Proceedings of the 2023 IEEE 11th International Conference on Systems and Control, Sousse, Tunisia, December 18-20, 2023

Diabetes Mellitus Prediction Based on Machine Learning Techniques

Eliana S.Omoora
Computer Systems And AI Department
University of Benghazi
Benghazi, Libya
eliana.omoora@uob.edu.ly

Hajer A.Altaweil
Computer Systems And AI Department
University of Benghazi
Benghazi, Libya
hajer.altaweil@uob.edu.ly

Tarek Nagem
Computer Systems And AI Department
University of Benghazi
Benghazi, Libya
tarek.nagem@uob.edu.ly

Kenz A. Bozed
Computer Systems And AI Department
University of Benghazi
Benghazi, Libya
kenz.bozed@uob.edu.ly

2023 IEEE 11th International Conference on Systems and Control (ICSC) | 979-8-3503-0488-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICSC58660.2023.10449831

***Abstract*—Diabetes is a common disease that can lead to dangerous health complications, including heart disease, oral health, nerve damage, vision, hearing, chronic kidney disease, and other problems in feet, and mental health. There are many causes of diabetes such as obesity, age, lifestyle, lack of exercise, hereditary diabetes, high blood pressure, poor diet, etc. Over time, people with diabetes have a high risk of diseases such as heart disease, stroke, kidney failure, nerve damage, eye issues, etc. Early diagnosis of diabetes is a very important factor in reducing the incidence of these complications. Machine learning can be used to develop models that can help Early diagnosis, leading to faster and better patient outcomes. This paper uses machine learning algorithms to train models on patients' data. The model was trained using two different datasets for detecting type-1 diabetes, type-2 diabetes, and whether the patient may have prediabetes. At the end of the paper, a proposal was made about developing the model to achieve cognitive artificial intelligence for the prediction of diabetes.**

***Index Terms*—Diabetes, Machine Learning (ML), Random Forest(RF), Extreme Gradient Boosting (XGBoost).**

I. Introduction

Diabetes is one of the most common diseases, as the World Health Organization stated that the number of infected people increased from 108 million in 1980 to 422 million in 2014, and it was the direct cause of death for 1.6 million deaths worldwide (World Health Organization, 2021), and it also stated that it is one of the causes of blindness, kidney failure, and seizures. heart disease, stroke, and lower limb amputations.

Symptoms of diabetes can appear suddenly and may take several years to be detected. Early diagnosis of diabetes leads to a lower risk of blindness, kidney failure, heart attack, and stroke, and a lower risk of death [1]. Symptoms of diabetes include [1]:

- Feeling very thirsty.
- Needing to urinate more often than usual.
- Blurred vision.
- Feeling tired.
- Losing weight unintentionally.

There are three main types of diabetes mellitus (DM): Type 1, Type 2, and Gestational diabetes. The body's failure to produce enough insulin results in Type 1 DM. Type 2 DM begins when cells fail to respond to insulin properly. As the disease advances, a deficiency in insulin production may also manifest. The third main form of DM occurs when pregnant women without a previous history of diabetes develop a high blood glucose level [2].

The latest World Health Organization (WHO) standard, the definitions of groups with a high risk of DM are as follows [3]:

- Age $>= 45$ and seldom exercising
- Body mass index (BMI) $>= 24kg/m * 2$
- Impaired fasting glucose (IFG) or impaired glucose tolerance (IGT)
- Family history of DM
- Lower high-density lipoprotein cholesterol or hypertriglyceridemia (HTG)
- Hypertension or cardiovascular and cerebrovascular disease
- Gestation female whose $age >= 30$

In this paper using a number of machine learning algorithms, the model was trained on datasets containing the above-mentioned criteria to predict three cases of diabetes: either the person has type-1 diabetes or type-2 diabetes, or prediabetes. Depending on the accuracy, The best classification technique for diabetes prediction was eXtreme Gradient Boosting (XGBoost). XGBoost is based on the gradient boosting framework, which involves recursively adding weak models to the ensemble and adjusting the weights of the training samples to reduce the loss function, making it an even better algorithm.

The paper is organized into six sections, as follows: section

II takes us into the main related work those were used for completing this paper. Section III deals with the supervised machine learning algorithms that were used to perform the learning process. Section IV discusses materials and methods which cover the proposed workflow and methodologies. Section V provides the results and compares the evaluation of the algorithms using different scales. Section VI briefs the conclusion and The last section shows a proposal for future work.

