HEALTH SYSTEMS
2023, VOL. 12, NO. 2, 243–254
https://doi.org/10.1080/20476965.2022.2141141

RESEARCH ARTICLE

Identification of the risk factors of type 2 diabetes and its prediction using machine learning techniques

Md. Merajul Islam[a,b], Md. Jahanur Rahman[a], Md. Menhazul Abedin[c], Benojir Ahammed[c], Mohammad Ali[c], N.A.M Faisal Ahmed[d] and Md. Maniruzzaman[c]

[a]Department of Statistics, University of Rajshahi, Rajshahi, Bangladesh; [b]Department of Statistics, Jatiya Kabi Kazi Nazrul Islam University, Mymensingh, Bangladesh; [c]Statistics Discipline, Khulna University, Khulna, Bangladesh; [d]Institute of Education and Research, University of Rajshahi, Rajshahi, Bangladesh

ABSTRACT

This study identified the risk factors for type 2 diabetes (T2D) and proposed a machine learning (ML) technique for predicting T2D. The risk factors for T2D were identified by multiple logistic regression (MLR) using p-value ($p<0.05$). Then, five ML-based techniques, including logistic regression, naïve Bayes, J48, multilayer perceptron, and random forest (RF) were employed to predict T2D. This study utilized two publicly available datasets, derived from the National Health and Nutrition Examination Survey, 2009-2010 and 2011-2012. About 4922 respondents with 387 T2D patients were included in 2009-2010 dataset, whereas 4936 respondents with 373 T2D patients were included in 2011-2012. This study identified six risk factors (age, education, marital status, SBP, smoking, and BMI) for 2009-2010 and nine risk factors (age, race, marital status, SBP, DBP, direct cholesterol, physical activity, smoking, and BMI) for 2011-2012. RF-based classifier obtained 95.9% accuracy, 95.7% sensitivity, 95.3% F-measure, and 0.946 area under the curve.

ARTICLE HISTORY
Received 12 October 2020
Accepted 20 October 2022

KEYWORDS
Risk factors; prediction; type 2 diabetes; machine learning

1. Introduction

Diabetes is a group of metabolic disorders or syndromes and is a serious public health concern worldwide. It is also a chronic non-communicable disease in which blood sugar or blood glucose levels are high and the body is unable to use energy from dietary sources (Ghosh, 2017). Due to high blood glucose, some symptoms, such as intense thirst and appetite, frequent urination, nausea and vomiting, and slow wound healing, all of which arise gradually in diabetes patients, Diabetes, if not properly treated, can affect the body's major organs and cause a variety of diseases, such as heart disease, stroke, blindness, kidney failure, and nerve damage (Habibi et al., 2015; Rawshani et al., 2018; Saedi et al., 2016). Diabetes affected 451 million people worldwide in 2017, and this figure is expected to rise to around 693 million by 2045 (Cho et al., 2018). According to the Centers for Disease Control and Prevention (CDC), 29.1 million people in the United States were diagnosed with diabetes in 2012 (Centers for Disease Control and Prevention, 2014). According to the World Health Organization, around 1.6 million people die owing to diabetes every year. The number of diabetic patients has been increased day by day. As a result, deaths are increasing day by day (Ma et al., 2014).

Type 1, type 2, and gestational diabetes are the most common kinds of diabetes. Among them, type 2 diabetes (T2D) begins with insulin resistance, and cells fail to produce enough insulin properly. T2D accounts for 90% to 95% of all diabetes diagnoses and has become a severe concern in both low- and middle-income countries, including USA, as its prevalence has increased dramatically (Canadian Diabetes Association, 2013). T2D is not a perfectly foreseeable or preventable disease. Diabetes is more likely to develop than any other chronic disease, especially if it is diagnosed late or not at all (Habibi et al., 2015; Rawshani et al., 2018).

Thus, early detection of diabetes is an important task to control and avoid its complications. A comprehensive guideline was issued for the prevention of diabetes, specifying lifestyle changes (Zhu et al., 2015). Various techniques have also been proposed to reduce the risk of diabetes (Woldesemayat et al., 2019). Naturally, prevention is preferable, but current treatment methods are not sufficient to achieve this goal. The tedious identification process results in visiting the patient at the diagnostic centre and consulting with the doctor. As a result, early diabetes detection is becoming more important and needed to propose an automated system. Therefore, detection and classification of diabetes at an early stage is increasingly gaining interest in the medical sciences. However, machine learning (ML)-based approaches address this critical issue. There were some studies available in the literature where the model was designed to only identify the risk factors

CONTACT Md. Maniruzzaman at monir.stat91@gmail.com Statistics Discipline, Khulna University, Khulna-9208, Bangladesh

for diabetes (Hamasaki, 2016; Saydah et al., 2014; Tan et al., 2019; Zafar et al., 2016). Despite the rapid development of theories for computational intelligence in ML-based classifiers, their application has been increased rapidly to predict diabetes disease in public health and other fields (Alanazi et al., 2018; Avorn, 2013; Bron et al., 2014; Das, 2014; Dong et al., 2013; Maguire & Dhar, 2013; Shankaracharya et al., 2012). However, several ML-based models have been used to predict and classify diabetes diseases in different countries using different diabetic datasets. Nevertheless, the model's performance needs to be improved (Dagliati et al., 2018; Kaur & Kumari, 2020; Kavakiotis et al., 2017; Nai-arun & Moungmai, 2015; Sneha & Gangil, 2019; Tsao et al., 2018; Zou et al., 2018). Nevertheless, the performance of the proposed existing ML-based predictive model is not at an as satisfactory level, which means model performance needs to be improved.

Thus, an attempt has been made in this paper to detect and predict T2D using ML-based classifiers. The hypothesis of this study is that the combination of a multiple logistic regression (MLR)-based risk factor identification method and an ML-based classifier with the highest classification accuracy yields enough information to identify potential T2D patients and thus improve diagnosis accuracy. This study aims to build several ML-based predictive models for risk factor prediction of T2D. Risk factor identification methods select the significant T2D risk factors while avoiding irrelevant factors in order to build an effective predictive learning model. Consequently, it can contribute to increasing the performance of the classifiers. In this study, we used an MLR model as a risk factor identification method. Five well-known and popular ML-based classifiers, namely logistic regression (LR), naïve Bayes (NB), J48, multilayer perceptron (MLP), and random forest (RF), have been included to predict T2D status based on the significant risk factors and their performance has been compared using accuracy (ACC), sensitivity (SE), F-measure (FM), and area under the curve (AUC).

2. Methods

2.1. Data source

In this study, the T2D dataset was derived from an existing public domain survey dataset, the 2009–2010 and 2011–2012 National Health and Nutrition Examination Survey (NHANES), which was nationally representative and freely available online (National Health and Nutrition Examination Survey (NHANES), 2009–2010; National Health and Nutrition Examination Survey (NHANES), 2011–2012). The datasets contained some extraneous information, such as do not know, refused, and missing values, which we excluded from our analysis. After excluding this extraneous information, the dataset consisted of 4922 respondents (diabetic: 387 vs. control: 4535) and 4936 respondents (diabetic: 387 vs. control: 4563) for 2009–2010 and 2011–2012. We reviewed some existing studies on diabetes before conducting this work. In this study, we selected the factors based on the existing studies that were closely related to diabetes and had an impact on diabetes. The risk factors were age (Bahour et al., 2022; Nanayakkara et al., 2021; Shamshirgaran et al., 2017; Suastika et al., 2012; Wang et al., 2021), sex (Derakhshan et al., 2014; Harreiter & Kautzky-Willer, 2018; Huebschmann et al., 2019), race (Cheng et al., 2019; Link & McKinlay, 2009; Spanakis & Golden, 2013), education (Flatz et al., 2015; Sil et al., 2020; Steele et al., 2017), marital status (Kposowa et al., 2021; Oliveira et al., 2020), occupation (Almeida et al., 2011; Carlsson et al., 2020; Nakazawa et al., 2022), systolic blood pressure (Aikens et al., 2017; Chen et al., 2015; S. W. Lee et al., 2017; H. S. Lee et al., 2013), diastolic blood pressure (Akalu & Belsti, 2020; Emdin et al., 2015), direct cholesterol (Bhowmik et al., 2018; Chen et al., 2015; Howard et al., 2000), total cholesterol (Bhowmik et al., 2018; Gimeno-Orna et al., 2005), physical activity (Ghaderpanahi et al., 2011; Hamasaki, 2016; F. Zhao et al., 2020), dirking alcohol (Holst et al., 2017; Kao, 2001; Li et al., 2016), smoking status (Chang, 2012; Maddatu et al., 2017; Yang et al., 2022), and body mass index (Abdissa et al., 2021; Bays et al., 2007; Q. Zhao et al., 2017; Maggio and Pi-Sunyer, 2003). The details and descriptions of these factors are presented in Table 2. The existing works also showed that exposure to mass media, sleep duration, diet, and so on were significant behavioral and lifestyle related individual risk factors of T2D. However, due to the unavailability of data for these variables in the NHANES database, we could not include themin our analysis.

