Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 09, No. 04, December 2023.

AUTOMATED DIABETES DISEASE PREDICTION SYSTEM BASED ON RISK FACTORS ASSESSMENT: TAKING CHARGE OF YOUR HEALTH

Nawal Sad-Houari[1], Hicham Reguieg[2], Chaimaa Bachiri[3] and Marwa Alioua[3]

(Received: 7-Sep.-2023, Revised: 4-Nov.-2023, Accepted: 18-Nov.-2023)

ABSTRACT

Diabetes is one of the most common diseases worldwide and its prevalence rate continues to rise. This increase is due to factors related to nutrition and lifestyle on the one hand and to genetic factors on the other hand, thus creating a real public-health problem. Therefore, it is crucial to identify diabetes early in order to allow rapid treatment, capable of slowing down the progression of the disease.

The objective of this work is to propose an automatic diabetes-prediction system based on the following machine-learning techniques: SVM, KNN, Decision Tree and Logistic Regression. Using risk factors specific to the Algerian environment, we constructed a new dataset that includes 823 patients, with 418 being diabetic and 405 being non-diabetic. In order to choose the relevant features and identify the most informative risk factors, we combined several feature-extraction methods, such as ANalysis of Variance (ANOVA) and Recursive Feature Elimination (RFE) and we used also the features proposed by the Pima Indian Diabetes Dataset (PIDD).

The results of this study provided valuable information on the comparative performance of different machine-learning models in the prediction of diabetes, as well as on the importance of the selected characteristics.

KEYWORDS

ANOVA, Diabetes, Feature extraction, Machine learning, Patients, Prediction, RFE.

1. INTRODUCTION

Artificial Intelligence (AI) has been integrated into the health field to improve the efficiency and quality of care [1]. It aims to create systems that can rival or even surpass human intellectual capacities. Thus, AI offers new perspectives for the optimization of medical practices and the well-being of patients. Techno- logical advancements have enabled the use of AI algorithms in many medical applications, covering areas, such as diagnosis, therapeutic decision-making, disease prediction, medical robotics, …etc. AI continues to evolve and develop, opening new perspectives to improve medical practices and patient well-being.

Currently, the health sector is facing various challenges in the diagnosis and treatment of diabetes, especially with regard to the increasing prevalence of this disease, which affects a large number of people of all ages and from various ethnic backgrounds [2]-[3]. According to statistics from the World Health Organization (WHO), approximately 463 million adults had diabetes in 2019, which corresponds to a global prevalence of 9.3%. In Algeria, the prevalence of diabetes continues to increase to reach 14.4% of the population between the ages of 18 and 69 years, which is equivalent to approximately 4 million people with diabetes in 2018 [4]. These statistics highlight the magnitude of the problem of diabetes and accentuate the importance of finding effective solutions for early detection, prevention and management of this disease.

In this context, early detection of diabetes can be a challenge due to the complexity of risk factors and underlying patterns in health data. Traditional screening and diagnostic approaches may have limitations in terms of accuracy and efficiency. Two cases of misdiagnosis of diabetes can be identified. The first case involves false negatives, where an individual with symptoms or risk factors

1. N. Sad-Houari is with Laboratoire d'Informatique Oran (LIO), Département du Vivant et de l'Environnement, Faculté des Sciences de la Nature et de la Vie, Université des Sciences et de la Technologie d'Oran Mohamed Boudiaf, USTO-MB, BP 1505, El M'naouer, Oran, 31000, Algeria. Email: nawal.sadhouari@univ-usto.dz
2. H. Reguieg is with SIMPA Laboratory, Department of Computer Science, Faculty of Mathematics and Computer Science, University of Science and Technology of Oran Mohamed Boudiaf, Algeria.
3. C. Bachiri and M. Alioua are with University of Science and Technology of Oran Mohamed Boudiaf, Algeria. Emails: bachirichaima656@gmail.com and marwa.alioua.me@gmail.com

for diabetes is not diagnosed as having diabetes. The second case involves false positives, where a person is misdiagnosed as diabetic when he/she does not actually have the disease. Indeed, automated diabetes-prediction systems using artificial intelligence and machine or deep learning, can analyze large amounts of health data to identify early indicators of diabetes. Our problematic consists of the following questions:

- What are the most important, influential and relevant Algerian risk factors to take into account to create our diabetes-prediction system?
- Which machine-learning algorithms should be used to obtain the best prediction results?

The objective of this paper is to develop a diabetes-prediction system using machine learning, to improve early detection and prediction accuracy. We seek to exploit machine-learning models, such as Support Vector Machine (SVM), Logistic Regression (LG), Decision Trees (DTs) and K-Nearest Neighbors (KNN). In terms of feature selection, we will use statistical methods, such as ANalysis of Variance (ANOVA), Recursive Feature Elimination (RFE) and comparison with the PIDD dataset to identify the most informative variables and reduce the dimensionality of the data. Data used in this study will include clinical measures, such as fasting blood sugar, Body Mass Index (BMI), blood pressure, cholesterol level, as well as information on the lifestyle and medical history of patients.

The article is organized as follows: Section 2 provides an overview of diabetes. Section 3 presents some related works and highlights the proposed contributions. In Section 4, our proposed approach is explained in detail. This section is followed by a discussion of the obtained experimental results in Section 5. Finally, Section 6 provides the conclusion of this paper as well as perspectives.

2. DIABETES

Diabetes is a chronic disease caused by genetic or acquired defects in the production of insulin by the pancreas or the fact that this insulin is not active enough [5]-[6]. This defect leads to an increase in blood sugar. The particularity of this disease is that it often progresses silently, without developing symptoms [7]-[8].

Insulin is a hormone produced by the pancreas that regulates blood-sugar levels in the body [9]. Our body converts the food that we eat into a form of sugar called glucose and insulin transports it from the bloodstream into cells where it can be used for energy [10]. Insulin also helps store glucose in the liver and muscle tissue for later use. Without enough insulin, glucose builds up in the bloodstream, leading to high blood sugar, a condition known as diabetes. Insulin therapy is a common treatment for diabetes, helping regulate blood-sugar levels by increasing glucose uptake into cells [11]-[12]. There are several types of diabetes, that are: type-1 diabetes, type-2 diabetes and gestational diabetes [13]-[14]. There is no specific and exclusive cause for diabetes, but there are a set of contributing factors that can affect everyone, including [15]-[16]:

- Genetics: A family history of diabetes may increase the risk of developing the disease,
- Overweight and obesity: Excess weight, especially around the waist, can increase the risk of developing type-2 diabetes.
- Lack of physical activity: A sedentary lifestyle can increase the risk of developing type-2 diabetes,
- Unhealthy diet: Consuming a diet that is high in processed foods and added sugars may increase the risk of developing type-2 diabetes.
- Aging: The risk of developing type-2 diabetes increases as people age,
- Certain medical conditions: Such as polycystic ovary syndrome or a history of gestational diabetes,
- Certain medications: Such as steroids, some antidepressants and antipsychotics can develop diabetes,
- Pancreas damage: Damage to the pancreas caused by alcohol abuse or traumatic injury can lead to diabetes.

The symptoms of diabetes can include excessive thirst, frequent urination, fatigue or lack of energy, frequent hunger, …etc. [17][18][19]. We note that there are also differences depending on the type of diabetes. Diabetes disease must be treated urgently because, it can cause severe damage to vital

organs, such as the heart, kidneys, blood vessels, eyes and nerves [20][21][22][23][24][25][26].

3. RELATED WORKS

In the literature, several works have contributed to the creation of automatic systems for diabetes detection. In the following, we will present the most relevant works.