II. Related work

In 2020, The Procedia Computer Science Journal published a research paper titled "Prediction of Type 2 Diabetes using Machine Learning Classification Methods" by Neha Prerna Tigga and Shruti Garg from the Department of Computer Science and Engineering at Birla Institute of Technology in India. The paper applied a number of machine learning algorithms to predict the probability of Type 2 diabetes in individuals by considering their lifestyle and familial background. The authors collected data from 952 individuals through a questionnaire consisting of 18 questions. Logistic Regression, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Naïve Bayes, Decision Tree (DT), and Random Forest are applied machine learning algorithms. The Random Forest Classifier was the most accurate algorithm for their dataset and the Pima Indian Diabetes database. The paper concludes that machine learning algorithms can be used to accurately predict the risk of diabetes and that individuals can use this information to self-assess their risk and take preventive action [4].

In 2021, a paper entitled "Diabetes Prediction using Machine Learning Algorithms with Feature Selection and Dimensionality Reduction" was presented at the 7th International Conference on Advanced Computing and Communication Systems (ICACCS). The paper discusses the application of machine learning algorithms for the early prediction of diabetes to prevent severe consequences. The paper used two machine learning algorithms: Support Vector Machine (SVM) and Random Forest (RF), and feature selection techniques to identify the most influential factors for prediction. It also applied Principle Component Analysis (PCA) for dimensionality reduction. Various machine learning algorithms were discussed, including decision trees, Naïve Bayes, and K-Nearest Neighbor (KNN), highlighting their strengths and applications in disease prediction [5].

In 2022, the Mathematical Biosciences and Engineering (MBE) journal published a research paper titled "Risk Prediction of Diabetes and Pre-diabetes Based on Physical Examination Data". The paper discusses the importance of early diagnosis and prediction of diabetes in order to prevent and control the disease and its complications. The paper collected physical examination data from the Beijing, China, Physical Examination Center and divided the population into three groups: normal fasting plasma glucose (NFG), mildly impaired fasting plasma glucose (IFG), and type 2 diabetes mellitus(T2DM). Four classification models were constructed to distinguish between the three groups. including eXtreme Gradient Boosting (XGBoost), random forest (RF), logistic regression (LR), and fully connected neural networks (FCN). Additionally, binary classification models were established to discriminate between the three groups [6].

All previous studies used supervised machine learning algorithms, which are commonly used in disease prediction for several reasons [7]:

1) They require labeled training data.
2) Provide insights into feature importance.
3) Aim to achieve high prediction accuracy.
4) Handle both classification and regression tasks.
5) Offer interpretability in some cases.

On the other side, the Prediction of disease in general and diabetes, in particular, requires historical data with known disease outcomes to train the model. It also requires identifying which features are most influential in predicting the disease outcome. predicting the disease outcome must be high accuracy to be reliable. When the goal is to assign a discrete label to a data instance, such as classifying a patient as either having a disease or not. And when the target variable is continuous, such as predicting disease severity or estimating the progression of a disease, Regression algorithms are used. Providing interpretability is sometimes required, allowing researchers and clinicians to understand the decision-making process of the model.

In the first and the second papers, the highest accuracy was achieved by Random Forest, and in the last one, the best accuracy was achieved by XGBoost.

III. DIABETES PREDICTION ALGORITHMS

In this paper, five supervised machine learning algorithms were used, which were chosen from a group of algorithms used in the Related work, including Random Forest and XGBoost. In the end, the accuracy was compared to find out which of them will achieve the highest accuracy on our datasets.

A. Support Vector Machines (SVM)

They are supervised learning models that work together with relevant learning algorithms to perform classification and regression analysis on the results [8]. In this project, SVM was used to predict the probability of developing diabetes based on various medical features. The SVM algorithm performed well in predicting diabetes.

B. Naive Bayes (NB)

It is a fast classification algorithm, which works very well in the case of real-time prediction. It is also considered one of the most common classification algorithms. It is used to classify data based on calculating conditional probability values[9]. Experiments were conducted on three datasets and the results showed the effectiveness of this algorithm.

C. Decision Tree

Decision tree is one of the most powerful and widespread algorithms. It can be used for both classification and regression problems, and it's two types of nodes, the decision-node and the leaf-node. The decision node represents the tests, while the leaf node represents the outcome of these tests [9].

D. Random Forest

RF is one of the most popular and effective classification algorithms. The random forest algorithm consists of a set of decision trees. The random forest technique assesses each instance individually and provides the prediction with the highest number of votes [9].

