Medicine® OPEN

Early prediction of diabetes by applying data mining techniques

A retrospective cohort study

Mohammed Zeyad Al Yousef, MBBS[a,*], Adel Fouad Yasky, MBBS[a], Riyad Al Shammari, PhD[b,c], Mazen S. Ferwana, PhD[d]

Abstract

Background: Saudi Arabia ranks 7th globally in terms of diabetes prevalence, and its prevalence is expected to reach 45.36% by 2030. The cost of diabetes is expected to increase to 27 billion Saudi riyals in cases where undiagnosed individuals are also documented. Prevention and early detection can effectively address these challenges.

Objective: To improve healthcare services and assist in building predictive models to estimate the probability of diabetes in patients.

Methods: A chart review, which was a retrospective cohort study, was conducted at the National Guard Health Affairs in Riyadh, Saudi Arabia. Data were collected from 5 hospitals using National Guard Health Affairs databases. We used 38 attributes of 21431 patients between 2015 and 2019. The following phases were performed: (1) data collection, (2) data preparation, (3) data mining and model building, and (4) model evaluation and validation. Subsequently, 6 algorithms were compared with and without the synthetic minority oversampling technique.

Results: The highest performance was found in the Bayesian network, which had an area under the curve of 0.75 and 0.71.

Conclusion: Although the results were acceptable, they could be improved. In this context, missing data owing to technical issues played a major role in affecting the performance of our model. Nevertheless, the model could be used in prevention, health monitoring programs, and as an automated mass population screening tool without the need for extra costs compared to traditional methods.

Abbreviations: ADA = American Diabetes Association, BC = Bayesian classifier, BMI = body mass index, BN = Bayesian network, CART = classification and regression tree, CBC = complete blood count, DA = discriminant analysis, DBP = diastolic blood pressure, DM = diabetes miletus, EGFR = estimated glomular filtration rate, FBS = fasting blood sugar, FINDRISC = Finnish diabetes risk score, HgA1c = hemoglobin A1c, IHME = Institute for Health Metrics and Evaluation, KACST = King Abdulaziz City for Science and Technology, KAIMRC = King Abdullah International Medical Research Center, KNN = K-nearest neighbors, LR = logistic regression, MRN = medical record number, NGHA = National Guard Health Affairs, QALY = quality adjusted life year, RBC = red blood cells, RBS = random blood sugar, RMSE = root mean square error, ROC = river operating characteristic, RTF = The Random Tree Forest, SBP = systolic blood pressure, STOME = synthetic minority oversampling technique, SVM = support vector machine.

Keywords: data mining, diabetes, diabetes prevention

1. Introduction

Diabetes is a major health problem in Saudi Arabia, with the second-highest rate of diabetes in the Middle East and the seventh highest in the world, with an estimated population of 7 million living with diabetes and more than 3 million with pre-diabetes.[1] The prevalence of type 2 diabetes in Saudi Arabia is 32.8%; however, it is predicted to reach 35.37% in 2020, 40.37% in 2025, and 45.36% in 2030.[2]

Based on data from the Institute for Health Metrics and Evaluation (IHME) in Saudi Arabia, the estimated cost of diabetes in 2014 was 17 billion riyals (US $4.5 billion), with the

This study was approved by the KAIMRC Institutional Review Board (RC 17/248/R) and funded by King Abdulaziz City for Science and Technology (KACST) (Grant number: 1-17-03-025-0003).

Mohammed Zeyad Al Yousef and Adel Fouad Yasky worked as physicians in King Abdulaziz Medical City, Riyadh. Riyadh, Saudi Arabia, is a professor in the Department of Health Informatics, College of Public Health and Health Informatics, King Saud bin Abdul-Aziz University for Health Sciences. Mazen S Ferwana worked as a consultant, in family medicine, and the primary healthcare department in King Abdulaziz Medical City, Riyadh, Kingdom of Saudi Arabia.

The data that support the findings of this study are available from the King Abdullah International Medical Research Center (KAIMRC), but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available.

[a] Family Medicine, King Abdulaziz Medical City, King Abdullah International Medical Research Center, Riyadh, Kingdom of Saudi Arabia, [b] Department of Health Informatics, College of Public Health and Health Informatics, King Saud bin Abdul-Aziz University for Health Sciences, Riyadh, Saudi Arabia, [c] Centre of Excellence in Health Informatics, Riyadh, Saudi Arabia, [d] Family Medicine and Primary Healthcare Department, King Abdulaziz Medical City, Riyadh, Kingdom of Saudi Arabia.

**Correspondence: Mohammed Zeyad Al Yousef, Family Medicine, King Abdulaziz Medical City/King Abdullah International Medical Research Center, Ar Rimayah, Riyadh 14812, Kingdom of Saudi Arabia (e-mail: m.z.alyousef@gmail.com).*

How to cite this article: Al Yousef MZ, Yasky AF, Al Shammari R, Ferwana MS. Early prediction of diabetes by applying data mining techniques: a retrospective cohort study. Medicine 2022;101:29(e29588).

Received: 14 December 2021 / Received in final form: 7 April 2022 / Accepted: 28 April 2022

http://dx.doi.org/10.1097/MD.0000000000029588

expectation that it will increase to 27 billion riyals (US $7.2 billion) in 2030 if undiagnosed people are documented. Moreover, if pre-diabetics were to become diabetic, the cost would increase to 43 billion riyals (USD 11.43 billion). These costs include medications, visits, and laboratory tests, which vary based on the patient's stage and complications.[3] Further, due to the high costs of treatment and the expected growth rate of diabetes, Saudi Arabia will encounter a health and financial dilemma in the near future. However, prevention and early detection can effectively address these challenges and decrease costs by preventing new cases and long-term complications.[4] To this end, the current evolution of information technology and the large amount of data available that could be used by applying data mining should be used.

