

## Development of an intelligent clinical decision support system for the early prediction of diabetic nephropathy

Mohammad Reza Afrash<sup>f</sup>, Fatemeh Rahimi<sup>a,\*\*</sup>, Hadi Kazemi-Arpanahi<sup>b,c</sup>,  
Mostafa Shanbezadeh<sup>d</sup>, Morteza Amraei<sup>e</sup>, Farkhondeh Asadi<sup>a,\*</sup>

<sup>a</sup> Department of Health Information Technology and Management, School of Allied Medical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran

<sup>b</sup> Department of Health Information Technology, Abadan University of Medical Sciences, Abadan, Iran

<sup>c</sup> Department of Student Research Committee, Abadan University of Medical Sciences, Abadan, Iran

<sup>d</sup> Department of Health Information Technology, Faculty of Paramedical, Ilam University of Medical Sciences, Iran

<sup>e</sup> Department of Health Information Technology, School of Allied Medical Sciences, Lorestan University of Medical Sciences, Khorramabad, Iran

<sup>f</sup> Department of Artificial Intelligence, Smart University of Medical Sciences, Tehran, Iran

### ARTICLE INFO

#### Keywords:

Diabetic nephropathies  
Data mining  
Risk factors  
Decision support system

### ABSTRACT

**Background:** Diabetic nephropathy (DN) is the most common microvascular complication of diabetes mellitus (DM) and is identified as a leading cause of the end-stage renal disease (ESRD). Considering the importance of early prediction of individuals at risk of this complication, the use of intelligent models through machine learning (ML) algorithms can be helpful. Therefore, this study aimed to identify the influential variables in predicting DN and fed them as inputs to develop an ML-based decision support system (DSS) for DN diagnosis.

**Methods:** The data of 327 patients with diabetes (types 1 and 2) were retrospectively analyzed. After data preparation, the genetic algorithm (GA) feature selection method was used to identify the predictor variables affecting DN. Then, several ML algorithms, including the support vector machine (SVM), decision tree (DT), K-nearest neighbors (KNN), and artificial neural networks (ANN) were used to train predictive models based on the selected features. Afterward, the performance of the developed models was evaluated using sensitivity, specificity, and accuracy criteria in 10 independent runs. Finally, the DSS was developed based on the best fit model in the C# programming language.

**Results:** Our findings illustrated that age, hemoglobin A1c (HbA1c) test, diastolic arterial pressure (DAP), systolic arterial pressure (SAP), fasting glycemia rate (FGR), and DM involvement time were the most important factors in predicting DN. Moreover, to predict the DN, GA combined with the DT algorithm obtained the highest performance in terms of accuracy, sensitivity, specificity, and area under the curve (AUC), equal to 98.9, 98.6, 99.2, and 98.9%, respectively.

**Conclusions:** The results revealed that GA combined with the DT classifier predicted DN with significant accuracy. Thus, the DSS developed based on DT can be considered a reliable tool to help physicians make decisions. Future studies are warranted to further validate the applicability of our model in clinical settings.

### 1. Introduction

According to the World Health Organization (WHO), the number of non-communicable diseases among the leading causes of global mortality has increased from four in 2000 to seven (74% of all deaths) in 2019. Among them, diabetes mellitus (DM) and kidney diseases are

ranked ninth and tenth, respectively. The mortality rate from diabetes has increased by 70% during this period, and kidney disease has changed from the 13th to the 10th cause of death. These fluctuations are especially conspicuous in low- and middle-income countries (LMICs) [1]. DM has several microvascular complications for various organs such as the kidneys, nerves, eyes, heart, skin, and brain [2,3]. These

\* Corresponding author. Department of Health Information Technology and Management, School of Allied Medical Sciences, Shahid Beheshti University of Medical Sciences, Darband St., Tehran, Iran.

\*\* Corresponding author.

E-mail addresses: [m.afrash@sbmu.ac.ir](mailto:m.afrash@sbmu.ac.ir) (M.R. Afrash), [Fatemeh\\_rahimi@sbmu.ac.ir](mailto:Fatemeh_rahimi@sbmu.ac.ir) (F. Rahimi), [Hadi.kazemi67@yahoo.com](mailto:Hadi.kazemi67@yahoo.com) (H. Kazemi-Arpanahi), [Mostafa.shanbezadeh@gmail.com](mailto:Mostafa.shanbezadeh@gmail.com) (M. Shanbezadeh), [asadifar@sbmu.ac.ir](mailto:asadifar@sbmu.ac.ir) (F. Asadi).

<https://doi.org/10.1016/j.imu.2022.101135>

Received 4 May 2022; Received in revised form 19 November 2022; Accepted 19 November 2022

Available online 24 November 2022

2352-9148/© 2022 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

complications are responsible for an annual 4 million deaths globally (9% of total deaths), and 10–20% of diabetic patients die due to renal complications [4]. Diabetic nephropathy (DN) takes place in about 25–40% of patients with type 2 diabetes and 30% of those with type 1 diabetes [5–7]. DN is characterized by albuminuria and decreasing glomerular filtration rate (GFR) [8]. It is one of the leading causes of end-stage renal disease (ESRD) and imposes heavy costs on the health-care system [9]. Therefore, it is essential to identify patients at risk of DN early to prevent related diseases.