E. Extreme Gradient Boosting (xgboost)

XGBoost is a commonly used machine learning algorithm for regression and classification tasks. It is based on the gradient boosting framework, which involves combining several weak models (such as decision trees) to generate a strong model.The XGBoost algorithm emerged from a research project conducted at the University of Washington and was introduced in 2016 by Tianqi Chen and Carlos Guestrin. XGBoost works by recursively training decision trees on the residuals (or errors) of previous trees, with the goal of minimizing the loss function [8]. XGBoost also includes several advanced features that make it particularly effective, such as [8]:

- **Decision Trees:** XGBoost uses decision trees as the weak models in the ensemble, and each tree is trained to predict the residual error of the previous trees.
- **Loss Function:** XGBoost supports several loss functions for different types of problems. The loss function is used to calculate the error of the model predictions and adjust the weights of the training samples.
- **Handling Missing Values:** XGBoost can handle missing values in the data by using a technique called "sparsity-aware split finding".

XGBoost is a powerful and flexible algorithm that can be used for a wide range of machine-learning tasks.

IV. MATERIALS AND METHODS OF DIABETES PREDICTION

In this paper, three models were built and trained on five Machine learning algorithms, as shown (figure 1):

A. Data set

Two different datasets were employed, partitioning one of them into two subsets. That built three models. Model A determines whether a person has diabetes or not, while Model B determines the type of diabetes whether it is type 1 or type 2, and Model C determines whether the patient may have prediabetes or not.

- **Dataset-1 :** available on Kaggle. It consisted of 253,680 patients, including 213,623 non-diabetics, 4,629 prediabetics, and 35,328 prediabetics (see figure 2). [10] it was extracted from another data set, The original dataset comprises data provided by 441,455 individuals and encompasses 330 characteristics. This dataset is well-organized and derived from the BRFSS 2015 dataset, available on Kaggle. The features in the dataset consist of either directly posed questions to participants or computed variables derived from individual responses but based on diabetes disease research regarding factors influencing diabetes disease and other chronic health conditions, only select features are included in this analysis... [11]. This group identifies factors that are important risk factors for diabetes, namely: blood pressure (high), cholesterol (high), smoking, diabetes, obesity, age, sex, race, diet, exercise, alcohol consumption, BMI, Household Income, Marital status, Sleep, Time since last checkup, Education, Healthcare coverage and mental health. Dataset-1 was used to train both models A and C.

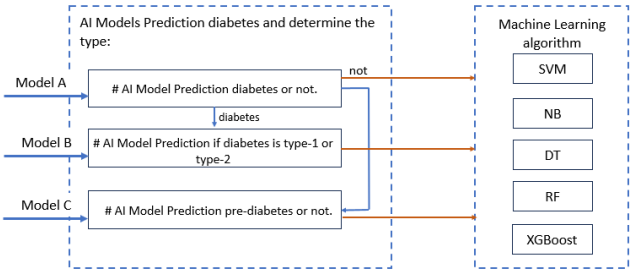

Fig. 1. Models of machine learning .

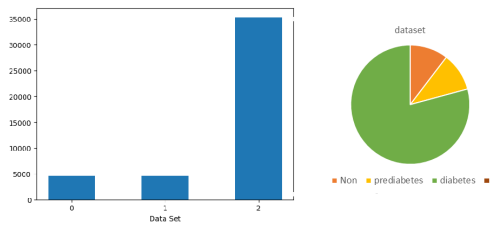

Fig. 2. 1st Dataset distribution

- **Dataset-2 :** available on Kaggle on May 21, 2023. [12]. This group identifies the factors most affecting diabetes, namely: pregnancy, age, glucose, blood pressure, skin thickness, insulin, body mass index, diabetes function, frequent eating, blurred vision, obesity, smoking, and high cholesterol. Dataset-2 was used to train model B.

B. Data Processing

To conduct exploratory statistical analysis and train models effectively, data processing is crucial. The extent of feature analysis and the quality of predictive results depend on the level of data processing during both the training and

testing phases. The mentioned processes were applied to both datasets [13]:

1. Finding Missing Values from the Dataset:

After checking both datasets, no missing values appeared

2. Making Categorical Variables in the Numeric Format:

The categorical variables have been converted into a numerical representation to make the data suitable for analysis using machine learning or statistical models. Additionally, that can help to uncover patterns and relationships in the data that may not be immediately apparent when the data is in its original categorical form. This processing was applied to the first dataset for the (gender) variable, and to the second dataset for a number of categorical variables (Polyphagia, Visualblurring, Obesity, and Gender).

