IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 6, June 2023

DOI: 10.17148/IJARCCE.2023.12657

An Intelligent Model for early prediction of Type 2 diabetes likelihood using human behaviors and biometrics among adults in Saudi Arabia

Samah Alzahrani[1]

Student, Information Systems Dept, King Abdulaziz University, Jeddah, Saudi Arabia[1]

Abstract: Diabetes is a chronic disease that spread over the past decades in abundance. It is a metabolic disease that may affect the entire body. Diabetes is classified are three types, which are type 1 diabetes (T1D), type 2 diabetes (T2D), and gestational diabetes (GD), where each type has specific causes. This research study aims to find out the most common behaviours that lead to diabetes and measure the relationship between human biometrics and the likelihood of behaving T2D.The study aimed to develop a machine learning prediction model by investigating five machine learning algorithms which are Support Vector Machine, Logistic Regression, K-Nearest Neighbour, Decision Tree, and Random Forest. This model was developed by Python using google colab, Random Forest algorithm outperformed in perform highly accurate behavioural prediction with 98% compared with other algorithms. The outcome from this research study would assist the medical practice and medical community with a tool that can early predict T2D.

Keywords: Behaviours, Diabetes mellitus, Machine Learning, Prediction, Type 2 diabetes.

I.INTRODUCTION

Diabetes is a chronic disease that affects the amount of insulin secreted by the pancreas, causing an abnormal fluctuation of sugar in the blood that could be lower or higher than the normal level. It requires continuous care and monitoring of an individual’s behaviour in terms of therapeutics, a balanced diet, and exercise [1]. As a number of countries witnessed a noticeable increment in diabetes incidences, A study stated that type 1 diabetes has increased in the United States of America in the past three decades [2], resulting in a rise in the rate of infection among young people from 9.0 cases per 100,000 people annually in 2002–2003 to 12.5 cases per 100,000 in 2011–2021 [2]. In addition to the Middle East, which witnessed an obvious increase in the rate of infections with type 2 diabetes (T2D), The Kingdom of Saudi Arabia and the State of Bahrain were listed among the global top ten rated countries according to diabetes in the world [1].

Therefore, early exploration and prediction of the likelihood of having diabetes will help to avoid negative outcomes that may destroy human life [3], Based on this, the integration of machine learning, deep learning, and artificial intelligence that has been proven to be used as advanced computing solutions is very useful for early prediction of diabetes, and the advances made by machine learning techniques in the field of medicine and healthcare today are more noticeable than they were in the past. These methods and techniques have been widely used in diabetes research regarding its diagnosis, complications, and the environmental and genetic background of the disease [3]–[5].

These artificial intelligence techniques such as expert systems and machine learning algorithms used to facilitate the understanding of human behaviors for the early prevention of diabetes and many infectious diseases. These techniques have made tremendous efforts to serve the field of medicine, to achieve therapeutic savings due to the accuracy of predictive results [6], [7]. Therefore, this research will focus on developing an intelligent system that makes early predictions of diabetes mellitus through analysing and studying human behaviors and biometrics using machine learning algorithms, as will be described in the following sections of this research.

II. LITERATURE REVIEW

A. Diabetes Causes

There are three types of diabetes: gestational, type 1, and type 2, each of which is caused by different factors. In brief, gestational diabetes is caused during pregnancy according to hormonal changes along with genes and lifestyle (behavioral factors). The causes of type 1 fall under genes and environmental factors, while type 2 is caused by genes and behavioral factors such as obesity and a lack of physical activity. This type is considered the most popular and most common type of diabetes [8].

DOI: 10.17148/IJARCCE.2023.12657

Gestational diabetes (GD) is high blood sugar that is developed during pregnancy as a result of placental hormones that prevent the body from using insulin effectively, which results in sugar remaining in the blood instead of being absorbed into cells [9]. There are major causes of this type of diabetes: overweight, having children at a later age, previous history of GD, family history of T2D, and ethnicity [9]. This type of diabetes can be treated by increasing physical activity and maintaining a healthy dietary intake [9].

Type 1 diabetes (T1D) is a body behavior that attacks and destroys -cells (hormones that control glucose levels in the blood and produce insulin) in the pancreas to prevent producing the proper insulin amount [10]. The incidence statistics of type 1 diabetes (T1D) mellitus vary among countries according to several factors. Firstly, the geographical variation, where the incidence rate of diabetes in North America and European countries is high to medium compared to the African countries, where it is ranked as medium and low in Asia [11]. Secondly, the ethnicity, gender, and age variables show that diabetes has a positive relationship with age. In addition, it is noted that the diagnosis of diabetes increases in males in countries where the incidence of infection is high [11]. The third factor is the temporal variation in the incidence of type 1 diabetes, which varies with different seasons, as the incidence rate increases in the winter and autumn while decreasing in the summer and spring [12]. In addition, etiological factors such as drugs, vaccinations, and others play a significant role in increasing the incidence of diabetes, along with genetic susceptibility [11]. In this type, there is no medical solution to cure it yet, but it could be controlled to prevent complications by managing blood sugar levels with insulin injections and lifestyle [13].

Type 2 diabetes (T2D) is a failure in the body's ability to utilize and set the glucose (blood sugar) amount as fuel for resistance. Due to a progressive lack of -cell insulin secretion [14]. In this case, the pancreas is unable to produce enough insulin. Thus, this case will increase blood sugar levels in the bloodstream. Ultimately, high blood sugar levels would lead to other consequences, including disorders in the immune system and circulatory system [15].

The incidence of type 2 diabetes (T2D) mellitus also varies according to the variation in prevalence according to geographic location, age, gender, and ethnicity [11]. The spread of obesity is also one of the reasons that led to the emergence of type 2 diabetes, where the most important factors are lack of physical activity, diet, and obesity [16].