Early detection of diabetic patients prone to DN and ESRD can help reduce the burden of diabetic kidney disease (DKD). Although a decrease in GFR is a good indicator of the progression of kidney disease, this change begins when kidney damage occurs [10]; therefore, it is not an early predictor. Some researchers have noted the importance of microalbuminuria in predicting DN [11–13], while others have considered microalbuminuria's sensitivity and specificity to be insufficient. In addition, the mere use of this factor in predicting DN has been questioned [14,15]. Note that DKD is observed in some patients without microalbuminuria [16]. Thus, it is interpreted that the lack of microalbuminuria and its progression to proteinuria does not always mean survival from kidney disease due to DKD [15].

Given this gap in the clinical area, many researchers in recent years have continued their efforts to identify predictive variables of DN and have tried to estimate the likelihood of people with diabetes developing DN. For example, we can refer to the study conducted by Yu et al. which identified and introduced 15 remarkable laboratory predictors [17]. Furthermore, Freedman et al. emphasized the influence of genetic factors on the incidence of DN [18]. However, despite advances in the clinical and genetics fields, there is a need to further investigate the proposed predictors [16]. Accordingly, it is suggested that new variables and models be considered in this regard [15]. Therefore, in recent years, the science of machine learning (ML) has been considered by various researchers to predict DN [19–24]. ML is a process in which valuable and interesting patterns are discovered from huge raw datasets and can support physician decision-making [25,26]. ML helps healthcare organizations to obtain patterns and trends from the mass of seemingly unrelated patient data. Based on ML, it is possible to predict the behaviors

of new patients in terms of diagnosis, prognosis, and treatment [27]. The present study first identified the most predictive risk factors affecting DN, and then fed them as inputs into ML algorithms; afterward, the performance of ML algorithms were compared, and finally, a DSS was developed to estimate the likelihood of developing the disease based on the best performing algorithm.

## 2. Methods

### 2.1. Study design and setting

In this retrospective study, data of 327 patients with DM (types 1 and 2) were collected through a checklist from an electronic medical record (EMR) database at Imam Khomeini Hospital, Tehran, Iran. Of these, 150 patients had DN. The process of data preparation and modeling was performed in MATLAB 2018 b software. In addition, the 10-fold cross-validation technique was employed to classify the data into training and test sets to achieve the best performance of the algorithms. The procedure of the study is displayed in Fig. 1.

### 2.2. Study variables

Several variables have been collected for DN patients in the EMR database. The criteria for selecting the candidate variables related to DN were based on consulting with expert nephrologists and reviewing the literature. DN was selected as the outcome variable, and a binary value for the prediction model was defined as either the development or absence of nephropathy. The following covariates were extracted based on the literature review coupled with experts' opinions from the EMR database (see Table 1).

### 2.3. Data preprocessing

Data preprocessing plays a key role in optimal datasets before training ML algorithms. Some preprocessing techniques were applied after data collection. In this step, the rows of the dataset with missing values of greater than 70%, noisy data, and outliers or inconsistent data

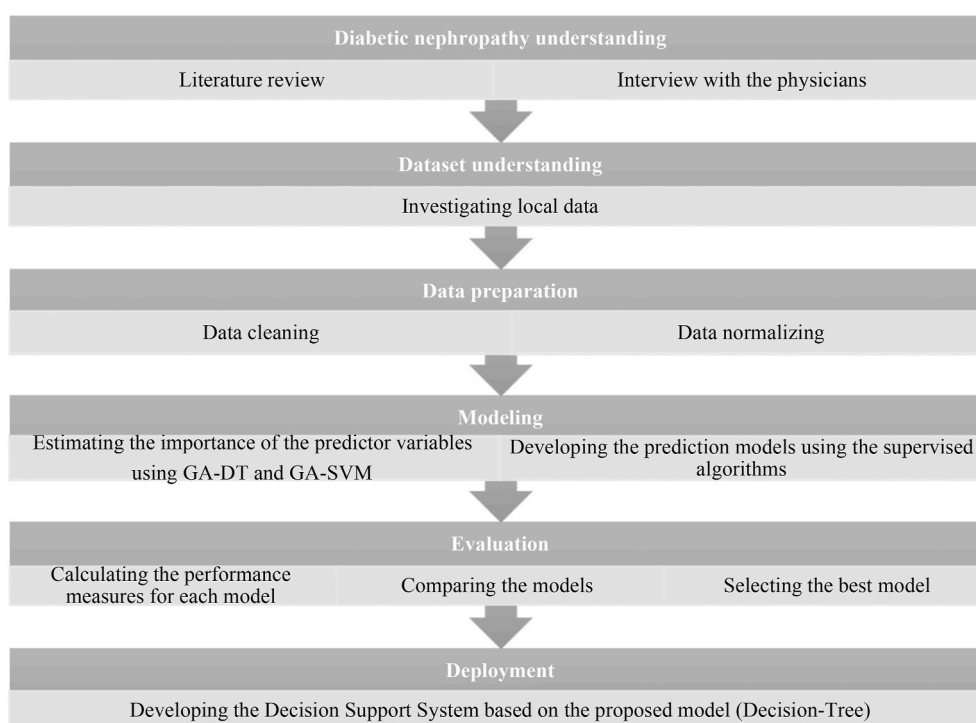


Fig. 1. Study roadmap.