3. Feature Scaling:

Some of the data have been normalized to ensure that the values of different variables are on a similar scale so that no variable dominates over the others. Especially for some machine learning algorithms that use distance-based methods to calculate similarities between data points. This processing was applied to the first dataset for a number of variables (BMI, Age, GenHlth, Education, Income, and PhysHlth), and to the second dataset for a number of categorical variables (Age, Glucose, blood pressure, SkinThickness, and Insulin).

4. Segmentation of the Dataset-1 to Tow Sub-Datasets:

All previous processing was done on both groups. The first dataset, as mentioned earlier, was used for models A and C. To do that it had a must to pre-process its content, which was: 0 person doesn't have diabetes, 1 has prediabetes, and 2 has diabetes. The goal was to make the output binary to facilitate training. Model A, the dataset has been converted all the results of 1 to 0 which means no diabetes, and then 2 to 1 which means it has diabetes. Model C has been only trained on data with results: 0 with diabetes or 1 with prediabetes.

5. Training and Validation Datasets:

To effectively train machine learning models, it is necessary to partition the data into training and testing sets. To accomplish this, the datasets were split into 80% for training and 20% for testing purposes. by using The train_test_split function, which enables the evaluation of a model's performance on previously unseen data. By randomly splitting a dataset into a training subset and a testing subset, the function allows the model to be trained on a subset of the data and tested on a completely independent subset. This helps to prevent overfitting. After the split, The model was then examined by subsets testing to ensure that it was accurate. The testing accuracy represents the procedure's overall testing accuracy as an average.

TABLE I
SPLITTING THE DATASET OF MODEL A

Dataset	Total	Percentage
Training	56300	80%
Testing	14076	20%

1

TABLE II
SPLITTING THE DATASET OF MODEL B

Dataset	Total	Percentage
Training	614	80%
Testing	154	20%

2

TABLE III
SPLITTING THE DATASET OF MODEL C

Dataset	Total	Percentage
Training	7249	80%
Testing	1813	20%

3

V. EVALUATION RESULT

This section shows the prediction accuracy obtained from each algorithm for each of the three models. In addition, predictions were evaluated using different scales. include:

Accuracy: It measures the total number of accurate predictions of a model, and the higher the accuracy, the better the model. Where, In the case of TP, the model correctly predicted a positive outcome when the actual class was also positive. For TN, the model accurately predicted a negative outcome when the actual class was negative. However, in the case of FP, the model incorrectly predicted a positive outcome when the actual class was negative. And finally, for FN, the model erroneously predicted a negative outcome when the actual class was positive. The accuracy is defined as shown below :

$$accuracy = (TP+TN)/(TP+TN+FP+FN)100 \quad (1)$$

Sensitivity: It is a measure that evaluates the model based on the true positive prediction. The Sensitivity is defined as shown below :

$$Sensitivity = (TP)/(TP+FN)100 \quad (2)$$

Specificity: It is a measure that evaluates the model based on a true negative prediction. The Specificity: is defined as shown below :

$$Specificity = (TN)/(TN+FP)100 \quad (3)$$

Precision: It is the proportion of true positives to all the positives. The Precision is defined as shown below :

$$Precision = (TP)/(TP + FP)100 \quad (4)$$

In addition to the aforementioned measures, the confusion matrix was presented for the models created in this system.

A. Evaluation results for Model A

where:

- If the probability of the system is that the patient has diabetes and the patient has diabetes, then the prediction is TP.
- If the probability of the system is that the patient has diabetes and the patient does not, then the prediction is FP.
- If the system probability is that the patient does not have diabetes and the patient does not have diabetes, then the prediction is TN.
- If the probability of the system is that the patient does not have diabetes and the patient has diabetes, then the prediction is FN.

TABLE IV
CONFUSION MATRIX FOR MODEL A

Algorithms	TP	FP	TN	FN
SVM	6976	2388	8292	3457
NB	7289	3159	7521	3144
DT	18837	9563	15003	5860
RF	7089	2575	8105	3344
XGBoost	24230	7372	27874	10900

4

TABLE V
PERFORMANCE RESULTS FOR MODEL A

Algorithms	Accuracy	Sensitivity	Specificity	Precision
SVM	72.31	66.86	77.64	74.49
NB	70.14	69.86	70.42	69.76
DT	68.69	76.27	61.07	66.32
RF	71.96	67.94	75.88	73.35
XGBoost	74.03	68.97	79.08	76.67

5

B. Evaluation results for Model B

where:

- If the system probability is that the patient has type 2 diabetes and the patient has type 2 diabetes, then the prediction is TP.
- If the system probability is that the patient has type 2 diabetes and the patient does not, then the prediction is FP.
- If the system probability is that the patient has type 1 diabetes and the patient has type 1 diabetes, then the prediction is TN.
- If the system probability is that the patient has type 1 diabetes and the patient has type 2 diabetes, then the prediction is FN.