This type of diabetes is mostly common in older adults, but in recent years, with a changing lifestyle that led to several consequences, including the spread of obesity among children, the incidence rate of type 2 diabetes has increased among young people [17]. This is supported by noticing the incidence rate in rural areas, where individuals did not change their lifestyle significantly. Type 2 prevalence is low in rural areas compared to developed countries, as is its prevalence in ethnic communities [11]. Additionally, the temporal variation affected the rate of type 2 diabetes, as the U.S. data proved that the incidence increased by five times that which was diagnosed in 1980–2015, from 5.5 million cases to 23.4 million cases [16].

Based on the above causes, lifestyle modalities would help to prevent this type, including healthy eating habits, exercising, and losing weight [18]. Otherwise, patients may need to balance their insulin levels by using prescribed diabetes medications [14].

The literature shows various studies that investigated type 2 diabetes, including genes and other factors [9], [19], [20], Those studies stated that individual behaviors are considered one of the main factors that may profoundly contribute to prediabetes and type 2 diabetes, including exercise, exertion, sleep quality [21], and dietary intake that led to poor glucose regulation in the blood [22]. In addition, biometrics can be used to assess some parameters that contribute to predicting type 2 diabetes (T2D), such as weight, body mass index (BMI), blood pressure, blood cholesterol, and blood sugar [22]. All the discussed causes that led to diabetes occurrence are required to be controlled through a focus on predicting the incidence likelihood of diabetes based on individual behaviors and biometrics [22], [23].

B. Some Efforts to Control Diabetes.

Behavioral Solutions

Looking precisely at human behaviors, there are several perspectives on detecting the incidence and likelihood of some diseases [23]. Thus, predicting and analyzing human behaviors is one of the common practices in medicine. These behaviors, including physical exercise and lifestyle interventions such as commitment to balanced dietary intake to manage weight, sleep quality and duration, and anxiety mitigation and management, are considered among the most effective practices to manage, mitigate, and prevent type 2 diabetes [24]. In addition, self-control contributes to avoiding some behaviors that may increase the risk of developing type 2 diabetes, including depression, smoking, and stress.

DOI: 10.17148/IJARCCE.2023.12657

Machine Learning Solutions

Applying intelligent solutions in the health and medical fields is currently indispensable, which in turn works to transform data into valuable knowledge, as the health field is currently facing remarkable progress in several areas, such as biotechnology, then it was in the past, and this appears in the production of clinical and genetic information data extracted from large electronic health records (EHRs) [25]. The numerous options of digital technologies and methods are widely used in diabetes research regarding complications, prognosis, health care, and the genetic and environmental background of this disease [26], [27].

Early detection and prediction of diabetes in people at high risk of developing this disease limit its spread. In the western region of Saudi Arabia, a cross-sectional questionnaire-based study was conducted using traditional risk factors for diabetes. The most important risk factors for type 2 diabetes were examined and analyzed by performing a Chi-Squared test and binary logistic regression. Cross-sectional data were balanced using the Synthetic Minority Over-sampling (SMOTE) technique [28]. In addition, the study considered some variables, including region, gender, smoking, dietary intake, blood pressure (BP), and body mass index (BMI) [28]. Furthermore, in Saudi Arabia, a study was conducted using the Random Forest algorithm and the logistic regression algorithm to predict diabetes based on 18 risk factors [29]. The Random Forest algorithm outperformed logistic regression to better predict diabetes [29]. Machine learning is one of the leading areas that is used to measure the variability in blood sugar (HbA1c) and lipids to predict complications and mortality from diabetes [30]. Similarly, in Malaysia, a cross-sectional study was conducted at the University of Kebangsaan Malaysia Medical Center (UKMMC) that aimed to determine the prevalence of poor glycemic control and its relationship to the biological, psychological, and social factors of diabetic patients through measures of psychological factors, personality traits, and quality of life (QOL). This study conducted several questionnaires measuring these factors and multivariate binary logistic regression to determine predictors of poor glycemic control in 300 patients. The prevalence of poor glycemic control in the study (HbA1C ≥ 7.0%) was 69%, with a median HbA1C of 7.6% (IQR = 2.7). The study predicted that neglecting medication adherence and the longest duration of diabetes are factors that contribute to poor glycaemic control, and psychological factors are not associated with poor glycaemic control [6].

C. Hypotheses

This section will present the study hypotheses that shown in figure 1.

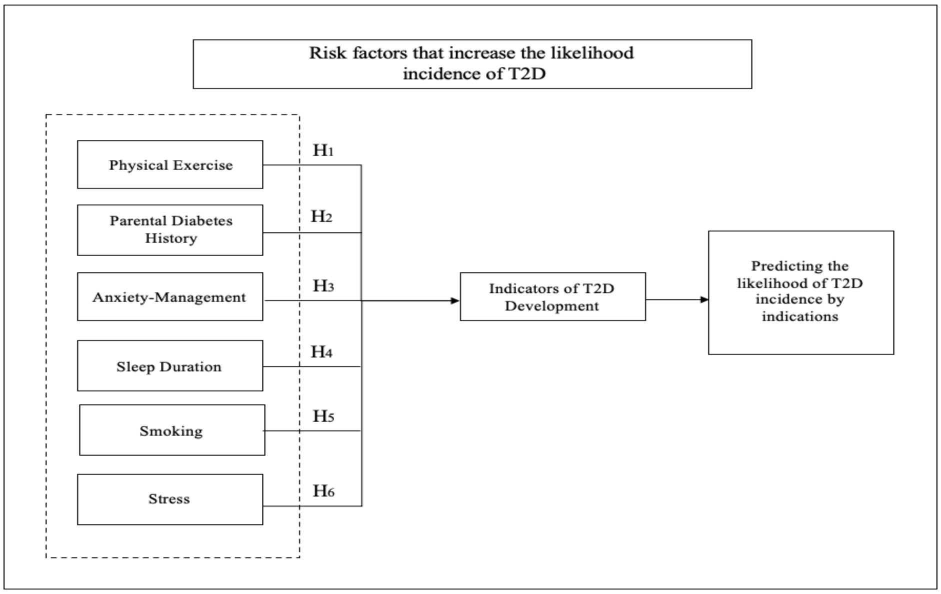
Risk factors that increase the likelihood incidence of T2D
Physical Exercise
Parental Diabetes History
Anxiety-Management
Sleep Duration
Smoking
Stress
H1
H2
H3
H4
H5
H6
Indicators of T2D Development
Predicting the likelihood of T2D incidence by indications