The work presented in [27] proposed a dynamic prediction system of glycemia by the use of deep-learning multi-series. The work proposed MT-LSTM (Multi-Time-series Long Short-Term Memory) for accurate and efficient prediction of blood-glucose concentration in 112 users. The proposed model takes current readings from Continuous Glucose Monitoring (CGM) devices and a few external factors (meals, medications, insulin, physical activity and quality of sleep) as inputs and gets as output the personalized glucose concentration within the next hour.

In the work of [28], a blood-glucose prediction system has been proposed for patients with type-1 diabetes based on deep convolutional neural network (CNN). To do this, the following steps were followed: (1) Data collection from 6 patients with type-1 diabetes and wearing Continuous Glucose Monitoring to capture data every 5 minutes for 8 weeks, (2) Treatment of missing data, (3) Training of the model and (4) Evaluation.

In [29], a medical decision-support system has been proposed to predict the diabetes disease based on machine-learning and deep learning techniques. For machine learning, Support Vector Machine (SVM) and Random Forest (RF) were used and for deep learning, the authors used Convolutional Neural Network (CNN).

The objective of the study presented in [30] is to propose a predictive model for three major diabetes-related complications in Indonesia and identify the significant risk factors related with them. To do this, the authors followed three steps, that are: data pre-processing, data-mining analysis and rule generation. The machine-learning algorithms used were: Naive Bayes Tree, C4.5 decision tree and k-means.

The objective of the work presented in [40] is to propose a diabetes-prediction pipeline model for a better classification of this disease which includes few external factors responsible for diabetes as well as regular factors, such as glucose, body mass index, age, insulin, …etc. Classification accuracy is improved with the use of a new dataset compared to the existing dataset. The authors followed 4 steps that are: (1) Data collection from diabetic and non-diabetic patients, (2) Grouping patients into two classes (diabetic and non-diabetic) with the K-means algorithm, (3) Model creation with the following machine-learning algorithms: Support Vector Classifier, Random Forest, Decision Tree, Extremely Random Tree Classifier, AdaBoost Boosting Algorithm, Perceptron, Linear Discriminant Analysis Algorithm, Logistic Regression, K-Nearest Neighbors Classifier, Naive Gaussian Bayes, Bagging Algorithm, Gradient Boosting, Support Vector Linear Classification and (4) Evaluation of the model.

The work presented in [43] focused on the early prediction of type-2 diabetes with the use of effective techniques to reduce the mortality rate, using unsupervised deep neural networks. The authors used F1-score feature selection on the Pima Indian Dataset after the pre-processing phase to reduce noise and empty entries.

In the work of [33], a system has been proposed to explore predictive analytics for early detection of diabetes using several machine-learning algorithms, such as K-Nearest Neighbors, Logistic Regression, Support Vector Machine, Naive Bayesian, Random Forest, Decision Tree on Pima Indian Dataset which contains 768 patients from an Indian population.

In order to identify misdiagnosed type-1 diabetes patients from patients with a prior type-2 diabetes diagnosis, the XGBoost machine learning model has been used in [34]. The model utilized a total of 932 variables, which were initially broken down into various metrics for prediction. Following recursive feature elimination (RFE), the model achieved an optimal balance between precision and complexity by selecting the top 250 variables. These 250 variables were then used to generate the model's outputs.

In order to predict *Diabetes mellitus* in Nigeria at an early stage, a new approach has been proposed

in [35]. To accomplish their objective, authors collect data *via* oral interviews with patients and performed pre-processing and transformation using label encoder and standard scaler, respectively. After that, three supervised learning algorithms (K-Nearest Neighbors, Decision Trees and Artificial Neural Networks) were applied on nine attributes in order to find the best one.

The aim of the work presented in [36] is to employ several machine-learning techniques in order to explore age adaptation in predicting the risk of *Diabetes mellitus*. The used machine-learning algorithms are: Linear Regression, Logistic Regression, Polynomial Regression, Neural Network, Support Vector Machines, Random Forest and XGboost. To examine the impact of age, the authors divided the dataset into six age groups, evenly distributed, where individuals aged 81 were excluded from the analysis.

The authors in [37] derived several datasets from PIMA dataset by applying various pre-processing techniques in order to implement four machine-learning algorithms for diabetes prediction. The used algorithms are: K-Nearest Neighbors, Decision Tree, Random Forest and Support Vector Machines.

Given the advantages offered by ontologies, the authors of [39] proposed an approach based on ontologies and machine-learning algorithms. After pre-processing and cleaning the PIDD dataset, six machine-learning algorithms; namely, SVM, KNN, ANN, Naive Bayes, Logistic Regression and Decision Tree, were tested in order to choose the best one. In parallel, they extracted generated rules from the Decision Tree algorithm and imported them to Protegé using the SWRLTab plugin.

A new dataset was created in [40], where new diabetes risk factors were used. The authors followed five steps that are: Dataset Collection, Data Pre-processing, Clustering, Model Creation and Evaluation. In the Clustering step, the K-means algorithm was applied on the most correlated attributes (glucose and age) in order to classify each patient into diabetic or non-diabetic. For the model creation step, the following algorithms were tested: Support Vector Classifier, Random Forest Classifier, Decision Tree Classifier, Extra Tree Classifier, Ada Boost algorithm, Perceptron, Linear Discriminant Analysis algorithm, Logistic Regression, K-Nearest Neighbour, Gaussian Naïve Bayes, Bagging algorithm and Gradient Boost Classifier.

An ontology-based machine learning was proposed in [41] in order to predict diabetes. To achieve their goal, the authors followed three phases that are: Data pre-processing, Creation of the Decision Tree by using J4.8 classifier algorithm in Weka and finally the creation of the ontology. After creating the ontology and importing the dataset, the SWRL rules were extracted from the Decision Tree for reasoning. Then, Pellet reasoner was applied to infer whether the patient is diabetic or not.

In order to predict diabetes using machine-learning algorithms on Hadoop cluster, a new architecture is proposed in [42]. This architecture includes three modules that are: Diabetes Data Collection Module, Diabetes Prediction Module and Prediction Results Module. The first module is responsible for the collection of data from different hospitals. The second module is composed of different sub-modules that are: Features extraction, Features selection by using Information gain method and Hadoop Cluster that apply several machine-learning algorithms (Neural network, Support vector machine, Decision tree, Naive Bayes and Random forest).

In order to predict gestational *Diabetes mellitus* in Singapore, the authors of [43] combined coalitional game theory concepts with machine learning. Shapley values and the SHapley Additive exPlanations (SHAP) framework was combined with CatBoost tree ensembles for feature selection. Logistic Regression, Support Vector Machine, CatBoost Gradient Boosting and Multilayer Perceptron Artificial Neural Network were applied on the collected dataset.

In the same context, a machine-learning model was proposed in [44] in order to prevent the progression of gestational *Diabetes mellitus* to type-2 diabetes. For this purpose, Shapley values were combined with CatBoost tree ensembles to perform feature selection and Logistic Regression, Support Vector Machine, CatBoost Gradient Boosting and Artificial Neural Network were implemented and tested in order to choose the best algorithm. Table 1 presents a comparison among the presented works in terms of method and data.

Table 1. Comparison among the related works.