TABLE VI
CONFUSION MATRIX FOR MODEL B

Algorithms	TP	FP	TN	FN
SVM	80	24	34	16
NB	90	33	25	6
DT	294	74	120	49
RF	143	38	36	14
XGBoost	470	44	224	30

6

TABLE VII
PERFORMANCE RESULTS FOR MODEL B

Algorithms	Accuracy	Sensitivity	Specificity	Precision
SVM	74.02	83.33	58.62	76.92
NB	74.67	93.75	43.10	73.17
DT	77.09	85.71	61.85	79.89
RF	77.48	91.08	48.64	79.00
XGBoost	90.36	94.00	83.58	91.43

7

C. Evaluation results for Model C

where:

- If the system probability is that the patient is prediabetic and the patient has prediabetic, then the prediction is TP.
- If the system probability is that the patient has prediabetic and the patient does not, then the prediction is FP.
- If the system probability is that the patient does not have prediabetic and the patient does not have prediabetic, then the prediction is TN.
- If the system's probability is that the patient does not have prediabetic and the patient does have prediabetic, then the prediction is FN.

TABLE VIII
CONFUSION MATRIX FOR MODEL C

Algorithms	TP	FP	TN	FN
SVM	900	374	987	458
NB	978	580	781	380
DT	1738	919	2251	1435
RF	890	390	971	468
XGBoost	3094	1034	3497	1437

8

TABLE IX
PERFORMANCE RESULTS FOR MODEL C

Algorithms	Accuracy	Sensitivity	Specificity	Precision
SVM	69.40	66.27	72.52	70.64
NB	64.69	72.01	57.38	62.77
DT	62.88	54.77	71.00	65.41
RF	68.44	65.53	71.34	69.53
XGBoost	72.73	68.28	77.17	74.95

9

These results proved the effectiveness of the algorithm XGBoost on all models the The accuracy results of the three

models proved the effectiveness of the algorithm XGBoost on all models established in this system.

VI. CONCLUSIONS

Although the medical field is a very sensitive aspect and relying on systems in it is not easy, the availability of data in any field on a large scale makes it easier to build and develop an artificial intelligence system for it. As in this case, the availability of diabetes data was the main factor that led to the achievement of the goals of this work was achieved, which are: promoting early prediction of diabetes so that it includes its various types (type-1, type-2, pre-diabetes) which were achieved through the three models, and also contributing to exploring practical challenges to contribute to generating more accurate forecasts. Finally, this paper confirmed that the XGBoost is superior to the rest of the algorithms used.

This paper serves as a prelude to a proposal for building a diabetes prediction cognitive artificial intelligence (CAI) system, which will be further developed in the coming period. The future work section includes a brief explanation along with a block diagram of the proposed system.

VII. FUTURE WORK

Most AI research seeks to develop intelligent systems that perform optimal decision-making procedures. In this paper, a proposal for a development plan is presented to transform the diabetes prediction system from AI to a CAI system, as Figure 3 shows. In addition to the three models presented in the research, a knowledge base, two GUI parts, and a paraphrase subsystem will be added in this proposal. The working mechanism will be as follows:

1) Part-1 of GUIs: It will be available to the end user (patient), where it will display the symptoms that a diabetic patient may have and the medical analysis required of him, and then display the result that will be determined by the models, and also tips that will be determined by knowledge base after both are passed on the paraphrase subsystem.
2) Part-2 of GUIs: It will be available to authorized persons only, such as WHO, it will show them some queries that will be determined by the knowledge base and also receive their responses for storage.
3) Knowledge Base: It will be responsible for storing the datasets used in training the models, and the results of the testing that will be done on it, in addition to the data that will be entered through the GUIs, as well as the data that will be displayed on the GUIs, such as symptoms, medical advice, and inquiries. After that, it makes statistics on all the data it has stored and outputs new datasets that include the features really affecting diabetes based on the statistics and inquiries from experienced people. After that, it will be retraining the three models by the dataset that resulted from the statistics.