Fig 1. Hypotheses framework

H_1 : “Physical exercises decreased the likelihood risk of incidence of type 2 diabetes (T2D)”.
H_2 : "Parental diabetes history increased the likelihood of the incidence of type 2 diabetes (T2D)".
H_3 : "Anxiety management contributes to decreasing the likelihood of type 2 diabetes (T2D)".
H_4 : "Having a sufficient sleep duration would minimize the likelihood of the incidence of type 2 diabetes (T2D)".
H_5 : "Smoking contributes to increasing the likelihood of type 2 diabetes (T2D).".
H_6 : “ Stress contributes to increase the likelihood of type 2 diabetes (T2D)”.

DOI: 10.17148/IJARCCE.2023.12657

III. METHODOLOGY

A. Sampling Method

The target population is all adults who are citizens or residents of the Kingdom of Saudi Arabia. The target population was identified in this research to predict the likelihood of having T2D. A convenience sampling method, non-probability sampling, was chosen to be representative of the target population [35]. A convenience sampling method was used to generate a potential hypothesis for each case in the population [35]. Finally, 532 responses were collected.

B. Data Collection

Data collection is a significant research portion. Thus, an online questionnaire was used to collect data related to human behavior and some biometrics that would increase the likelihood of developing T2D among adults in Saudi Arabia. The questionnaire is a research instrument that aims to collect data from target respondents using ethical approval standards to test chosen hypotheses. The questions asked in the questionnaire to collect data were validated and supported by studies. Social media platforms were utilized to collect data to reach the largest possible sample through an online Microsoft Forms questionnaire. The questionnaire contains 26 questions that were categorized into 4 demographic questions, 7 multiple-choice questions, 13 closed-ended questions, and 2 open-ended questions. The questionnaire consists of three main sections.

The collected data were analyzed by using Microsoft Excel to be ready to fit into the machine learning model. Different descriptive and inferential statistical tech-niques were applied, including the average mean, median, mode, and standard deviation that were used in descriptive statistics to provide potential relationships between variables and basic features of variables in the data, while ANOVA tests was used in inferential statistics to explore the relationship strength between selected variables. After comparing several machine learning models in this field as shown in table 1 previously, an intelligent model was developed to predict the likelihood of developing type 2 diabetes through analyzing human behaviors and biometrics by using machine learning platform. A machine learning model was developed by Python using google colab. Different supervised machine learning algorithms were applied to prediction process, such as Support Vector Machine, Logistic Regression, K-Nearest Neighbor Algorithm, Decision Tree algorithm, and Random Forest classification.

IV. RESULT AND ANALYSIS

A. Demographics

This section of the questionnaire sought to collect demographic data on the research sample to obtain a clear background on the participants' gender, age, weight, and height. In addition, this stage aims to explore the BMI by manually calculating the height and weight.

TABLE 1 Demographics information

Classification	Frequency (N=532)	
	n	%
Gender		
Female	377	71%
Male	155	29%
Age Group		
18-30	212	40%
31-40	147	28%
41-50	114	21%
51-60	50	9%
>60	9	2%
Weight (kg)		
<=40	11	2%
41-50	63	12%
51-60	116	22%
61-70	114	21%
71-80	96	18%
81-90	71	13%
91-100	35	7%
>100	26	5%
Height (cm)		

<=150	44	8%
151-160	219	41%
161-170	186	35%
171-180	67	13%
181-190	16	3%

The presented data above in Table 1 presents the demographic information of the 532 responses. All responses were from Saudi Arabia. This study collected data based on gender, age groups, and BMI that was manually inferred according to weight and height. In addition, some other behaviors and biometrics were collected, which are discussed in detail in the next sections. According to their age, the participants were organized to be classified into five age groups. Participants of the age group from 18 to 30 were clearly dominating the sample with 212 (40%) participations, followed by the 31 to 40 and 41 to 50 groups with 147 (28%) and 114 (21%), respectively. On the other hand, the remaining groups were minority groups with 51 to 60 and more than 60 with 50 (9%) and 5 (2%) participations, respectively.

Regarding participants' weight, the study was organized to be classified into eight groups: more than or equal to 40, 41 to 50, 51 to 60, 61 to 70, 71 to 80, 81 to 90, 91 to 100, and heavier than 100 kilograms. As shown in table 1, the 51 to 60, 61 to 70, and 71 to 80 groups were the majority with 116 (22%), 114 (21%), and 96 (18%), respectively, followed by the 81 to 90 and 41 to 50 groups with 71 (13%) and 63 (12%), respectively. The other groups, 91 to 100, more than 100, and more than or equal to 40, were partially close to each other by 35 (7%), 26 (5%), and 11 (2%).

Also, the height was organized to be classified into five groups: less than or equal to 150, 151 to 160, 161 to 170, 171 to 180, and 181 to 190 centimeters. As presented in Table 1, the 151–160 group was the highest at 219 (41%), followed by the 161–170 group at 186 (35%). The remaining groups were relatively close, except for the 181–190 group, which was the minority at 16 (3%).