Work	Used dataset	Data size	Selected variables	Class	Limits
[27]	New	112	Data from CGM, meal, medication, insulin, physical activity, quality of sleep and glucose	Type-1 Diabete	Number of data
[28]	Ohio T1DM dataset	6	Glucose level, insulin, carbohydrate intake and mass index	/	The high rate of CGM missing values and the training-set length
[29]	PIDD	768	Pregnancies, Age, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Class	Diabetic and Non-diabetic	The performance is adversely affected by missing or unknown values in the datasets
[30]	New	158	Gender, BMI, Family history of diabetes, Blood pressure, duration of diabetes suffers, Blood-glucose level, patient complication diseases	Retinophaty, Neurophaty and Nephrophaty	Number of data
[40]	PIDD	768	Pregnancies, Age, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Class	Diabetic and Non-diabetic	The performance is adversely affected by missing or unknown values in the datasets
[43]	New	800	Age, genetic function, pregnancy number, glucose, blood pressure, skin thickness, BMI, insulin, type of labor and output	Diabetic and Non-diabetic	Number of data
[33]	PIDD	768	Pregnancies, Age, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Class	Diabetic and Non-diabetic	The performance is adversely affected by missing or unknown values in the datasets
[34]	IQVIA Ambulatory Electronic Medical Records	737,776	250	Type-1 and type 2 diabetes	The use of one algorithm
[35]	New	255	Age, sex, number of pregnancies, glucose level, blood pressure level, body mass index, height, weight and how regularly they exercise	Diabetic and Non-diabetic	Number of data
[36]	PIDD	768	Pregnancies, Age, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Class	Diabetic and Non-diabetic	The performance is adversely affected by missing or unknown values in the datasets
[37]	PIDD	768	Pregnancies, Age, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Class	Diabetic and Non-diabetic	The performance is adversely affected by missing or unknown values in the datasets
[39]	PIDD	768	Pregnancies, Age, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Class	Diabetic and Non-diabetic	The performance is adversely affected by missing or unknown values in the datasets
[40]	New	800	Number of Pregnancies, Glucose Level, Blood Pressure, Skin Thickness(mm), Insulin, BMI, Age, Job Type(Office-work/Fieldwork/Mac work) and Outcome	Diabetic and Non-diabetic	Number of data

[41]	PIDD	768	Pregnancies, Age, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Class	Diabetic and Non-diabetic	The performance is adversely affected by missing or unknown values in the datasets
[42]	National Institute of Diabetes	75,664	13 variables	Diabetic and Non-diabetic	The number of node hadoop cluster
[43]	New	909	Prepregnancy obesity, family history of diabetes, previous history of GDM, previous delivery of a macrosomic baby and Indian ethnicity	Diabetic and Non-diabetic	Number of data
[44]	New	561	Demographics, medical or obstetric history, physical measures, lifestyle information and GDM diagnosis	Diabetic and Non-diabetic	Number of data

3.1 Our Contribution

The objective of this work is to propose an automatic diabetes-prediction system, in order to raise patient awareness and avoid serious complications that can occur in the long term. Our system makes it possible to identify people at high risk of developing diabetes in order to put in place early preventive measures and slow down the progression of the disease, which results in the improvement of the quality of life. We can summarize our contribution in the following points:

- The construction of our own dataset from scratch by collecting Algerian data. The goal behind this initiative is to collect the data that influences the diabetes disease in Algeria and to focus on the risk factors of this last. Creating our own dataset was a crucial step, because the publically available datasets don't represent all the important risk factors of diabetes; so, they do not reflect the reality of the disease. For this, we have collected all the features that concern the patient; namely, his/her demographic, anthropometric, genetic, lifestyle, medical and laboratory information.
- The application of pre-processing techniques to improve the quality of the collected medical data,
- Exploratory data analysis with the intention of identifying the most significant features and the correlation between them. To do that, we applied several feature-extraction techniques, such us: ANOVA, RFE and the features proposed by the PIDD dataset.
- Application of several machine-learning algorithms, to compare the obtained results and choose the best among them,
- Drawing possible predictive conclusions in terms of the obtained results and deciding whether the patient is diabetic or non-diabetic.

4. PROPOSED APPROACH

Diabetes mellitus is a major and prevalent pathology that affects many people. Several factors can increase the risk of developing *Diabetes mellitus*, such as age, obesity, physical inactivity, family history, lifestyle, unbalanced diet, high blood pressure, …etc. Figure 1 presents the architecture of our diabetes prediction system.

In order to predict the diabetes disease, we will follow the following steps:

4.1 Data Collection

Data collection is a crucial step in any scientific research study. In our case, we collected real data in Algeria, in the city of Oran, from diabetic and non-diabetic patients, using a questionnaire administered at the Oran University hospital. The purpose was to collect the largest amount of data on these patients.

The questionnaire was validated by experts in the field in order to guarantee its quality and relevance. The questions asked were adapted to the studied population and allowed an in-depth analysis of the various risk factors associated with diabetes. The answers obtained were carefully processed and

used to feed our diabetes-prediction system.

Our internship took place in two departments; namely, the Internal Medicine department reserved for diabetic patients and the Endocrinology department to collect data from non-diabetic patients affected by other conditions that may cause diabetes. On the other hand, we created and shared an online questionnaire *via* social-media networks (Facebook, Twitter, Linkedin, …etc.), to gather more data and enrich our dataset. We have implemented rigorous protocols to guarantee the quality of the collected data, in particular by ensuring the confidentiality and anonymity of patients. Table 2 shows the collected data, with an indication of the start and end period of the collection and the total number of patients taken.

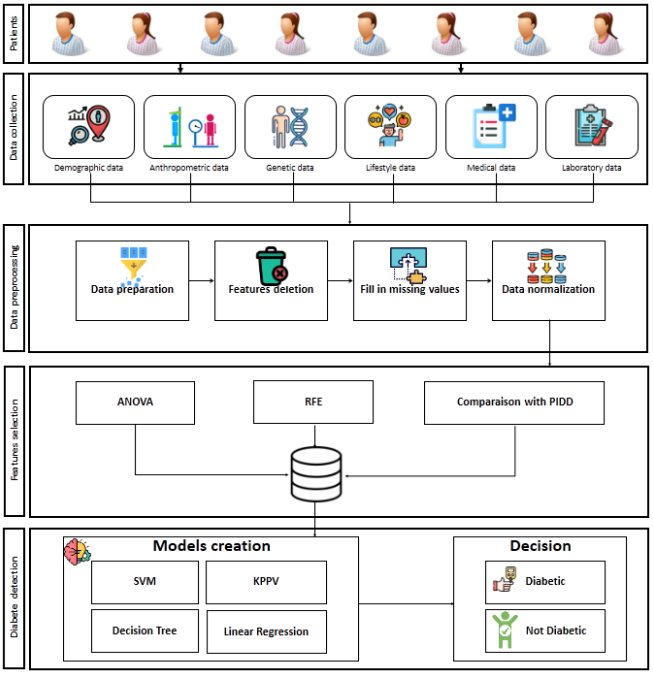

Figure 1. Proposed architecture.

Table 2. Our dataset.

Start date of data collection	End date of data collection	Dataset size	Number of diabetic patients	Number of non-diabetic patients
09 February 2023	17 May 2023	823	418	405

The general population statistics are presented in Tables 3, 4 and 5.

Table 3. Population description.

Features	**Min.**	**Max.**	**Average**	**Mode**	**Median**
Age	16	82	42	21	33
Height	120	190	164	160	165
Weight	44	160	71	65	71

Table 4. Distribution of the population according to sex.

Women	65%
Men	35%

Table 5. Number of individuals with specific diseases.