Data mining is a concept that emerged in the 1990s as a new method for data analysis and knowledge discovery. One definition of data mining is the analysis of large observational datasets to discover unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner. Another definition is the process of finding previously unidentified patterns and trends in databases and using this information to build predictive models. Data mining has evolved from its beginning to include pattern recognition, clustering, classification, machine learning, artificial intelligence, and visualization.[5,6]

Unfortunately, the real application of and information on data mining applications in healthcare are usually not found in the scientific literature but on vendor websites as success stories that do not provide the complete technical details of which data mining algorithms were used and how. Examples of success stories include those on healthcare fraud prevention and the maximization of medical revenues by the detection of underdiagnosed patients.[5] Additionally, we found that data mining is being used in healthcare research. In this context, there are 2 major types of predictive analytics: supervised and unsupervised. Supervised learning uses known data or information on a specific problem and produces a predictive model, whereas unsupervised learning does not require previously known data on a certain problem to train its model, but defines clusters or groups.[7,8]

In clinical medicine, data mining can be used to extract knowledge from large complex datasets assessing disease risks, supporting clinical decisions, and predicting disease development.[5,9–12] A systematic review of the application of data mining techniques in the field of diabetes research, which included 17 articles, concluded that data mining is a valuable asset for diabetes researchers because it can unearth hidden knowledge from a large amount of data. Thus, it can significantly help diabetes research and, ultimately, improve the quality of healthcare for diabetic patients.[13]

Multiple studies have applied data-mining techniques to health data to construct predictive models for different diseases. Harper depicted the use of multiple healthcare datasets to compare classification algorithms and predict different health problems.[14] He applied discriminant analysis (DA), regression models (multiple and logistic), Classification and Regression tree-based algorithms (CART), and artificial neural networks to build the model; CART (a decision tree algorithm) achieved the best overall accuracies. Another study was conducted on 395 colorectal cancer patients to predict the 5-year survival rate using 17 variables and comparing 18 algorithms. The results showed an area under the curve of more than 0.9; in other words, the accuracy of the model's performance was greater than 90%.[15]

Sayad and Halkarnikar built a model to predict heart disease in patients with a sample size of 170 records and 13 attributes. They applied a multilayer perceptron neural network as a training algorithm. Their model achieved 94% accuracy with a sensitivity of 92% and specificity of 92.5%.[16] A study by Daghestani and Alshammari built a diabetes prediction model using 18 attributes from the NGHA database and found 66325 instances (diabetics 64.47%, and non-diabetics 35.53%); however, they excluded pre-diabetics.

In another chronic silent disease, such as hypertension, a study in Qatar found that it is possible to use noninvasive predictors through machine learning to achieve a predictive model that can achieve the targeted screening tool for such diseases.[17]

In this context, calculators for predicting and estimating the risk of developing diabetes are becoming more widely used. We found many such tools, but the most frequently mentioned in the literature were the American Diabetes Association (ADA) risk calculator, Cambridge Diabetes Risk Score, and Finnish Diabetes Risk Score (FINDRISC).[18–20]

Diabetes is one of the main topics of medical research because of its longevity and high cost in the healthcare system. We aimed to build a model to predict diabetic and pre-diabetic patients using demographic information and laboratory tests without the use of diagnostic tests for diabetes, hemoglobin A1c (HbA1c), random blood sugar (RBS), and fasting blood sugar (FBS), while considering that the attributes and variables are related to diabetes and are available in the database. We used a large sample size to include as much information as possible, along with multiple measures to assess the performance of our model.

2. Materials and Method

This study was conducted at the National Guard Hospital in Riyadh, Saudi Arabia. Data sets were collected from 5 hospitals that store information in the NGHA databases in the 3 highest-populated regions: the central region (Riyadh city), the western region (Jeddah and Al Madinah cities), and the eastern region (Al Ahsa and Dammam cities). This was a chart review and retrospective cohort study. All National Guard employees, dependents, and other eligible patients with available data in the NGHA database from January 2015 to January 2019 were included. Several phases were required to achieve the study objective: data collection, data preparation, data mining, model building, and model evaluation and validation.

2.1. Data collection and attribute selection

Initially, we identified the attributes that needed to be extracted from the database by reviewing the literature on diabetes (Fig. 1). The requested attributes are listed in Table 1. Complete blood count (CBC) and basic screening tests were added because almost all patients had available test results in their records. HgA1c, FBS, and RBS were requested only to classify patients as diabetic, pre-diabetic, or non-diabetic. Subsequently, they were removed from the dataset before building the model. However, other important attributes related to diabetes were unavailable (family history of diabetes, hypertension, smoking and alcohol history, and an active or sedentary lifestyle). Initially, we retrieved 1,256,898 records for demographic data (sex, age, and region), 972,239 for vital signs (systolic blood pressure [SBP], diastolic blood pressure [DBP], body mass index [BMI]), and 2,598,103 records for laboratory tests, all of which were in separate files and sheets. Table 1 summarizes the attributes used in this study.

2.2. Data preparation, cleaning, and preprocessing

Data pre-processing is an essential step that affects the prediction quality. It includes missing values, smooth noisy data, identifying or removing outliers, normalization, and transformation. Initially, we combined all data based on the medical record number (MRN) and dates when the demographic and laboratory data were entered, which resulted in the gathering