TABLE 2 The average mean and standard deviation

Variables	Mean	Median	Mode	Std
Parental diabetes history	0.64	1.0	0	0.47
Physical exercises	0.61	1.0	1.0	0.48
Anxiety-management	0.76	1.0	1.0	0.42
Sleep duration	0.55	0	0	0.47
Smoking	0.13	0	0	0.34
Stress	1.14	1.0	2.0	0.80

Table 2 shown above also presents a parental diabetes history was with means and standard deviations of 0.64 and 0.47, respectively. The Physical exercises was with means and standard deviations of 0.61 and 0.48, respectively. While the anxiety-management was with means and standard deviations of 0.76 and 0.42.The sleep duration was with means and standard deviations of 0.55 and 0.47. In addition to, smoking was with means and standard deviations of 0.13 and 0.34, respectively. Furthermore, the stress was with means and standard deviations of 1.14 and 0.80.

The collected data was organized and classified into four ranges of BMI, as presented in Table 3. In Table 3 below, there are four ranges of BMI: underweight, normal, overweight, and obesity. If the BMI is less than 18.5 kilograms, it is considered underweight. If the BMI is between 18.5 and 24.9 kilograms, it is considered to be in the normal weight range. If the BMI is between 25 and 29.9 kilograms, it is considered overweight. If the BMI is between 30 kilograms and higher, it is considered to be in the obesity range. The weight of 51.7 kilograms, was highly dominant compared with other ranges, which is within the status of obesity and clearly increases the likelihood of developing type 2 diabetes, as proven in previous studies [36]–[38], followed by the majority of values that were more or less close to each other and within the status of normal, overweight, or obesity.

TABLE 3 Range of BMI

Status	Range of BMI values (Kg)
Underweight	<= 18.5
Normal	<=24.9
Overweight	<+29.9
Obesity	>=30

B. Medical Records

Discovering the medical history of participants was the second stage of the questionnaire, which focused on exploring the status of most chronic diseases, especially type 2 diabetes. Most of the respondents 388 participants (73%) did not suffer from any chronic disease. On the other hand, 144 participants (27%) in the group who suffered from a chronic disease were asked to specify it. Followed by the other group of participants who suffered from more than one chronic disease, such as arthritis, thyroid disease, cholesterol, immune thrombocytopenia, and others.

C. Behaviors

Behavioral analysis was the third stage of the questionnaire, which aimed to capture common behaviors that would contribute to increasing or decreasing the likelihood of developing type 2 diabetes. This section focused on physical exercises, duration and quality of sleep, anxiety, stress, mood swings, smoking, eating unhealthy food, drinking water, pressures, anger, and depression, as the literature highlights their possible effect on developing type 2 diabetes. Participants were asked during this section of the questionnaire whether they did some or all of the specific behaviors and noticed any changes. More than half of the sample were doing their physical exercise, with 327 participants (61%), while 205 participants (39%) didn't do exercise anyway. The frequency of physical exercise among participants. In a relatively large sample, 189 (42%) preferred to do physical exercise rarely, 111 of participants (24%) would rather do physical exercise 3-5 times a week, and 100 of participants (22%) did physical exercise once a week. While 54 of the participants (12%) did daily physical exercise. This finding indicates that the majority of participants have sufficient awareness of the importance of physical exercise, which contributes to the prevention of diseases such as diabetes, heart disease, and others.

The following subsections discuss the findings of the explored behaviors.

Anxiety-management

The study participants were asked when they felt anxious about something and felt physically tired or lethargic. 406 participants (76%) actually feel lethargic and tired when anxious, while 126 participants (24%) do not have these symptoms.

Sleep Disorder and Duration

Suffering from sleep disorders was taken into consideration among the behaviors that increase the likelihood of developing T2D. More than half of the sample's 325 participants (61%) suffered from sleep disturbances, whereas 207 participants (39%) did not suffer from sleep disorders. After that, they were asked in detail regarding their sleep duration. The sample was divided into three groups, which are less than 6 hours, 6 to 8 hours, and more than 8 hours. The 6 to 8 hours group was the top with 291 participants (55%), followed by the less than 6 hours group with 187 participants (35%), and then the more than 8 hours group was the lowest with 57 participants (10%). Furthermore, the participants were asked, based on the number of hours they slept, whether they felt lethargic and physically tired when they did some simple tasks the next day. The majority of the sample said "yes" with 342 participants (64%), unlike 190 participants (36%), who said "no.".

Mood Swing

Coupled with these behaviors are the mood swings. When the participants were asked whether they have rapid and severe mood swings, the highest rate of participants (360, or 68%) said "yes,", while 172, or 32%) said "no.".

Smoking

In addition to smoking, the results of the collected data were relatively positive: only 71 participants (13%) were smokers, the number of smoking times of the participants who smoke daily was 58 (64%), followed by the participants who smoke 3 to 5 times a week: 7 participants (8%), who smoke once a week: 22 participants (24%), and who smoke twice a week: 4 participants (4%).

Unhealthy Diet

An unhealthy diet was among the behaviors that were focused on. Asked whether they followed an unhealthy diet, the participants were mostly balanced: 270 participants (51%) said "yes" and 262 participants (49%) said "No". Also, the participants were asked about the rate of their unhealthy food intake. Participants who eat unhealthy food once a week were the highest of the sample with 188 participants (35%), followed by those who eat unhealthy food 3 to 5 times a week with 142 participants (27%). Then, those who eat unhealthy food on a daily basis had 96 participants (18%), while those who eat unhealthy food once a month came in second with 65 participants (12%), and the lowest were those who do not eat unhealthy food with 41 participants (8%).