2.2. Statistical analysis

The demographic and baseline characteristics of T2D patients are reported as frequency (%) and mean standard deviation (SD) for categorical and continuous data. We have employed the chi-square test and the paired t-test for categorical and continuous data to investigate the association between various factors and diabetes. All statistical analyses were performed using Stata version 14 and Ri86 3.6.1.

2.3. Risk factor identification and machine learning techniques

The identification of risk factors is a crucial task for reducing the computational complexity of ML systems. As a result, model performance improves while construction time and costs are reduced.

Various risk factor identification methods are available in the literature, such as: MLR (Maniruzzaman et al., 2020; Xie et al., 2019), analysis of variance (ANOVA; Elssied et al., 2014), mutual information (MI; Shrivastava et al., 2017), RF (Gregorutti et al., 2017), and principal component analysis (PCA; Adhao & Pachghare, 2020). In the current study, the MLR model was adopted to identify the prominent risk factors of T2D using a p-value (<0.05). Then, the dataset is divided into two sets: 90% dataset as a training set and 10% dataset as a test set. Five predictive models that are more popular and applicable in the literature, including LR (Islam et al., 2020), NB (Maniruzzaman et al., 2018), J48 (Sisodia & Sisodia, 2018), MLP (Mohapatra et al., 2019), and RF (Kumar et al., 2019), are applied to the training set and predict T2D on the test set. We tuned the hyperparameters of the ML-based models, and the final results were reported based on the optimum hyperparameters for RF (100 trees), MLP (50 sizes), and J48 (0.025 confidence threshold (C): 0.025 and minimum instances per leaf (M): 25). An overview of the proposed ML framework is presented in Figure 1.

2.4. Performance evaluation metrics

The performance of ML-based models is evaluated by four evaluation metrics, including ACC, SE, FM, and AUC. The values of evaluation metrics have been computed from the confusion matrix by four quantities: true positive (t_p), the number of cases correctly predicted as positive; false positive (f_p), the number of cases incorrectly predicted as positive; true negative (t_n), the number of cases correctly predicted as negative; and false negative (f_n), the number of cases incorrectly predicted as negative. The confusion matrix is shown in Table 1. These t_p, f_p, t_n, and f_n were used to calculate ACC, SE, and FM, which are mathematically defined as follows:

2.4.1. Accuracy

It measures the proportion of true results, either t_p, or t_n against the total population. Mathematically, ACC is defined as:

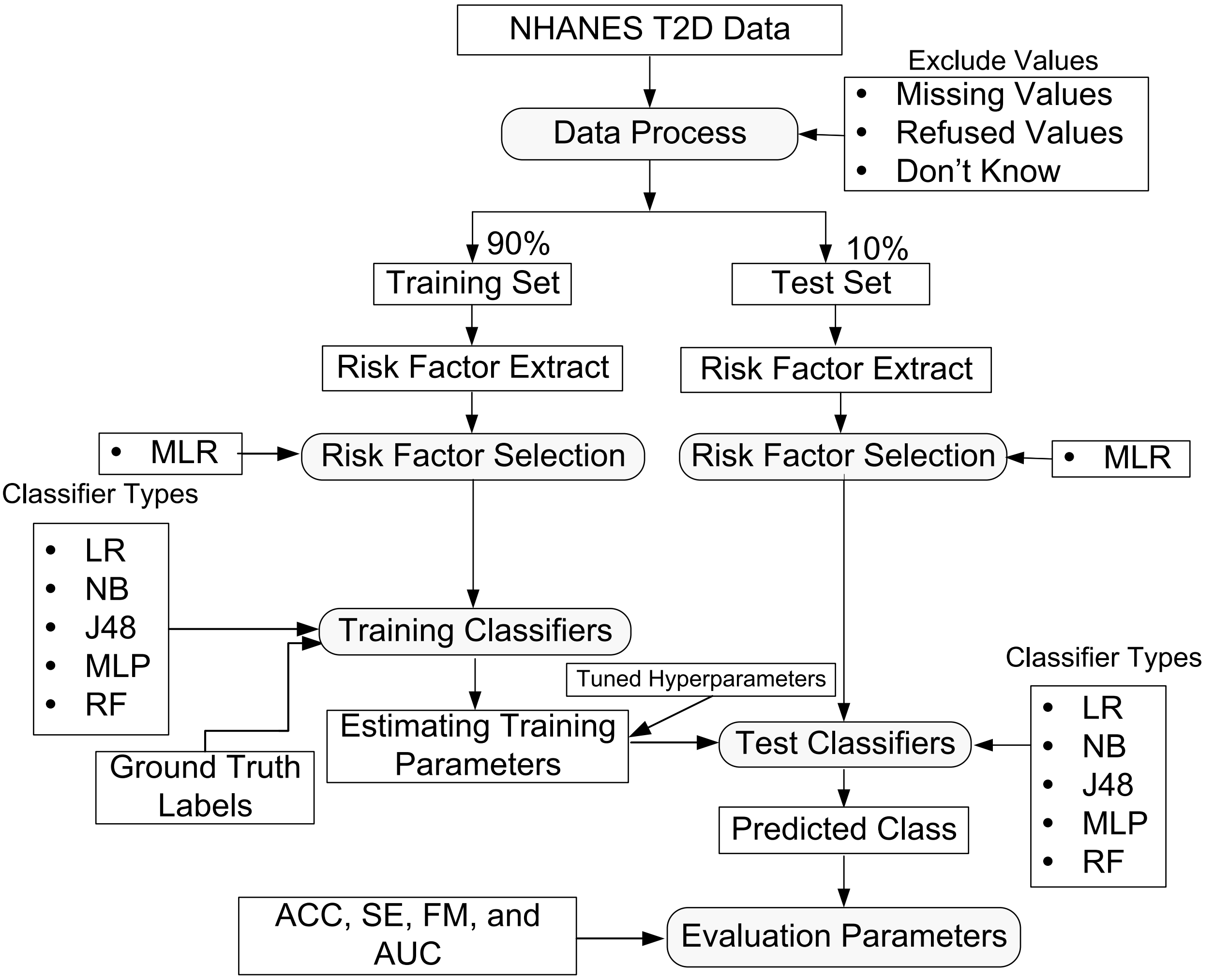


Figure 1. Overview of the proposed ML-based framework.

Table 1. Confusion matrix.

	Total population (p + n)	Predicted class	
		Diabetic (p)	Control (n)
Actual class	Diabetic (p)	True positive (t_p)	false negative (f_n)
	Control (n)	False positive (f_p)	true negative (t_n)

$$\mathrm{ACC}(\%)=\frac{\mathrm{t_p+t_n}}{\mathrm{t_p+t_n+f_p+f_n}}\times 100 \tag{1}$$

2.4.2. Sensitivity

It measures the proportion of t_p cases against the predicted positive cases. Model with high SE indicates few f_n. It is also called recall or true positive rate. Mathematically, SE is defined as:

$$\mathrm{SE}(\%)=\frac{\mathrm{t_p}}{\mathrm{t_p+f_n}}\times 100 \tag{2}$$

2.4.3. F-measure

It is a harmonic mean of precision and SE. Mathematically, FM is defined as:

$$\mathrm{FS}(\%)=\frac{\mathrm{2t_p}}{\mathrm{2t_p+f_p+f_n}}\times 100 \tag{3}$$

2.4.4. Area under the curve

The area under the curve (AUC) is defined as an integral of the receiver operating characteristic (ROC) function over the given range and used to assess the quality of the constructed predictive model. The mathematical formula of AUC as follows

$$AUC = \int_{x=0}^{1} TPR(FPR^{-1}(x))dx \tag{4}$$

A ROC curve is a plot of true positive rate (TPR) or *sensitivity* on the y axis against false-positive rate (FPR) or *1*-specificity on the x axis for different cut-off values. ROCs generate an AUC value from 0 to 1.