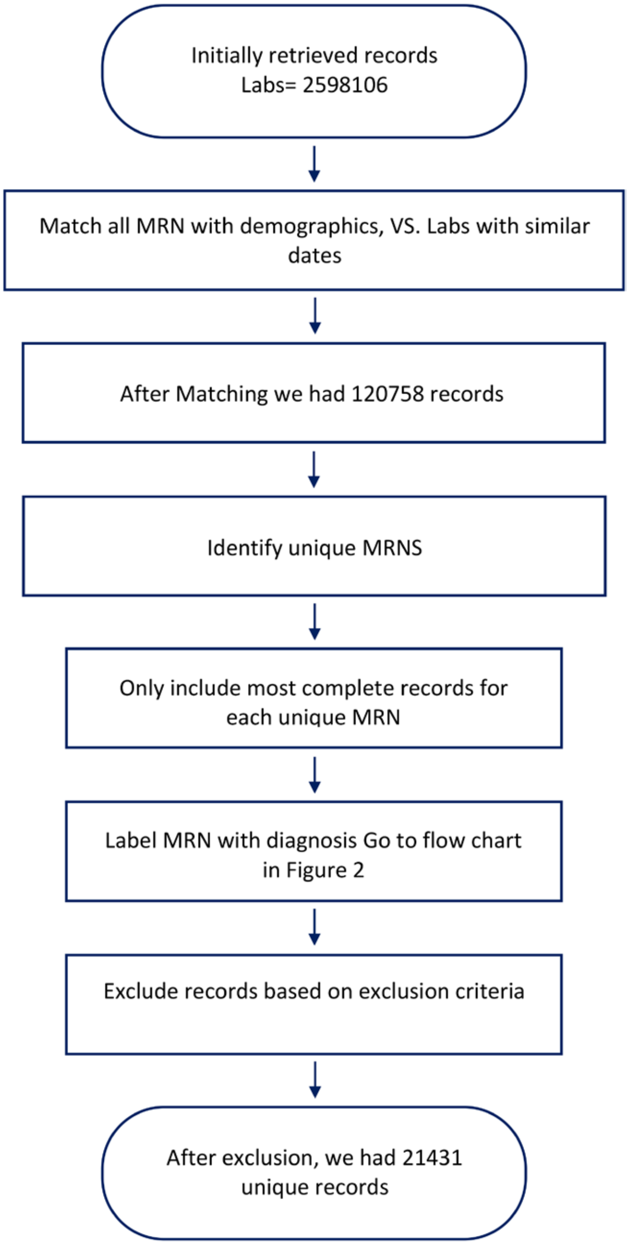
Initially retrieved records
Labs= 2598106
Match all MRN with demographics, VS. Labs with similar dates
After Matching we had 120758 records
Identify unique MRNS
Only include most complete records for each unique MRN
Label MRN with diagnosis Go to flow chart in Figure 2
Exclude records based on exclusion criteria
After exclusion, we had 21431 unique records

Figure 1. Flow chart showing the process of preparing the data.

of 120,758 records. Healthcare data are usually not well organized and have a lot of missing data and noise; therefore, they must be prepared first. To this end, we identified the number of unique MRNs by removing duplicates from the dataset; we had 45,365 unique MRNs. We then extracted a single record with the most complete attributes for each unique MRN from the 120,758 records. The missing data included either demographic information or laboratory tests that were not retrieved properly from the database due to technical issues or were not performed or requested by the medical team for the patient. Demographic information accounted for 15% (6 out of 41) of the attributes (region, sex, age, BMI, SBP, and DBP) in this study. Patients who had 0 of 6 demographic attributes were excluded.

2.3. Labeling the patient with the diagnosis

We used the ADA guidelines to diagnose diabetes and pre-diabetes status (Fig. 2),[21] and HgA1c as the main diagnostic test to label patients with diabetes, pre-diabetic, or non-diabetic. If HgA1c was missing, we used it as a diagnostic test. If FBS was missing, we used the RBS, and if RBS was missing, we labeled the patient as "non-diabetic." The flow charts in Figures 1 and 2 demonstrate how the data were retrieved, cleaned, and prepared for the model, and Tables 2 and 3 show the distribution of the demographic variables and statistical values. Additionally, manual inspection of the data was performed to ensure that the data were consistent and accurate. The MRNs were removed after preparing the data so that no patients could be identified. After preparing the data, we obtained a sample size of 21431 unique records that included 41 attributes. Table 4 presents the number of missing values for each attribute.

Table 1

Attributes used in our study.

Study	Region	RBC	WBC	ALK_Phos	Sodium	HDL
Current Study (41 attributes)	Gender	Hgb	MCHC	Adj_Ca	CO_2	Triglyceride
	Age	MPV	Mg	AGAP	Potassium	LDL
	SBP	HCT	Phosphorus	Creatinine	BUN	FBS
	DBP	MCH	Uric Acid	ALT	Chloride	A1c
	BMI	RDW	T Bili	AST	eGFR	RBS
	Platelet	MCV	Albumin	Ca	Cholesterol	

A1c = glycated hemoglobin, ADJ_Ca = adjusted calcium, AGAP = anion gap, ALK_Phos = alkaline phosphatase, ALT = alanine transaminase, AST = aspartate aminotransferase, BMI = body mass index, BUN = blood urea nitrogen, DBP = diastolic blood pressure, eGFR = estimated glomerular filtration rate, FBS = fasting blood sugar, HCT = hematocrit, HDL = high-density lipoprotein, Hgb = hemoglobin, LDL = low-density lipoprotein, MCH = mean corpuscular hemoglobin, MCHC = mean cell hemoglobin concentration, MCV = mean corpuscular volume, MPV = mean platelet volume, RBC = red blood cells, RBS = random blood sugar, RDW = red cell distribution width, SBP = systolic blood pressure, T Bili = total bilirubin, WBC = white blood cells.

2.4. Missing values

Missing values were replaced with the mean values of the associated attributes. This step aims to reduce the number of values for continuous attributes. This is achieved by splitting the range of the continuous attribute into intervals. Furthermore, discretization reduces the time needed to build the prediction model and improve the prediction results.[22] The age was thus discretized into 4 groups: group 1, 0–36; group 2, 36–54; group 3, 54–72; and group 4, 72–92.

2.5. Sampling

The dataset used in this study consisted of 21,431 unique records with 12,791 who experienced diabetes; however, 4567 did not have diabetes, and 4073 had pre-diabetics. The most common metric used to evaluate machine-learning techniques is accuracy; however, this measure will not work here because of the imbalanced nature of the 3 classes. In general, there are 2 ways to address this issue: oversampling the minority class or sampling the majority class. In this study, we used both undersampling and oversampling to solve the imbalanced data problem and compared the performance of both techniques. To this end, we used the Synthetic Minority Oversampling Technique (SMOTE).[23] The percentage of synthetic examples generated by SMOTE from class "None Diabetes Miletus (DM)" and "Pre-Diabetic" was 100% for both classes, meaning that the number of instances of "None DM" and "Pre-Diabetic" were 9134 and 8146, respectively.

