Human-Centric Intelligent Systems (2022) 2:44–54
https://doi.org/10.1007/s44230-022-00003-1

RESEARCH ARTICLE

Identifying Heterogeneity of Diabetics Mellitus Based on the Demographical and Clinical Characteristics

Linta Islam[1] · Md Rafiqul Islam[2] · Shanjita Akter[3] · Md Zobaer Hasan[4] · Mohammad Ali Moni[5] · Mohammed Nasir Uddin[1]

Received: 13 July 2021 / Accepted: 25 October 2021 / Published online: 12 May 2022

Abstract

Background: Diabetes is a long-term disease, which is characterised by high blood sugar and has risen as a public health problem worldwide. It may prompt a variety of serious illnesses, including stroke, kidney failure, and heart attacks. In 2014, diabetes affected approximately 422 million people worldwide and it is expected to hit 642 million people in 2040. The aim of this study is to analyse the effect of demographical and clinical characteristics for diabetics disease in Bangladesh.

Methods: This study employs the quantitative approach for data analysis. First, we analyse differences in variables between diabetic patients and controls by independent two-sample t-test for continuous variables and Pearson Chi-square test for categorical variables. Then, logistic regression (LR) identifies the risk factors for diabetes disease based on the odds ratio (OR) and the adjusted odds ratio (AOR).

Results: The results of the t-test and Chi square test identify that the factors: residence, wealth index, education, working status, smoking status, arm circumference, weight and BMI group show statistically ($p < 0.05$) significant differences between the diabetic group and the control group. And, LR model demonstrates that 2 factors ("working status" and "smoking status") out of 13 are the significant risk factors for diabetes disease in Bangladesh.

Conclusions: We believe that our analysis can help the government to take proper preparation to tackle the potentially unprecedented situations in Bangladesh.

Keywords Diabetes detection · Quantitative analysis · Logistic regression · Significance p-value · Odds ratio · Adjusted odds ratio

1 Introduction

Diabetes, also known as Diabetes Mellitus, has risen as a significant public health problem worldwide, particularly in lower and moderate-income countries Diabetes is a familiar disease as 80% of the people of the world is suffered from it according to the International Diabetes Fedaration, Belgium [1, 2]. This disease is an assemblage of metabolic disorders that are distinguished by high blood sugar. Diabetes occurs when the body can not process the food we eat

✉ Linta Islam
lislamcsejnu@gmail.com

✉ Md Rafiqul Islam
rafiqulislam.cse24@gmail.com

Shanjita Akter
sanjitaprome@gmail.com

Md Zobaer Hasan
mdzobaer.hasan@monash.edu

Mohammad Ali Moni
m.moni@uq.edu.au

Mohammed Nasir Uddin
nasir.jnu.cse@gmail.com

[1] Department of CSE, Jagannath University, Dhaka, Bangladesh

[2] Advanced Analytics Institute (AAi), University of Technology Sydney (UTS), Sydney, Australia

[3] Department of CSE, Islamic University of Technology (IUT), Gazipur, Bangladesh

[4] School of Science, Monash University Malaysia, Selangor D.E. Subang Jaya, Selangor D. E., Malaysia

[5] School of Health and Rehabilitation Sciences, The University of Queensland, St Lucia, Australia

and increase the sugar level [3]. This disorder can prompt many serious long-term, complicated diseases; concurrently, it can cause blindness, stroke, heart attacks, kidney failure, and lower limb amputation. Diabetes and Cardiovascular diseases (CVD) are two of the most prevalent and frequent chronic diseases which lead to death in the United States [4]. In 2015, approximately 9% of the United States population had been determined to have diabetes, while an additional 3% were undiscovered [5]. Moreover, nearly 34% had prediabetes [6]. Furthermore, around 90% of adults with prediabetes were unaware of their condition. The number of diabetic patients increased from 122 million to 422 million between 1980 and 2014 [7]. The estimation will be struck to 642 million approximately in 2040 [8]. In 2016, an approximation of 1.6 million deaths was caused by diabetes which is a great range, while 2.2 million passings were inferable from high blood glucose in 2012 [9]. Moreover, the frequency of diabetes in Kuwait, Bahrain, Jordan and the United Arab Emirates were 12.8%, 14.9%, 17.1%, and 20.1%, respectively [10, 11]. Like other countries, Bangladesh is highly affected by diabetes. The IDF (International Diabetes Federation) Diabetes Atlas 5th edition projected that diabetes prevalence in Bangladesh would increment to over half by the next 15 years, setting Bangladesh as the eighth-most highest diabetic crowded country in the world [12]. WHO ranked diabetes as the seventh influential reason of death in 2016 [9]. Young adults are likewise profoundly influenced by diabetes. In another statistical report, the global predominance of diabetes among adults has doubled from 4.7 to 8.5% within the year 1980–2014 [12]. Therefore, the number of diabetes patients is expanded daily with a noteworthy range; subsequently, deaths also increase daily, which is alarming to us.

Diabetes can be of different categories: Type I, Type II and Gestational diabetes. Type I diabetes happens to people when an insulin-producing cell named beta cell is attacked by our immune system [13]. As 90% of the beta cells are permanently damaged through this attack, the pancreas can not produce enough insulin. For this reason, it is also called insulin-dependent diabetes. Among all the diabetes patients, only 5–10% of people have type I diabetes. This type of diabetes is also called juvenile-onset diabetes, as it may occur at the early stage of our life. The next type of diabetes is Type II diabetes which is called non-insulin-dependent diabetes [14]. As it does not respond to the insulin produced by the pancreas, the body becomes resistant to insulin. This type of disease is also known as adult-onset diabetes, as it affects young and older people. Approximately 90% of people have type II diabetes. The last type of diabetes is gestational diabetes. It generally occurs to the women during their pregnancy [15]. However, it affects the baby's growth rather than affecting the mother. 2–10% of women might have gestational diabetes during their pregnancy [16]. Generally, the disease goes away after pregnancy, but it may also cause type II diabetes.