3. Results

3.1. Baseline characteristics of the patients

The objective of this section was to observe a bivariate association between the included risk factors and diabetes status. The results of the association have been shown in Table 2. The average age of T2D patients in 2009–2010 and 2011–2012 were 60.1 ± 13.8 years and 58.4 ± 15.5 years, respectively. The prevalence of T2D patients was almost 8% in both 2009–2010 and 2011–2012 datasets. In both datasets, the majority of T2D patients were male. In 2009–2010, 9.7%~10% of T2D patients were graduated from college, while 11.1%~11% in 2011–2012. Table 2 also shows that ~11% of T2D patients were married. All factors except gender and DBP were significantly associated with T2D in 2009–2010 dataset, whereas, except for gender, TC, and alcohol, all factors were also significantly associated with T2D in 2011–2012 (see Table 2).

3.2. Identification of risk factors of T2D using MLR

The objective of this section was to identify the prominent risk factors for T2D. The prominent risk factors have been identified by odds ratio (OR) and p-value (<0.05). The value of OR is 1, greater than 1, and less than 1. If the value of OR = 1, then that factors had no risk of T2D; OR >1, the factors have a higher risk of T2D, and vice-versa. The results of MLR for the identification of risk factors for T2D are summarised in Table 3. It is to be noticed that, in 2009–2010, 9–11th grade (OR: 2.86, CI: 2.04–3.00; p = 0.044), high school (OR: 1.48, CI: 1.10–2.01; p = 0.01) educated respondents were more likely to have T2D than 8th grade, whereas 8th grade educated respondents were more likely to have T2D than college (OR: 0.73, CI: 0.54–0.97; p = 0.007). Married respondents (OR: 1.82, CI: 1.07–3.08; p = 0.030) had more likely to have T2D than divorced respondents. The participants who smoked (OR: 1.87, CI: 1.18–2.97; p = 0.007) had a higher risk of T2D than non-smoker. It is also observed that age (OR: 1.05, CI: 1.04–1.07; p < 0.001), SBP (OR:

Table 2. Baseline characteristics of the T2D patients.

			2009–2010			2011–2012	
Factors	Descriptions	Diabetic	Control	p-value[a]	Diabetic	Control	pvalue[a]
Total, n (%)		387 (7.9)	4535 (92.1)		373 (7.6)	4563 (92.4)	
Age (years)	Age in years	60.1 ± 13.8	35.4 ± 21.9	<0.001	58.4 ± 15.5	35.5 ± 21.9	<0.001
Gender, Male n (%)	Gender	205 (8.4)	2234 (91.6)	0.161	198 (8.0)	2272 (92.0)	0.222
Race, White n (%)	Race	232 (7.3)	2,958 (92.7)	0.035	218 (7.0)	2882 (93.0)	0.008
Education, College n (%)	Education status	177 (9.7)	1757 (90.3)	<0.001	205 (11.1)	1634 (88.9)	<0.001
Marital status, Married n (%)	Marital status	223 (10.9)	1826 (89.1)	<0.001	203 (10.7)	1693 (89.3)	<0.001
Occupation, Working, n (%)	Occupation	143 (6.1)	2207 (93.9)	<0.001	151 (6.7)	2112 (93.3)	<0.001
SBP (mm Hg)	Systolic blood pressure	126.5 ±20.3	116.7 ±16.5	<0.001	129.4 ±18.4	117.7 ±17.0	<0.001
DBP (mm Hg)	Diastolic blood pressure	65.7 ±13.5	66.8 ± 14.3	0.157	70.5 ±14.3	68.1 ±14.4	0.003
DC (mg/dL)	Direct cholesterol	1.3 ±0.4	1.4 ±0.4	<0.001	1.2 ±0.4	1.4 ±0.4	<0.001
TC (mg/dL)	Total cholesterol	4.8 ±1.1	4.9 ±1.1	0.001	4.8 ±1.2	4.8 ±1.1	0.286
PA, Yes, n (%)	Physical activity	142 (6.4)	2135 (93.8)	<0.001	143 (6.0)	2226 (94.0)	<0.001
Alcohol, Yes, n (%)	Drinking alcohol	255 (10.0)	2298 (90.0)	0.001	263 (9.9)	2394 (90.1)	0.245
Smoking, Yes, n (%)	Smoking status	56 (7.3)	715 (92.7)	<0.001	50 (7.2)	648 (92.8)	<0.001
BMI (kg/m^2)	Body mass index	32.6 ±8.0	26.4 ±7.4	<0.001	32.5 ±7.7	26.0 ±7.0	<0.001

[a] p-value is obtained from paired t-test for continuous and Chi-Square test for categorical data.

Table 3. Identification of risk factors of T2D using MLR.

	2009–2010		2011–2012	
Factors	OR (95% CI)	p-value	OR	p-value
Age (years)	1.05(1.04–1.07)	<0.001	1.04(1.02–1.06)	<0.001
Race				
White®	1.00	-	1.00	-
Hispanic	1.42(0.50–4.00)	0.512	1.22(0.46–3.26)	0.680
Mexican	1.34 (0.62–2.87)	0.458	2.13(1.03–4.42)	0.040
Black	1.58(0.80–2.98)	0.159	2.56(1.44–4.55)	0.001
Other's	1.46(0.53–4.00)	0.459	2.21(1.06–4.61)	0.034
Education				
8 grade®	1.00	-	1.00	-
9–11th grade	2.86(2.04–3.00)	0.044	0.72(0.32–1.61)	0.422
High school	1.48(1.10–2.01)	0.010	1.11(0.51–2.39)	0.794
College	0.73(0.54–0.97)	0.007	1.51(0.79–2.88)	0.209
Marital status				
Divorced®	1.00	-	1.00	-
Married	1.82 (1.07–3.08)	0.030	1.82(1.04–3.18)	0.034
Unmarried	2.20(0.95–5.09)	0.060	2.37(1.08–5.21)	0.031
Separated	2.32 (0.65–8.19)	0.190	0.52(0.13–2.05)	0.355
Widow	1.62 (0.75–3.50)	0.220	2.42(1.05–5.58)	0.038
Occupation				
Unemployed®	1.00	-	1.00	-
Employed	1.25(0.80–1.95)	0.317	1.57(1.02–2.41)	0.060
SBP	1.25(1.20–1.94)	<0.001	1.01(0.80–1.04)	0.020
DBP	–	–	0.98(0.97–1.00)	0.001
DC	0.44(0.27–0.74)	0.286	0.34(0.19–0.62)	<0.001
TC	0.96(0.81-.14)	0.659		
PA				
Yes®	1.00	-	1.00	-
No	1.33(0.89–1.97)	0.160	2.07(1.36–3.16)	0.001
Alcohol				
No®	1.00	-	-	-
Yes	1.07(0.61–1.87)	0.805		
Smoking				
No®	1.00	-	1.00	-
Yes	1.87(1.18–2.97)	0.007	1.85(1.20–2.86)	0.005
BMI	1.07(1.04–1.11)	<0.001	1.26(1.04–1.60)	<0.001

1.25, CI: 1.20–1.94; p < 0.001), BMI (OR: 1.07, CI: 1.04–1.11; p < 0.001) were the high-risk factors for developing T2D. In 2011–2012, the participants who came from Mexican (OR: 2.13, CI: 1.03–4.42; p = 0.04), black (OR: 2.56, CI: 1.44–4.55, p = 0.001), other's (OR: 2.21, CI: 1.06–4.61; p = 0.034) race had a higher risk of T2D than white participants. Divorce respondents were less likely to have T2D than married (OR: 1.82, CI: 1.04–3.18; p = 0.034), unmarried (OR: 2.37, CI: 1.08–5.27; p = 0.031), and widow (OR: 2.42, CI: 1.05–5.58; p = 0.038). Participants had a greater chance of becoming T2D compared to their counterparts if they were physically inactive (OR: 2.07, CI: 1.36–3.16; p = 0.001) or had a smoking habit (OR: 1.85, CI: 1.20–2.86; p = 0.005). The other factors for T2D included age (OR: 1.04, CI: 1.02–1.06; p < 001), SBP (OR: 1.01, CI: 0.80–1.04; p = 0.020), DBP (OR: 0.98, CI: 0.97–1.00; p = 0.001), DC (OR: 0.34, CI: 0.19–0.62; <0.001), BMI (OR: 1.26, CI: 1.04–1.60; p < 0.001) were the risk factors of T2D. These six significant risk factors (age, education, marital status, SBP, smoking, and BMI) and nine factors (age, race, marital status, SBP, DBP, direct cholesterol, physical activity, smoking, and BMI) for T2D in both 2009–2010 and 2011–2012 datasets. These significant risk factors were fed into the ML-based predictive models for the classification of diabetic or control subjects.

