Journal of Information Hiding and Multimedia Signal Processing ©2025 ISSN 2073-4212
Ubiquitous International Volume **16**, Number **1**, March **2025**

Innovative Machine Learning Approaches for Identifying Pre-diabetes in Patients

Marwa Hussien Mohamed*

Computer Technology Engineering Department
Engineering Technologies College, Al-Esraa University
Baghdad, 10081 IRAQ
maraw@esraa.edu.iq; eng_maroo1@yahoo.com.

Mohamed Elkholy

Faculty of Computer Science & Engineering
Alamein International University
Alamein, Egypt
melkholy@Aiu.edu.eg.

Marwa. A.Marzouk

Faculty of Computers and Artificial Intelligence
Matrouh University
Matrouh,Egypt
Marwa.Abdel.Azim@mau.edu.eg; mabdelazeem@nctu.edu.eg.

*Corresponding author: Marwa Hussien Mohamed
Received November 22, 2024, revised January 9, 2025, accepted January 11, 2025.

Abstract. *The healthcare industry has greatly benefited from technological advancements, which have resulted in the development of various technologies for disease prediction. One such condition that has been rapidly increasing across all age groups is diabetes mellitus, with multiple contributing factors and consequences. In this study, these factors are treated as independent features. Machine learning algorithms are employed to predict type 2 diabetes in patients. Specifically, this study utilizes two machine learning techniques: decision trees and random forests. The diabetes datasets, sourced from the Mendeley Data for diabetic patients , contain 1,000 records, including diabetic, Pre-diabetic, and non-diabetic entries. While other researchers have focused on binary classifications, this study uses a three-class system. We introduce a new algorithm that emphasizes three key features— age, gender, and BMI—aimed at predicting the patients. The algorithm demonstrates high accuracy, achieving by decision Tree 99.3% while maintaining a balanced dataset.*
Keywords: Diabetes Mellitus (DM), Machine Learning, Decision Tree, Random Forest, Type 2 diabetes.

1. **Introduction.** Diabetes, commonly diabetes mellitus is the medical term for this condition [1]is a chronic disease that affects your body while useing glucose (blood sugar). Glucose is important to give energy to body cells. However, too much glucose in the bloodstream can cause various health issues. Diabetes has two categories: Diabetes mellitus type 1 [2]: A rises when the immune system of the body targets and destroys pancreatic cells that generate insulin, a hormone that regulates blood sugar levels. Children and young people diabetic patients, which requires lifetime treatment with insulin therapy. Diabetes mellitus type 2 [2] develops when the body develops insulin resistance or fails to

produce enough insulin to keep blood sugar levels stable. Obesity, lack of exercise, and poor diet are all lifestyle factors that have been linked to type 2 diabetes. Although it can also happen to kids, it is more common in adults. Diabetes symptoms might change depending on the level of blood glucose[3]. A few people with type 2 diabetes or prediabetes, in particular, may not develop symptoms at all [4] [5].

Prediabetes, also known as borderline diabetes, is characterized by high blood sugar levels yet not high enough to be diagnosed with diabetes. Diabetic of other types includes gestational diabetes, which occurs during pregnancy, and rare forms of diabetes caused by genetic defects or other medical conditions.

Diabetes with type 1 symptoms appears earlier and is more chronic: Ketones are present in the urine , thirst increases , Urinating frequently , Frequently losing weight , Fatigue, Vision issues. Also, Infections that repeat frequently, such as vaginal infections and gum or skin infections and A BMI of more than 25 is considered obese.

Diabetes can may affect a variety of issues, like Heart illness, nerve damage, kidney damage, visual problems, and foot problems are all possibilities [6]. However, people with diabetes can live healthy and productive lives with proper management. Medications, lifestyle modifications, and routine blood sugar testing are frequently used as treatments.

In recent years, many researchers using concepts of data-miming techniques and machine learning to predict diabetes patients early to reduce the risk factors for these patients. Some research used to predict if the patients are diabetes patients or not diabetic based on some data features. Also, limited research talks about pre-diabetic patients[7].

The most common techniques in data mining [8] used for prediction are random forest (RF), decision trees (DTs), logistic regression (LR), XGBoost (XGB), gradient boosting (GB) Extra Trees, and light gradient boosting machines (LGBM), Naive Bayes (NB), Support Vector Machine (SVM), C4.5 Decision Tree (DT), K-Nearest Neighbour (KNN), Artificial Neural Networks (ANNs), Bayesian Networks. This classification algorithm has many advantages to getting high-accuracy results.