**Table 1**  
Statistical characteristics of the variables.

ID	Risk factor	Measuring Unit	Range	DN	Diabetic patient	Average	STD. deviation
1	Age	Year	30–59	53	110	54.7	3.8
			60–83	97	67	68.23	5.6
2	Fasting glycemia rate	mg/dl	<125 mg/Dl	7	22	120.9	3.4
			>125 mg/Dl	143	155	141.2	8.15
3	Body mass index (BMI)	kg/m2	16–33	–	–	24.90	5.32
4	Hemoglobin A1c (HbA1c)	%	6.2–7.9	–	–	7.3	1.89
5	Diastolic arterial pressure (DAP)	mmHg	73–95	–	–	85.29	13.19
6	Systolic arterial pressure (SAP)	mmHg	114–148	–	–	132.65	16.57
7	Diabetes mellitus time involvement	Year	7–27	18.37	13.21	15.21	5.10
8	Smoking		Yes	16	29	–	–
			No	134	148	–	–
9	Gender		Female	69	82	–	–
			Male-	81	95	–	–
10	Alcoholism		Yes	11	14	–	–
			No	139	163	–	–
11	Diet		Yes	26	31	–	–
			No	124	146	–	–
12	Regular physical		Yes	12	17	–	–
			No	138	160	–	–
13	Outcome variable	Patient	–	150	177	–	–

were selected by two health-information management specialists and two nephrologists to enhance the classification algorithms.

#### 2.4. Feature selection

Feature selection is the processing of selection features that are more effective and relevant in the prediction tasks, and excludes other unrelated or duplicate features from the analysis [28]. Various techniques can be used for this purpose, including the genetic algorithm (GA) [29, 30]. The GA is an evolutionary algorithm to solve optimization problems [31]. It is a feature selection algorithm based on the mechanics of natural selection and natural genetics that can fully consider the relationships between variables and identify the most suitable ones [32–36]. GA is a random optimization technique inspired by the Darwinian theory of natural evolution and survival. Evolution usually starts from a population of randomly generated individuals, and is an iterative process, with the population in each iteration called a generation. In each generation, the fitness of every individual in the population is evaluated; fitness is usually the value of the objective function in the optimization problem being solved. The new generation of candidate solutions is then used in the next iteration of the algorithm. The GA operates on many solutions at various sequential generations to select superior offspring based on the survival of the fittest principles [37,38]. This algorithm was used in this study to identify the predictors of DN.

#### 2.5. ML algorithms

Based on the selected features, six supervised algorithms were used to train the predictive models of DN, including radial basis function network (RBF), fuzzy neural network, pattern recognition network (PatternNET), support vector device (SVM), decision tree (DT), and K-nearest neighbors (KNN). The first three algorithms are artificial neural network (ANN)-based algorithms.

**ANN:** It is an example of computational algorithms modeled on brain neurons' behavior in the learning process to solve problems. It has at least three layers, including the input layer, the hidden layer(s), and the output layer. In neural networks, each neuron communicates with the neurons of its subsequent layers using a weighted connection. Each time a new piece of data is entered, the algorithm learns how to adjust its parameters and estimates the output variable(s)' value from the input variables to solve nonlinear problems such as prediction problems [39–41]. The ANNs used in the modeling of the present study included the RBF network, the fuzzy neural network, and the PatternNET).

**SVM:** It is a powerful approach to binary classification problems.

Depending on the number of features (dimensions of data space), a line, plane, or hyperplane is used to separate data of different classes. This algorithm solves linear and nonlinear problems (especially when the data are highly intertwined) [42,43].

**DT:** This algorithm, used to solve classification and regression problems, creates a tree-like structure. In this structure, each path from the root to one of the leaves of a tree expresses a rule. Rules are used to classify any new piece of data [44].

**KNN:** This algorithm is straightforward but practical in classification. It detects the class of each piece of data according to the class of the K data that are in the feature space in the nearest neighborhood of that piece of data [45].

#### 2.6. Performance evaluation of the developed ML models

The sensitivity, specificity, and accuracy of each predictive model were calculated in 10 independent runs, and the averages of the evaluation criteria were compared in different models.

- (1) Specificity =  $TN / (TN + FP)$
- (2) Sensitivity (Recall) =  $TP / (TP + FN)$
- (3) Accuracy =  $(TP + TN) / (TP + TN + FP + FN)$

**TP (true positive):** People with DN who were correctly placed in the class of DN by the algorithm.

**TN (true negative):** People without DN who were correctly placed in the class of people without DN by the algorithm.

**FP (false positive):** People without DN who were incorrectly placed in the class of people with DN by the algorithm.

**FN (false negative):** People with DN who were incorrectly placed in the class of people without DN by the algorithm.

For the training phase, the performance of prediction algorithms was assessed according to a 10-time-iteration of the 10-fold cross-validation method. In the testing phase, the performance of the prediction algorithms was independently assessed. To use the ML classifiers more accurately and sensitively, we searched the optimal hyper-parameters providing the highest validation accuracy, sensitivity, and specificity (Table 2).