2.6. Data mining and building the model

2.6.1. Feature selection. Feature selection is an important aspect of building a high-performance machine learning model, and one of its main benefits is reducing the data dimensionality,

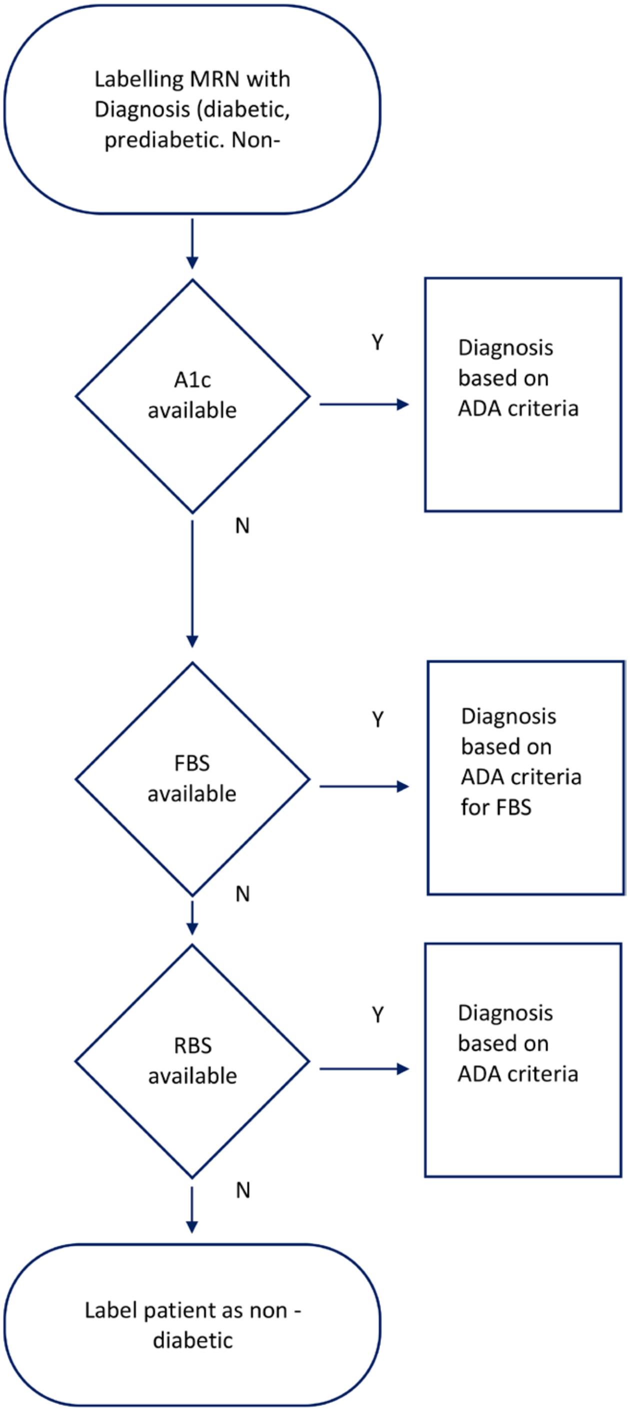
Labelling MRN with Diagnosis (diabetic, prediabetic. Non-
A1c available
Y
N
Diagnosis based on ADA criteria
FBS available
Y
N
Diagnosis based on ADA criteria for FBS
RBS available
Y
N
Diagnosis based on ADA criteria
Label patient as non - diabetic

Figure 2. Flow chart showing the process of labeling each MRN with the diagnosis. MRN = medical record number.

which in turn reduces the number of features used in building the model, and subsequently, the time needed to build the model.[24] This can be achieved by choosing the most important attributes that improve the prediction accuracy. We used an automated R-based machine learning feature selection algorithm that ranks the attributes based on their information gain.[25] It evaluates the importance of an attribute by measuring the entropy gain with respect to the outcome and then ranks the attributes based on their individual evaluations. Figure 3 shows the attributes after feature selection.

In this study, we compared the following 6 different machine learning algorithms:

(1) K-nearest neighbors (KNN) are identified from the neighbors with K similar points in the training data that are closest to the test observation. These are then classified by estimating the conditional probability of belonging to each class and choosing the class with the highest probability.[26]
(2) The random tree forest (RTF) is a classification algorithm that works by forming multiple decision trees during training and outputting the class that is the mode of the classes (classification) at testing.[27] Decision trees work by learning simple decision rules extracted from data features. The deeper the tree, the more complex are the decision rules and fit of the model. However, random decision forests overcome the problem of overfitting decision trees.
(3) Support vector machine (SVM) represents the instances as a set of points of 2 types in an N-dimensional place and generates an (N − 1)-dimensional hyperplane to separate those points into 2 groups.[28] SVM attempts to find a straight line that separates those points into 2 types and is situated as far as possible from all those points.
(4) The naïve Bayesian classifier (BC) is based on the Bayesian theorem and is particularly suited when the dimensionality of the inputs is high.[29] despite its simplicity, naïve Bayes often outperforms more complex machine learning techniques such as SVM.
(5) Bayesian network (BN) is a simple probabilistic classifier that is considered a generalization of the naive BC that removes the dependencies between variables.[30] BN is designed for modeling under uncertainty where the nodes represent variables, and arcs represent direct connections between them. The BN model allows probabilistic beliefs about variables to be updated automatically as new information becomes available.
(6) Logistic regression (LR) is a type of regression that is used to predict the outcome of the categorical dependent variable. (i.e., categorical variables have a limited number of categorical values) based on 1 or more independent variables.[31]

All machine learning algorithms were conducted using WEKA software (version 3.8) and R-based machine learning packages (version 3.3.1).[32,33]

2.7. Model evaluation and validation

The model was evaluated using the hold-out method,[34] in which the dataset was partitioned into 2 separate datasets: one for training the machine learning model and another for testing it. For the holdout method, 2 data splits were used: training with 70% of the dataset and testing with 30% of the dataset.