DOI: 10.17148/IJARCCE.2023.12657

Drinking Water

Besides this, the participants were asked regarding about rate of drink water, the water drinking rate of the majority was less than 8 glasses of water with 418 participants (79%), while the rest of sample 114 participants (21%) were drink 8 or more glasses of water.

Stress

According to the literature, work pressure is one of the behaviors that could have a negative impact and lead to type 2 diabetes [43]. Participants were asked whether they faced pressure in their work. A large group of the participants, with 215 (40%) did not work anyway, followed by the group of participants with 177 (33%) who faced pressure, and the last group of participants with 140 (26%) did not face any pressure.

As well as that depression, where the participants were asked whether they suffer from it, the result was positive: the majority of 407 participants (77%) do not suffer, while the minority of 125 participants (23%) do.

D. ANOVA Test

ANOVA is an abbreviation for "analysis of variance," and it is a statistical technique used for testing whether the model is significant or not.

TABLE 4 Regression output and confidence interval

Variables	coefficients	std. error	t (df=525)	p-value	95% lower	95% upper
Intercept	31.2203					
Parental diabetes history	0.9045	0.5359	1.688	0.0920	-0.1482	1.9572
Physical exercises	-1.0926	0.5274	-2.072	0.0388	-2.1288	-0.0565
Anxiety management	0.0503	0.6008	0.084	0.9333	-1.1300	1.2306
Sleep duration	-1.5674	0.4111	-3.812	0.0002	-2.3751	-0.7597
Smoking	-0.2120	0.7561	-0.280	0.7793	-1.6974	1.2734
Stress	-0.7446	0.3194	-2.331	0.0201	-1.3721	-0.1171

From table 4, there is no statistically significant effect of parental diabetes history on BMI (P = 0.0920); it is more than α = 0.05.

There is a statistically significant effect of physical exercise on BMI where P = 0.0388 is less than α = 0.05 and looking at the value of the coefficient = -1.0926, the greater the physical exercise by 1 unit, the more BMI decreases by 1.0926.

There is a statistically significant effect of sleep duration on BMI, where P = 0.0002 is less than α = 0.05 and looking at the value of the coefficient = -1.5674, the greater the sleep duration by 1 unit, the more BMI decreases by 1.5674.

There is a statistically significant effect of stress on BMI, where P = 0.0201 is less than α = 0.05 and looking at the value of the coefficient = -0.7446, the greater the stress by 1 unit, the more BMI decreases by 0.7446.

There is no statistically significant effect of Anxiety management on BMI, P = 0.933, it is more than α = 0.05. Also, there is no statistically significant effect of Smoking on BMI, P = 0.7793, it is more than α = 0.05.

E. Evaluation metrics

Table 5 depicts different evaluation metrics, which are accuracy, recall, and precision. On the other hand, K-Nearest Neighbor, Decision Tree, and Random Forest were moderately precise, scoring 1.0 compared with Support Vector Machine and Logistic Regression.

The statistics computed for the evaluation metrics of the recall score presented comparable trends. For the recall, the values inferred were 0.97, 0.93, 0.86, 0.74, and 0.71 for Random Forest, Decision Tree, K-Nearest Neighbor, Support Vector Machine, and Logistic Regression, respectively. As shown in Table 5, all algorithms show considerable performance.

DOI: 10.17148/IJARCCE.2023.12657

TABLE 5 Evaluation metrics

Classifier	Accuracy	Precision	Recall
Support Vector Machine (SVM)	0.85	0.98	0.74
Logistic Regression (LR)	0.82	0.90	0.71
K-Nearest Neighbour (KNN)	0.92	1.00	0.86
Decision Tree (DT)	0.96	1.00	0.93
Random Forest (RF)	0.98	1.00	0.97

Figure 2 presents the performance of the ML algorithms applied in this research in a chart. Using evaluation criteria, namely accuracy, precision, and recall, one chart has been created to facilitate the comparison between the five models; thus, the performance of algorithms from the highest to the lowest is RF, DT, KNN, SVM, and LR. It was observed that the Random Forest algorithm performs highly accurate behavioral prediction with 98% accuracy compared with other algorithms. Followed by the decision tree and K-nearest neighbor by 96% and 92%, respectively. While the support vector machine and logistic regression achieved 85% and 82%, respectively.

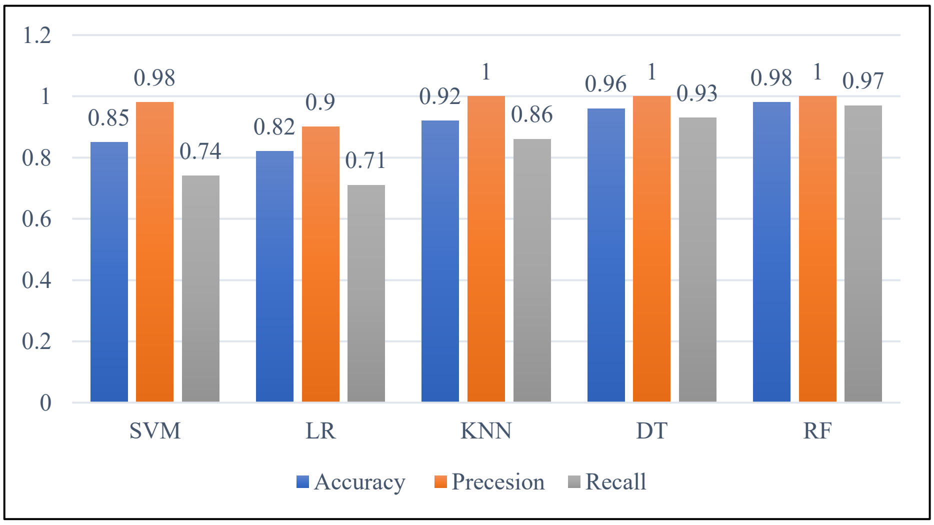

Fig. 2 Performance of all ML algorithms

V. DISSCUSION

The findings clearly present the presence of a U-shaped association between parental diabetes history, practicing wrong behaviors such as eating unhealthy food or sleeping a few hours, feeling anxious and developing the likelihood of T2D. Thus, the observed association between these behaviors and the likelihood of T2D may contributes to early prediction. Based on the regression output and confidence interval of variables that were applied in the ANOVA, there is no statistically significant effect of parental diabetes history, anxiety management, or smoking on BMI. But there is a statistically significant effect of physical exercise, sleep duration, and stress on BMI.