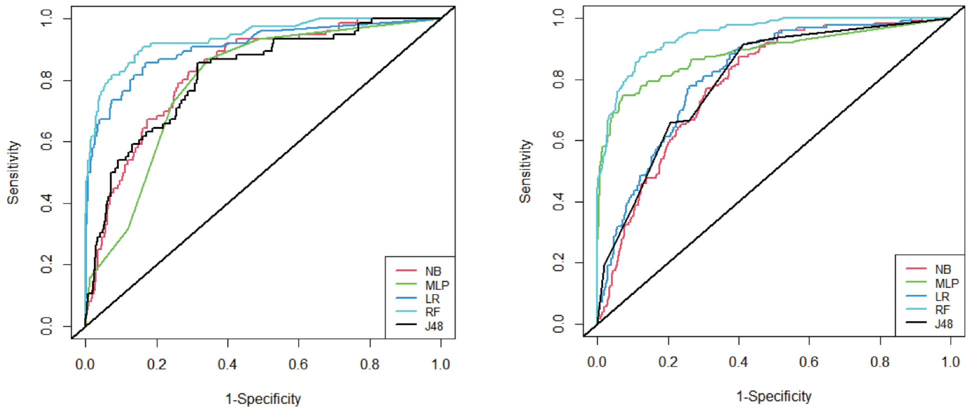
3.3. Performance comparison of five ML-based predictive models

The objective of this section was to compare the performance of five ML-based predictive models. The performance comparison results of ML-based predictive models between 2009–2010 and 2011–2012 datasets have been presented in Table 4. The classification accuracy and AUC of five ML-based predictive models were between 89.8%-95.9% and 0.778–0.946. The RF-based predictive model achieved the highest ACC of 94.8% with an AUC of 0.917 in 2009–2010, whereas NB gave 89.8% ACC and J48 gave 0.812 AUC. Among five predictive models, RF-based model provided the highest scores of other performance metrics as SE of 94.3% and FM of 94.1% for 2009–2010. Alternatively, the RF-based predictive model also attained the highest ACC of 95.9% and AUC of 0.946 for the 2011–2012 dataset. The RF-based model also achieved the highest SE of 95.7% and FM of 95.3% for 2011–2012. The ROC curve of five ML-based predictive models is presented in Figure 2. Figure 2 also confirmed that the RF-based model accomplished the highest AUC. Therefore, the RF-based model was the best predictive model for the prediction of T2D patients for both datasets.

Table 4. Performance scores of five ML-based classifiers for T2D.

	2009–2010				2011–2012			
Classifiers	ACC (%)	SE (%)	FM (%)	AUC	ACC (%)	SE (%)	FM (%)	AUC
LR	92.1	92.1	89.0	0.852	92.5	89.8	89.4	0.778
NB	89.8	89.8	89.5	0.848	90.2	90.0	90.1	0.839
J48	92.4	92.1	90.8	0.812	93.7	92.6	92.8	0.864
MLP	92.2	89.5	89.7	0.836	92.9	91.3	91.4	0.811
RF	**94.8**	**94.3**	**94.1**	**0.917**	**95.9**	**95.7**	**95.3**	**0.946**

Bold values indicate the proposed method results



(a) 2009-2010

(b) 2011-2012

Figure 2. ROC curves of five ML classifiers for T2D in: (a) 2009–2010; (b) 2011–2012.

3.4. Validation of the proposed system

We have considered the Bangladesh Demographic Health Survey (BDHS) diabetes-related dataset to validate our proposed system/combination (National Institute of Population Research and Training (NIPORT) and ICF, 2020). The dataset consisted of 6965 patients (Diabetic: 1047 and Control: 5918). The validation results of our proposed system are presented in Table 5. The results revealed that our proposed system was enabled to correctly classify T2D and obtained an ACC of 84.9%, SE of 84.8%, FM of 78.9%, and AUC of 0.677. Therefore, our proposed system is validated for both the NHANES and BDHS diabetes datasets.

4. Discussion

Based on the NHANES of T2D, 2009–2010 and 2011–2012 datasets, the goal of the current study was to identify the significant risk factors of T2D and predict patients as diabetic and control using ML techniques based on the significant risk factors and eventually propose a suitable combination to predict T2D patients. Risk factors of T2D were identified by MLR using p-values ($p < 0.05$). The current study showed that there were 6 and 9 significant risk factors of T2D in 200–2010 and 2011–2012, respectively (see Table 3). The age of the respondents was a highly significant risk factor for T2D for both datasets 2009–2010 and 2011–2012. The results have coincided with the previous settings (Park et al., 2003; Rahman et al., 2020; Xia et al., 2021; Zhang et al., 2017). Our findings also showed that Mexican, black, and other race respondents were more at risk of T2D than the white race in 2011–2012, which is corroborated with the previous literature (Park et al., 2003; Xie et al., 2019; Zhang et al., 2017). In 2009–2010, respondents who had completed 8 grade and high school had a higher risk of T2D, which contradicted the previous studies (Hu et al., 2017; Tripathy et al., 2017; Zafar et al., 2016), whereas college educated respondents had a lower risk of T2D. This is because educated respondents were more conscious of their lifestyles and food habits.

Table 5. Validation of the proposed model using BDHS diabetes dataset.

Classifier types	ACC (%)	SE (%)	FM (%)	AUC
LR	81.7	81.8	78.0	0.603
NB	84.6	84.7	78.9	0.674
J48	80.7	80.7	78.0	0.576
MLP	84.8	84.8	78.3	0.652
RF	**84.9**	**84.8**	**78.9**	**0.677**

Bold values indicate the proposed method results

Married and widow respondents had a higher risk of T2D, which was consistent with the findings of the previous study (Ramezankhani et al., 2019). In contrast to previous studies, the current study found that physical activity was also a high-risk factor for T2D (Hu et al., 2017; Tripathy et al., 2017; Zafar et al., 2016). PA improves glycaemic and reduces the risk of CVD and mortality in T2D patients (Hamasaki, 2016). T2D can be managed if people engage in regular physical activity. It was also noted that SBP, DBP, DC, PA, smoking, and BMI (Hu et al., 2017; Tripathy et al., 2017; Wei et al., 2021; Xie et al., 2019; Zafar et al., 2016) were also the risk factors of T2D. Furthermore, for both the 2009–210 and 2011–2012 datasets, five ML-based techniques (LR, NB, J48, MLP, and RF) were implemented to predict T2D. The current study was also compared to previous studies in the literature, which are described in the section that follows.

4.1. Distinction between the current study against similar existing studies

The objective of this section was to compare the performance of our study against previous studies in literature. We reviewed existing studies that were mainly conducted on diabetes and implemented ML-based algorithms for the prediction of T2D. The summarisation of these existing works is presented in Table 6. We mainly highlighted the performance of their proposed method and marked in bold which were shown in column 3 of Table 6. Moreover, the classification accuracy (presented in %) and AUC were presented in column 4 of Table 6. For example, Nai-arun and Moungmai (2015) employed five well-known ML-based techniques: decision tree (DT), artificial neural network (ANN), LR, NB, and RF for the prediction of diabetic patients, and their findings showed that RF-based model could more accurately classify diabetic patients and obtained the highest ACC of 85.6%. Zheng et al. (2017) utilised six ML-based models: k-nearest neighbourhood (k-NN), NB, DT, RF, support vector machine (SVM), and LR for accurate risk prediction of T2D, and they mentioned that these three models performed with better accuracy compared to others. Maniruzzaman et al. (2017) proposed a Gaussian process classification (GPC) model for the prediction of diabetes and its performance was compared with three techniques: linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), NB, and GPC, obtaining an ACC of 82.0%. The ACC of ML-based models will be improved by selecting the important features or factors. For example, Semerdjian and Frank (2017) extracted the important risk factors for diabetes using an RF-based model and then utilised five techniques (LR, KNN, RF, SVM, and gradient boosting (GB)) for prediction of diabetes, and GB achieved the higher AUC of 0.84.

Maniruzzaman et al. (2018) proposed an ML-based system by the combination of six risk factor identification methods (RF, LR, MI, PCA, ANOVA, and FDR) and 10 ML techniques (LDA, QDA, NB, ANN, GPC, SVM, Adaboost (AB), LR, DT, and RF) for prediction of diabetes. Zou et al. (2018) also proposed an ML system by the combination of two feature selection methods (PCA and minimum redundancy of maximum relevance (mRMR)) and three models (DT, ANN, and RF) for the prediction of diabetes patients. Both studies suggested that the highest accuracy was obtained by RF (Maniruzzaman et al., 2018; Zou et al., 2018). Dagliati et al. (2018) imputed the missing values by RF and proposed an LR-based model for the prediction of diabetes complications (retinopathy, neuropathy, or nephropathy) and their proposed model provided the highest accuracy of 83.8%. Mohapatra et al. (2019) and Xiong et al. (2019) adopted MLP for the prediction of diabetes and obtained the highest classification accuracy. Xie et al. (2019) also introduced LR for the identification of risk factors of diabetes and employed four (LR, SVM, DT, and NN) techniques to classify diabetic patients. NN obtained ACC of 82.4%. Islam et al. (2020) used six ML-based techniques, including SVM, RF, LDA, LR, KNN, and bagged classification and regression tree (Bagged-

Table 6. Distinction between the current study against previous studies.