To implement the DSS under Windows, the C # programming language was used as the front end, and the MySQL software was used to implement and develop the DSS database.

**Table 2**  
Best hyper-parameters of all the trained algorithmsNum

	Data mining Models	Hyper-parameters
1	Decision tree	–
2	RBF network	12-3142-2, Spread = 120
3	SVM	Kernel function = linear kernel
4	KNN	K = 3
5	Pattern recognition network	12-8-5-2
6	FNN	Using Matlab ToolBox

### 3. Results

#### 3.1. Patient characteristics

Of the 327 patients with diabetes, 150 patients developed critical DN. The median age of the patients who progressed to DN was higher than those who did not (66 vs. 53 years). The median duration of DM involvement time for DN is shown in Table 1.

#### 3.2. Results of feature selection

The feature set of this paper consisted of 12 variables. After combining the GA with the two DT and SVM algorithms, six eligible variables were used for modeling: age, HbA1C, DAP, SAP, FGR, and DM involvement time (Table 2). The results of the feature selection algorithm revealed that these six variables had played a crucial role in predicting DN (Table 3).

#### 3.3. Results of k-fold cross-validation

Based on the results of the feature selection phase, we determined the six most important factors including age, DM involvement time, DAP, SAP, HbA1C, and FGR as inputs to the model to evaluate whether diabetic patients would develop NP or not. We tried six ML algorithms for modeling, including DT, Pattern Net, RBF network, SVM with linear kernel, KNN (K = 3), and FNN.

Table 4 shows the 10-fold cross-validation with the mean for sensitivity, specificity, and accuracy. The sensitivity and accuracy of DT were much higher than those of the other five ML models. The predictive algorithm established by KNN (k = 3) exhibited the worst performance among the six ML algorithms. Therefore, the algorithm constructed by the DT model was the optimum tool for DN prediction and was used for developing a Windows-based intelligence system.

Fig. 2 displays the ROC curve of the DT. The area under the ROC curve for the DT was 98.9, which indicates the remarkable performance of this model.

#### 3.4. Windows-based application of prediction models

Using the best-achieved ML algorithm, a Windows-based application was developed for easy accessibility and simple use of the ML models. The system calculates the probability of DN, which is computed according to the best model. A view of the user interface of the system designed based on this model can be observed below (Fig. 3).

### 4. Discussion

This study designed a DSS to predict DN using ML techniques. Initially, GA-based feature selection was performed to select important

**Table 3**  
Common predictors among the 10 runs.

ID	Algorithm	Predictive variables
1	GA-DT	Age, DM time involvement, DAP, SAP, HbA1C
2	GA-SVM	Age, DM time involvement, DAP, FGR

**Table 4**  
Comparing the mean performance of the used algorithms in 10 independent runs.

Algorithm	Sensitivity	Specificity	Accuracy
DT	98.6	99.2	98.9
Pattern Net	97.5	99.0	98.3
RBF network	97.5	99.8	97.8
SVM with linear kernel	96.7	98.4	97.6
KNN(K = 3)	88.9	96.3	92.6
FNN	93.4	100	92.6

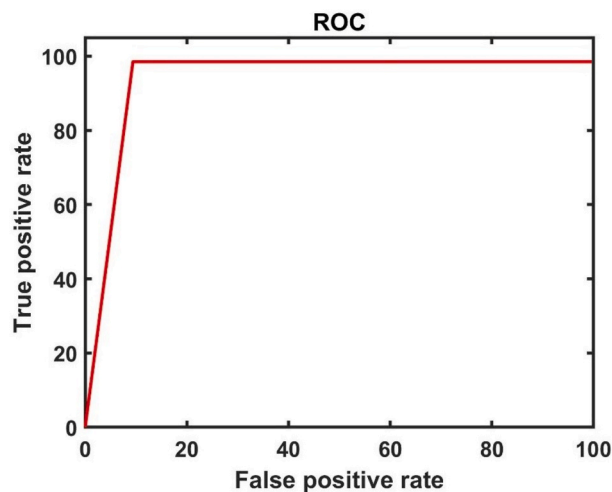


Fig. 2. Performance evaluation of the decision tree (ROC curve).

variables affecting DN. Among a total of 12 primary variables, a set of six variables, including age, HbA1C, DAP, SAP, FGR, and DM involvement time were identified as the most important predictors of DN. The combination of GA with the DT classifier achieved the highest ability in diagnosing DN with a sensitivity of 98.6%, specificity of 99.2%, and accuracy of 98.9%.