The analysis of diabetes data is difficult because a large portion of the clinical information is nonlinear, non-normal, relationship organised, and complex [17]. On the other hand, the traditional clinical methods to detect diabetes are time-consuming and inconvenient. Therefore, healthcare professionals need time-efficient, simple and open diabetes detection systems that can precisely identify the stage of diabetes which the patient holds. In the world of constantly increasing data, where hospitals are gradually embracing big data systems [18], there are incredible advantages to utilise advanced technology in the medical sector to give experiences, augment diagnosis, improve results, and decrease costs [19, 20].

Machine learning has significant impacts on medical healthcare. This technology enhances the efficiency of the medical health care system [21]. A considerable amount of research has been done to predict or identify diabetes using a machine learning system. To detect diabetes mellitus, Maniruzzaman et al. [22] applied four classifiers and LR (logistic regression) models as the most risk-identifiable factors. Their model provides 94.25% accuracy with a combination of LR-based features and Random Forest (RF) based classifiers. Dinh et al. [23] used 123 variables and the eXtreme Gradient Boost (XGBoost) model, which helps to achieve 86.2% without using laboratory data and 95.7% with laboratory data. In addition to machine learning, deep learning algorithms have also been used to detect diabetes [24]. The authors used a combination of convolutional neural network (CNN) and long short-term memory (LSTM) and gained an accuracy of 95.7%. However, these approaches might provide the inaccurate and biased result if we apply the algorithms on the dataset without analysing them properly [25]. Therefore, we need to analyse and adjust our data according to the standard model. In this case, statistical analysis will assist us in analysing our data and providing direction in achieving a better result in diabetes prediction.

Thus, in this study, the purpose of applying the statistical approach is to obtain the significant features that may prompt improved accuracy in detecting diabetes. This research aims to identify the most significant factors and treat people accurately diagnosed with diabetes. Analysing the above research work, we aim to reveal the impact of various factors on diabetes by performing a quantitative analysis. We look to achieve the following research objectives:

- We investigated both demographic and clinical characteristics to explain the estimation of diabetes.
- We analyse and determine the most significant risk factors of diabetes diseases using the Logistic Regression.

The remainder of the paper is organised as follows. Section 2 presents related work. The methodology is presented in Sect. 3. Result analysis and Discussion are illustrated in Sect. 4. Section 5 finally concludes the paper.

2 Related Work

Diabetes is a chronic disease where a person suffers from an extensive blood glucose level in their body and may cause many complications. Several researchers have adopted machine learning (ML) techniques in healthcare analysis in the last few decades, especially cancer, brain tumour, and diabetes detection. The primary motivation behind the engagement of ML models is to improve the healthcare systems [26–28]. For example, Sneha [29] proposed an ML model that aims to focus on attribute selection, which helps detect diabetes mellitus using predictive analysis. Their performance returned 98.20% highest accuracy with the combination of the decision tree algorithm and Random forest. Sisodia [30] sketched a model that can forecast the likelihood of diabetes in patients with the highest accuracy. Contrastingly, Ozcift [31] proposed an ensemble approach named "Rotation forest", merging 30 machine learning methods. Han et al. [32] preferred another ML-based model, which alters the SVM prediction rules. Therefore, we included the listing of existing ML work in Table 1.

Although existing studies show various ML-based models for diagnosing diabetes diseases that successfully demonstrated and estimated their performance, however, none of them were able to achieve 100% accuracy in terms of diabetes prediction. One of the main reasons behind this argument is that the dataset was not adequately analysed. Introducing statistical analysis in diabetes prediction might lessen the biasness and increase the improvement of the result. For example, Afroz et al. [40] assessed the diabetic patient of Bangladesh. However, the authors applied only the odds ratio to determine the factors causing diabetes. Camara et al. [41] evaluated the reason for diabetes in Cameroon and Guinea. Nevertheless, the authors only applied a non-adjusted odds ratio to find the significant factors that cause diabetes. Therefore, our current study has researched a diabetes dataset, and our goal is to identify the significant factors that cause diabetes by applying logistic regression and find the odds ratio as well as adjusted odds ratio of the factors. Additionally, we have calculated the p-value for selecting critical features to ameliorate the detection accuracy and determine the relationship between the predictor and the result class. In our last stage, we have determined the importance of factors by using Artificial Neural Network (ANN).

3 Research Methodology

Diabetes is a major metabolic disorder that affects the entire body system skeptically [42]. However, medical datasets are often larger in dimensions with complex redundant features, which increases the possibility of noise and dependency among the features and reduces the performance and accuracy. Hence data preprocessing plays a significant role in performing machine learning tasks with medical datasets [30, 43]. The process of reducing the dimensionality would be either feature selection or feature extraction. This work aims to demonstrate proper features which are significant risk factors for determining diabetes diseases.

In this study, we have used a diabetic dataset in Bangladesh. The dataset is freely available. After the data collection, we present our performing model to investigate the effect of selected factors on diabetics. The process flow of the proposed framework is represented in Fig. 1. We have divided our proposed framework into three segments: Data Preprocessing, Data Analysis and Outcome. In the data preprocessing segment, we remove the null values and unevenly distributed factors. In the next segment, we analyse our data based on demographical and clinical characteristics. Then, we calculate each factor's odds ratio and adjusted odds ratio using logistic regression. At the final stage, we define the significant factors that produce an impact on diabetes. Finally, we determine the importance of the factors using the ANN.