Disease	**Non-diabetic**	**Diabetic**	**Total**
Hypertension	188	126	314
Thyroid	52	10	62
Heart disorder	74	50	124
Osteoarthritis	34	17	51

Allergies	102	0	102
Bone problem	15	0	15
Cholesterol	25	18	43
Sight problem	56	23	79
Asthma	29	0	29
Anemia	57	36	93
Depression	34	51	85
Kidney problem	27	28	55
Headaches	42	0	42
Weakness	0	36	36

We divided the collected features into 6 categories, that are:

1) **Demographic data:** The patient demographic data includes: city, age, sex, family situation, profession, state of pregnancy, number of pregnancies and number of children,
2) **Anthropometric data:** The patient anthropometric data includes 4 characteristics, that are: size, weight, BMI and obesity.
3) **Genetic data:** The genetic data includes: familial genetic load of diabetes and diabetes family tree function.
4) **Lifestyle data:** The lifestyle data comprises: sport, anxious nature, type of consumed sugar, balanced diet, consumption of fruits and vegetables, consumption of fast food, fasting rhythm, smoking and alcohol.
5) **Medical data:** The medical data encompasses: diagnosis of diabetes during pregnancy, history of health problems, the type of disease, antihypertensive medication, hypoglycemia diagnosis, diabetes, duration, type of diabetes, treatment, health problems and complications after diabetes and the type of complication disease.
6) **Laboratory data:** The data related to patients' blood tests includes 12 features, which are: Blood sugar (in g/l), Total cholesterol (in g/l), Urea (in g/l), HbA1c (in percentage (%)), Triglycerides (in g/l), Creatinine (in mg/l), TGO/ASAT (in U/l), TGP/ALAT (in U/l), TSH/TSHus (in U/l), T4/FT4 (in pg/ml), T3/FT3 (in pg/ml) and Uric Acid (in mg/l).

4.2 Data Pre-processing

After describing all the data that we have collected, we will proceed to data visualization and analyze the different characteristics, then we move on to cleaning and correcting the records to obtain reliable data, ready to be used for diabetes prediction.

The goal of this phase is to improve the quality of the collected data, eliminate errors and deal with inconsistencies and missing data, to make more informed and accurate decisions. In order to prepare our data for training predictive models, we will follow the following steps:

4.2.1 Data Preparation

The first step is the detection and correction of the typing errors, as they are common in medical data. Next, we transformed all the categorical data into numerical data, such as Pregnant = 1 and Not-pregnant = 0 or Single = 0, Married = 1, Divorced = 2 and Widowed = 3. We also removed records that have empty values in most columns.

4.2.2 Feature Deletion

In this step, we decided to delete some column in order to take only the relevant features. For this, we visualized the missing values in our dataset represented by the white color in Figure 2. We notice that the data of the blood tests has a very high rate of missing values, such as FT3, FT4, THus, …etc.

After missing-data visualization, we made a percentage of the missing values for each attribute in order to categorize them into 3 groups, as follows:

- Group 1: Features that have between 0% and 40% missing values,
- Group 2: Features that have between 40% and 80% missing values,
- Group 3: Features that have between 80% and 100% missing values.

After that, we removed group-3 attributes, as they have a large percentage of missing values. We deleted also the irrelevant features, like the city attribute (it is specific to the geographical location and may not provide a direct correlation with the diabetes-prediction study). We removed also height and weight (they have a correlated relationship with the BMI column).

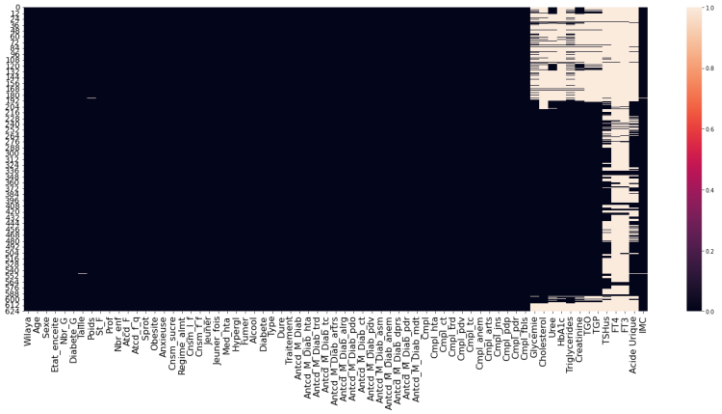

Figure 2. Visualization of missing data.

4.2.3 Fill in Missing Values

To fill in the missing values, we replaced the missing data with the value 0 or the median value according to the attribue. The median consists of replacing missing or outlying values in the data with the median of the entire dataset. It is a robust measure of the central tendency of the data that is not influenced by extreme values or outliers.

4.2.4 Data Normalization

We carried out data normalization to put all the variables on the same scale. This ensures that certain variables do not dominate others when building the model.

4.3 Feature Selection

Feature selection is the process of selecting a sub-set of the most relevant features in the dataset to describe the target variable. It improves computation time, generalization performance and interpretation problems. There are several types of feature-selection techniques, in our case, we used:

4.3.1 Uni-varied Selection

This technique consists in evaluating each characteristic independently of the other characteristics by estimating the correlation or the dependence between its characteristics and the target variable, in order to choose the variables that have the strongest correlation or target dependence. Among these univariate selection techniques, we have the ANOVA method.

ANOVA (ANalysis of Variance) is a well-known statistical method to determine whether there is a difference in means between two groups. In our study, the ANOVA test was used to select the significant numerical characteristics in the prediction of diabetes. The F-statistic is used for feature ranking. The larger the value of the F-statistic, the better the discriminative ability of the feature [38]. The F-value is calculated as follows (see formula 1):

$$F = \frac{SSB}{df_b} \div \frac{SSW}{df_w} \tag{1}$$

where SSB (sum of squares between groups) is the variation of the means of the groups from the total grand mean and SSW (sum of squares within the groups) is the sum of the squared deviations from the means of the groups and at each sighting. The degrees of freedom for the mean square between and within groups are defined by df_b and df_w, respectively. For all numerical features in the dataset, the F-value was calculated using the previous equation 1 and the features with the highest values were selected [38].

Our goal was to assess the importance of each feature in our database based on its contribution to the variance in the data, so that we could use this information to select the most important features. The

best ANOVA selection is 32 out of 68 features (see Figure 3).

The visualization of the selected features and the F-score provide a visual overview of the most important variables for prediction, thereby helping better understand the influence of each feature on the final outcome (see Figure 4).

- Age
- Nbr_G
- Diabete_G
- St_F
- Prof
- Nbr_enf
- Atcd_f_q
- Sprot
- Obesite
- Cnsm_l_f
- Jeuner_fois
- Med_hta
- Hypergl
- Fumer
- Antcd_M_Diab
- Antcd_M_Diab_hta
- Antcd_M_Diab_tc
- Antcd_M_Diab_ct
- Cmpl
- Cmpl_hta
- Cmpl_ct
- Cmpl_tc
- Cmpl_anem
- Cmpl_pdp
- Cmpl_pdr
- Cmpl_fbls
- Glycemie
- HbA1c
- Triglycerides
- Creatinine
- TGP
- Acide Urique

Figure 3. The features selected by ANOVA.

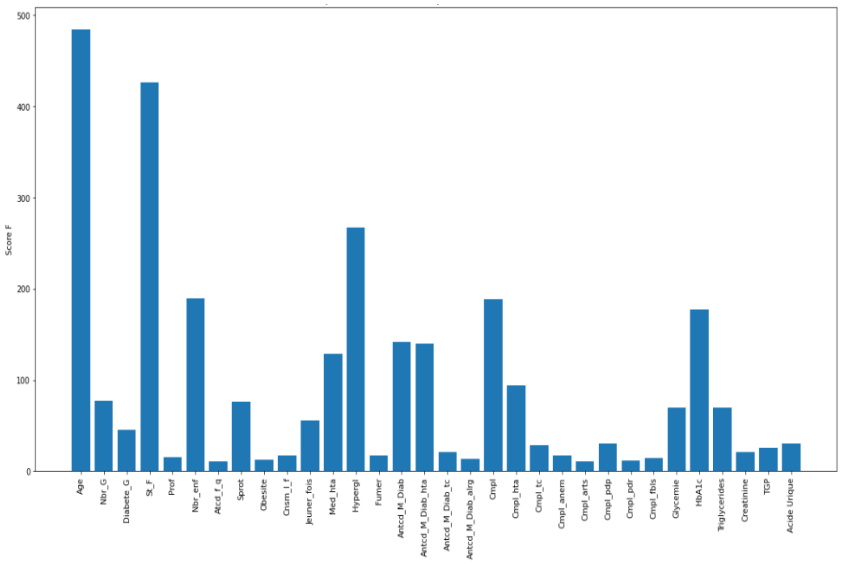

Figure 4. The importance of the features selected by ANOVA and the F-score.