This study employs machine learning to forecast diabetes mellitus. This work contributes significantly in the following ways:

• A key contribution of this work is the publication of a unique diabetes mellitus dataset covering 1000 samples. In this paper, a private dataset was collected from the Mendeley website. We collected 11 features from 14 individuals, excluding irrelevant variables such as patient ID and number.

• Another addition of this work is data pre-processing to remove outliers and missing values before training and testing.

• SMOTE techniques are used to reduce the issue of class imbalance. Also, the data must be balanced among the three types.

• This technique aids in interpreting which features were employed and how they affected the accuracy results in predicting the three types of diabetic patients.

• The unique aspect of this work is that it predicts pre-diabetes individuals from the data, whereas other articles only predict diabetic or non-diabetic patients.

This Research will be structured as follows: Section Two includes a related work, Section three will discuss machine learning techniques, section four the datasets used for our experimental results, section five the proposed algorithm, section six the model testing and experimental results and section seven conclusions and future research.

2. **RELATED WORK.**

Machine learning technology have been increasingly employed in recent years to predict disease risk, especially for chronic diseases[9][10] . Numerous

studies have been conducted on machine learning and data mining classification algorithms. Classification and Diabetes prediction with artificial intelligence, machine learning, and deep learning techniques have received much attention.

The following researchers employed machine learning to forecast diabetes DM disease with different datasets. Khaleel and Al-Bakry et al. [11] developed a model to identify whether or not a person has diabetes illness detection procedures by using machine learning (ML). The study uses the PIMA dataset. Logistic regression (LR), Naive Bayes (NB), and K-Nearest Neighbour (KNN) are the algorithms employed. These algorithms yielded 94%, 79%, and 69% accuracy, respectively. Measures including precision,recall, accuracy, and F-measure are considered.

Sumbal Malik et al. [12] Machine learning techniques are used in multiple domains of healthcare. They employ machine learning to uncover hidden patterns in datasets and how to select the most appropriate features to make it important to detect the disease. The dataset used is the Frankfurt Hospital (Germany) dataset. The authors applied ten different machine learning experiments into practice: Naive Bayes, Bayes Net, Decision Tree, Random Forest, AdaBoost, Bagging, K-Nearest Neighbour, Support Vector Machine, Logistic Regression, and Multi-Layer Perceptron. The results indicate that K-Nearest Neighbour, Random Forest, and Decision Tree are the most accurate, 98.62%, 98.8%, and 93.88%, respectively.

Navya Pratyusha Miriyala et al. [13] have developed a decision-support system for diabetes mellitus diagnosis (DM).They used their experimental using Pima Indians Diabetes dataset, to train and test the model. They start with data analysis while building their model and using a pre-processing step to select the important features by using six machine learning techniques, including Naive Bayes, KNN, Random Forest, Logistic Regression, Decision Tree, and Extreme gradient boosting The Extreme Gradient Boosting produces high accuracy outcomes with 88.2%. The decision tree is at 85.3% compared with the other machine learning algorithms.

Othmane Daanouni et al. [14] We know that some individuals have type 1 and type 2 diabetes in these militias. However, the authors used four machine learning approaches to predict just type 2 diabetic patients. (Decision Tree, K-Nearest Neighbors, Artificial Neural Network, and Deep Neural Network) all these algorithms were tested on two different datasets Frankfurt Hospital (Germany) and the Pima Indian dataset. These data sets need pre-processing steps like noisy data and missing data authors prepare this data before training the prediction model. It will give more accurate results. The deep neural network algorithm has 98%, and K Nearest Neighbours have 97% accuracy.

Nadia Mahmood Ali et al. [9] This work collects data manually from the Iraqi population society to detect diabetes using machine learning techniques. They compare the output results of the classifier models decision trees and k-nearest neighbor (KNN) with the use of k-cross validation while computing the accuracy and divide the data into different partitioning (60%, 70%, 80%) to train the model and get high results.KNN algorithms gain an accuracy of 84.75%, 89.33%, and 92.45%, and random forest algorithms gain an accuracy of 99.5%, 99.66%, and 99.75%. When partitioning the train datasets larger than the test, the results for the accuracy are better and higher than using the smallest train data set, like 60% partitioning and random forest has more accurate results with this model.