3.1 Data Collection and Processing

The diabetes dataset used in this study were collected from Bangladesh Demographic and Health Survey (BDHS), which is available on The DHS Program website [44]. The data was consists of 1564 participants having nominal and ordinal factors. In this dataset, two patients have zero cm of arm circumference, systolic blood pressure is zero in three patients, two patients have zero diastolic blood pressure, and two patients have zero kg of weight. These zero values are represented as missing or null values. Thus, we have to eliminate the null values in the preprocessing step. Thus, our final dataset has 1555 participants' data. The dataset has 132 diabetic patients and 1423 controls. The descriptions of the attributes and brief statistical summary are shown in Tables 2 and 3, respectively.

From Table 2, we observe that the attributes: age, systolic blood pressure, diastolic blood pressure, height, weight, and arm circumference has numeric values. Thus, we do not need to encode these attributes. However, we had to encode the other features as they have categorical

Table 1 State of the art of diabetes detection models

Reference	Objective	Variable	Approach
Maniruzzaman et al. [22]	Develop ML-based system for predicting diabetes diseases	Age, gender, BMI, height, weight	Classification and Prediction model
Hossain et al. [33]	Detect Autism Spectrum Disorder by exploring the most significant feature selection method	Gender, ethnicity, jaundice, Family ASD, Residence and ASD class	Benchmark ML-algorithm
Cantin-Garside et al. [34]	Create a continuous and smart SIB monitoring system which provide high accuracy, efficiency and specificity	Age, SIB, classic and fine behaviours	Classification algorithm (NN, KNN, SVM)
Dinh et al. [23]	Develop machine-learned models dependent on study survey can provide an automated identification mechanism of diabetes and cardiovascular diseases	family history, age, gender, race and ethnicity, weight, height, waist circumference, BMI, hypertension, smoking and alcohol use	Classification algorithm (LR, SVM, RF, GB)
Kavakiotis et al. [35]	Uphold various Machine Learning and Data Mining approaches for identifying DM	Medical datasets	Supervised, Unsupervised and Reinforcement learning
Kaur and Kumari. [36]	Implemented a model using supervised ML-techniques in R for understanding patterns to identify diabetes	Age, DBP, BMI	SVM-linear, K-NN, ANN, MDR
Kumar Dewangan and Agrawal [37]	Develop an ensemble model using Bayesian classification and Multilayer perceptron of diagnosis of diabetes-mellitus	Age, BMI, DBP and health condition	Bayesian classification, Multilayer perceptron
Larabi-Marie-Sainte et al. [38]	Created a unified repository for analysing the best ML and DL technique-based diabetes predictions	Age, family history, regular diet, DBP and BMI	ANN, SVM, DT
Islam et al. [39]	Automated system for detecting and classifying diabetes using machine learning	Region, residence, electricity, wealth index, age, AC, SBP, DBP and BMI	SVM, RF, KNN, LR, RT and LDA
Afroz et al. [40]	Assessed glycemic control in Bangladeshi diabetic people	Age, Gender, Education Level, and BMI	Odds ratio, Chi-square Test
Camara et al. [41]	Investigated the predictors of diabetes in Cameroon and Guinea	Age, Sex, Employment, Education Level, BMI, SBP and DBP	Odds ratio

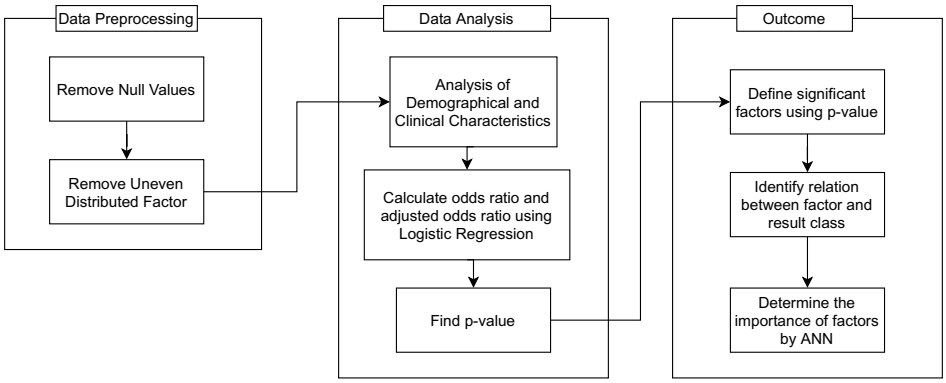

Fig. 1 Overview of the Proposed Statistical Model

Table 2 Description of attributes of the dataset

Attribute	Type	Description
Age	Number	Age in years
Gender	String	Male or female
Area	String	Rural or Urban
Education	String	Various type of education stage, e.g. primary, secondary, higher or no education
Electricity	Boolean (yes/no)	Whether the residential area has electricity
Working status	Boolean (yes/no)	Whether any of them is unemployed
Smoking status	Boolean (yes/no)	whether any of them is non-smoker
Systolic blood pressure (SBP)	Number	SBP in numbers
Diastolic blood pressure (DBP)	Number	DBP in numbers
Height	Decimal	Height in decimal number
Take medicine	Boolean (yes/no)	Whether any of them ignore medicine
Weight	Decimal	Weight in decimal numbers
Wealth index	String	Three types of wealth condition consists of poorer, middle and richest
BMI group	String	Normal or underweight or overweight
Arm circumference	Number	Arm circumference in numbers
Class	String (diabetic or control)	Whether the objective identified as diabetic or normal

values like string or boolean values. To find the difference between diabetic patients and control patients, we conducted an independent two-sample t-test for continuous variables and a Pearson Chi-square test for categorical variables, illustrated in Table 3. From this table, we can see that most of the patients were from Dhaka. However, the percentage of rural people were more than the urban. We have also categorised the age into five classes based on the paper [45]. Our dataset contains the data of the people having age limits from 35 to 54 years old. We also removed the gender feature as the data percentage of the female class is 0.26% only. We also removed the Electricity and Take medicine feature as the data were unevenly distributed. Conclusively, our final dataset is consists of 1555 data, 13 features and one target variable.