Authors	Sample size	Classifier types	Performance
Nai-arun and Moungmai (2015)	30,122	ANN, DT, LR, **RF**, NB,	ACC: 85.6%
Zheng et al. (2017)	123,241	KNN, NB, **DT, RF, SVM**, LR	AUC: 0.98
Semerdjian et aand Frank (2017)	5515	LR, KNN, RF GB,SVM	AUC: 0.84
Maniruzzaman et al. (2017)	768	LDA, QDA, NB, **GPC**	ACC: 82.0%
Maniruzzaman et al. (2018)	768	LDA, QDA, NB, ANN, GPC, SVM, AB, LR, DT, **RF**	ACC: 92.3%
Zou et al. (2018)	220,680	DT, **RF**, NN	ACC: 80.8%
Dagliati et al. (2018)	973	LR	ACC: 83.8%
Mohapatra et al. (2019)	768	MLP	ACC: 77.5%
Sneha and Gangil (2019)	768	SVM, RF, **NB**, DT, and KNN	ACC: 82.3%
Xie et al. (2019)	138,146	LR, SVM, DT, **NN**	ACC: 82.4%
Xiong et al. (2019)	11,845	**MLP**, AD, RF, SVM,GTB	ACC: 87.0%
Islam et al. (2020)	1569	SVM, RF, LDA, LR, KNN, **CART**	ACC: 94.3%
Kaur and Kumari (2020)	768	SVM (linear+RBF), KNN,ANN, MDR	ACC: 89.0%
Adua et al. (2021)	438	**NB**, KNN, SVM, and DT	ACC: 93.0%
Current study	4922 & 4936	LR, NB, J48, MLP, **RF**	ACC: 95.6%

Bold values indicate the proposed method results.

CRT) for predicting the risk of T2D. Among six techniques, Bagged-CART attained the highest performance score with ACC of 94.3%. Kaur and Kumari (2020) implemented five predictive models as SVM with linear and RBF kernels, k-NN, ANN, and multifactor dimensionality reduction (MDR) to classify diabetes patients. SVM-linear performed with a higher ACC of 89.0%. Sneha and Gangil (2019) and Adua et al. (2021) implemented NB-based model for accurate risk predictions of diabetes disease and attained the highest classification ACC of 93.0%.

Our study illustrated that the RF-based predictive model achieved the highest performance score with ACC of 95.6% compared to previous studies (see, Table 6). Finally, it may be concluded that RF-based is the best predictive model for classification and prediction of T2D diseases. The RF-based predictive model performs better. Some of the key reasons are as follows: (i) it is appropriate for both nonlinear and non-normal data; (ii) it avoids over-fitting of the data and provides robustness to noise; (iii) it is adaptable to both categorical and continuous data; and (iv) it fits well for data imputation and cluster analysis.

4.2. Limitation and extension of the study

The dataset used in this study was cross-sectional, with only 14 attributes. Despite the fact that the current study's goal was to identify risk factors of T2D using only MLR. Different feature extraction techniques could be used as PCA, information gain (IG), relief, and so on. This study presented only ML-based techniques for predicting T2D. This study can be extended to any deeper models like deep learning, and multiobjective reinforcement learning (MORL) for the prediction of T2D and compared their performance, especially with this current study.

5. Conclusion

Diabetes is a leading cause of chronic and noncommunicable diseases worldwide, with its prevalence increasing over time. In this paper, an attempt has been made to build an ML-based system using a combination of MLR and five different ML-based predictive models for the prediction of T2D. MLR results show that age, education, marital status, SBP, smoking, and BMI were the prominent risk factors in T2D for 2009–2010, whereas age, race, marital status, SBP, DBP, DC, PA, smoking, and BMI in 2011–2012. Our experimental results exhibit that among the five predictive models; the RF-based model provides excellent performance scores for the prediction of T2D. Based on our findings, we can easily create a web-based tool that will assist physicians in making an initial decision about identifying T2D patients and controlling the diabetes disease at an early stage, ultimately reducing the burden on the health system.

Acknowledge

The authors acknowledge the reviewers and editor for their valuable suggestions and comments, which helps us to improve the manuscript.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The authors do not receive any funds for conducting this research.

ORCID

Md. Maniruzzaman http://orcid.org/0000-0001-6151-8071

References

Abdissa, D., Dukessa, A., & Babusha, A. (2021). Prevalence and associated factors of overweight/obesity among type 2 diabetic outpatients in Southwest Ethiopia. *Heliyon*, *7*(2), e06339. https://doi.org/10.1016/j.heliyon.2021.e06339

Adhao, R., & Pachghare, V. (2020). Feature selection using principal component analysis and genetic algorithm. *Journal of Discrete Mathematical Sciences and Cryptography*, *23*(2), 595–602. https://doi.org/10.1080/09720529.2020.1729507

Adua, E., Kolog, E. A., Afrifa-Yamoah, E., Amankwah, B., Obirikorang, C., Anto, E. O., , and Tetteh, A. Y. (2021 ()). *Predictive Model and Feature Importance for Early Detection of Type II Diabetes Mellitus*, 6(1), 1–15. https://doi.org/10.1186/s41231-021-00096-z

Aikens, R. C., Zhao, W., Saleheen, D., Reilly, M. P., Epstein, S. E., Tikkanen, E., Salomaa, V., & Voight, B. F. (2017). Systolic blood pressure and risk of type 2 diabetes: A Mendelian randomization study. *Diabetes*, *66*(2), 543–550. https://doi.org/10.2337/db16-0868

Akalu, Y., & Belsti, Y. (2020). Hypertension and its associated factors among type 2 diabetes mellitus patients at Debre Tabor general hospital, northwest Ethiopia. *Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy*, *13*, 1621. https://doi.org/10.2147/DMSO.S254537

Alanazi, H. O., Abdullah, A. H., Qureshi, K. N., & Ismail, A. S. (2018). Accurate and dynamic predictive model for better prediction in medicine and healthcare. *Irish Journal of Medical Science*, *187*(2), 501–513. https://doi.org/10.1007/s11845-017-1655-3

Almeida, V. D. C. F. D., Zanetti, M. L., Almeida, P. C. D., & Damasceno, M. M. C. (2011). Occupation and risk factors for type 2 diabetes: A study with health workers. *Revista Latino-Americana de Enfermagem*, *19*(3), 476–484. https://doi.org/10.1590/S0104-11692011000300005

Avorn, J. (2013). The promise of pharmacoepidemiology in helping clinicians assess drug risk. *Circulation*, *128*(7), 745–748. https://doi.org/10.1161/CIRCULATIONAHA.113.003419

Bahour, N., Cortez, B., Pan, H., Shah, H., Doria, A., & Aguayo-Mazzucato, C. (2022). Diabetes mellitus correlates with increased biological age as indicated by clinical biomarkers. *GeroScience*, *44*(1), 415–427. https://doi.org/10.1007/s11357-021-00469-0

Bays, H. E., Chapman, R. H., & Grandy, S., & SHIELD Investigators' Group. (2007). The relationship of body mass index to diabetes mellitus, hypertension and dyslipidaemia: Comparison of data from two national surveys. *International Journal of Clinical Practice*, *61*(5), 737–747. https://doi.org/10.1111/j.1742-1241.2007.01336.x

Bhowmik, B., Siddiquee, T., Mujumder, A., Afsana, F., Ahmed, T., Mdala, I. A., Do V. Moreira, N., Khan, A., Hussain, A., Holmboe-Ottesen, G., & Omsland, T. K. (2018). Serum lipid profile and its association with diabetes and prediabetes in a rural Bangladeshi population. *International Journal of Environmental Research and Public Health*, *15*(9), 1944. https://doi.org/10.3390/ijerph15091944

Bron, M., Guerin, A., Latremouille-Viau, D., Ionescu-Ittu, R., Viswanathan, P., Lopez, C., & Wu, E. Q. (2014). Distribution and drivers of costs in type 2 diabetes mellitus treated with oral hypoglycemic agents: A retrospective claims data analysis. *Journal of Medical Economics*, *17*(9), 646–657. https://doi.org/10.3111/13696998.2014.925905

Canadian Diabetes Association. (2013). *Diabetes: Canada at the tipping point 2011*.