Roshi Saxena et al. [15] use Pima Indians datasets with different classifier models (multi-layer perceptron, decision trees, K-nearest neighbor, and random forest). They remove the missing data values and the outliers from the data using Weka 3.9 program. This model results from accuracy was high with multi-layer perceptron is 77.60%, for

decision trees is 76.07%, for K-nearest neighbor is 78.58%, and for a random forest is 79.8%, so the highest results with random forest.

Sasmita Padhy et al. [16] use the new technology Internet of Things (IOT) with mobile applications To recognize diabetes early using an innovative non-invasive self-care system based on IOT and machine learning (ML) to analyze blood sugar plus the most common critical signs. This application follows the patient by building a hybrid machine learning ensemble combining bagging and boosting approaches to predict the patient. They collect their data online via a questionnaire about the people's lifestyle, health, and family history. If they have diabetes in their family, the total number of responses to the questionnaire is 10221 responses. They use various machine learning techniques (logistic regression, K-nearest neighbor, Support vector machine).This model helps people to predict if they are diabetic or not at an early stage using IOT technology. Also, they compare the model with the Pima Indian datasets, and this model has 98.4% accuracy results.

3. **Machine learning techniques.** Supervised machine learning algorithms are used[17]. We used two different ways decision trees and random forest to determine whether a patient has diabetic, pre-diabetic, and non-diabetic patients with machine learning.

A. Decision Trees(DT) This approach uses supervised learning [18]. It functions with continuous and categorical input and output variables. Regression or classification processes are indicated by using it [19]. The several categories of DTs are ID3, ID 4.5, CART, and CHAID. The DT-related measures are standard deviation, Gini index, and entropy.

B. Random Forest(RF) The results of different Decision trees[20] are combined with the Random Forest to produce a single outcome. Row and Column selection is employed, with decision trees as the basis. The variance may decline if the base learner population grows or vice versa. K is a valid option for cross-validation. It is regarded as a necessary bagging technique [21].

4. **Diabetes Data sets.** The dataset was obtained from Mendeley website contains three diabetic dataset groups: non-diabetic (n = 103), pre-diabetic (n = 53), and diabetic (n = 844)[22] (https://data.mendeley.com/datasets/wj9rwkp9c2/1).

The 14 independent variables (predictors) are included in ID, No of Patient, Gender, Age, Urea, Creatinine ratio (Cr), Fasting lipid profile, which includes total low-density lipoprotein (LDL), very-low-density lipoprotein (VLDL), and glycosylated hemoglobin (HBA1C), Triglycerides (TG) and High-Density Lipoprotein (HDL), Cholesterol (Chol), Body Mass Index (BMI) [23]. The Class attribute (target or dependent variable) is meticulous. The diabetes disease class for the patient may be Diabetic (Y), Predicted (Probable) -Diabetic (P), or Non-Diabetic (N). The dataset has 14 features collected from patients, as listed in Table 1.

The dataset containing :

1. a person's risk factors, including age, gender, and Body Mass Index (BMI).
2. laboratory data such as HbA1c, creatinine ratio.
3.lipid profile data Like, as age grows, the diagnostic efficiency of glycosylated hemoglobin (HbA1c) for diabetes decreases because of decreasing Red Blood Cell (RBC) count.

5. **Proposed algorithm.** The early detection of diabetes class supports the prevention and treatment in the diabetic, pre-diabetic and non-diabetic classes, respectively. Numerous effective techniques are available to avoid type 2 diabetes as well as the complications and early death that can arise from any type of diabetes.

Policies and practices across people and environments are examples of such methods. Many lifestyles public health programs are addressed to prevent type 2 diabetes focusing

Table 1. Dataset Features

No	*Attribute*	*Description*
1	ID	patient ID
2	No Patient	Patient Number
3	Gender	Male "M" or Female "F"
4	Age	In years (Min: 20, Max: 79)
5	Body Mass Index (BMI)	(Min: 19, Max: 47.75)
6	Cr	Creatinine ratio
7	Urea	Urea
8	Chol	Cholesterol
9	LDL	Low-Density Lipoprotein
10	VLDL	Very Low-Density Lipoprotein
11	HDL	High-Density Lipoprotein
12	TG	Triglycerides
13	HBA1C	Glycosylated Haemoglobin
14	CLASS Patient's diabetes	(Diabetic, Predict-Diabetic, or Non-Diabetic)

on eating and physical activity habits for long-term regulation of energy balance. These therapies reduce the chance of developing obesity and type 2 diabetes later in life.