3.2 Model Specification

There are various machine learning based works for identifying diabetes disease. However in this study, our main concern is identifying the risk factors along with diabetes detection. As our dependent variables are binary and discrete, we use a logistic regression model. The logistic regression model can be specified as follows:

$$P(Y=1|X) = f\left(x_1, x_2, x_3, ...x_n\right) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+...+\beta_n X_n)}}$$

This can also be expressed as:

Table 3 Demographic and clinical characteristics of the study population

Factors	Overall (n = 1555)		Diabetic (n = 132)		Control (n = 1423)		p-value
	Freq	Percent (%)	Freq	Percent (%)	Freq	Percent (%)	
Region							0.490
Barisal	147	9.5	15	11.4	132	9.3	
Chittagong	207	13.3	17	12.9	190	13.4	
Dhaka	342	22.0	28	21.2	314	22.1	
Khulna	188	12.1	15	11.4	173	12.2	
Rajshahi	253	16.3	21	15.9	232	16.3	
Rangpur	234	15.0	14	10.6	220	15.5	
Sylhet	184	11.8	22	16.7	162	11.4	
Residence							0.002
Rural	992	63.8	68	51.5	924	64.9	
Urban	563	36.2	64	48.5	499	35.1	
Wealth Index							< 0.0005
Poorest	302	19.4	19	14.4	283	19.9	
Poorer	286	18.4	19	14.4	267	18.8	
Middle	273	17.6	10	7.6	263	18.5	
Richer	317	20.4	24	18.2	293	20.6	
Richest	377	24.2	60	45.5	317	22.3	
Age Group							0.157
35–39	516	33.2	34	25.8	482	33.9	
40–44	504	32.4	46	34.8	458	32.2	
45–49	535	34.4	52	39.4	483	33.9	
Education							< 0.0005
No Education	494	31.8	29	22.0	465	32.7	
Primary	449	28.9	38	28.8	411	28.9	
Secondary	372	23.9	27	20.5	345	24.2	
Higher	240	15.4	38	28.8	202	14.2	
Working Status							0.001
No	15	1.0	5	3.8	10	0.7	
Yes	1540	99.0	127	96.2	1413	99.3	
Smoking Status							0.003
No	1307	84.1	123	93.2	1184	83.2	
Yes	248	15.9	9	6.8	239	16.8	
Arm Circumference	26.95±2.65		27.56±2.90		26.90±2.62		0.006
SBP (mm Hg)	114.47±16.00		115.45±17.07		114.38±15.90		0.461
DBP (mm Hg)	76.27±11.78		77.29±12.78		76.18±11.68		0.299
Height (m)	1.63±0.06		1.62±0.07		1.63±0.06		0.706
Weight (Kg)	55.84±10.03		58.95±11.32		55.55±9.86		0.001
BMI Group							<0.0005
Underweight	355	22.8	19	14.4	336	23.6	
Normal	991	63.7	80	60.6	911	64.0	
Overweight	195	12.5	32	24.2	163	11.5	
Obese	14	0.9	1	0.8	13	0.9	

The continuous variables are expressed as mean±SD and the categorical variables are expressed as n(%)

$$Logit(p) = log\frac{1}{1-p} = \left(\beta_0 + \beta_1 X_1 + ... + \beta_n X_n\right)$$

where β_0 is the intercept, X_i are a set of predictor variables, and β_i are the regression coefficients associated with the ith predictor. In the above, p is the probability of a change in life insured status and $\frac{1}{1-p}$ is known as the odds ratio. β_i ives an estimate of change in the logodds associated with a unit change in the predictor variable. The parameters in the model are estimated using the method of maximum likelihood estimation (MLE).

4 Result and Discussion

In the following subsections, we briefly present the various outcomes which were identified during the quantitative phase of the study. We then determine the effect of selected features on the diabetics using logistic regression.

4.1 Result of Quantitative Analysis

The dataset contains both continuous and categorical data. The demographical characteristics of the participants contain only categorical data, while the clinical characteristics of the participants are composed of both continuous and categorical data. Here, the continuous variables are expressed as mean ± SD, and the categorical variables are expressed as n(%). We calculated the p-value for both continuous variable and categorical variable by performing two different tests. In this study, we refer to the level of significance as 5% ($p < 0.05$).

To find the difference between diabetic patients and control patients, we illustrated the t-test and Pearson Chi-Square test result of categorical and continuous values in Table 3. From this table, it is notable that the factors Residence, Wealth Index, Education, Working Status, Smoking Status, Arm Circumference, Weight and BMI group has a p-value of less than 0.05. This result proves that these factors are statistically significant as there are significant differences in these factors between the diabetic group and the control group. However, we can notice that the factors Region, Age Group, SBP, DBP and Height are not statistically significant as their p-value is more than 0.05. This statement indicates no significant differences in these factors between the diabetic and control groups. To better understand the relationship between factors along this dimension, we computed the similarity score between the factors. Figure 2 presents the correlation matrix, which illustrates the association between the factors. From this figure, we can easily interpret the correlated and associated factors that assist us in predicting diabetes disease. The systolic and diastolic blood pressures and the weight with