In recent years, various researchers have attempted to use ML techniques to predict the development of diabetes. However, given the different characteristics of individuals in different communities, the model proposed for one community may not be appropriate for another. Some studies have examined both genetic factors and laboratory routine factors [21,22], while others have focused solely on laboratory routine factors [19]. After training nine ML models to predict DKD on a dataset with 410 cases and 18 attributes, David et al. found that the DT classifier with an accuracy of 93.65 and a higher K value (0.8731) yielded the best performance [46]. Priyanka et al. evaluated the performance of four ML algorithms (KNN, SVM linear, SVM polynomial, and DT) to predict diabetic retinopathy. Finally, the DT model with 90% AUC showed the best performance [47]. Joshi et al. assessed the capability of the DT model for predicting type 2 diabetes. Their proposed model with an accuracy of 78.26% demonstrated reasonable performance [48]. Furthermore, the combination of logistic regression (LR)-based feature selection and DT-based classifier with 0.95% AUC attained the best performance for predicting diabetes in the study by Maniruzzaman et al. [49]. Pei et al. (2020) proposed a simple and reliable predictive model based on the 14 risk factors associated with 657 diabetic and 5904 non-diabetic data by using the J48 DT classifier. Ultimately, the developed model achieved an accuracy of 90.3%, a precision of 89.7%, and a recall of 90.3% [50]. Eyasu et al. conducted a retrospective study on a dataset with 2640 records and concluded that the J48 DT model was the best-performing model with an accuracy of 97.34% for diagnosis and classification of diabetes [51]. Pei et al. also analyzed the capability of five well-known ML classifiers to identify individuals with diabetes

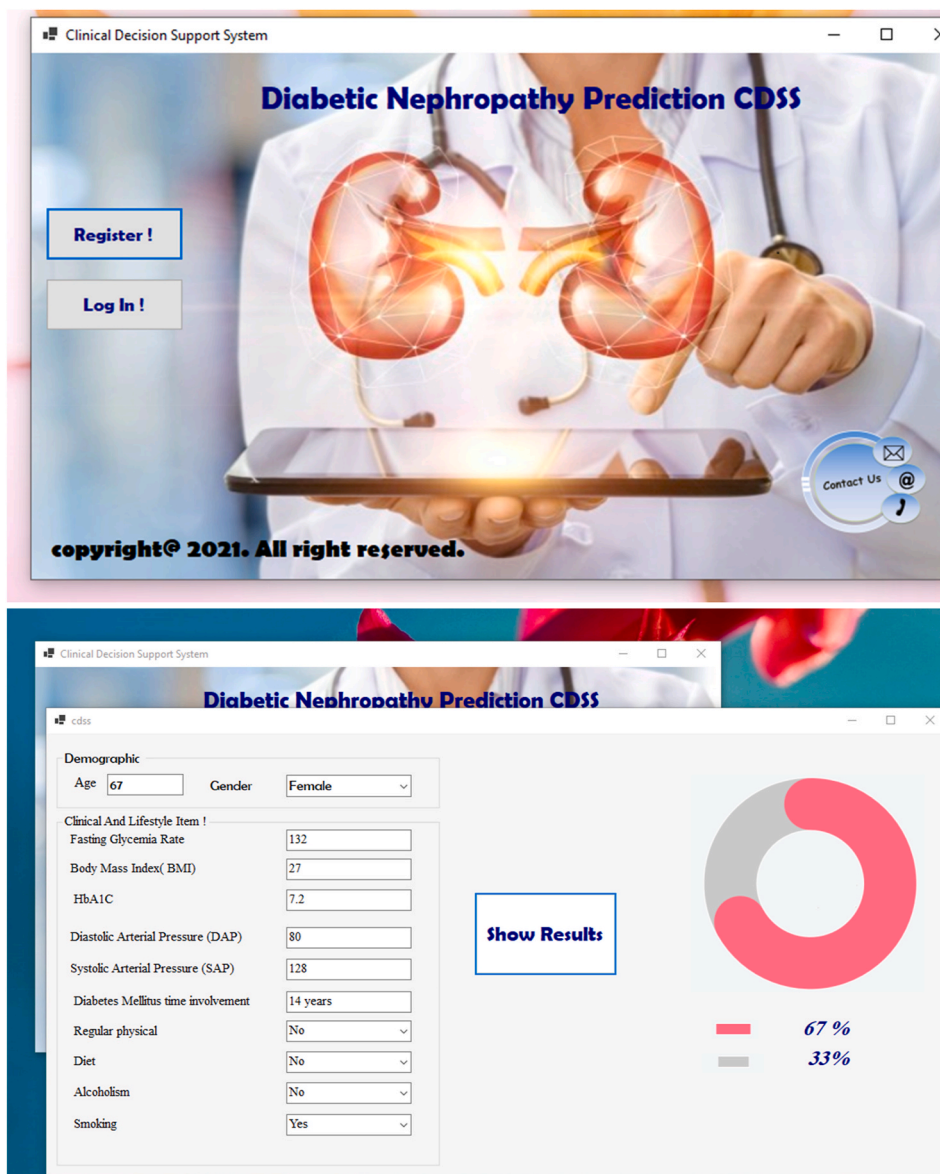


Fig. 3. User Interface of the software.

based on nine clinical variables of 4205 diabetic and non-diabetic cases. The experimental results showed that the J48 DT classifier has the best performance with an accuracy of 0.9503, precision of 0.950, recall of 0.950, an F-measure of 0.948, and an AUC of 0.964 [52]. Similarly, Ye et al. developed eight ML techniques on 10237 patients' data, and finally, the DT model performed best (AUC 0.74, 95% CI 0.71–0.76) [53]. Liu et al. analyzed the data of 3025 diabetic and non-diabetic individuals; the results showed that the DT achieved the highest accuracy (91.21%) in the prediction of diabetes [54]. Stefanopoulos et al. used the DT classifier and diagnosed the diabetic foot ulcer (DFU) with 80.6% sensitivity and 78.3% specificity [54].