For all classifiers, the following evaluation metrics were calculated: precision, recall, F-score root, mean squared error, and receiver operating characteristic (ROC) curve.

3. Results

The final sample size was 21,431 patients. Finally, 15 out of 41 attributes were used to build the final model, which was chosen based on the information gained by the initial model (Fig. 3). Moreover, in our sample, 46.9% were female, 34.74% were male, and 18.36% were of unknown sex (Table 2).

Table 3 shows pre-diabetic, diabetic, and non-diabetic percentages for all genders, as well as the relevant trends. However, some of the records retrieved had missing attribute values. The 3 highest missing values out of the 15 attributes used in building the model were RBC, triglyceride, and eGFR (Table 4).

Table 2

Distribution and statistical values of gender based on the region.

			Central region		Eastern region		Western region	
Gender	**N**	**N %**	**N**	**N%**	**N**	**N%**	**N**	**N%**
Female	10,051	46.90%	6989	55.69%	1147	30.98%	1915	36.96%
Male	7446	34.74%	5045	40.2%	1081	29.20%	1320	25.48%
None available gender	3934	18.36%	515	4.1%	1474	39.81%	1945	37.54%
Total	21,431	100%	12,549	100%	3702	100%	5180	100%

Table 3

Distribution and statistical values of diagnosis based on gender and region.

	Diabetics		Prediabetics		None diabetics		Total	
Gender	**N**	**N%**	**N**	**N%**	**N**	**N%**	**N**	**N%**
Female	6132	47.94%	1881	18.71%	2038	20.27%	10,051	46.89%
Male	4554	35.60%	1554	20.87%	1338	17.96%	7446	34.74%
None available gender	2105	16.46%	638	16.21%	591	15.02%	3934	18.36%
Total	12,791	59.93%	4073	19.00%	4567	21.31%	21,431	100%

	Diabetics		Prediabetics		None diabetics		Total	
Region	**N**	**N%**	**N**	**N%**	**N**	**N%**	**N**	**N%**
Central	6989	54.63%	2640	64.81%	2920	63.93%	12,549	58.56%
Eastern	1993	15.58%	880	21.6%	829	18.15%	3702	17.27%
Western	3809	29.77%	553	13.57%	818	17.91%	5180	24.17%
Total	12,791	59.68%	4073	19.01%	4567	21.31%	21,431	100%

Table 4

The attributes and the number of missing values from the 21431 patients.

Attribute	Missing	Missing %	Attribute	Missing	Missing %	Attribute	Missing	Missing %
Region	0	0.00%	MCV	3229	15.07%	AST	5548	25.89%
Gender	2934	13.69%	WBC	3667	17.11%	Ca	6645	31.01%
Age	2933	13.69%	MCHC	3260	15.21%	Sodium	3011	14.05%
SBP	161	0.75%	Mg	6159	28.74%	CO2	1927	8.99%
DBS	161	0.75%	Phosphors	7297	34.05%	Potassium	1860	8.68%
BMI	861	4.02%	Uric Acid	8107	37.83%	BUN	1849	8.86%
Platelet	3187	14.87%	T Bili	5954	27.78%	Chloride	1937	9.04%
RBC	4139	19.31%	Albumin	6218	29.01%	eGFR	3463	16.16%
Hgb	3164	14.76%	ALK_Phos	5732	26.75%	Cholesterol	3561	16.62%
MPV	3171	14.80%	Adj_Ca	8164	38.09%	HDL	2894	13.50%
HCT	3173	14.81%	AGAP	1967	9.18%	Triglyceride	3581	16.71%
MCH	3216	15.01%	Creatinine	2389	11.15%	LDL	5803	27.08%
RDW	4123	19.24%	ALT	5356	24.99%			

ADJ_Ca = Adjusted Calcium, AGAP = anion gap, ALK_Phos = Alkaline phosphatase, ALT = alanine transaminase, AST = aspartate aminotransferase, BMI = body mass index, BUN = blood urea nitrogen, DBP = diastolic blood pressure, eGFR = estimated glomerular filtration rate, HCT = hematocrit, HDL = high density lipoprotein, Hgb = hemoglobin, LDL = low-density lipoprotein, MCH = mean corpuscular hemoglobin, MCHC = mean cell hemoglobin concentration, MCV = mean corpuscular volume, MPV = Mean Platelet Volume, RBC = red blood cells, RDW = red cell distribution width, SBP = systolic blood pressure, T Bili = total bilirubin, WBC = white blood cells.

Table 5 presents a comparison of the performances of the different classification models with and without using the SMOTE sampling method. The results of this experiment show that BN outperforms all the other classifiers in terms of precision, recall, area under the curve, F-score, root mean square error (RMSE), and accuracy.

Table 6 shows the same comparison but with the SMOTE sampling method. The results of this experiment show better results and that BN outperforms all other classifiers in terms of precision, recall, AUC, F-score, RMSE, and accuracy.

4. Discussion

One of the goals of the Saudi Vision 2030 strategic framework in the healthcare sector is to focus on primary care, preventive medicine, and tackling chronic diseases. However, when considering mass screening for diabetes to identify undiagnosed or at-risk individuals, the costs are always considered. In this context, a study was conducted in Brazil on a population screening program for type 2 diabetes, in which 22 million capillary glucose tests were performed in individuals aged 40 years and older. They concluded that the screening program will yield a large health benefit, but higher costs compared to no screening resulted in US$ 31,147 per quality-adjusted life-year (QALY) gained.[35] Using prediction models on available data from electronic health records to aid in diagnosing and identifying people at risk could be useful for early intervention and cost-effectiveness, as data are already available, and retrieving it will cost much less than conventional mass screening.