In this research, we build our algorithm based on three classes diabetes, pre-diabetes, and non-diabetes. To predict early the patients in the first stage of this disease, we apply our experimental results to the diabetes datasets collected from Mendeley Data mentioned above in the dataset section. This Medical data and test results were taken from the patient's medical file and entered into the database from the data attributes. The next figure 1 shows the new proposed diabetes prediction framework.

In the next section, we will discuss the steps for building our new algorithm and the steps applied to prepare the datasets accordingly to the requirements to get high-accuracy results and reduce the data errors found during the running phase and experimental results.

A. Data Analysis and Visualization:

We need to prepare the datasets by removing records not affected by the results, so we drop the two columns ID and No Patient. We develop the code for our approach in Python. The describe( ) method only works with numeric data, not categorical values. We find in the data sets some of the data are numeric and others categorical, like gender has F (female), M (male) and class category Diabetic (Y), Predicted (Probably) -Diabetic (P), or Non-Diabetic (N). On the datasets, we utilize the describe ( ) method in Python and use count, average, mean, Standard Deviation, minimum value, and maximum value, with 25%, 50%, and 75% representing the percentile/quartile of each feature. This quartile data assists us in detecting outliers in the data values. The statistics produced by the describe () method as shown in table 2

The Descriptive Statistics show that the datasets have outliers based on the maximum and minimum values so that variables in the datasets are not normally distributed. A

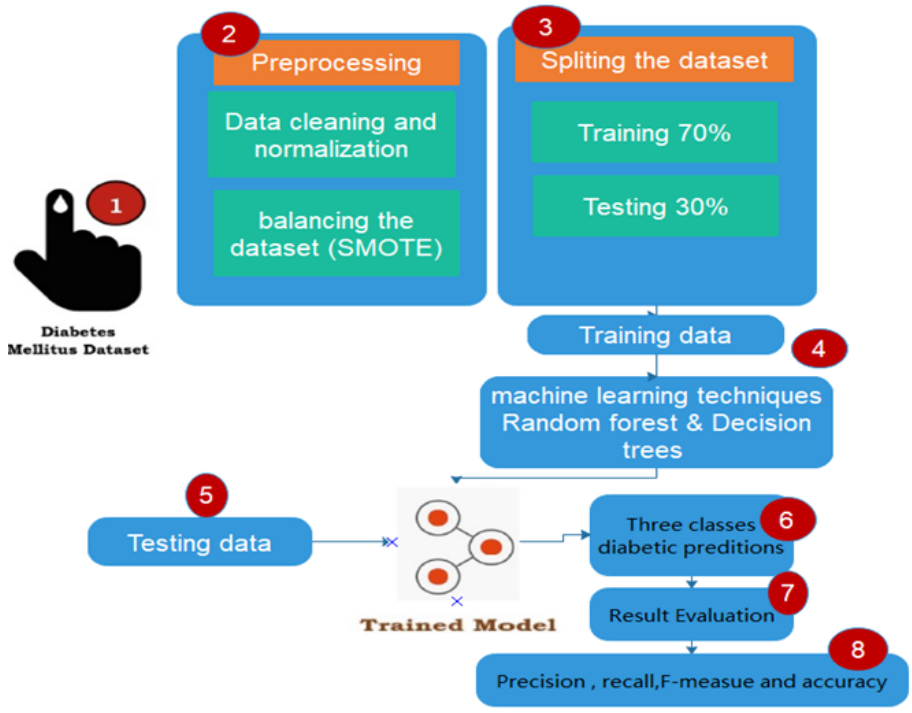


FIGURE 1. Diabetes prediction proposed Architecture

TABLE 2. The output from the described method

−−	*AGE*	*Urea*	*Cr*	*HbA1c*	*Chol1*	*TG*	*HDL*	*LDL*	*VLDL*	*BMI*
Count	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
Mean	53.52	5.12	68.9	8.28	4.86	2.34	1.20	2.6	1.85	29.5
Std	8.79	2.93	59.9	2.53	1.30	1.4	0.66	1.11	3.66	4.96
Min	20.0	0.50	6.0	0.90	0.00	0.3	0.2	0.3	0.10	19.0
25%	51.0	3.70	48.0	6.5	4.0	1.6	0.9	1.8	0.7	26.0
50%	55.0	4.6	60.0	8.0	4.8	2.0	1.1	2.5	0.9	30.0
75%	59.0	5.7	73.0	10.2	5.6	2.9	1.3	3.3	1.5	33.0
max	79.0	38.9	800	16.0	10.3	13.8	9.9	9.9	35.0	47.7