As an evaluation criterion, sensitivity shows how well an algorithm can identify true positive cases. The importance of solid sensitivity relates to correctly identifying cases that are more likely to develop DN. As a result, a physician can constantly monitor these patients' conditions. The model proposed in this study correctly diagnosed 98.6% of cases with a high probability of developing the disease. In addition, Pattern Net, RBF, SVM, KNN, and KNN algorithms had the highest accuracy, respectively. In this study, age, HbA1C, DAP, SAP, FBS, and duration of involvement with diabetes were identified as important factors in

predicting DN. Among these, the patient's DAP, SAP, and FBS are factors that can be improved and controlled to some extent by lifestyle management and self-care. This, in turn, can reduce the risk of developing DN in the future. Some studies have shown that genetic factors significantly affect the prediction of DN [22], and the combination of genetic and non-genetic factors can significantly predict the DN risk [22].

Due to the limited access to genetic data in the present study, the issue has been examined only from the perspective of non-genetic data (laboratory and non-laboratory). In the present study, the predictive models were developed using easily available routine features. Therefore, screening DM patients to find those who are susceptible to developing DN is possible in primary clinics. Another strength of the present study is that among the ML techniques used, the DT has obtained the most robust response in terms of accuracy and sensitivity. Since the method of inference in DT algorithms is visually quite clear, the proposed model can be considered a reliable and user-friendly model for physicians; on the other hand, models such as the neural network act as a so-called "black box," and its interpretation is not clear to the physician.

This study also had some limitations that must be addressed. The lack of access to genetic data was a limitation that we could turn into a

strength. In developing the models, the focus was only on routine and easily available factors collected for diabetic patients in healthcare centers. Another limitation of the present study was that the data of diabetic patients who did not develop DN were taken from the records of those referred to the hospital. Therefore, these people may not be good representatives of people with diabetes.

Although the proposed DSS showed a significant performance for the present data, it is better to test it using the data of other healthcare centers as well. This was not possible due to the limitations of data access in the present study. The significance point of this study is that despite the presence of GFR and microalbuminuria in individuals, this variable is generally removed from the dataset and is not involved in modeling because the purpose is to predict the likelihood of DN preceding kidney damage. However, changes in GFR and microalbuminuria begin when kidney damage begins.

## 5. Conclusion

GA combined with the DT algorithm attained the highest accuracy among all the ML algorithms. Thus, the proposed DSS based on this combination helps detect DM patients prone to DN in an early stage. Physicians and health authorities can use the developed system to inform their interventions for decreasing mortality and morbidity of DN and avoiding the burden on the healthcare system. Future studies are warranted to further validate the applicability of our model in clinical settings.

## Ethical considerations

The present study was approved by the Ethics Committee of Shahid Beheshti University of Medical Sciences (ethics code [IR.SBMU.RETECH.REC.1398.628](#)). In addition, all data revealing the patient's identity were deleted, and the data were used only for the purpose collected in the study.

## Financial support

There is no financial support.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgement

Hereby, the authors would like to thank the Imam Khomeini hospital for their cooperation in data gathering.