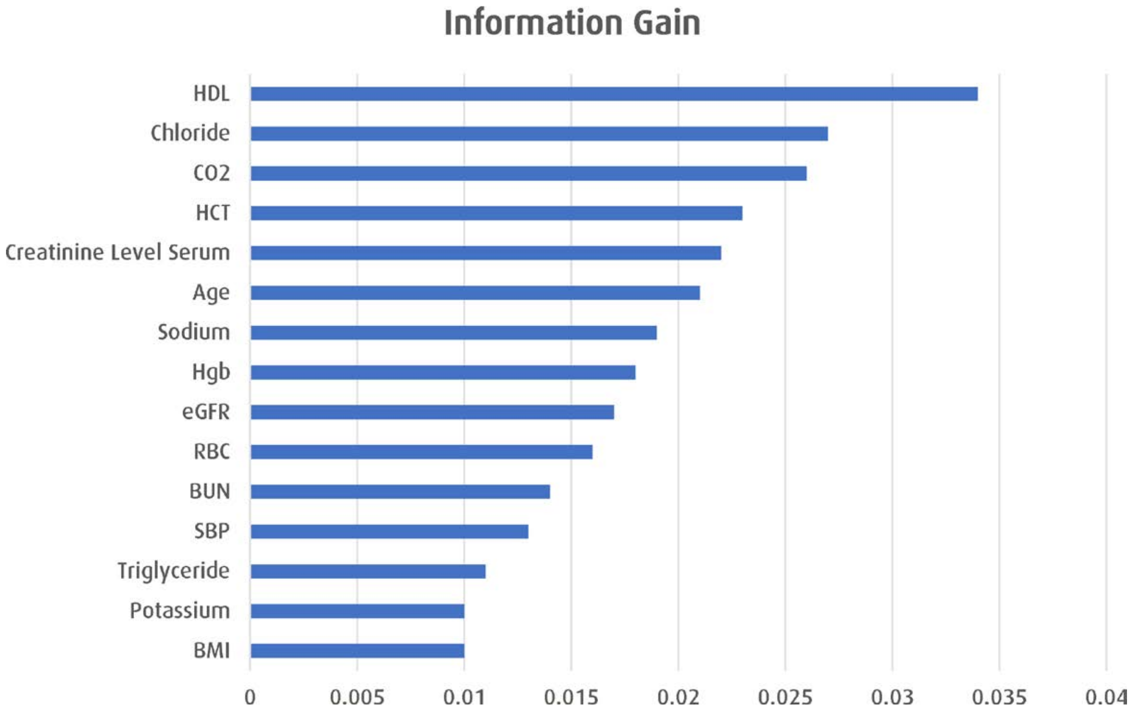

Figure 3. The selected attributes according to their information gain measures.

Table 5

Comparison of the performance of the different classification models without using the synthetic minority oversampling technique.

						KNN		
Measures	**RF**	**SVM**	**LR**	**BC**	**BN**	**K = 1**	**K = 10**	**K = 50**
Precision	56%	–	57%	56%	59%	50%	52%	–
Recall	60%	61%	62%	59%	63%	55%	62%	60%
AUC	0.67	0.53	0.67	0.70	0.71	0.55	0.62	0.62
F-score	54%	–	53%	56%	60%	52%	51%	–
RMSE	0.45	0.44	0.41	0.44	0.41	0.55	0.44	0.44
Accuracy	53%	60%	62%	59%	63%	55%	62%	60%

Table 6

Comparison of the performance of the different classification models using the synthetic minority over-sampling technique.

						KNN		
Measures	**RF**	**SVM**	**LR**	**BC**	**BN**	**K = 1**	**K = 10**	**K = 50**
Precision	60%	53%	54%	66%	62%	51%	53%	49%
Recall	28%	61%	56%	59%	66%	53%	58%	60%
AUC	0.64	0.58	0.66	0.70	0.75	0.56	0.60	0.59
F-score	22%	54%	55%	56%	61%	52%	53%	48%
RMSE	0.50	0.44	0.43	0.44	0.42	0.56	0.44	0.43
Accuracy	28%	61%	56%	59%	66%	53%	58%	60%

We used 5 measures to assess the performance of all the classifiers. We then compared the 6 classifiers with and without SMOTE. As shown in Tables 5 and 6, BN was the best classifier with an AUC of 0.71 and 0.75, RMSE of 0.41 and 0.42, the precision of 59% and 62%, recall of 63% and 66%, the accuracy of 63% and 66%, and F-score of 60% and 61%, respectively. Thus, the results show a fair model despite the missing data in our dataset.

Most prior studies on the prediction of diabetes have focused on discussing the technical aspects of machine learning, data mining, and prediction models. However, clinicians and patients are more concerned about the clinical aspects of the models, such as why the authors chose the attributes, the relevance of attributes to diabetes, and the availability of the attributes in all electronic health records, so it can be generalized to a different population.[4]

In comparison, our study included attributes that are relevant to diabetes or usually available in almost every electronic medical record. We excluded any diagnostic laboratory test that could diagnose diabetes, which would aid the model in identifying patients with the disease and eventually lead to high-performance bias. This was not the case in the other studies examined. Nevertheless, according to ADA guidelines, there are cut-off points for FBS and RBS levels to diagnose patients as diabetic, pre-diabetic, or non-diabetic, and a model will be able to find patterns similar to the cut-offs, which would increase its accuracy. [18]

Furthermore, most studies on prediction models do not include an important class: pre-diabetics. In our model, we included all 3 groups—diabetic, pre-diabetic, and non-diabetic—which ensures that no patient at risk is left out. Moreover, the data used in most models were from a specific population, whereas our data were collected from multiple hospitals in different cities. Many studies have confirmed that prediction models are a promising method for use in healthcare because of their good performance in prediction. However, not all models use a measure that can assure the significance of their model performance, as well as no bias, under, or overfitting. Table 7 summarizes a comparison of the diabetes prediction models mentioned above.