Carlsson, S., Andersson, T., Talbäck, M., & Feychting, M. (2020). Incidence and prevalence of type 2 diabetes by occupation: Results from all Swedish employees. *Diabetologia*, *63*(1), 95–103. https://doi.org/10.1007/s00125-019-04997-5

Centers for Disease Control and Prevention. (2014). *National diabetes statistics report, 2014*.

Chang, S. A. (2012). Smoking and type 2 diabetes mellitus. *Diabetes & Metabolism Journal*, *36*(6), 399–403. https://doi.org/10.4093/dmj.2012.36.6.399

Cheng, Y. J., Kanaya, A. M., Araneta, M. R. G., Saydah, S. H., Kahn, H. S., Gregg, E. W., Fujimoto, W. Y., & Imperatore, G. (2019). Prevalence of diabetes by race and ethnicity in the United States, 2011-2016. *Jama*, *322*(24), 2389–2398. https://doi.org/10.1001/jama.2019.19365

Chen, L., Pei, J. H., Kuang, J., Chen, H. M., Chen, Z., Li, Z. W., & Yang, H. Z. (2015). Effect of lifestyle intervention in patients with type 2 diabetes: A meta-analysis. *Metabolism*, *64*(2), 338–347. https://doi.org/10.1016/j.metabol.2014.10.018

Cho, N., Shaw, J. E., Karuranga, S., Huang, Y., da Rocha Fernandes, J. D., Ohlrogge, A. W., & Malanda, B. (2018). IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. *Diabetes Research and Clinical Practice*, *138*, 271–281. https://doi.org/10.1016/j.diabres.2018.02.023

Dagliati, A., Marini, S., Sacchi, L., Cogni, G., Teliti, M., Tibollo, V., Bellazzi, R., Chiovato, L., & Bellazzi, R. (2018). Machine learning methods to predict diabetes complications. *Journal of Diabetes Science and Technology*, *12*(2), 295–302. https://doi.org/10.1177/1932296817706375

Das, S. K. (2014). Integrating transcriptome and epigenome: Putting together the pieces of the type 2 diabetes pathogenesis puzzle. *Diabetes*, *63*(9), 2901–2903. https://doi.org/10.2337/db14-0757

Derakhshan, A., Sardarinia, M., Khalili, D., Momenan, A. A., Azizi, F., Hadaegh, F., & Taheri, S. (2014). Sex specific incidence rates of type 2 diabetes and its risk factors over 9 years of follow-up: Tehran Lipid and Glucose Study. *PloS one*, *9*(7), e102563. https://doi.org/10.1371/journal.pone.0102563

Dong, X., Bahroos, N., Sadhu, E., Jackson, T., Chukhman, M., Johnson, R., and Hynes, D. (2013). Leverage Hadoop framework for large scale clinical informatics applications. AMIA Joint Summits on Translational Science proceedings. *AMIA Joint Summits on Translational Science, 2013*, 53. https://pubmed.ncbi.nlm.nih.gov/24303235/

Elssied, N. O. F., Ibrahim, O., & Osman, A. H. (2014). A novel feature selection based on one-way anova f-test for e-mail spam classification. *Research Journal of Applied Sciences, Engineering and Technology*, *7*(3), 625–638. https://doi.org/10.19026/rjaset.7.299

Emdin, C. A., Anderson, S. G., Woodward, M., & Rahimi, K. (2015). Usual blood pressure and risk of new-onset diabetes: Evidence from 4.1 million adults and a meta-analysis of prospective studies. *Journal of the American College of Cardiology*, *66*(14), 1552–1562. https://doi.org/10.1016/j.jacc.2015.07.059

Flatz, A., Casillas, A., Stringhini, S., Zuercher, E., Burnand, B., & Peytremann-Bridevaux, I. (2015). Association between education and quality of diabetes care in Switzerland. *International Journal of General Medicine*, *8*, 87. https://doi.org/10.2147/IJGM.S77139

Ghaderpanahi, M., Fakhrzadeh, H., Sharifi, F., Badamchizade, Z., Mirarefin, M., Ebrahim, R. P., Ghotbi, S., Nouri, M., & Larijani, B. (2011). Association of physical activity with risk of type 2 diabetes. *Iranian Journal of Public Health*, *40*(1), 86–93. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3481723/

Ghosh, S. (2017). Implementation of Proper Nutrition in Managing Diabetes-A Review Article. *J Food Sci Res*, *2*(1), 103. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3481723/

Gimeno-Orna, J. A., Faure-Nogueras, E., & Sancho-Serrano, M. A. (2005). Usefulness of total cholesterol/HDL-cholesterol ratio in the management of diabetic dyslipidaemia. *Diabetic Medicine*, *22*(1), 26–31. https://doi.org/10.1111/j.1464-5491.2004.01341.x

Gregorutti, B., Michel, B., & Saint-Pierre, P. (2017). Correlation and variable importance in random forests. *Statistics and Computing*, *27*(3), 659–678. https://doi.org/10.1007/s11222-016-9646-1

Habibi, S., Ahmadi, M., & Alizadeh, S. (2015). Type 2 diabetes mellitus screening and risk factors using decision tree: Results of data mining. *Global Journal of Health Science*, *7*(5), 304. https://doi.org/10.5539/gjhs.v7n5p304

Hamasaki, H. (2016). Daily physical activity and type 2 diabetes: A review. *World Journal of Diabetes*, *7*(12), 243. https://doi.org/10.4239/wjd.v7.i12.243

Harreiter, J., & Kautzky-Willer, A. (2018). Sex and gender differences in prevention of type 2 diabetes. *Frontiers in Endocrinology*, *9*, 220. https://doi.org/10.3389/fendo.2018.00220

Holst, C., Becker, U., Jørgensen, M. E., Grønbæk, M., & Tolstrup, J. S. (2017). Alcohol drinking patterns and risk of diabetes: A cohort study of 70,551 men and women from the general Danish population. *Diabetologia*, *60*(10), 1941–1950. https://doi.org/10.1007/s00125-017-4359-3

Howard, B. V., Robbins, D. C., Sievers, M. L., Lee, E. T., Rhoades, D., Devereux, R. B., Howard, B. V., Cowan, L. D., Gray, R. S., Welty, T. K., Go, O. T., &

Howard, W. J. (2000). LDL cholesterol as a strong predictor of coronary heart disease in diabetic individuals with insulin resistance and low LDL: The Strong Heart Study. *Arteriosclerosis, Thrombosis, and Vascular Biology*, *20*(3), 830–835. https://doi.org/10.1161/01.ATV.20.3.830

Huebschmann, A. G., Huxley, R. R., Kohrt, W. M., Zeitler, P., Regensteiner, J. G., & Reusch, J. E. (2019). Sex differences in the burden of type 2 diabetes and cardiovascular risk across the life course. *Diabetologia*, *62*(10), 1761–1772. https://doi.org/10.1007/s00125-019-4939-5

Hu, M., Wan, Y., Yu, L., Yuan, J., Ma, Y., Hou, B., Shang, L., Otani, H., Pazour, G. J., Hsu, V. W., Minami, Y., & Takumi, T. (2017). Prevalence, awareness and associated risk factors of diabetes among adults in Xi'an, China. *Scientific Reports*, *7*(1), 1–9. https://doi.org/10.1038/s41598-016-0028-x

Islam, M. M., Rahman, M. J., Roy, D. C., & Maniruzzaman, M. (2020). Automated detection and classification of diabetes disease based on Bangladesh demographic and health survey data, 2011 using machine learning approach. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, *14*(3), 217–219. https://doi.org/10.1016/j.dsx.2020.03.004

Kao, W. L. (2001). Alcohol consumption and the risk of type 2 diabetes mellitus: Atherosclerosis risk in communities study. *American Journal of Epidemiology*, *154*(8), 748–757. https://doi.org/10.1093/aje/154.8.748

Kaur, H., & Kumari, V. (2020). Predictive modelling and analytics for diabetes using a machine learning approach. *Applied Computing and Informatics*, 18(1/2), 90–100. https://doi.org/10.1016/j.aci.2018.12.004

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. *Computational and Structural Biotechnology Journal*, *15*, 104–116. https://doi.org/10.1016/j.csbj.2016.12.005

Kposowa, A. J., Ezzat, D. A., & Breault, K. (2021). Diabetes mellitus and marital status: Evidence from the National Longitudinal Mortality Study on the effect of marital dissolution and the death of a spouse. *International Journal of General Medicine*, *14*, 1881. https://doi.org/10.2147/IJGM.S307436

Kumar, N. K., Vigneswari, D., Krishna, M. V., & Reddy, G. P. (2019). An optimized random forest classifier for diabetes mellitus Kacprzyk, Janusz. In *Emerging technologies in data mining and information security* 1st. Vol. Advances in Intelligent Systems and Computing 813. Vol.813. (pp. 765–773. Springer. ISSN:9811314985. doi:10.1007/978-981-13-1498-8_67.