box plot is a graphical representation of numerical data groups progressing through their quartiles. The Box expands from the data's Q1 to Q3 quartile values, with a line at the median value (Q2). The lines are no more than 1.5 * IQR (IQR = Q3 - Q1) inches from the Box's edges to show range data; the output of this step shows that all variables have many outliers except "HbA1c" and "BMI," which show fewer outliers' figure 2 and figure 3 shows the box plot results for some features of the datasets.

All outliers [24] (values falling outside the minimum and maximum values of the Box) are handled by the quantile-based flooring and capping method to optimize the original diabetes dataset. Also, the distribution of "CLASS" show that the data is imbalanced shown in Figure 4.

B. DATA PREPARATION (pre-processing)

The preparation of data for analysis is the most important step in data mining and machine learning research.

Data Cleaning - Data Imputation

Data cleaning[25] is an important step in any machine learning method. Various statistical

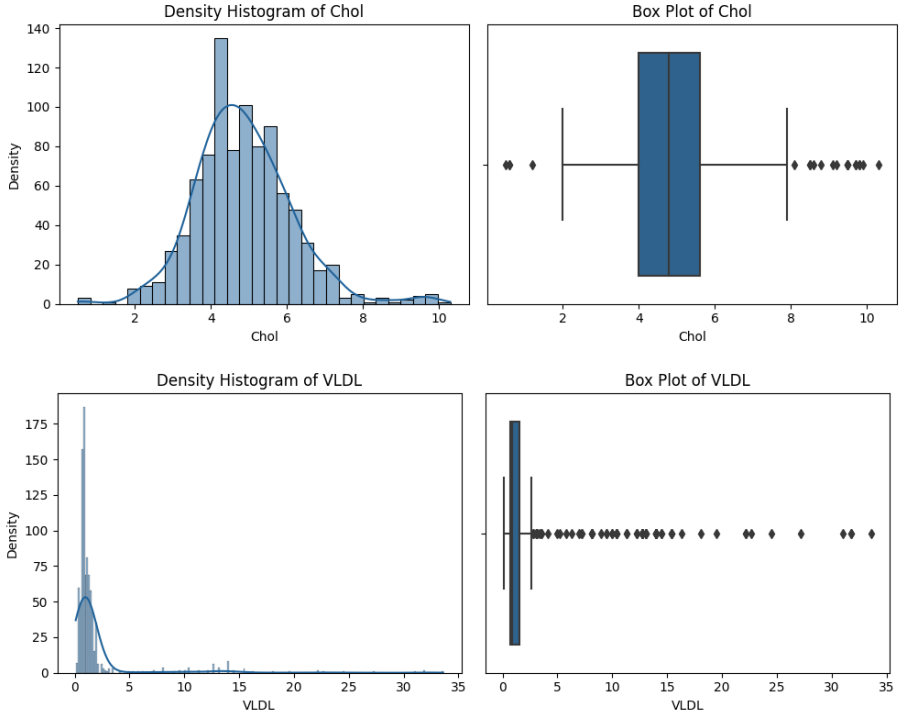


FIGURE 2. Box plot of Chol and Vldl Features

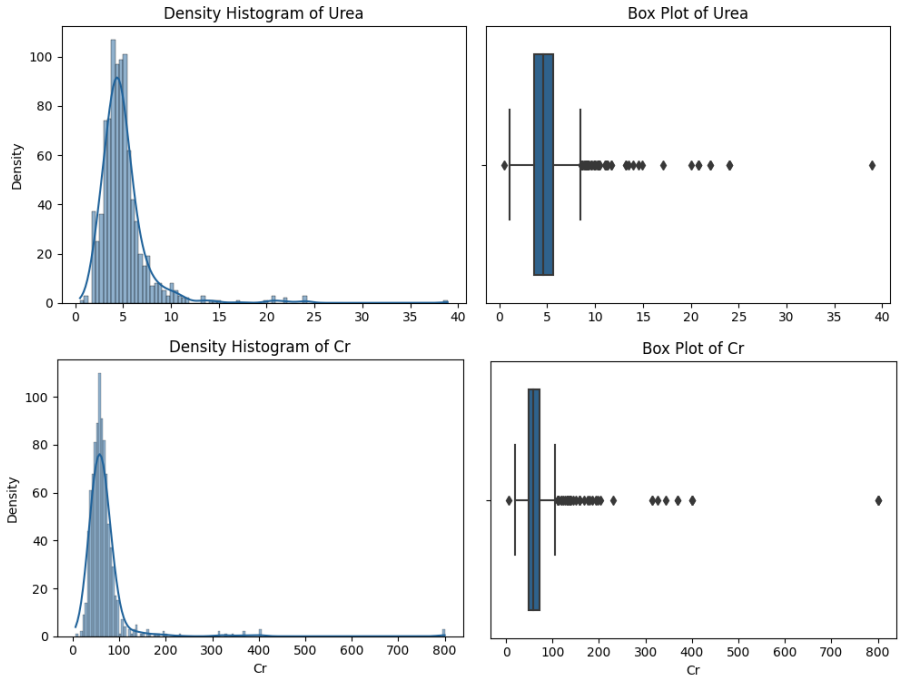


FIGURE 3. Box plot of Urea and Cr Features

analysis and data visualization techniques used in the data exploration phase of tabular data help to discover the necessary data cleaning processes.

Firstly, we drop the two un-useful data columns, like id and patient id. Secondly: a poor labeling of the data was observed for the variables 'Gender' and 'CLASS'. The unique values explored for 'Gender' were ('F,' M,' f'), and for 'CLASS' were ('N,' N, " P,' Y and Y '). Such errors were handled by capitalizing the 'F' value of 'Gender' and removing

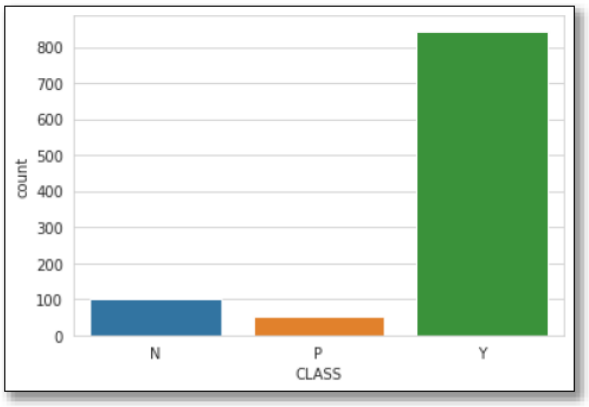


FIGURE 4. The variable class records in the data

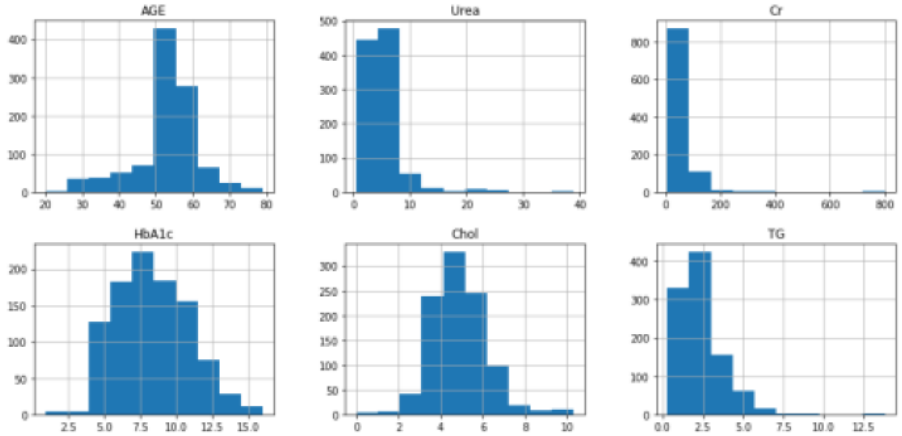


FIGURE 5. Histograms (Distribution of the Numerical Variables)

the space from the values of 'CLASS' and replacing them with the values (F,'M') and (N,'P,'Y'), respectively.