Comparing our results with the risk calculator tools, we found that the ADA risk calculator had a sensitivity of 79%, specificity of 67%, and positive predictive value of 10%, and was validated in a Chinese population with an AUC of 0.725.[18,36,37] The Cambridge Diabetes risk score had a specificity of 72%, a sensitivity of 77%, and a likelihood ratio of 2.76. The area under the receiver-operating characteristic curve was 80% and was validated in different populations, but mostly in European countries; the results were lower with an AUC ranging from 63% to 74.5%,[19,38–41] FINDRISC had a sensitivity of 81%, a specificity of 76%, and a positive predictive value of 0.05. The tool was validated in other populations, and the results showed an AUC of 0.724 and 0.75.[20,42,43] All risk calculator tools were validated and showed good results in identifying at-risk patients. Some of the questions were found to be similar in all risk calculator tools, indicating their importance in identifying at-risk patients. Unfortunately, although we know the importance of these questions and how they will increase the performance of our model, we could not retrieve some of these data because they were not available in the system. On the other hand, comparing the results of identifying patients in our model had an AUC of 0.71 and 0.75, which is promising and similar to the validated risk tool measures; thus, our model could be improved if we added important missing attributes.

Moreover, all risk calculator tools need to be filled manually by a physician, nurse, or person who speaks English and would require either an Internet connection to access it or have it printed. This would lead to increased costs due to time loss, papers used, availability of manpower, and internet connectivity to run these tools. This increase would affect the number of patients assessed using the risk tools. Nevertheless, it is noteworthy that our model was capable of screening 21,431 patients in a couple of minutes at no cost.

Integrating an automated prediction model, such as our model, into the electronic health system would give the healthcare provider a live identification of those at risk of developing diabetes. It can be sent automatically or after the patient's health information has been reviewed by a healthcare provider, so that proper diagnostic screening tests can be performed. This eventually leads to early identification, intervention, and management for better health outcomes and reduction of disease complications and costs.

4.1. Limitations

Because of the unavailability and missing data, as shown in Table 4, our accuracy was affected, which consequently affected the feature selection for the attributes. Nevertheless, more attributes that are clinically related to the disease can be added to increase the possibility of prediction. However, because these data were unavailable in the database, they would either have to be collected manually or the stakeholders in our hospital and the health information system department would have to be contacted to implement some changes to the system that would ensure that these data are available in the future.

Furthermore, we recommend identifying technical problems with the information technology department regarding missing data, adding more attributes that are related and known to increase the risk of developing diabetes from the literature to the prediction model, collaborating with other hospitals to cover more population, increasing the sample size, excluding pediatrics, and identifying better methods for imputing missing data.

5. Conclusion

The results from our model are acceptable but can be improved. Missing and unavailable data owing to technical issues play a major role in affecting the performance of our model. Nevertheless, it could still be used in preventive and health monitoring programs and as an automated mass population screening tool to identify diabetics and individuals at high risk of developing diabetes with the available data in the database and without the need for extra costs as compared to traditional methods. Moreover, with the use of better data entry and sorting methods in our health attribution datasets, the implementation of such prediction models would be more effective in the long run, although we have achieved similar outcomes to those of similar models.

Table 7

A comparison summary of the models.

Results	Performance measures	Validation and training	# of attributes	class	# of records	Data set	Author
RF: Recall (90%), precision (68%)	Sensitivity (recall) and PPV (precision)	training/testing, percentage not mentioned	18	Dm, Non-DM	66,325	NGHA 2013-2015	Daghistani T[4]
PNN: accuracy 81.49%	Accuracy	76% training/25% testing	9	Dm, Non-DM	768	PIMA	Soltani Z[18]
clustering + SVM: Accuracy 98.93, sensitivity 99.33, specificity 98.73% a and AUC of 0.97	Accuracy, sensitivity, specificity, AUC	Cross validation	9	Dm, Non-DM	768	PIMA	Ilango B[19]
clustering + C4.5: accuracy 92.38 %, sensitivity (90.38), and specify (93.29).	Accuracy, sensitivity, specificity	Cross validation	9	Dm, Non-DM	768	PIMA	Patil B[20]
BN: Precision (62%), Recall (66%), AUC (0.75), F-Score (61%), RMSE (0.42), and Accuracy (66%)	Sensitivity (recall), PPV (precision), AUC, F score, RMSE, accuracy	70% training/ 30% testing	41	Dm, Pre Dm, Non-DM	18,181	NGHA 2015–2018	Current work

BN = Bayesian Network, Dm = Diabetic, Non-DM = Non-Diabetic, PNN = probabilistic neural network, Pre Dm = Pre-diabetic, RF = Random Forest, SVM = support vector machine.

Acknowledgments

The authors would like to thank KAIRMC and KACST for their cooperation and support shown by them during the study. The authors would also like to acknowledge and value the input of Prof. Sherif Sakr as he passed away recently and could not witness the final result.

References

[1] Robert AA, Al Dawish MA, Braham R, et al. Type 2 diabetes mellitus in Saudi Arabia: major challenges and possible solutions. Curr Diabetes Rev. 2017;13:59–64.

[2] Meo SA. Prevalence and future prediction of type 2 diabetes mellitus in the Kingdom of Saudi Arabia: a systematic review of published studies. J Pak Med Assoc. 2016;66:722–5.

[3] El Bcheraoui C, Basulaiman M, Tuffaha M, et al. Status of the diabetes epidemic in the Kingdom of Saudi Arabia, 2013. Int J Public Health. 2014;59:1011–1021.

[4] Daghistani T, Alshammari R. Diagnosis of diabetes by applying data mining classification techniques. Int J Adv Comput Sci Appl. 2016;7.

[5] Yoo I, Alafaireet P, Marinov M, et al. Data mining in healthcare and biomedicine: a survey of the literature. J Med Syst. 2012;36:2431–48.

[6] Kincade K. Data mining: digging for healthcare gold. Ins Technol. 1998;23:IM2–IM7.

[7] Eckerson W. Predictive analytics. Transforming data with intelligence. 2007 Oct 5. Available at: https://tdwi.org/articles/2007/05/10/predictive-analytics.aspx

[8] Finlay S. Predictive analytics. Data mining and big data. London: Palgrave Macmillan UK. 2014.