4.3.2 Recursive Feature Elimination (RFE)

Recursive feature elimination is a recursive procedure to select features based on model accuracy. The metric determines the discriminative ability of the features. At each iteration, the rank score metric is calculated and lower-ranked features are eliminated. The recursive procedure is repeated until the desired number of features is reached. In this study, RFE was used as the final step in the feature- selection procedure [38].

We could use recursive feature elimination (RFE) to identify the most important features one by one. We used this technique with the logistic regression algorithm to evaluate their performance on the selected characteristics. According to this method, the optimal number of data selections is 20 among 68 features (see Figure 5).

The visualization of features selected by RFE shows the positive and negative relationships depending on the effect of each variable on the target variable (see Figure 6).

- Etat_enceite
- Diabete_G
- St_F
- Prof
- Atcd_f_q
- Cnsm_sucre
- Cnsm_l_f
- Jeuner_fois
- Hypergl
- Fumer
- Antcd_M_Diab
- Antcd_M_Diab_hta
- Antcd_M_Diab_alrg
- Antcd_M_Diab_asm
- Antcd_M_Diab_pdr
- Antcd_M_Diab_mdt
- Cmpl
- Cmpl_pdp
- Cholesterol
- HbA1c

Figure 5. The features selected by recursive feature elimination (RFE).

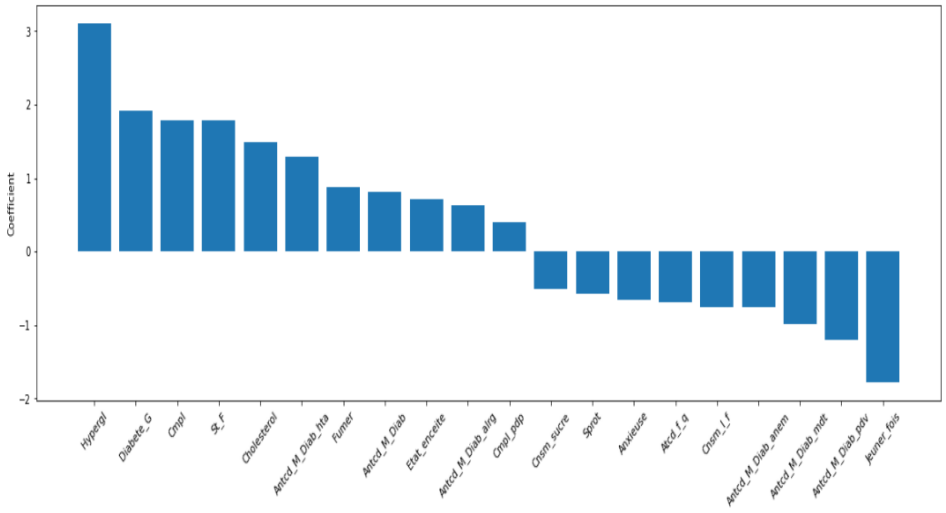

Figure 6. The importance of the features selected by recursive feature elimination.

4.3.3 Comparison with the Pima Indian Diabetes Dataset

The Pima Indian Diabetes Dataset (PIDD) is a well-known dataset used for diabetes prediction. It contains medical information from 768 female Pima Indians, including 9 features, such as glucose level, blood pressure, BMI and age. The target variable indicates whether the individual has diabetes (1) or not (0). This dataset is widely used in machine learning and data-mining research for predicting the risk of diabetes based on the given features.

It is important to note that this dataset is imbalanced, as there are significantly more non-diabetic samples compared to diabetic samples (500 samples belonging to non-diabetic persons and 268 samples belonging to diabetic persons).

In the dataset, six attributes contain zero values for some samples. The occurrences of zero values in these attributes are as follows: 111 for pregnancy, 5 for glucose, 35 for blood pressure, 227 for skin thickness, 374 for insulin and 11 for BMI. The presence of these zero values might be due to errors during data collection or missing data.

These zero values can potentially impact the performance of the classifier negatively. To create a reliable prediction model, it is essential to handle these missing values appropriately and fill them with suitable values. This process ensures that the model can make accurate predictions and avoid any potential biases caused by incomplete or incorrect data.

For a comparative analysis, we compared our database with the famous Pima Indians Diabetes Dataset, which is widely used in the scientific literature for the study of diabetes. This approach will allow us to evaluate and improve the study to fully understand the differences and similarities between the data collected in Algeria and the reference data in the scientific literature. So, we selected the same available features in our dataset according to PIDD. Table 6 shows a summary of the selected features according to the PIMA dataset.

Table 6. Summary of the selected futures according to the PIMA dataset.

Pima Indian Diabetes Dataset	**Our Dataset**
Glucose	Blood sugar
Pregnancies	Number of pregnancies
Blood pressure	Antihypertensive medication
Skin thickness	/
Insulin	/
BMI	BMI
Diabetes pedigree Function	Diabetes family tree function
Âge	Âge
Outcome	Diabete

Figure 7 presents the correlation matrix of the selected features.

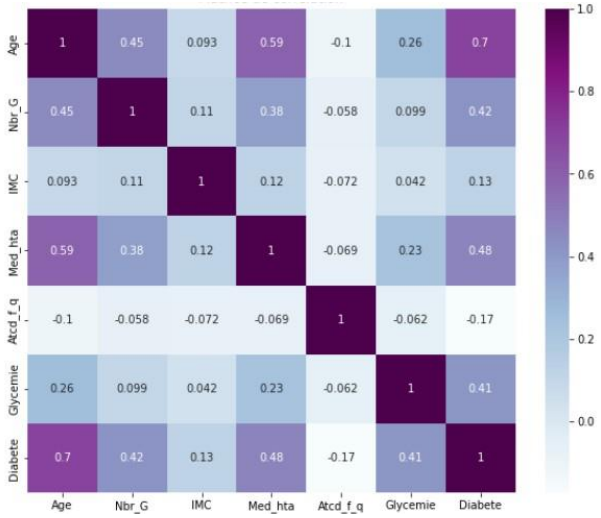

Figure 7. Ccorrelation matrix of the selected features.

4.4 Diabetes Prediction

We split the dataset into 70% for the training set and 30% for the test set. The training phase is a crucial part in the creation of any machine-learning and deep-learning model, including for diabetes prediction, in order to optimize the parameters so that the model is capable of improving performance. For the creation of the model, we used several machine-learning algorithms to evaluate their performance and choose the most robust one. This step also includes finding the best hyper-parameters for each algorithm and training the model on the training data. Once the model is trained, it can be used to predict output values for new data.

The algorithms used in this step are: SVM, KNN, Decision Tree and Logistic Regression. Once the models are trained, they are saved to use them in the test phase or to deploy them to predict new cases of patients.

The next phase is testing and evaluation, which is an important step in the prediction process, because it allows to evaluate the performance of the model on the test data, it includes several steps, which are:

1) Retrieving test data,
2) Deploying the machine-learning model from the database,
3) Evaluating performance by estimating performance metrics, such as accuracy, precision, …etc.
4) Analyzing the results to understand the strengths and weaknesses of the created model. This may include prediction visualizations,
5) Improving the model created from the previous analysis in order to adjust the hyperparameters again or make modifications in the previous steps. This phase can be repeated until the model reaches high performance.

5. EXPERIMENTATIONS

To carry out our experiments and train our prediction models, we used a single personal computer with the following characteristics (see Table 7):

Table 7. The specifications of the used machine.