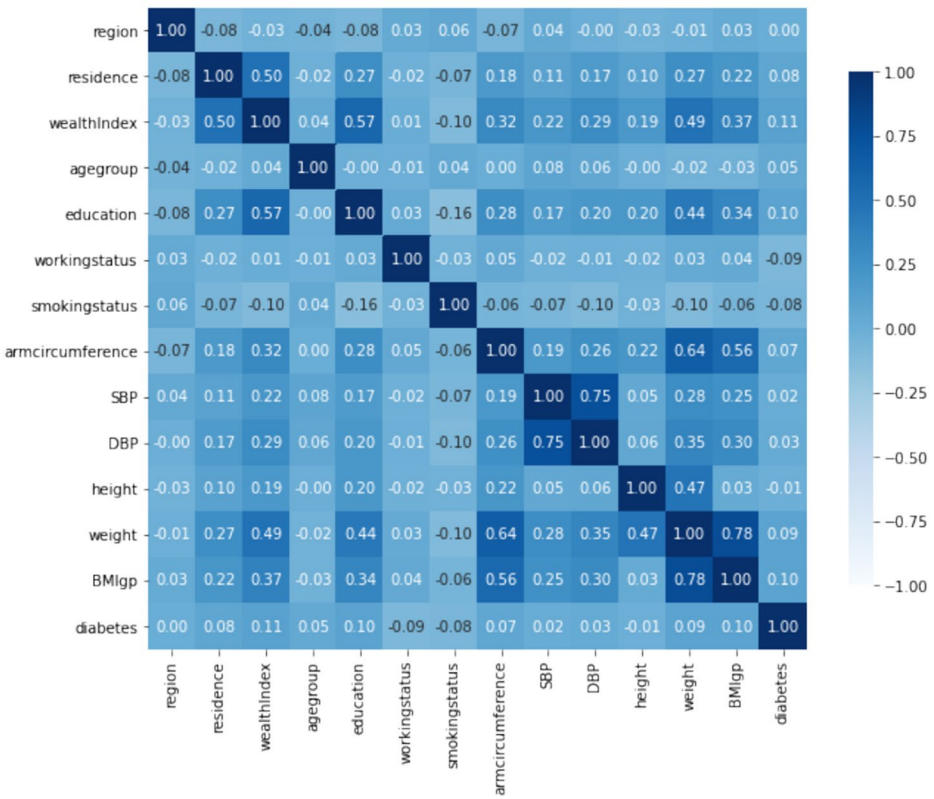

Fig. 2 Correlation between Factors using Similarity Score

Table 4 Effect of selected factors on the diabetes using logistic regression

Factors	OR	p-value	95% CI for OR		AOR	p-value	95% CI for AOR	
			[Lower]	[Upper]			[Lower]	[Upper]
Region								
Barisal (ref)	1.000				1.000			
Chittagong	0.787	0.520	0.380	1.632	0.685	0.331	0.320	1.468
Dhaka	0.785	0.471	0.406	1.517	0.697	0.323	0.341	1.425
Khulna	0.763	0.480	0.360	1.616	0.621	0.243	0.279	1.383
Rajshahi	0.797	0.522	0.397	1.598	0.779	0.502	0.376	1.614
Rangpur	0.560	0.135	0.262	1.197	0.672	0.336	0.299	1.510
Sylhet	1.195	0.615	0.596	2.395	1.107	0.789	0.526	2.333
Residence								
Rural (ref)	1.000				1.000			
Urban	1.743	**0.002**	1.218	2.494	0.931	0.762	0.585	1.481
Wealth Index								
Poorest (ref)	1.000				1.000			
Poorer	1.060	0.862	0.549	2.046	1.072	0.840	0.543	2.117
Middle	0.566	0.155	0.259	1.240	0.526	0.126	0.231	1.197
Richer	1.220	0.532	0.654	2.276	1.135	0.726	0.559	2.306
Richest	2.819	**<0.0005**	1.642	4.839	2.107	0.066	0.953	4.658
Age Group								
35–39 (ref)	1.000				1.000			
40-44	0.655	0.066	0.418	1.028	1.412	0.156	0.876	2.274
45–49	0.933	0.744	0.615	1.415	1.554	0.065	0.973	2.484
Education								
No Education (ref)	1.000				1.000			
Primary	0.332	**<0.0005**	0.199	0.552	1.225	0.471	0.706	2.127
Secondary	0.491	**0.004**	0.304	0.794	0.823	0.552	0.434	1.562
Higher	0.416	**0.001**	0.247	0.702	1.392	0.352	0.693	2.793
Working Status								
No (ref)	1.000				1.000			
Yes	0.180	**0.002**	0.061	0.534	0.133	0.001	0.041	0.430
Smoking Status								
No (ref)	1.000				1.000			
Yes	0.362	**0.004**	0.182	0.723	0.419	0.017	0.205	0.854
Arm Circumference	1.094	**0.006**	1.026	1.166	1.015	0.755	0.923	1.117
SBP (mm Hg)	1.004	0.461	0.993	1.015	0.998	0.797	0.980	1.016
DBP (mm Hg)	1.008	0.299	0.993	1.023	0.990	0.447	0.965	1.016
Height (m)	0.531	0.663	0.031	9.156	0.035	0.144	0.000	3.144
Weight (Kg)	1.031	**<0.0005**	1.015	1.048	1.021	0.412	0.972	1.072
BMI Group								
Underwei-ght (ref)	1.000				1.000			
Normal	1.553	0.094	0.927	2.601	1.125	0.746	0.551	2.296
Overweight	3.472	**<0.0005**	1.910	6.311	1.403	0.598	0.399	4.934
Obese	1.360	0.772	0.169	10.952	0.410	0.524	0.026	6.349

Ref stands for reference. Statistically significant results at $p<0.05$ are in bold

the BMI group are highly interconnected to each other. Moreover, the education with wealth index; and the weight with the arm circumference are moderately interconnected. However, the smoking status with both education and wealth index are inappreciably associated with one another.