Lee, S. W., Kim, H. C., Lee, J. M., Yun, Y. M., Lee, J. Y., & Suh, I. (2017). Association between changes in systolic blood pressure and incident diabetes in a community-based cohort study in Korea. *Hypertension Research*, *40*(7), 710–716. https://doi.org/10.1038/hr.2017.21

Lee, H. S., Lee, S. S., Hwang, I. Y., Park, Y. J., Yoon, S. H., Han, K., Park, Y.-J., Son, J.-W., Ko, S.-H., Park, Y. G., Yim, H. W., Lee, W.-C., & Park, Y. M. (2013). Prevalence, awareness, treatment and control of hypertension in adults with diagnosed diabetes: The Fourth Korea National Health and Nutrition Examination Survey (KNHANES IV). *Journal of Human Hypertension*, *27*(6), 381–387. https://doi.org/10.1038/jhh.2012.56

Link, C. L., & McKinlay, J. B. (2009). Disparities in the prevalence of diabetes: Is it race/ethnicity or socioeconomic status? Results from the Boston Area Community Health (BACH) survey. *Ethnicity & Disease*, *19*(3), 288–292. doi:PMID: 19769011; PMCID: PMC3706078.

Li, X. H., Yu, F. F., Zhou, Y. H., & He, J. (2016). Association between alcohol consumption and the risk of incident type 2 diabetes: A systematic review and dose-response meta-analysis. *The American Journal of Clinical Nutrition*, *103*(3), 818–829. https://doi.org/10.3945/ajcn.115.114389

Maddatu, J., Anderson-Baucum, E., & Evans-Molina, C. (2017). Smoking and the risk of type 2 diabetes. *Translational Research*, *184*, 101–107. https://doi.org/10.1016/j.trsl.2017.02.004

Maggio, C. A., & Pi-Sunyer, F. X. (2003). Obesity and type 2 diabetes. *Endocrinology and Metabolism Clinics*, *32*(4), 805–822. https://doi.org/10.1016/S0889-8529(03)00071-9

Maguire, J., & Dhar, V. (2013). Comparative effectiveness for oral anti-diabetic treatments among newly diagnosed type 2 diabetes: Data-driven predictive analytics in healthcare. *Health Systems*, *2*(2), 73–92. https://doi.org/10.1057/hs.2012.20

Ma, R. C., Lin, X., & Jia, W. (2014). Causes of type 2 diabetes in China. *Lancet Diabetes Endocrinol*, *2*(12), 980–991. https://doi.org/10.1016/S2213-8587(14)70145-7

Maniruzzaman, M., Kumar, N., Abedin, M. M., Islam, M. S., Suri, H. S., El-Baz, A. S., & Suri, J. S. (2017). Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm. *Computer Methods and Programs in Biomedicine*, *152*, 23–34. https://doi.org/10.1016/j.cmpb.2017.09.004

Maniruzzaman, M., Rahman, M., Ahammed, B., & Abedin, M. (2020). Classification and prediction of diabetes disease using machine learning paradigm. *Health Information Science and Systems*, *8*(1), 1–14. https://doi.org/10.1007/s13755-019-0095-z

Maniruzzaman, M., Rahman, M. J., Al-MehediHasan, M., Suri, H. S., Abedin, M. M., El-Baz, A., & Suri, J. S. (2018). Accurate diabetes risk stratification using machine learning: Role of missing value and outliers. *Journal of Medical Systems*, *42*(5), 1–17. https://doi.org/10.1007/s10916-018-0940-7

Mohapatra, S. K., Swain, J. K., & Mohanty, M. N. (2019). Detection of diabetes using multilayer perceptron. In *International conference on intelligent computing and applications* (pp. 109–116). Springer, Singapore.

Nai-arun, N., & Moungmai, R. (2015). Comparison of classifiers for the risk of diabetes prediction. *Procedia Computer Science*, *69*, 132–142. https://doi.org/10.1016/j.procs.2015.10.014

Nakazawa, S., Fukai, K., Furuya, Y., Kojimahara, N., Hoshi, K., Toyota, A., & Tatemichi, M. (2022). Occupations associated with diabetes complications: A nationwide-multicenter hospital-based case-control study. *Diabetes Research and Clinical Practice*, *186*, 109809. https://doi.org/10.1016/j.diabres.2022.109809

Nanayakkara, N., Curtis, A. J., Heritier, S., Gadowski, A. M., Pavkov, M. E., Kenealy, T., Owens, D. R., Thomas, R. L., Song, S., Wong, J., Chan, J. C. N., Luk, A. O. Y., Penno, G., Ji, L., Mohan, V., Amutha, A., Romero-Aroca, P., Gasevic, D., Magliano, D. J., … Zoungas, S. (2021). Impact of age at type 2 diabetes mellitus diagnosis on mortality and vascular complications: Systematic review and meta-analyses. *Diabetologia*, *64*(2), 275–287. https://doi.org/10.1007/s00125-020-05319-w

National Health and Nutrition Examination Survey (NHANES). 20092010.

National Health and Nutrition Examination Survey (NHANES). 20112012.

National Institute of Population Research and Training (NIPORT), and ICF. (2020). *Bangladesh Demographic and health survey 2017-18*. NIPORT and ICF.

Oliveira, C. M., Viater Tureck, L., Alvares, D., Liu, C., Horimoto, A. R. V. R., Balcells, M., Pereira, A. C., Pereira, A. C., & de Oliveira Alvim, R. (2020). Relationship between marital status and incidence of type 2 diabetes mellitus in a Brazilian rural population: The Baependi heart study. *PLoS one*, *15*(8), e0236869. https://doi.org/10.1371/journal.pone.0236869

Park, Y. W., Zhu, S., Palaniappan, L., Heshka, S., Carnethon, M. R., & Heymsfield, S. B. (2003). The metabolic syndrome: Prevalence and associated risk factor findings in the US population from the Third National Health and Nutrition Examination Survey, 1988-1994. *Archives of Internal Medicine*, *163*(4), 427–436. https://doi.org/10.1001/archinte.163.4.427

Rahman, M. N., Alam, S. S., Zobayed, A., Hasan, M. M., Nisha, S., Hossian, M., andIslam, K. (2020). Prevalence of diabetes mellitus and associated risk factors among adult individuals in selected areas of Bangladesh. *American Journal of Public Health*, *8*(6), 209–213. http://pubs.sciepub.com/ajphr/8/6/5

Ramezankhani, A., Azizi, F., Hadaegh, F., & Shimosawa, T. (2019). Associations of marital status with diabetes, hypertension, cardiovascular disease and all-cause mortality: A long term follow-up study. *PloS one*, *14*(4), e0215593. https://doi.org/10.1371/journal.pone.0215593

Rawshani, A., Sattar, N., Franzén, S., Rawshani, A., Hattersley, A. T., Svensson, A. M., Gudbjörnsdottir, S., & Gudbjörnsdottir, S. (2018). Excess mortality and cardiovascular disease in young adults with type 1 diabetes in relation to age at onset: A nationwide, register-based cohort study. *The Lancet*, *392*(10146), 477–486. https://doi.org/10.1016/S0140-6736(18)31506-X

Saedi, E., Gheini, M. R., Faiz, F., & Arami, M. A. (2016). Diabetes mellitus and cognitive impairments. *World Journal of Diabetes*, *7*(17), 412–422. https://doi.org/10.4239/wjd.v7.i17.412

Saydah, S., Bullard, K. M., Cheng, Y., Ali, M. K., Gregg, E. W., Geiss, L., & Imperatore, G. (2014). Trends in cardiovascular disease risk factors by obesity level in adults in the United States, NHANES 1999-2010. *Obesity*, *22*(8), 1888–1895. https://doi.org/10.1002/oby.20761

Semerdjian, J., & Frank, S. (2017). An ensemble classifier for predicting the onset of type II diabetes. *arXiv preprint arXiv:1708.07480*. doi:10.48550/arXiv.1708.07480.