PC	CPU	RAM	GPU	Storage space
DELL	Intel® Core™ i7-8550U CPU	8.00 Go	Mesa Intel® UHD Graphics 620 (KBL GT2)	1 TB

We have implemented several machine-learning algorithms, such as SVM, KNN, Decision Tree and Logistic Regression. Our objective was to compare these algorithms with each other and select the one that offers the best performance in terms of recall, precision, accuracy, F1-measurement, RMSE, MSE and MAE. In addition, we will examine the results obtained by each algorithm on a dataset, using 3 feature-extraction techniques; namely, Anova, RFE and comparison with the PIMA dataset.

5.1 SVM

The results obtained with the SVM algorithm are presented in Table 8, highlighting the performance of the model.

Table 8. Diabetes-prediction results with the SVM algorithm.

	Accuracy	Recall	Precision	F1- measure	RMSE	MSE	MAE
ANOVA	96%	96.97%	95.52%	96.24%	0.2	0.04	0.024
RFE	97.60%	100%	95.65%	97.78%	0.16	0.024	0.024
Comparison with PIMA	89.6%	84.85%	94.92%	89.6%	0.32	0.10	0.09

After analyzing the results, we found that the use of the SVM algorithm for the prediction of diabetes does not show a large significant difference in accuracy between the use of RFE and ANOVA. The difference observed is minimal, with only 1% variation between the two techniques.

5.2 KNN

The results obtained with the KNN algorithm are presented in Table 9.

Table 9. Diabetes-prediction results with the KNN algorithm.

	Accuracy	Recall	Precision	F1- measure	RMSE	MSE	MAE
ANOVA	97.60%	96.97%	98.49%	97.71%	0.16	0.02	0.02
RFE	95.20%	92.42%	98.39%	95.31%	0.22	0.05	0.05
Comparaison with PIMA	92.80%	90.91%	95.24%	93.02%	0.27	0.07	0.07

From the results presented in Table 9, it is clear that using the KNN algorithm in combination with the Anova method for feature selection gives superior performance compared to other feature selection techniques. Indeed, the model achieves an accuracy of 97.60%, indicating its ability to accurately predict diabetes.

5.3 Decision Tree

The results obtained with the decision-tree algorithm are presented in Table 10.

Table 10. Diabetes-prediction results with the decision-tree algorithm.

	Accuracy	Recall	Precision	F1- measure	RMSE	MSE	MAE
Anova	95.20%	93.94%	96.88%	95.38%	0.22	0.05	0.05
RFE	97.60%	98.48%	97.01%	97.74%	0.16	0.02	0.02
Comparaison with PIMA	93.60%	90.91%	96.77%	93.75%	0.25	0.06	0.06

The decision-tree algorithm produced promising results in our study. When we used the features extracted by RFE, we obtained an accuracy of 97.60%. However, using features selected by ANOVA, the accuracy decreased, reaching 95.20%.

5.4 Logistic Regression

The results obtained with the logistic-regression algorithm are shown in Table 11.

Table 11. Diabetes-prediction results with the logistic-regression algorithm.

	Accuracy	Recall	Precision	F1- measure	RMSE	MSE	MAE
Anova	96%	93.94%	98.41%	96.12%	0.20	0.04	0.04
RFE	97.60%	100%	95.65%	97.78%	0.15	0.02	0.02
Comparaison with PIMA	90.40%	83.33%	98.21%	90.16%	0.31	0.09	0.09

The results obtained demonstrate that there are no significant differences between the used feature-selection methods. However, it is important to note that the RFE algorithm achieved an accuracy of 97.60% and this is the best-performing method.

5.5 Discussion

To facilitate the understanding and analysis of the results, we have divided the results into three separate tables, corresponding to each feature-selection method used in our study. This allows us to compare the performances of different methods and identify the best combination of algorithm and future-selection method for diabetes prediction.

Each table presents the key performance measures, such as precision, recall and F-measure, obtained for each combination of algorithm and feature-selection method. By examining these tables, we will be able to evaluate the performance of each selection method and determine which one offers the best results in terms of predicting diabetes.

Table 12. Results of the ANOVA selection method for the prediction of diabetes.

	SVM	KNN	Decision Tree	Logistic Regression
Accuracy	96%	97.60%	95.20%	96%
Recall	96.97%	96.97%	93.94%	93.94%
Precision	95.52%	98.49%	96.88%	98.41%
F1-measure	96.24%	97.71%	95.38%	96.12%
RMSE	0.2	0.16	0.22	0.20
MSE	0.04	0.02	0.05	0.04
MAE	0.024	0.02	0.05	0.04

Table 13. Results of the RFE selection method for the prediction of diabetes.

	SVM	KNN	Decision Tree	Logistic Regression
Accuracy	97.60%	95.20%	97.60%	97.60%
Recall	100%	92.42%	98.48%	100%
Precision	95.65%	98.39%	97.01%	95.65%
F1-measure	97.01%	95.31%	97.74%	97.78%
RMSE	0.16	0.22	0.16	0.15
MSE	0.024	0.05	0.02	0.02
MAE	0.024	0.05	0.02	0.02

Table 14. Results of the comparative method with PIMA for the prediction of diabetes.

	SVM	KNN	Decision Tree	Logistic Regression
Accuracy	89.6%	92.80%	93.60%	90.40%
Recall	84.85%	90.91%	90.91%	83.33%
Precision	94.92%	95.24%	96.77%	98.21%
F1-measure	89.6%	93.02%	93.75%	90.16%
RMSE	0.32	0.27	0.25	0.31
MSE	0.10	0.07	0.06	0.09
MAE	0.09	0.07	0.06	0.09

From the results of the previous tables, we can identify the best machine-learning algorithm for diabetes prediction. The most efficient algorithms are:

- The KNN algorithm with the ANOVA feature-extraction method. When we use the ANOVA method, we obtain an accuracy rate of 97.60%.
- The Logistic Regression algorithm performs well, using the RFE feature-extraction method. When we apply the RFE method, we obtain an accuracy rate of 97.60%.
- By using the SVM algorithm with the RFE method, we manage to obtain a remarkable accuracy rate of 97.60%,
- The Decision Tree algorithm, combined with the RFE method for feature selection, provides an accuracy rate of 93.60%.

We note that during the experiments, we sought to optimize the performance of our algorithms by carrying out exhaustive tests with different parameter configurations. The objective was to find the parameters that would allow us to obtain the most efficient results. The results of each parameter configuration were carefully noted and compared. As an example, we set the number of clusters in the KNN algorithm to 7 and used Euclidean distance as a similarity measure. In the SVM algorithm, we chose kernel = 'linear'. For the decision tree, we took entropy as a division criterion.

In order to evaluate the performance of our classification models, in particular with regard to their abilities to distinguish between positive and negative classes, we plotted the ROC (Receiver Operating Characteristic) curve and the area under the curve (AUC) which are presented in Figure 8. These metrics are commonly used in the evaluation of classification models, particularly in the field of diabetes prediction, where the distinction between true positives and false positives is of critical importance. From the presented results, the AUC of our models is closer to 1.0 which is considered efficient.