Applying logistic regression, we calculated the odds ratio, adjusted odds ratio, p-value, and confidence interval for each factor illustrated in Table 4. These values will be crucial to identify the key risk factors of having diabetes. We choose the low-frequency variables or negative

classifiers as our reference group. Thus, for example, for the first factor, "Region", we choose Barisal as our reference group, and we determine the odds ratio and adjusted odds ratio based on this reference group. From the table, it is visible that a person from the Sylhet region is 1.1 times more likely to be diabetic than the reference region Barisal. However, a person from the Khulna region is 0.6 times less prone to be diabetic than the reference region Barisal. Moreover, Urban people are 0.9 times less expected to be diabetic compared with reference residence group rural. Besides, diabetes is more inclined to the richest category and less possible among the middle category compared to the reference category, "poorest". It is also seen that the probability of having diabetes is increasing with age and also education. Also, The odds of having diabetes disease are 0.133 times lesser for working people than non-working people. However, we notice that the odds of having diabetes disease is 0.419 times lower for a smoker than for a non-smoker.

For the clinical characteristics, all the predictors are continuous except the BMI group. A person with normal and overweight is more likely to be diabetic, and an obese person is less likely to be diabetic compared with the reference group underweight. Since the adjusted odds ratio of arm circumference and weight are more than one, we can state that an increase in these factors leads to an increased probability of having diabetes. From Table 4, it can be concluded that "Working Status" and "Smoking Status" are statistically significant factors for diabetes disease at a 5% level of significance because for both OR and AOR the p-values are smaller than 0.05.. The rest of the 11 factors are insignificant in diabetes detection as the p-values in AOR were greater than 0.05. Finally, in Figure 3 we have shown the importance of the factors which are determined by ANN.

4.2 Discussion

In this research, we have tried to determine the factors that cause diabetes. The factors of our dataset are divided according to the demographic and clinical characteristics of the study population. We can suggest that a person's working and smoking status are more critical in prompting diabetes from the experiments. However, the arm circumference, weight and the BMI of a person also confers particular importance in this research. Overall, the characteristics mentioned above are essential in efficient diabetes detection though the working and smoking status work as the best factor when we measure based on a definite objective.

Among the demographic characteristics, residence, wealth index, education, working status and smoking status has more importance than the other factors. The other factors like region and age group have a p-value of more than 0.05, which clearly explains they are nonsignificant in diabetes detection. The risk of possessing diabetes is higher in people with higher education [46]. Moreover, a study in the UK found that people with a lower wealth are prone to diabetes [47]. Thus we can say these factors are associated with people's lifestyle and can influence diabetes. The impact of the clinical characteristic is enormous. Arm circumference, Weight and BMI are significantly essential for resulting in diabetes in a person. The previous study found that BMI has a strong association with diabetes and obese individuals are more likely to have diabetes [48]. However, the blood pressure and height has a p-value greater than .05. Thus, these factors have no significant difference between the diabetic and control group.

This research has several advantages over previous researches. We have developed a statistical model to identify

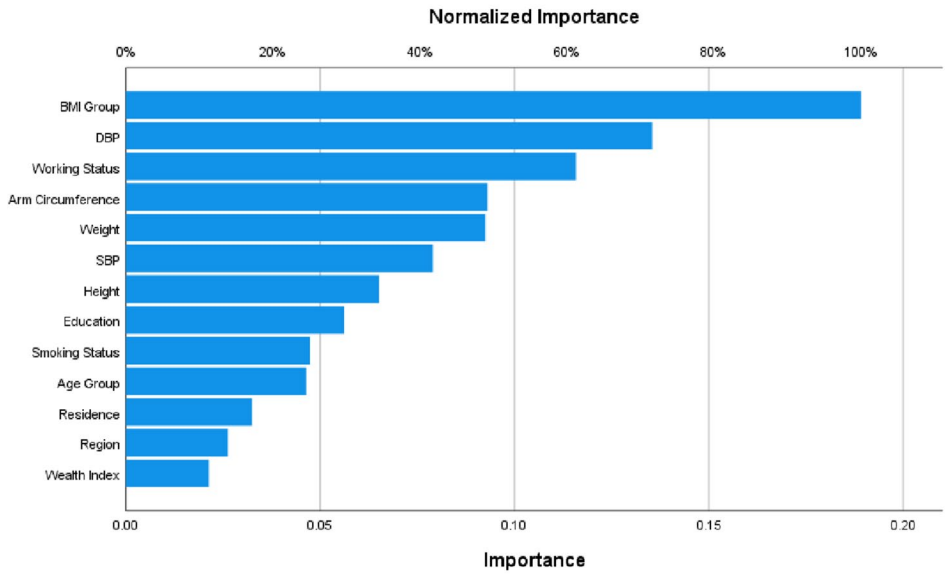

Fig. 3 Importance of factors calculated by Artificial Neural Network (ANN)

the factors causing diabetes efficiently at the early stage of the disease. Our paper's main objective was to analyse the diabetes dataset from quantitative points of view. For this reason, our proposed statistical framework will be significant towards the statistics of the diabetes analysis. Moreover, we have found several key factors that directly lead to diabetes. More wealthy people and working people are more likely to have diabetes. Another interesting finding from our research is that people who smoke are less likely to have diabetes which contradicts the previous studies [49]. Consequently, our findings illustrate that different characteristics and factors are related to a different frequency of diabetes. Furthermore, our work is not confined to determining the association of the predictors with the diabetes class; we also discover the correlation among the factors by calculating the similarity score. In future, we have to choose the factors carefully while prognosticating diabetes.