[9] Dong W, Huang Z, Ji L, et al. A genetic fuzzy system for unstable angina risk assessment. BMC Med Inform Decis Mak. 2014;14.

[10] Zhang Y, Guo S-L, Han L-N, et al. Application and exploration of big data mining in clinical medicine. Chin Med J. 2016;129:731–8.

[11] Rastgarpour M, Shanbehzadeh J. A new kernel-based fuzzy level set method for automated segmentation of medical images in the presence of intensity inhomogeneity. Comput Math Methods Med. 2014;2014:978373.

[12] Sato F, Shimada Y, Selaru FM, et al. Prediction of survival in patients with esophageal carcinoma using artificial neural networks. Cancer. 2005;103:1596–605.

[13] Marinov M, Mosa AS, Yoo I, et al. Data-mining technologies for diabetes: a systematic review. J Diabetes Sci Technol. 2011;5:1549–56.

[14] Harper PR. A review and comparison of classification algorithms for medical decision making. Health Policy. 2005;71:315–31.

[15] Pourhoseingholi MA, Kheirian S, Zali MR. Comparison of basic and ensemble data mining methods in predicting 5-year survival of colorectal cancer patients. Acta Inform Med. 2017;25:254–8.

[16] Sayad A, Halkarnikar P. Diagnosis of heart disease using neural network approach. Int J Adv Sci Eng Technol. 2014;2.

[17] AlKaabi LA, Ahmed LS, Al Attiyah MF, et al. Predicting hypertension using machine learning: findings from Qatar Biobank study. PLoS One. 2020;15:e024.

[18] Bang H, Edwards AM, Bomback AS, et al. Development and validation of a patient self-assessment score for diabetes risk. Ann Intern Med. 2009;151:775–83.

[19] Griffin SJ, Little PS, Hales CN, et al. Diabetes risk score: towards earlier detection of type 2 diabetes in general practice. Diabetes Metab Res Rev. 2000;16:164–71.

[20] Lindström J, Tuomilehto J. The diabetes risk score: a practical tool to predict type 2 diabetes risk. Diabetes Care. 2003;26:725–31.

[21] American Diabetes Association. 2. Classification and diagnosis of diabetes: standards of medical care in diabetes-2018. Diabetes Care. 2018;41(suppl 1):S13–27.

[22] Kurgan L, Swiercz W, Cios K. Semantic mapping of XML tags using inductive machine learning. Proceedings of International Conference on Artificial Intelligence. CSREA Press. 2002.

[23] Chawla NV. Data mining for imbalanced datasets: an overview. In: Maimon O, Rokach L, eds. Data mining and knowledge discovery handbook. Boston, MA: Springer 2009; 875–886.

[24] Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res. 2003;3:1157–82.

[25] Kent JT. Information gain and a general measure of correlation. Biometrika. 1983;70:163–73.

[26] Mullick SS, Datta S, Das S. Adaptive learning-based k-nearest neighbor classifiers with resilience to class imbalance. IEEE Trans Neural Netw Learn Syst. 2018;29:5713–25.

[27] Ho TK Random decision forests. Proceedings of 3rd International Conference on Document Analysis and Recognition IEEE. Wiley-IEEE Press; 2002.

[28] Hearst MA, Dumais ST, Osuna E, et al. Support vector machines. IEEE Intell Syst. 1998;13:18–28.

[29] Webb GI, Boughton JR, Wang Z. Not so naive Bayes: Aggregating one-dependence estimators. Mach Learn. 2005;58:5–24.

[30] Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers. Mach Learn. 1997;29:131–63.

[31] Balakrishnan N. Handbook of the logistic distribution. New York: Marcel Dekker. 1991.

[32] Machine Learning Group at the University of Waikato. WEKA. 2020. Available at: http://www.cs.waikato.ac. nz/ml/weka/

[33] The R Foundation. The R project for statistical computing. 2020 Oct 29. Available at: https://www.r-project.org/

[34] Hawkins DM. The problem of overfitting. J Chem Inform Comput Sci. 2004;44:1–12.

[35] Toscano CM, Zhuo X, Imai K, et al. Cost-effectiveness of a national population-based screening program for type 2 diabetes: the Brazil experience. Diabetol Metab Syndr. 2015;7.

[36] Poltavskiy E, Kim DJ, Bang H. Comparison of screening scores for diabetes and prediabetes. Diabetes Res Clin Pract. 2016;118:146–53.

[37] Woo YC, Lee CH. Fong CHY, et al. Validation of the diabetes screening tools proposed by the American diabetes association in an aging Chinese population. PLoS One. 2017;12:e0184840.

[38] Kengne AP, Beulens JW, Peelen LM, et al. Noninvasive risk scores for prediction of type 2 diabetes (EPIC-InterAct): a validation of existing models. Lancet Diabetes Endocrinol. 2014;2:19–29.

[39] Spijkerman AM, Yuyun MF, Griffin SJ, et al. The performance of a risk score as a screening test for undiagnosed hyperglycemia in ethnic minority groups: data from the 1999 health survey for England. Diabetes Care. 2004;27:116–22.

[40] Rahman M, Simmons RK, Harding AH, et al. A simple risk score identifies individuals at high risk of developing type 2 diabetes: a prospective cohort study. Fam Pract. 2008;25:191–6.

[41] Park PJ, Griffin SJ, Sargeant L, et al. Performance of a risk score in predicting undiagnosed hyperglycemia. Diabetes Care. 2002;25:984–8.

[42] Makrilakis K, Liatis S, Grammatikou S, et al. Validation of the Finnish diabetes risk score (FINDRISC) questionnaire for screening for undiagnosed type 2 diabetes, dysglycaemia, and metabolic syndrome in Greece. Diabetes Metab. 2011;37:144–51.

[43] Zhang L, Zhang Z, Zhang Y, et al. Evaluation of Finnish diabetes risk score in screening undiagnosed diabetes and prediabetes among U.S. adults by gender and race: NHANES 19992010. PLoS One. 2014;9:e97865.