Based on the regression output and confidence interval of variables that were applied in the ANOVA, there is no statistically significant effect of parental diabetes history, anxiety management, or smoking on BMI. But there is a statistically significant effect of physical exercise, sleep duration, and stress on BMI.

By considering the significance level at alpha = 0.05, if the significant value in the ANOVA table (P-value) is less than 0.05, it can be said that the model is significant, while if it is higher than 0.05, it can be said that the model is not significant.

Thus, there is a statistically significant effect of physical exercise on BMI where P = 0.0388 is less than α = 0.05. where more than half of the sample were doing their physical exercise, with 327 participants (61%). Thus, they have a lower likelihood of T2D based on [39], [40]. It is indicated that this hypothesis is an alternative hypothesis and has been accepted.

DOI: 10.17148/IJARCCE.2023.12657

H_{1-1}: "Physical exercises decreased the likelihood of the incidence of type 2 diabetes (T2D)".
On the other hand, there is no statistically significant effect of parental diabetes history on BMI (P = 0.0920), which is more than α = 0.05. It is indicated that this hypothesis is a null hypothesis and should be rejected.

H_{2-0} : "Parental diabetes history increases the likelihood of the incidence of type 2 diabetes (T2D)".
Also, there is no statistically significant effect of anxiety management on BMI (P = 0.933), which is more than α = 0.05. This means this hypothesis is a null hypothesis.

H_{3-0} : "Anxiety management contributes to decreasing the likelihood of type 2 diabetes (T2D)".
There is a statistically significant effect of sleep duration on BMI, where P = 0.0002 is less than α = 0.05. In this aspect, the sample was divided into three groups, which are less than 6 hours, 6 to 8 hours, and more than 8 hours. The 6 to 8 hours group was the top with 291 participants (55%), followed by the less than 6 hours group with 187 participants (35%), and the more than 8 hours group was the lowest with 57 participants (10%).

In addition, the participants were asked, based on the number of hours they slept, whether they felt lethargic and physically tired when they did some simple tasks the next day. The majority of the sample said "yes," with 342 participants (64%). This indicates a relationship between sleep duration and the likelihood of the incidence of type 2 diabetes based on [41], [42]. This indicates that this hypothesis is an alternative hypothesis.

H_{4-1} : "Having a sufficient sleep duration would minimize the likelihood of the incidence of type 2 diabetes (T2D)".
And there is no statistically significant effect of smoking on BMI, P = 0.7793; it is more than α = 0.05, which leads to a null hypothesis being rejected.

H_{5-0} : "Smoking contributes to increasing the likelihood of type 2 diabetes (T2D)".
There is a statistically significant effect of Stress on BMI, where P= 0.0201 is less than α = 0.05. So, according to the literature, work pressure is one of behaviors what could bring negative impact and would lead to type 2 diabetes [43]. Participants were asked whether they faced pressures in their work. A large group of the participants with 215 (40%) did not work anyway, followed by the group of participants with 177 (33%) they face a pressure, and the last group of participants with 140 (26%) did not face any pressure. Which indicates a relationship between Stress and the likelihood of T2D. that means this hypothesis is an alternative hypothesis.

H_{6-1} : "Stress contributes to increasing the likelihood of type 2 diabetes (T2D)".

VI. CONCLOSION AND FUTURE WORK

This study addresses the problem of the prevalence of type 2 diabetes among adults in Saudi Arabia. Therefore, one of the main contributions of our work was to build an intelligent model that predicts the likelihood of developing T2D by analyzing the behaviors and biometrics of the individual for use in medical practice.

A discussion on different aspects of diabetes causes, and some medical, behavioral, and computing efforts to control it was provided. An approach based on five supervised machine learning algorithms, which are Support Vector Machine, Logistic Regression, K-Nearest Neighbour, Decision Tree, and Random Forest, was introduced to provide prediction of the likelihood of developing type 2 diabetes among adults based on the given information (behaviors and biometrics).

Thus, the random forest algorithm outperformed in perform highly accurate behavioral prediction with 98%. Finally, ANOVA was conducted to measure the relationship between the selected factors.

The following ideas can be tested:

It would be interesting to consider more factors in the model with different concerns, such as environmental factors that may increase the likelihood of developing T2D. This idea would, for instance, aid in deep prediction through the use of factors related to this disease, whose effect is stronger and clearer on human health; thus, prediction will be more accurate.

It is tempting to choose optimization algorithms (e.g., gradient descent or stochastic gradient descent) to optimize the performance of a model. It would also be important to conduct the study on a larger sample thus, its help researchers to detect the outliers in data and provide less margins of error.

DOI: 10.17148/IJARCCE.2023.12657

REFERENCES

[1] M. Abu-Farha, J. Tuomilehto, and J. Abubaker, ‘Editorial: Diabetes in the Middle East’, Front. Endocrinol. (Lausanne)., vol. 12, no. February, pp. 10–12, 2021, doi: 10.3389/fendo.2021.638653.