This work has few shortcomings too. The dataset does not provide any additional information about the family or ancestor of the diabetic person. This additional information might be a critical factor in diabetes prediction. Moreover, we have to ignore the gender feature as the gender data is unevenly distributed. According to the studies, men have higher incidences of diabetes than women [50]. Thus, gender could be another key feature that we ignored in our study due to the uneven distribution. Furthermore, we have particularly used regression analysis to detect diabetes and prove our hypotheses. This methodological limitation might be solved if we could have applied the classification-based analysis.

To conclude, we should analyse both demographic and clinical characteristics in explaining the estimation of diabetes. The clinical characteristics are the most significant predictors in describing the emergence of diabetes. Moreover, demographic characteristics are likewise necessary for predicting diabetes. Smoking status, working status, weight and BMI plays a vital role in determining the disease accurately. Thus, non-smoker, non-working, and obese people hold a high probability of having diabetes.

5 Conclusion

Diabetes is a familiar disease as 80% of the people of the world is suffered from it as reported by the International Diabetes Fedaration, Belgium. It is essential to discover the causes behind diabetes and understand the pattern of the disease. In this paper, we studied the demographical and clinical characteristics of 1555 participants from Bangladesh and found the significant factors that cause diabetes. Using logistic regression, we computed the odds ratio and the adjusted odds ratio, and found the relation between the factors and the outcome. We found that notable significant characteristics (smoking status and working status) demonstrated the diabetes class as the p-value for both these factors were less than 0.05 in the OR and AOR calculation. Moreover, the odds ratio for smoking class was 0.419 times lower and the working class was 0.133 times lower than the non-smoking and non working class respectively. Finally, this study presents the diabetes pattern and will be able to assist the healthcare professionals in prompt decision-making.

Declarations

Conflict of Interest There is no conflict of interest.

References

1. Edition IDAS. International diabetes federation. Brussels: Belgium; 2013.
2. Islam SMS, Purnat TD, Phuong NTA, Mwingira U, Schacht K, Fröschl G. Non-communicable diseases (ncds) in developing countries: a symposium report. Glob Health. 2014;10:1–8.
3. Mohapatra SK, Swain JK, Mohanty MN. Detection of diabetes using multilayer perceptron. In: International Conference on intelligent computing and applications, Springer, 2019; p. 109–116.
4. Leon BM, Maddox TM. Diabetes and cardiovascular disease: epidemiology, biological mechanisms, treatment recommendations and future research. World J Diabetes. 2015;6:1246.
5. for Disease Control C, Prevention et al. National diabetes statistics report, 2020. Atlanta, GA: Centers for Disease Control and Prevention, US Department o Health and Human Services , 2020; p. 12–15.
6. Dall TM, Yang W, Gillespie K, Mocarski M, Byrne E, Cintina I, Beronja K, Semilla AP, Iacobucci W, Hogan PF. The economic burden of elevated blood glucose levels in 2017: diagnosed and undiagnosed diabetes, gestational diabetes mellitus, and prediabetes. Diabetes Care. 2019;42:1661–8.
7. Collaboration NRF, et al. Trends in adult body-mass index in 200 countries from 1975 to 2014: a pooled analysis of 1698 population-based measurement studies with 19· 2 million participants. The Lancet. 2016;387:1377–96.
8. Zimmet P, Alberti KG, Magliano DJ, Bennett PH. Diabetes mellitus statistics on prevalence and mortality: facts and fallacies. Nat Rev Endocrinol. 2016;12:616.
9. Lütkebohle, I. BWorld Robot Control Software. https://www.who.int/health-topics/diabetes. Accessed 23 May 2021.
10. Khattab M, Khader YS, Al-Khawaldeh A, Ajlouni K. Factors associated with poor glycemic control among patients with type 2 diabetes. J Diabetes Complic. 2010;24:84–9.