## References

- [1] Stefanopoulos S, Ayoub S, Qiu Q, Ren G, Osman M, Nazzal M, et al. Machine learning prediction of diabetic foot ulcers in the inpatient population. *Vascular* 2021;17085381211040984.
- [2] Xu Y, Ju L, Tong J, Zhou C-M, Yang J-J. Machine learning algorithms for predicting the recurrence of stage IV colorectal cancer after tumor resection. *Sci Rep* 2020;10:1–9.
- [3] Kazerouni F, Bayani A, Asadi F, Saeidi L, Parvizi N, Mansoori Z. Type2 diabetes mellitus prediction using data mining algorithms based on the long-noncoding RNAs expression: a comparison of four data mining approaches. *BMC Bioinf Aug* 27 2020;21:372.
- [4] Rabieenia E, Jalali R, Mohammadi M. Prevalence of nephropathy in patients with type 2 diabetes in Iran: a systematic review and meta-analysis based on geographic information system (GIS). *Diabetes Metabol Syndr: Clin Res Rev* 2020;14:1543–50.
- [5] Moradi Y, Baradaran HR, Djalalinia S, Chinekesh A, Khamseh ME, Dastoorpoor M, et al. Complications of type 2 diabetes in Iranian population: an updated systematic review and meta-analysis. *Diabetes Metabol Syndr: Clin Res Rev* 2019;13:2300–12.
- [6] Huang G-M, Huang K-Y, Lee T-Y, Weng JT-Y. An interpretable rule-based diagnostic classification of diabetic nephropathy among type 2 diabetes patients. In: *BMC bioinformatics*; 2015. p. 1–10.
- [7] Gheith O, Farouk N, Nampoory N, Halim MA, Al-Otaibi T. Diabetic kidney disease: world wide difference of prevalence and risk factors. *J. Nephropharmacol.* 2016;5:49.
- [8] Bjornstad P, Cherney D, Maahs DM. Early diabetic nephropathy in type 1 diabetes—new insights. *Curr Opin Endocrinol Diabetes Obes* 2014;21:279.
- [9] Global report on diabetes," *World Health Organization* 2016.
- [10] Dabla PK. Renal function in diabetic nephropathy. *World J Diabetes* 2010;1:48.
- [11] Viberti GC, Jarrett RJ, Mahmud U, Hill RD, Argyropoulos A, Keen H. Microalbuminuria as a predictor of clinical nephropathy in insulin-dependent diabetes mellitus. *Lancet* 1982;319:1430–2.
- [12] Mogensen CE, Christensen CK. Predicting diabetic nephropathy in insulin-dependent patients. *N Engl J Med* 1984;311:89–93.
- [13] Parving HH, Oxenboll B, Svendsen PA, Christiansen JS, Andersen AR. Early detection of patients at risk of developing diabetic nephropathy. A longitudinal study of urinary albumin excretion. *Acta Endocrinol* 1982;100:550–5.
- [14] Tabaei BP, Al-Kassab As Fau - Ilag LL, Ilag Ll Fau - Zawacki CM, Zawacki Cm Fau - Herman WH, Herman WH. Does microalbuminuria predict diabetic nephropathy?. 2001.
- [15] MacIsaac RJ, Ekinci EI, Jerums G. Progressive diabetic nephropathy. How useful is microalbuminuria?: contra. *Kidney Int* 2014;86:50–7.
- [16] Radcliffe NJ, Seah Jm, Clarke M, MacIsaac RJ, Jerums G, Ekinci EI. Clinical predictive factors in diabetic kidney disease progression. *J. Diabetes. Investig.* 2017;8:6–18.
- [17] Yu DA-O, Shang J, Cai Y, Wang Z, Zhao B, Zhao Z, et al. A low-cost laboratory-based method for predicting newly diagnosed biopsy-proven diabetic nephropathy in people with type 2 diabetes. 2019.
- [18] Freedman BI, Bostrom M, Daeihagh P, Bowden DW. Genetic factors in diabetic nephropathy. *Clin J Am Soc Nephrol* 2007;2:1306.
- [19] 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* 2008;42:37–53.
- [20] Fiarni C, Sipayung EM, Maemunah S. Analysis and prediction of diabetes complication disease using data mining algorithm. *Procedia Comput Sci* 2019;161:449–57.
- [21] G.-M. Huang, K.-Y. Huang, T.-Y. Lee, and J. T.-Y. Weng, "An interpretable rule-based diagnostic classification of diabetic nephropathy among type 2 diabetes patients," *J. King Saud Univ.* pp. 1-10.
- [22] Blech I, Katzenellenbogen M, Katzenellenbogen A, Wainstein J, Rubinstein A, Harman-Boehm I, et al. Predicting diabetic nephropathy using a multifactorial genetic model. *PLoS One* 2011;6:e18743.
- [23] 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.* 2019;12:519–28.
- [24] Marini S, Trifoglio E, Barbarini N, Sambo F, Di Camillo B, Malovini A, et al. A Dynamic Bayesian Network model for long-term simulation of clinical complications in type 1 diabetes. *J Biomed Inf* 2015;57:369–76.
- [25] C. Clifton. (May 4). Data mining. Available: <https://www.britannica.com/technology/data-mining>.
- [26] Arani LA, Hosseini A, Asadi F, Masoud SA, Nazemi E. Intelligent computer systems for multiple sclerosis diagnosis: a systematic review of reasoning techniques and methods. *Acta Inf Med Dec* 2018;26:258–64.
- [27] Milovic B, Milovic M. Prediction and decision making in health care using data mining. *Int J Publ Health Sci* 2012;1.
- [28] Afrash MR, Erfanniya L, Amraei M, Mehrabi N, Jelvas S, Shanbehzadeh M. Machine Learning-Based Clinical Decision Support System for automatic diagnosis of COVID-19 based on the routine blood test. *J. Epidemiol. Biostat.* 2022;8:77–89.
- [29] Neeraj KN, Maurya V. A review on machine learning (feature selection, classification and clustering) approaches of big data mining in different area of research. *J. Crit. Rev.* 2020;7:2610–26.
- [30] Alqudah AM, Al-Hashem M, Alqudah A. Reduced number of parameters for predicting post-stroke activities of daily living using machine learning algorithms on initiating rehabilitation. *Informatica* 2021;45.
- [31] November 26 Genetic algorithm. Available: <https://www.techopedia.com/definition/17137/genetic-algorithm>.
- [32] Mónica JC, Melin P, Sánchez D. Genetic optimization of ensemble neural network architectures for prediction of COVID-19 confirmed and death cases. *Fuzzy Logic Hybrid Extensions of Neural and Optimization Algorithms: Theory Appl.* 2021;940:85.
- [33] Shukla PK, Sandhu JK, Ahirwar A, Ghai D, Maheshwary P, Shukla PK. Multiobjective genetic algorithm and convolutional neural network based COVID-19 identification in chest X-ray images. *Math Probl Eng* 2021. 2021.
- [34] Aggarwal D, Bali V, Mittal S. An insight into machine learning techniques for Predictive Analysis and Feature Selection. *Int J Innovative Technol Explor Eng* 2019;8:342–9.
- [35] Subramani P, K S, B KR, R S, B DP. Prediction of muscular paralysis disease based on hybrid feature extraction with machine learning technique for COVID-19 and post-COVID-19 patients. *Personal Ubiquitous Comput Mar* 3 2021:1–14.
- [36] Sun L, Mo Z, Yan F, Xia L, Shan F, Ding Z, et al. Adaptive feature selection guided deep forest for COVID-19 classification with chest CT. *IEEE J Biomed Health Inform Oct* 2020;24:2798–805.
- [37] Curtis F, Li X, Rose T, Vazquez-Mayagoitia A, Bhattacharya S, Ghiringhelli LM, et al. Gator: a first-principles genetic algorithm for molecular crystal structure prediction. *J Chem Theor Comput* 2018;14:2246–64.