[2] J. Liu et al., ‘Trends in the incidence of diabetes mellitus: results from the Global Burden of Disease Study 2017 and implications for diabetes mellitus prevention’, BMC Public Health, vol. 20, no. 1, pp. 1–12, 2020, doi: 10.1186/s12889-020-09502-x.

[3] T. Mahboob Alam et al., ‘A model for early prediction of diabetes’, Informatics Med. Unlocked, vol. 16, no. July, p. 100204, 2019, doi: 10.1016/j.imu.2019.100204.

[4] H. Kaur and V. Kumari, ‘Predictive modelling and analytics for diabetes using a machine learning approach’, Appl. Comput. Informatics, 2019, doi: 10.1016/j.aci.2018.12.004.

[5] U. A. Zia and N. Khan, ‘Predicting Diabetes in Medical Datasets Using Machine Learning Techniques’, Int. J. Sci. Res. Eng. Trends, vol. 5, no. 2, pp. 1538–1551, 2019, [Online]. Available: http://www.ijser.org.

[6] M. F. I. L. Bin Abdullah et al., ‘How Much Do We Know about the Biopsychosocial Predictors of Glycaemic Control? Age and Clinical Factors Predict Glycaemic Control, but Psychological Factors Do Not’, J. Diabetes Res., vol. 2020, 2020, doi: 10.1155/2020/2654208.

[7] D. J. Hemanth, Human behaviour analysis using intelligent systems, vol. 6. Springer International Publishing, 2020.

[8] ‘What is diabetes? | CDC’. https://www.cdc.gov/diabetes/basics/diabetes.html (accessed Oct. 02, 2021).

[9] H. D. McIntyre, P. Catalano, C. Zhang, G. Desoye, E. R. Mathiesen, and P. Damm, ‘Gestational diabetes mellitus’, Nat. Rev. Dis. Prim., vol. 5, no. 1, 2019, doi: 10.1038/s41572-019-0098-8.

[10] H. Lee et al., ‘Beta Cell Dedifferentiation Induced by IRE1α Deletion Prevents Type 1 Diabetes’, Cell Metab., vol. 31, no. 4, pp. 822-836.e5, 2020, doi: 10.1016/j.cmet.2020.03.002.

[11] E. J. Mayer-Davis et al., ‘Incidence Trends of Type 1 and Type 2 Diabetes among Youths, 2002–2012’, N. Engl. J. Med., vol. 376, no. 15, pp. 1419–1429, 2017, doi: 10.1056/nejmoa1610187.

[12] J. N. Harvey, R. Hibbs, M. J. Maguire, H. O’Connell, and J. W. Gregory, ‘The changing incidence of childhood-onset type 1 diabetes in Wales: Effect of gender and season at diagnosis and birth’, Diabetes Res. Clin. Pract., vol. 175, p. 108739, 2021, doi: 10.1016/j.diabres.2021.108739.

[13] S. Rathod, ‘Novel Insights into the Immunotherapy-Based Treatment Strategy for Autoimmune Type 1 Diabetes’, Diabetology, vol. 3, no. 1, pp. 79–96, 2022, doi: 10.3390/diabetology3010007.

[14] A. Artasensi, A. Pedretti, G. Vistoli, and L. Fumagalli, ‘Type 2 diabetes mellitus: A review of multi-target drugs’, Molecules, vol. 25, no. 8, pp. 1–20, 2020, doi: 10.3390/molecules25081987.

[15] A. Berbudi, N. Rahmadika, A. I. Tjahjadi, and R. Ruslami, ‘Type 2 Diabetes and its Impact on the Immune System’, pp. 442–449, 2020, doi: 10.2174/1573399815666191024085838.

[16] N. G. Forouhi and N. J. Wareham, ‘Epidemiology of diabetes’, Med. (United Kingdom), vol. 47, no. 1, pp. 22–27, 2019, doi: 10.1016/j.mpmed.2018.10.004.

[17] G. Twig et al., ‘Adolescent obesity and early-onset type 2 diabetes’, Diabetes Care, vol. 43, no. 7, pp. 1487–1495, 2020, doi: 10.2337/dc19-1988.

[18] H. Shakoor et al., ‘Effect of Calorie Restriction and Exercise on Type 2 Diabetes’, Prilozi, vol. 42, no. 1, pp. 109–126, 2021, doi: 10.2478/prilozi-2021-0010.

[19] M. Abdul et al., ‘Epidemiology of Type 2 Diabetes – Global Burden of Disease and Forecasted Trends’, J. Epidemiol. Glob. Health, vol. 10, pp. 107–111, 2020.

[20] S. I. Asahara, H. Inoue, and Y. Kido, ‘Regulation of Pancreatic β-Cell Mass by Gene-Environment Interaction’, Diabetes Metab. J., vol. 46, no. 1, pp. 38–48, 2022, doi: 10.4093/DMJ.2021.0045.

[21] X. Liu et al., ‘A Novel Risk Score for Type 2 Diabetes Containing Sleep Duration: A 7-Year Prospective Cohort Study among Chinese Participants’, J. Diabetes Res., vol. 2020, 2020, doi: 10.1155/2020/2969105.

[22] S. S. Bhat and G. A. Ansari, ‘Predictions of diabetes and diet recommendation system for diabetic patients using machine learning techniques’, 2021 2nd Int. Conf. Emerg. Technol. INCET 2021, no. May, 2021, doi: 10.1109/INCET51464.2021.9456365.

[23] N. Bharti, ‘Linking human behaviors and infectious diseases’, Proc. Natl. Acad. Sci. U. S. A., vol. 118, no. 11, pp. 3–5, 2021, doi: 10.1073/pnas.2101345118.