These additions will achieve the cognitively because the final system provides the basic CAI Properties:

1) Learning from experience.
2) Acquiring knowledge from experienced people in the field and consulting them.
3) Making a decision.
4) Providing appropriate advice and treatment to the patient.

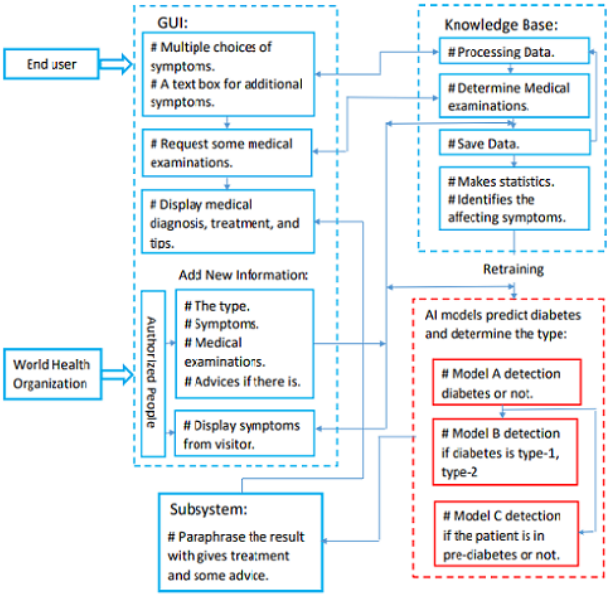

Fig. 3. Cognitive System for Detection Diabetes's Type and Pre-diabetes.

REFERENCES

[1] World Health Organization.(April 2023). diabetes[Online] Available: https://www.who.int/ar/news-room/fact-sheets/detail/diabetes.

[2] Kumar, R., Saha, P., Kumar, Y., Sahana, S., Dubey, A., & Prakash, O. (2020). A Review on Diabetes Mellitus: Type1 & Type2. World Journal of Pharmacy and Pharmaceutical Sciences, 9(10), 838-850.

[3] Wu, H., Yang, S., Huang, Z., He, J., & Wang, X. (2018). Type 2 diabetes mellitus prediction model based on data mining. Informatics in Medicine Unlocked, 10, 100-107.

[4] Tigga, N. P., & Garg, S. (2020). Prediction of type 2 diabetes using machine learning classification methods. Procedia Computer Science, 167, 706-716.

[5] Sivaranjani, S., Ananya, S., Aravinth, J., & Karthika, R. (2021, March). Diabetes prediction using machine learning algorithms with feature selection and dimensionality reduction. In 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS) (Vol. 1, pp. 141-146). IEEE.

[6] Han, Y. M., Yang, H., Huang, Q. L., Sun, Z. J., Li, M. L., Zhang, J. B., ... & Lin, H. (2022). Risk prediction of diabetes and pre-diabetes based on physical examination data. Math Biosci Eng, 19, 3597-608.

[7] Nasteski, V. (2017). An overview of the supervised machine learning methods. Horizons. b, 4, 51-62.

[8] Ali,Z.A.,Abduljabbar,Z.H.,Taher,H.A.,Sallow, A.B.,Almufti,S. M. (2023). Exploring the Power of eXtreme Gradient Boosting Algorithm in Machine Learning: a Review. Academic Journal of Nawroz University, 12(2), 320-334.

[9] Chaki, J., Ganesh, S. T., Cidham, S. K., Theertan, S. A. (2022). Machine learning and artificial intelligence-based Diabetes Mellitus detection and self-management: A systematic review. Journal of King Saud University-Computer and Information Sciences, 34(6), 3204-3225.

[10] A.Teboul. (2021). Diabetes Health Indicators Dataset[Online]. Available: https://www.kaggle.com/diabetes-health-indicators-dataset.
[11] Alex Teboul.(2022). Diabetes Health Indicators Dataset Notebook [Online]. Available: https://www.kaggle.com/code/alexteboul/diabetes-health-indicators-dataset-notebook/notebook.
[12] tmleyncodes. (2023). Diabetes and its type Prediction (Notebook 2.0): diabetes dataset [Online]. Available: https://www.kaggle.com/diabetes-and-it-s-type-prediction-notebook-2-0/notebook.
[13] AlSadi, K.,& Balachandran, W.(2023). Prediction Model of Type-2 Diabetes Mellitus for Oman Prediabetes Patients Using Artificial Neural Network and Six Machine Learning Classifiers. Applied Sciences, 13(4), 2344.