11. Al-Khawaldeh OA, Al-Hassan MA, Froelicher ES. Self-efficacy, self-management, and glycemic control in adults with type 2 diabetes mellitus. J Diabetes Complic. 2012;26:10–6.
12. Atlas D. International diabetes federation. idf diabetes atlas. Brussels: International Diabetes Federation 2015.
13. DiMeglio LA, Evans-Molina C, Oram RA. Type 1 diabetes. The Lancet. 2018;391:2449–62.
14. Chatterjee S, Khunti K, Davies MJ. Type 2 diabetes. The Lancet. 2017;389:2239–51.
15. Khajehei M, Assareh H. Temporal trend of diabetes in pregnant women and its association with birth outcomes, 2011 to 2017. J Diabetes Complic. 2020;34: 107550.
16. Kramer CK, Campbell S, Retnakaran R. Gestational diabetes and the risk of cardiovascular disease in women: a systematic review and meta-analysis. Diabetologia. 2019;62:905–14. https://doi.org/10.1007/s00125-019-4840-2.
17. Maniruzzaman M, Kumar N, Abedin MM, Islam MS, Suri HS, El-Baz AS, Suri JS. Comparative approaches for classification of diabetes mellitus data: machine learning paradigm. Comput Methods Programs Biomed. 2017;152:23–34.
18. Gans D, Kralewski J, Hammons T, Dowd B. Medical groups' adoption of electronic health records and information systems. Health Aff. 2005;24:1323–33. https://doi.org/10.1377/hlthaff.24.5.1323.
19. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. Health Inf Sci Syst. 2014;2:3.
20. Vijayan VV, Anjali C. Decision support systems for predicting diabetes mellitus-a review. In: 2015 Global Conference on communication technologies (GCCT), IEEE, 2015; p. 98–103.
21. Islam MR, Liu S, Wang X, Xu G. Deep learning for misinformation detection on online social networks: a survey and new perspectives. Soc Netw Anal Min. 2020;10:1–20. https://doi.org/10.1007/s13278-020-00696-x.
22. Maniruzzaman M, Rahman MJ, Ahammed B, Abedin MM. Classification and prediction of diabetes disease using machine learning paradigm. Health Inf Sci Syst. 2020;8:7.
23. Dinh A, Miertschin S, Young A, Mohanty SD. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. BMC Med Inf Decis Mak. 2019;19:211.
24. Swapna G, Vinayakumar R, Soman K. Diabetes detection using deep learning algorithms. ICT Express. 2018;4:243–6.
25. Olteanu A, Castillo C, Diaz F, Kıcıman E. Social data: biases, methodological pitfalls, and ethical boundaries. Front Big Data. 2019;2:13. https://doi.org/10.3389/fdata.2019.00013.
26. Islam MR, Miah SJ, Kamal ARM, Burmeister O. A design construct of developing approaches to measure mental health conditions. Australas J Inf Syst. 2019;23.
27. Islam MR, Kabir MA, Ahmed A, Kamal ARM, Wang H, Ulhaq A. Depression detection from social network data using machine learning techniques. Health Inf Sci Syst. 2018;6:1–12.
28. Islam MR, Kamal ARM, Sultana N, Islam R, Moni MA. et al. Detecting depression using k-nearest neighbors (knn) classification technique. In: 2018 International Conference on computer, communication, chemical, material and electronic engineering (IC4ME2), IEEE, 2018; p. 1–4.
29. Sneha N, Gangil T. Analysis of diabetes mellitus for early prediction using optimal features selection. J Big data. 2019;6:13.
30. Sisodia D, Sisodia DS. Prediction of diabetes using classification algorithms. Proc Comput Sci. 2018;132:1578–85.
31. Ozcift A, Gulten A. Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms. Comput Methods Programs Biomed. 2011;104:443–51.
32. Han L, Luo S, Yu J, Pan L, Chen S. Rule extraction from support vector machines using ensemble learning approach: an application for diagnosis of diabetes. IEEE J Biomed Health Inf. 2014;19:728–34.
33. Delowar Hossain M, Ashad Kabir M, Anwar A, Zahidul Islam M. Detecting autism spectrum disorder using machine learning. arXiv e-prints , arXiv–2009, 2020.
34. Cantin-Garside KD, Kong Z, White SW, Antezana L, Kim S, Nussbaum MA. Detecting and classifying self-injurious behavior in autism spectrum disorder using machine learning techniques. J Autism Dev Disord. 2020;50:1–14.
35. Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine learning and data mining methods in diabetes research. Comput Struct Biotechnol J. 2017;15:104–16.
36. Harleen Kaur VK. Predictive modelling and analytics for diabetes using a machine learning approach. Appl Comput Inform. 2018.
37. Kumar Dewangan A, Agrawal P. Classification of diabetes mellitus using machine learning techniques. Int J Eng Appl Sci. 2015;2(5):145–8.
38. Larabi-Marie-Sainte S, Aburahmah L, Almohaini R, Saba T. Current techniques for diabetes prediction: review and case study. Appl Sci. 2019;9:4604.
39. Islam M, Rahman J, Roy DC et al. Automated detection and classification of diabetes disease based on Bangladesh demography and health survey data, 2011 using machine learning approach. Diabetes Metab Syndr Clin Res Rev. 2020.
40. Afroz A, Chowdhury H A, Shahjahan M, Hafez M A, Hassan M N, Ali L. Association of good glycemic control and cost of diabetes care: experience from a tertiary care hospital in Bangladesh. Diabetes Res Clin Pract. 2016;120:142–8.
41. Camara A, Baldé NM, Sobngwi-Tambekou J, Kengne AP, Diallo MM, Tchatchoua AP, Kaké A, Sylvie N, Balkau B, Bonnet F, et al. Poor glycemic control in type 2 diabetes in the south of the Sahara: the issue of limited access to an hba1c test. Diabetes Res Clin Pract. 2015;108:187–92.
42. Han J, Rodriguez JC, Beheshti M. Diabetes data analysis and prediction model discovery using rapidminer. In: 2008 Second International Conference on future generation communication and networking, IEEE, 2008; p. 96–99.
43. Satu MS, Atik ST, Moni MA. A novel hybrid machine learning model to predict diabetes mellitus. In: International Joint Conference on computational intelligence, Springer, 2019; p. 453–465.
44. Program TD. Survey datasets files, Bangladesh: Standard dhs, 2011. https://dhsprogram.com/data/dataset/Bangladesh_Standard-DHS_2011.cfm. Accessed 23 May 2021.
45. Mahendran K, Patel S, Sproat C. Psychosocial effects of the covid-19 pandemic on staff in a dental teaching hospital. Br Dent J. 2020;229:127–32.
46. Seiglie JA, Marcus ME, Ebert C, Prodromidis N, Geldsetzer P, Theilmann M, Agoudavi K, Andall-Brereton G, Aryal KK, Bicaba BW, et al. Diabetes prevalence and its relationship with education, wealth, and bmi in 29 low-and middle-income countries. Diabetes Care. 2020;43:767–75.
47. Tanaka T, Gjonça E, Gulliford MC. Income, wealth and risk of diabetes among older adults: cohort study using the English longitudinal study of ageing. Eur J Public Health. 2012;22:310–7.
48. Al-Goblan AS, Al-Alfi MA, Khan MZ. Mechanism linking diabetes mellitus and obesity. Diabetes Metab Syndr Obes Targets Therapy. 2014;7:587.
49. Campagna D, Alamo A, Di Pino A, Russo C, Calogero A, Purrello F, Polosa R. Smoking and diabetes: dangerous liaisons and confusing relationships. Diabetol Metab Syndr. 2019;11:1–12. https://doi.org/10.1186/s13098-019-0482-2
50. Siddiqui MA, Khan MF, Carline TE. Gender differences in living with diabetes mellitus. Mater Socio-med. 2013;25:140.