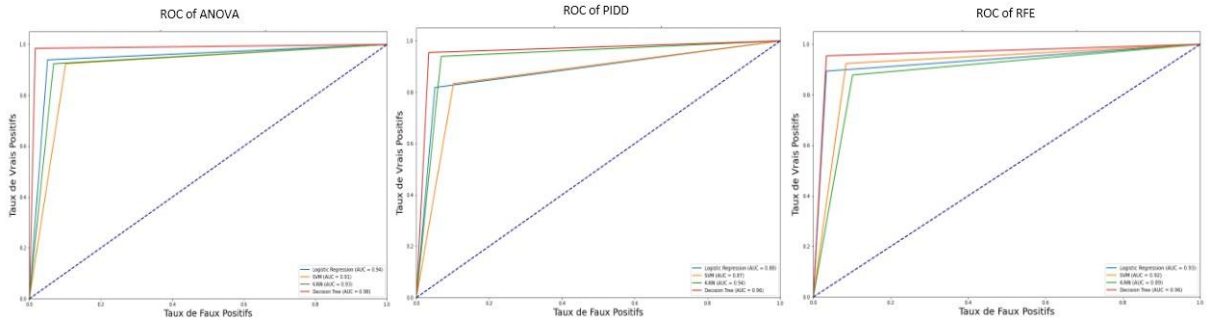

Figure 8. The ROC and AUC results.

The limitation of this work lies in the number of collected data, since the volume of data is a critical factor in data analysis and modeling projects, especially in the medical field. Due to time, resource and data-access constraints, we had to work with a limited sample of data. This may lead to potential biases in our results, as a larger sample could have provided a more complete and representative view of the problem.

6. CONCLUSION

Diabetes prediction has become a growing area of interest due to the increasing prevalence of diabetes, advances in technology and the growing importance of health data. Research efforts in this area aim to develop accurate predictive tools that can aid in the prevention, early diagnosis and effective management of diabetes, thereby improving the health and well-being of individuals affected by this disease.

In this work, we implemented and evaluated several algorithms, such as SVM, KNN, Logistic Regression and Decision Tree, by comparing their performances and analyzing the obtained results. Our results demonstrated that the KNN algorithms proved to be particularly effective in the prediction of diabetes, because it took advantage of its nearest neighbor-based approach to achieve good results. It is also important to emphasize the importance of feature selection in the prediction of diabetes. We found that different selection methods influenced the performance of the algorithms,

highlighting the importance of a well-considered approach, such as RFE or ANOVA, to choose the most relevant features.

In terms of future perspectives, there are several aspects to consider to improve and further develop the diabetes-prediction system:

- In terms of the dataset, it would be interesting to extend our study to large data on other wilayas to reach the entire Algerian population,
- Collecting other risk factors for diabetes,
- Developing a useful online website or mobile application and deploying it on Play store.

REFERENCES

[1] S. M. H. Mahmud et al., "Machine Learning Based Unified Framework for Diabetes Prediction," Proc. of the 2018 Int. Conf. on Big Data Engineering and Technology (BDET 2018), pp. 46–50, 2018.

[2] A. Singh, A. Dhillon, N. Kumar, M. S. Hossain, G. Muhammad and M. Kumar, "eDiaPredict: An Ensemble-based Framework for Diabetes Prediction," ACM Transactions on Multimedia Computing, Communications and Applications, vol. 17, no. 2, pp 1–26, 2021.

[3] L. Xu, J. He and Y. Hu, "Early Diabetes Risk Prediction Based on Deep Learning Methods," Proc. of the 4th Int. Conf. on Pattern Recognition and Artificial Intelligence, pp. 282-286, Yibin, China, 2021.

[4] M. Belhadj et al., "BAROMÈTRE Algérie: Enquête Nationale sur la prise en Charge des Personnes Diabétiques," Médecine des Maladies Métaboliques, vol. 13, no. 2, pp. 188-194, 2019.

[5] D. S. Sisodia and R. Agrawal, "Data Imputation-based Learning Models for Prediction of Diabetes," Proc. of the 2020 Int. Conf. on Decision Aid Sciences and Application (DASA), pp. 966-970, 2020.

[6] H. Song and S. Lee, "Implementation of Diabetes Incidence Prediction Using a Multilayer Perceptron Neural Network," Proc. of the IEEE Int. Conf. on Bioinformatics and Biomedicine, pp. 3089-3091, Houston, USA, 2021.

[7] Z. Punthakee, R. Goldenberg and P. Katz, "Definition, Classification and Diagnosis of Diabetes, Prediabetes and Metabolic Syndrome," Canadian Journal of Diabetes, vol. 42, no. 1, pp. 10-15, 2018.

[8] Y. Wei, W. Guo, B.W. Ling, Y. Dai and Q. Liu, "Both Forward Approach and Backward Approach for Performing Both Regressions and Classifications Using the Histogram Information for Predicting the Baseline Screening Scores for Performing the Prognostic of the Diabetes," Signal, Image and Video Processing, vol. 17, pp. 3803–3809, 2023.

[9] A. Al-Sideiri, Z. B. C. Cob and S. B. M. Drus, "Machine Learning Algorithms for Diabetes Prediction: A Review Paper," Proc. of the 2019 Int. Conf. on Artificial Intelligence, Robotics and Control (AIRC '19), pp. 27–32, 2019.

[10] M. Komi, J. Li, Y. Zhai and X. Zhang, "Application of Data Mining Methods in Diabetes Prediction," Proc. of the 2017 2nd IEEE Int. Conf. on Image, Vision and Computing (ICIVC), pp. 1006-1010, Chengdu, 2017.

[11] A. H. Khan and J. E. Pessin, "Insulin Regulation of Glucose Uptake: A Complex Interplay of Intracellular Signalling Pathways," Diabetologia, vol. 45, pp. 1475–1483, 2002.

[12] S. V. Hemanth, S. Alagarsamy and T. D. Rajkumar, "Convolutional Neural Network-based Sea Lion Optimization Algorithm for the Detection and Classification of Diabetic Retinopathy," Acta Diabetologica, vol. 60, pp. 1377–1389, 2023.

[13] X. Li, M. Curiger, R. Dornberger and T. Hanne, "Optimized Computational Diabetes Prediction with Feature Selection Algorithms," Proc. of the 2023 7th Int. Conf. on Intelligent Systems, Metaheuristics and Swarm Intelligence (ISMSI '23), pp. 36–43, DOI: 10.1145/3596947.3596948, 2023.

[14] M. M. Hassan, Z. J. Peya, S. Mollick, M. A. Billah, M. M. Hasan Shakil and A. U. Dulla, "Diabetes Prediction in Healthcare at Early Stage Using Machine Learning Approach," Proc. of the 12th Int. Conf. on Computing Communication and Networking Technologies, pp. 01- 05, Kharagpur, India, 2021.

[15] J. M. Ekoé, Z. Punthakee, T. Ransom, A. P. H. Prebtani and R. Goldenberg, "Dépistage du Diabète de Type 1 et de Type 2," Canadian Journal of Diabetes, vol. 37, pp. S373-S376, 2013.

[16] P. J. Shermila, A. Ahilan, M. Shunmugathammal and J. Marimuthu, "DEEPFIC: Food Item Classification with Calorie Calculation Using Dragonfly Deep Learning Network," Signal, Image and Video Processing, vol. 17, pp. 3731–3739, 2023.

[17] S. Brahimi and Y. F. Drioua, Etude Rétrospective des Facteurs de Risques du Diabète au Niveau de l'EPSP EsSenia, M.Sc Thesis, University of Science and Technology of Oran Mohamed Boudiaf, USTO-MB oran, Algeria, 2021.

[18] F. Zafar et al., "Predictive Analytics in Healthcare for Diabetes Prediction," Proc. of the 2019 9th Int. Conf. on Biomedical Engineering and Technology (ICBET' 19), pp. 253–259, 2019.

[19] S. Mahajan, P. K. Sarangi, A. K. Sahoo and M. Rohra, "Diabetes Mellitus Prediction Using Supervised Machine Learning Techniques," Proc. of the 2023 Int. Conf. on Advancement in Computation and Computer Technologies, pp. 587-592, Gharuan, India, 2023.

[20] M. van der Schaar et al., "How Artificial Intelligence and Machine Learning Can Help Healthcare Systems Respond to COVID-19," Machine Learning, vol. 110, no. 1, pp. 1–14, 2021.