- [38] Kim NH, Yang DW, Choi SH, Kang SW. Machine learning to predict brain amyloid pathology in pre-dementia alzheimer's disease using QEEG features and genetic algorithm heuristic. *Front Comput Neurosci* 2021;15.
- [39] Abiodun OI, Jantan A, Omolara AE, Dada KV, Mohamed NA, Arshad H. State-of-the-art in artificial neural network applications: a survey. *Heliyon Nov* 2018;4: e00938.
- [40] Nasser IM, Abu-Naser SS. Lung cancer detection using artificial neural network. *Int. J. Eng. Inf. Syst.* 2019;3:17–23.
- [41] Afrash MR, Khalili M, Salekde MS. A comparison of data mining methods for diagnosis and prognosis of heart disease. *Int J Adv Intell Paradigms* 2020;16: 88–97.
- [42] Berbić J, Ocvirk E, Carević D, Lončar G. Application of neural networks and support vector machine for significant wave height prediction. *Oceanologia* 2017; 59:331–49.
- [43] Pisner DA, Schnyer DM. Support vector machine. In: *Machine learning*. Elsevier; 2020. p. 101–21.
- [44] Cherfi A, Nouira K, Ferchichi A. Very fast C4. 5 decision tree algorithm. *Appl Artif Intell* 2018;32:119–37.
- [45] Pandey A, Jain A. Comparative analysis of KNN algorithm using various normalization techniques. *Int J Comput Netw Inf Secur* 2017;11:36.
- [46] S. K. David, M. Rafiullah, and K. Siddiqui, "Comparison of different machine learning techniques to predict diabetic kidney disease," *J. Healthc. Eng.*, vol. 2022, 2022.
- [47] Priyanka R, Aravinth J. Comparative analysis of different machine learning classifiers for prediction of diabetic retinopathy. In: *2021 international conference on recent trends on electronics, Information, Communication & Technology (RTEICT)*; 2021. p. 233–9.
- [48] Joshi RD, Dhakal CK. Predicting type 2 diabetes using logistic regression and machine learning approaches. *Int J Environ Res Publ Health* 2021;18:7346.
- [49] Maniruzzaman M, Rahman M, Ahammed B, Abedin M. Classification and prediction of diabetes disease using machine learning paradigm. *Health Inf Sci Syst* 2020;8:1–14.
- [50] Pei D, Yang T, Zhang C. Estimation of diabetes in a high-risk adult Chinese population using J48 decision tree model. *Diabetes, Metab Syndrome Obes Targets Ther* 2020;13:4621.
- [51] K. Eyasu, W. Jimma, and T. Tadesse, "Developing a prototype knowledge-based system for diagnosis and treatment of diabetes using data mining techniques," *Ethiop. J. Health Sci.*, vol. 30, 2020.
- [52] Pei D, Gong Y, Kang H, Zhang C, Guo Q. Accurate and rapid screening model for potential diabetes mellitus. *BMC Med Inf Decis Making* 2019;19:1–8.
- [53] Y. Ye, Y. Xiong, Q. Zhou, J. Wu, X. Li, and X. Xiao, "Comparison of machine learning methods and conventional logistic regressions for predicting gestational diabetes using routine clinical data: a retrospective cohort study," *J Diabetes Res*, vol. 2020, 2020.
- [54] Y. Liu, Z. Yu, and Y. Yang, "Diabetes risk data mining method based on electronic medical record analysis," *J. Healthc. Eng.*, vol. 2021, 2021.