetic patients: a Meta-analysis. Int J Ophthalmol. Press of International Journal of Ophthalmology; 2021;14:141.
36. Verani RR. Obesity-associated focal segmental glomerulosclerosis: pathological features of the lesion and relationship with cardiomegaly and hyperlipidemia. Am J kidney Dis. Elsevier; 1992;20:629–634.
37. Kambham N, Markowitz GS, Valeri AM, Lin J, D'Agati VD. Obesity-related glomerulopathy: an emerging epidemic. Kidney Int. Elsevier; 2001;59:1498–1509.
38. Darouich S, Goucha R, Jaafoura MH, Zekri S, Maiz H Ben, Kheder A. Clinicopathological characteristics of obesity-associated focal segmental glomerulosclerosis. Ultrastruct Pathol. Taylor & Francis; 2011;35:176–182.
39. Huang W-H, Chen C-Y, Lin J-L, Lin-Tan D-T, Hsu C-W, Yen T-H. High body mass index reduces glomerular filtration rate decline in type II diabetes mellitus patients with stage 3 or 4 chronic kidney disease. Medicine (Baltimore). Wolters Kluwer Health; 2014;93.
40. Hovind P, Tarnow L, Rossing P, Graae M, Torp I, Binder C, et al. Predictors for the development of microalbuminuria and macroalbuminuria in patients with type 1 diabetes: inception cohort study. Bmj. British Medical Journal Publishing Group; 2004;328:1105.
41. Tapp RJ, Shaw JE, Zimmet PZ, Balkau B, Chadban SJ, Tonkin AM, et al. Albuminuria is evident in the early stages of diabetes onset: results from the Australian diabetes, obesity, and lifestyle study (AusDiab). Am J Kidney Dis. Elsevier; 2004;44:792–798.
42. De Boer IH, Rue TC, Cleary PA, Lachin JM, Molitch ME, Steffes MW, et al. Long-term renal outcomes of patients with type 1 diabetes mellitus and microalbuminuria: an analysis of the diabetes control and complications trial/epidemiology of diabetes interventions and complications cohort. Arch Intern Med. American Medical Association; 2011;171:412–420.
43. Araki S, Haneda M, Sugimoto T, Isono M, Isshiki K, Kashiwagi A, et al. Factors associated with frequent remission of microalbuminuria in patients with type 2 diabetes. Diabetes. Am Diabetes Assoc; 2005;54:2983–2987.
44. Hirsch IB, Brownlee M. Beyond hemoglobin A1c—need for additional markers of risk for diabetic microvascular complications. Jama. American Medical Association; 2010;303:2291–2292.
45. Pavkov ME, Bennett PH, Sievers ML, Krakoff J, Williams DE, Knowler WC, et al. Predominant effect of kidney disease on mortality in Pima Indians with or without type 2 diabetes. Kidney Int. Elsevier; 2005;68:1267–1274.
46. Skupien J, Warram JH, Smiles AM, Niewczas MA, Gohda T, Pezzolesi MG, et al. The early decline in renal function in patients with type 1 diabetes and proteinuria predicts the risk of end-stage renal disease. Kidney Int. Elsevier; 2012;82:589–597.
47. Coca SG, Nadkarni GN, Huang Y, Moledina DG, Rao V, Zhang J, et al. Plasma biomarkers and kidney function decline in early and established diabetic kidney disease. J Am Soc Nephrol. 2017;28:2786–93.