[21] M. H. Arnold, "Teasing out Artificial Intelligence in Medicine: An Ethical Critique of Artificial Intelligence and Machine Learning in Medicine," Journal of Bioethical Inquiry, vol. 18, no. 1, pp. 121–139, 2021.

[22] F. Ali et al., "An Intelligent Healthcare Monitoring Framework Using Wearable Sensors and Social Networking Data," Future Generation Computer Systems, vol. 114, pp. 23–43, 2021.

[23] P. Whig, K. Gupta, N. Jiwani, H. Jupalle, S. Kouser and N. Alam, "A Novel Method for Diabetes Classification and Prediction with Pycaret," Microsystem Technologies, vol. 29, pp. 1479-1487, 2023.

[24] B. Karaagac, K. M. Owolabi and E. Pindza, "A Computational Technique for the Caputo Fractal-fractional Diabetes Mellitus Model without Genetic Factors," Int. Journal of Dynamics and Control, vol. 11, pp. 2161–2178, 2023.

[25] E. Daniel, J. Johnson, U. A. Victor, G. V. Aditya and S. A. Sibby, "An Efficient Diabetes Prediction Model Using Machine Learning," Proc. of the 4th Int. Conf. on Electronics and Sustainable Communication Systems, pp. 1202-1208, Coimbatore, India, 2023.

[26] D. N. Katsarou et al., "Short Term Glucose Prediction in Patients with Type 1 Diabetes Mellitus," Proc. of the 44th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society, pp. 329-332, Glasgow, Scotland, United Kingdom, 2022.

[27] W. Gu, Z. Zhou, Y. Zhou, M. He, H. Zou and L. Zhang, "Predicting Blood Glucose Dynamics with Multi-time-series Deep Learning," Proc. the 15th ACM Conf. on Embedded Network Sensor Systems, pp. 1-2, DOI: 10.1145/3131672.3136965, 2017.

[28] T. Zhu, K. Li, P. Herrero, J. Chen and P. Georgiou, "A Deep Learning Algorithm For Personalized Blood Glucose Prediction," Proc. of the 27th Int. Joint Conf. on Artificial Intelligence and the 23rd European Conf. on Artificial Intelligence, pp. 1-5, 2018.

[29] A. Yahyaoui, A. Jamil, J. Rasheed and M. Yesiltepe, "A Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Techniques," Proc. of the 1st Inte. Informatics and Software Engineering Conf. (UBMYK), pp. 1-4, Ankara, Turkey, 2019.

[30] C. Fiarni, E. M. Sipayung and S. Maemunah, "Analysis and Prediction of Diabetes Complication Disease Using Data Mining Algorithm," Procedia Computer Science, vol. 161, pp. 449–457, 2019.

[31] A. Mujumdar and V. Vaidehi, "Diabetes Prediction Using Machine Learning Algorithms," Procedia Computer Science, vol. 165, pp. 292-299, 2019.

[32] P. B. M. Kumar et al., "Type 2: Diabetes Mellitus Prediction Using Deep Neural Networks classifier," Int. Journal of Cognitive Computing in Engineering, vol. 1, pp. 55-61, 2020.

[33] Himanshi, A. Agarwal, Sidharth and K. Middha, "Prediction of Diabetes Using Machine Learning Algorithms," Int. Research Journal of Modernization in Engineering Technology and Science, vol. 3, no. 12, pp. 1778-1782, 2021.

[34] R. Cheheltani et al., "Predicting Misdiagnosed Adult-onset Type 1 Diabetes Using Machine Learning," Diabetes Research and Clinical Practice, vol. 191, p. 110029, 2022.

[35] A. E. Evwiekpaefe and N. Abdulkadir, "A Predictive Model for Diabetes Mellitus Using Machine Learning Techniques (A Study in Nigeria)," The African Journal of Information Systems, vol. 15, no. 1, pp. 1-21, 2023.

[36] Y. Su, C. Huang, W. Yin, X. Lyu, L. Ma and Z. Tao, "Diabetes Mellitus Risk Prediction Using Age Adaptation Models," Biomedical Signal Processing and Control, vol. 80, p. 104381, 2023.

[37] S. C. Gupta and N. Goel, "Predictive Modeling and Analytics for Diabetes Using Hyper-parameter Tuned Machine Learning Techniques," Procedia Computer Science, vol. 218, pp. 1257–1269, 2023.

[38] H. M. Deberneh and I. Kim, "Prediction of Type 2 Diabetes Based on Machine Learning Algorithm," Int. Journal of Environmental Research and Public Health, vol. 18, no. 6, 2021.

[39] H. El Massari, Z. Sabouri, S. Mhammedi and N. Gherabi, "Diabetes Prediction Using Machine Learning Algorithms and Ontology," J. of ICT Standardization, vol. 10, no. 02, pp. 319–338, 2022.

[40] A. Mujumdar and V. Vaidehi, "Diabetes Prediction Using Machine Learning Algorithms," Procedia Computer Science, vol. 165, pp. 292–299, 2019.

[41] H. EL Massari, S. Mhammedi, Z. Sabouri and N. Gherabi, "Ontology-based Machine Learning to Predict Diabetes Patients," Proc. of the Int. Conf. on Information, Communication and Cybersecurity (ICI2C 2021), pp. 437–445, DOI: 10.1007/978-3-030-91738-8_40, 2022.

[42] N. Yuvaraj and K. R. SriPreethaa, "Diabetes Prediction in Healthcare Systems Using Machine Learning Algorithms on Hadoop Cluster," Cluster Computing, vol. 22, no. 1, pp. 1–9, 2019.

[43] M. Kumar et al., "Population-centric Risk Prediction Modeling for Gestational Diabetes Mellitus: A Machine Learning Approach," Diabetes Research and Clinical Practice, vol. 185, no. 01, pp. 01-11, 2022.

[44] M. Kumar et al., "Machine Learning-derived Prenatal Predictive Risk Model to Guide Intervention and Prevent the Progression of Gestational Diabetes Mellitus to Type 2 Diabetes: Prediction Model Development Study," JMIR Diabetes, vol. 07, no. 03, pp. 01-12, 2022.

ملخص البحث:

مرض السُّكَّر واحدٌ من أكثرِ الأمراضِ انتشاراً على مستوى العالم، كما أنّ معدل انتشاره آخذٌ في الارتفاع. والعواملُ التي تتسبّب في ذلك الانتشار المتزايد للمرض إنّما تتعلّقُ بالتّغذية ونمط الحياة من جانبٍ وبعواملٍ وراثيةٍ من جانبٍ آخر، الأمر الذي يشكل مُعضلة صحية عامة. لذا فإنّ الكشْف المبكّر عن المرض أمرٌ حاسِمٌ حتى يتمّ التّدخُّل لعلاجه بسرعةٍ وبالتّالي الحدُّ من تفاقمه وانتشاره.

يهدف هذا البحث الى اقتراح نظامٍ أوتوماتيكي لتوقُّع الإصابة بمرض السُّكَّر، باستخدام عددٍ من تقنيات تعلُّم الآلة. وعن طريق جمْع بياناتٍ عن المرض في البيئة الجزائرية، تمّ استخدام عددٍ من طُرق انتقاء السِّمات ومقارنة أداء كلٍّ منها بالسِّمات المقترحة من قبل مجموعة البيانات الهندية الخاصّة بمرض السُّكَّر، المعروفة باسم PIMA.

وقد أسفرت نتائج البحث عن معلوماتٍ قيّمةٍ حولَ المقارنة بين تقنيات تعلُّم الآلة المختلفة المستخدمة من حيث الأداء المرتبط بتوقُّع المرض وكذلك حول أهمّية السِّمات المختارة التي تمّت الاستفادة منها في النّموذج المقترح في هذه الدراسة.