Shamshirgaran, S. M., Mamaghanian, A., Aliasgarzadeh, A., Aiminisani, N., Iranparvar-Alamdari, M., & Ataie, J. (2017). Age differences in diabetes-related complications and glycemic control. *BMC Endocrine Disorders*, *17*(1), 1–7. https://doi.org/10.1186/s12902-017-0175-5

Shankaracharya, D. O., Samanta, S., Vidyarthi, A. S., & Vidyarthi, A. S. (2012). Computational intelligence-based diagnosis tool for the detection of prediabetes and type 2 diabetes in India. *The Review of Diabetic Studies: RDS*, *9*(1), 55. https://doi.org/10.1900/RDS.2012.9.55

Shrivastava, V. K., Londhe, N. D., Sonawane, R. S., & Suri, J. S. (2017). A novel and robust Bayesian approach for segmentation of psoriasis lesions and its risk stratification. *Computer Methods and Programs in Biomedicine*, *150*, 9–22. https://doi.org/10.1016/j.cmpb.2017.07.011

Sil, K., Das, B. K., Pal, S., & Mandal, L. (2020). A study on impact of education on diabetes control and complications. *Njmr*, *10*(1), 26–29. https://njmr.in/index.php/file/article/view/48

Sisodia, D., & Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. *Procedia Computer Science*, *132*, 1578–1585. https://doi.org/10.1016/j.procs.2018.05.122

Sneha, N., & Gangil, T. (2019). Analysis of diabetes mellitus for early prediction using optimal features selection. *Journal of Big Data*, *6*(1), 1–19. https://doi.org/10.1186/s40537-019-0175-6

Spanakis, E. K., & Golden, S. H. (2013). Race/ethnic difference in diabetes and diabetic complications. *Current Diabetes Reports*, *13*(6), 814–823. https://doi.org/10.1007/s11892-013-0421-9

Steele, C. J., Schöttker, B., Marshall, A. H., Kouvonen, A., O'Doherty, M. G., Mons, U., Saum, K.-U., Boffetta, P., Trichopoulou, A., Brenner, H., & Kee, F. (2017). Education achievement and type 2 diabetes—what mediates the relationship in older adults? Data from the ESTHER study: A population-based cohort study. *BMJ open*, *7*(4), e013569. https://doi.org/10.1136/bmjopen-2016-013569

Suastika, K., Dwipayana, P., Semadi, M. S., & Kuswardhani, R. T. (2012). Age is an important risk factor for type 2 diabetes mellitus and cardiovascular diseases. *Glucose Tolerance*, 67–80. http://dx.doi.org/10.5772/52397

Tan, S. Y., Wong, J. L. M., Sim, Y. J., Wong, S. S., Elhassan, S. A. M., Tan, S. H., Ling Lim, G. P., Rong Tay, N. W., Annan, N. C., Bhattamisra, S. K., & Candasamy, M. (2019). Type 1 and 2 diabetes mellitus: A review on current treatment approach and gene therapy as potential intervention. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, *13*(1), 364–372. https://doi.org/10.1016/j.dsx.2018.10.008

Tripathy, J. P., Thakur, J. S., Jeet, G., Chawla, S., Jain, S., Pal, A., Prasad, R., & Saran, R. (2017). Prevalence and risk factors of diabetes in a large community-based study in North India: Results from a STEPS survey in Punjab, India. *Diabetology & Metabolic Syndrome*, *9*(1), 1–8. https://doi.org/10.1186/s13098-017-0207-3

Tsao, H. Y., Chan, P. Y., & Su, E. C. Y. (2018). Predicting diabetic retinopathy and identifying interpretable biomedical features using machine learning algorithms. *BMC Bioinformatics*, *19*(9), 111–121. https://doi.org/10.1186/s12859-018-2277-0

Wang, T., Zhao, Z., Wang, G., Li, Q., Xu, Y., Li, M., Wang, T., Wang, G., Hu, R., Chen, G., Su, Q., Mu, Y., Tang, X., Yan, L., Qin, G., Wan, Q., Gao, Z., Yu, X., Shen, F., . . . Wang, W. (2021). Age-related disparities in diabetes risk attributable to modifiable risk factor profiles in Chinese adults: A nationwide, population-based, cohort study. *The Lancet Healthy Longevity*, *2*(10), e618–e628. https://doi.org/10.1016/S2666-7568(21)00177-X

Wei, Y., Liu, C., Lai, F., Dong, S., Chen, H., Chen, L., Shi, L., Zhu, F., Zhang, C., Lv, X., Peng, S., & Hao, G. (2021). Associations between serum total bilirubin, obesity and type 2 diabetes. *Diabetology & Metabolic Syndrome*, *13*(1), 1–7. https://doi.org/10.1186/s13098-021-00762-0

Woldesemayat, B., Amare, H., Ataro, Z., Gutema, G., Kidane, E., Belay, D., & Asrat, H. (2019). Prevalence of diabetes mellitus and associated risk factors among adults attending at feres meda health centre, Addis Ababa, Ethiopia, 2017. *J Biomed Mater Res*, *7*(1), 8–15. doi:10.11648/j.ijbmr.20190701.12.

Xia, M., Liu, K., Feng, J., Zheng, Z., & Xie, X. (2021). Prevalence and risk factors of type 2 diabetes and prediabetes among 53,288 middle-aged and elderly adults in China: A cross-sectional study. *Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy*, *14*, 1975. https://doi.org/10.2147/DMSO.S305919

Xie, Z., Nikolayeva, O., Luo, J., & Li, D. (2019). Peer reviewed: Building risk prediction models for type 2 diabetes using machine learning techniques. *Preventing Chronic Disease*, 16, E130–E139. doi:10.5888/2Fpcd16.190109.

Xiong, X. L., Zhang, R. X., Bi, Y., Zhou, W. H., Yu, Y., & Zhu, D. L. (2019). Machine learning models in type 2 diabetes risk prediction: Results from a cross-sectional retrospective study in Chinese adults. *Current Medical Science*, *39*(4), 582–588. https://doi.org/10.1007/s11596-019-2077-4

Yang, Y., Peng, N., Chen, G., Wan, Q., Yan, L., Wang, G., Qin, Y., Luo, Z., Tang, X., Huo, Y., Hu, R., Ye, Z., Qin, G., Gao, Z., Su, Q., Mu, Y., Zhao, J., Chen, L., Zeng, T., . . . Shi, L. (2022). Interaction between smoking and diabetes in relation to subsequent risk of cardiovascular events. *Cardiovascular Diabetology*, *21*(1), 1–12. https://doi.org/10.1186/s12933-022-01447-2

Zafar, J., Nadeem, D., Khan, S. A., Jawad Abbasi, M. M., Aziz, F., & Saeed, S. (2016). Prevalence of diabetes and its correlates in urban population of Pakistan: A Cross-sectional survey. *Journal of Pakistan Medical Association*, *66*(8), 922–927. https://jpma.org.pk/article-details/7850?article_id=7850

Zhang, N., Yang, X., Zhu, X., Zhao, B., Huang, T., & Ji, Q. (2017). Type 2 diabetes mellitus unawareness, prevalence, trends and risk factors: National Health and Nutrition Examination Survey (NHANES) 1999–2010. *Journal of International Medical Research*, *45*(2), 594–609. https://doi.org/10.1177/0300060517693178

Zhao, Q., Laukkanen, J. A., Li, Q., & Li, G. (2017). Body mass index is associated with type 2 diabetes mellitus in Chinese elderly. *Clinical Interventions in Aging*, *12*, 745. https://doi.org/10.2147/CIA.S130014

Zhao, F., Wu, W., Feng, X., Li, C., Han, D., Guo, X., & Lyu, J. (2020). Physical activity levels and diabetes prevalence in us adults: Findings from NHANES 2015–2016. *Diabetes Therapy*, *11*(6), 1303–1316. https://doi.org/10.1007/s13300-020-00817-x

Zheng, T., Xie, W., Xu, L., He, X., Zhang, Y., You, M., Yang, G., & Chen, Y. (2017). A machine learning-based framework to identify type 2 diabetes through electronic health records. *International Journal of Medical Informatics*, *97*, 120–127. https://doi.org/10.1016/j.ijmedinf.2016.09.014

Zhu, M., Li, J., Li, Z., Luo, W., Dai, D., Weaver, S. R., Fu, H., Luo, R., & Fu, H. (2015). Mortality rates and the causes of death related to diabetes mellitus in Shanghai Songjiang District: An 11-year retrospective analysis of death certificates. *BMC Endocrine Disorders*, *15*(1), 1–8. https://doi.org/10.1186/s12902-015-0042-1

Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting diabetes mellitus with machine learning techniques. *Frontiers in Genetics*, *9*, 515. https://doi.org/10.3389/fgene.2018.00515