[24] H. Bekele, A. Asefa, B. Getachew, and A. M. Belete, ‘Barriers and Strategies to Lifestyle and Dietary Pattern Interventions for Prevention and Management of TYPE-2 Diabetes in Africa, Systematic Review’, J. Diabetes Res., vol. 2020, 2020, doi: 10.1155/2020/7948712.

[25] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, ‘Machine Learning and Data Mining Methods in Diabetes Research’, Comput. Struct. Biotechnol. J., vol. 15, pp. 104–116, 2017, doi: 10.1016/j.csbj.2016.12.005.

[26] H. M. Deberneh and I. Kim, ‘Prediction of type 2 diabetes based on machine learning algorithm’, Int. J. Environ. Res. Public Health, vol. 18, no. 6, 2021, doi: 10.3390/ijerph18063317.

DOI: 10.17148/IJARCCE.2023.12657

[27] M. Shokrekhodaei, D. P. Cistola, R. C. Roberts, and S. Quinones, ‘Non-Invasive Glucose Monitoring Using Optical Sensor and Machine Learning Techniques for Diabetes Applications’, IEEE Access, vol. 9, pp. 73029–73045, 2021, doi: 10.1109/ACCESS.2021.3079182.

[28] A. H. Syed and T. Khan, ‘Machine learning-based application for predicting risk of type 2 diabetes mellitus (t2dm) in saudi arabia: A retrospective cross-sectional study’, IEEE Access, vol. 8, pp. 199539–199561, 2020, doi: 10.1109/ACCESS.2020.3035026.

[29] T. Daghistani and R. Alshammari, ‘Comparison of statistical logistic regression and randomforest machine learning techniques in predicting diabetes’, J. Adv. Inf. Technol., vol. 11, no. 2, pp. 78–83, 2020, doi: 10.12720/jait.11.2.78-83.

[30] S. Lee et al., ‘Glycemic and lipid variability for predicting complications and mortality in diabetes mellitus using machine learning’, BMC Endocr. Disord., vol. 21, no. 1, pp. 1–15, 2021, doi: 10.1186/s12902-021-00751-4.

[31] J. Li et al., ‘A tongue features fusion approach to predicting prediabetes and diabetes with machine learning’, J. Biomed. Inform., vol. 115, no. February, p. 103693, 2021, doi: 10.1016/j.jbi.2021.103693.

[32] T. Tagami, ‘An overview of thyroid function tests in subjects with resistance to thyroid hormone and related disorders’, Endocr. J., vol. 68, no. 5, pp. 509–517, 2021, doi: 10.1507/endocrj.EJ21-0059.

[33] I. C. Mason and G. K. Adler, ‘Impact of circadian disruption on glucose metabolism : implications for type 2 diabetes Chronobiology terminology’, pp. 462–472, 2020.

[34] Y. T. Wu et al., ‘Early Prediction of Gestational Diabetes Mellitus in the Chinese Population via Advanced Machine Learning’, J. Clin. Endocrinol. Metab., vol. 106, no. 3, pp. E1191–E1205, 2021, doi: 10.1210/clinem/dgaa899.

[35] S. J. Stratton, ‘Population Research: Convenience Sampling Strategies’, Prehosp. Disaster Med., vol. 36, no. 4, pp. 373–374, 2021, doi: 10.1017/S1049023X21000649.

[36] E. A. Silveira, L. P. de S. Rosa, A. S. e. A. de C. Santos, C. K. de S. Cardoso, and M. Noll, ‘Type 2 diabetes mellitus in class II and III obesity: Prevalence, associated factors, and correlation between glycemic parameters and body mass index’, Int. J. Environ. Res. Public Health, vol. 17, no. 11, pp. 1–13, 2020, doi: 10.3390/ijerph17113930.

[37] L. Ismail, H. Materwala, and J. Al Kaabi, ‘Association of risk factors with type 2 diabetes: A systematic review’, Comput. Struct. Biotechnol. J., vol. 19, pp. 1759–1785, 2021, doi: 10.1016/j.csbj.2021.03.003.

[38] J. Berumen et al., ‘Influence of obesity, parental history of diabetes, and genes in type 2 diabetes: A case-control study’, Sci. Rep., vol. 9, no. 1, pp. 1–15, 2019, doi: 10.1038/s41598-019-39145-x.

[39] W. Zhu, ‘Exercise is medicine for type 2 diabetes: An interview with Dr. Sheri R. Colberg’, J. Sport Heal. Sci., vol. 11, no. 2, pp. 179–183, 2022, doi: 10.1016/j.jshs.2022.01.006.

[40] M. Savikj and J. R. Zierath, ‘Train like an athlete: applying exercise interventions to manage type 2 diabetes’, Diabetologia, vol. 63, no. 8, pp. 1491–1499, 2020, doi: 10.1007/s00125-020-05166-9.

[41] M. Y. Baden, F. B. Hu, C. Vetter, E. Schernhammer, S. Redline, and T. Huang, ‘Sleep duration patterns in early to middle adulthood and subsequent risk of type 2 diabetes in women’, Diabetes Care, vol. 43, no. 6, pp. 1219–1226, 2020, doi: 10.2337/dc19-2371.

[42] ‘Diabetes and Sleep: Sleep Disturbances & Coping | Sleep Foundation’. https://www.sleepfoundation.org/physical-health/lack-of-sleep-and-diabetes (accessed Sep. 24, 2022).

[43] R. R. Ahmadi, Z. Majdi, M. Y. Shokouh, N. Masihipour, and S. A. Hos, ‘Clinical Studies & Medical Case Reports Review Article Psychological Stress and Type 2 Diabetes : A Review of the Bidirectional Link in the Onset and Progression of the Disease’, pp. 1–13, 2022, doi: 10.46998/IJCMCR.2022.20.000477.