Finally: For our dataset, there are no missing values for all features, but there is only one zero value for the variable of the 'Chol' level, which is an abnormal value and may be due to an error in the data entry process. These zero values are meaningless and are handled as missing values. We corrected this error by substituting the median value of the "Chol column" for the zero values.

C. Categorical Variables Encoding and numerical variables

Visualization of data helps to understand the data and also to explain the data to another person. A histogram is a good way to visualize the distribution of the variables in a dataset. Histograms figure 5 group values of each feature into bins and display a count of the data points whose values are in a particular bin. It also helps to identify outliers that will appear outside the overall pattern of distribution.

Encoding is necessary for ML algorithms [26] based on a mathematical equation and cannot handle categorical values. Our data set includes two categorical variables: 'Gender' and 'CLASS' diabetes. The Gender categorical column is binary with values of "F" for female will be "0" and "M" for male will be "1". These values must be converted into equivalent numerical representations so ML algorithms can process them. The class has three values, Y, P, and N (2,1,0).

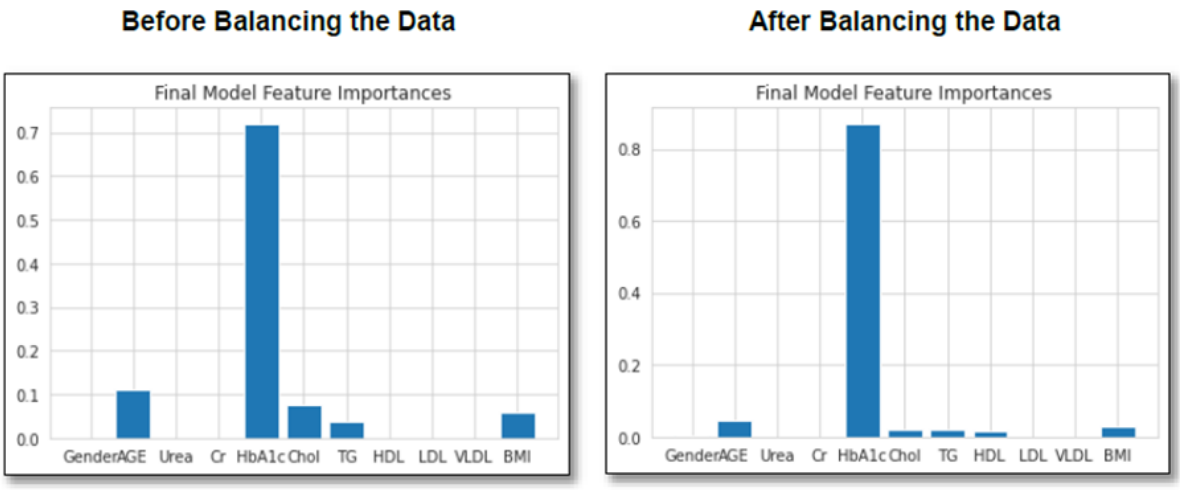


FIGURE 6. Model feature importance Decision Tree before and after data balancing

D. Data Balancing

Imbalanced classes are a common issue when developing and training machine learning classification models in which the classes are not equally represented across the data. The approach to address the imbalanced datasets was to oversampling the two minority classes through the over-sampling SMOTE technique (Synthetic Minority Over-Sampling Technique). SMOTE is an oversampling method that generates synthetic minority samples. Because the new examples are synthesized from existing examples, these synthetic samples provide no new information to the model[27]. Figure 6 show the decision tree model feature importance before and after applying data balancing.

E. Data Scaling (Normalization or Standardization)

Our datasets have many numerical variables with different scales and values. This difference between values will affect the data performance, so we use equation 1 to rescale the data between values 0 and 1[28]. Equation 1 and Equation 2 standardization Rescale the numerical features by changing the data to have a zero mean and a standard deviation of one unit.

equation 1:Shows the Normalization Formula

$$Newvalue(x') = (Originalvalue(x) - Min(x))/(Max(x) - Min(x)) \tag{1}$$

equation 2:Shows the Standardization Formula

$$Newvalue(x') = (Originalvalue(x) - Mean(x))/(SD(x)) \tag{2}$$

Standard Scaler [23] scales data by setting the mean to 0 and standard deviations to 1. Data after Scaling by Standardization is shown in Figure 7.

F. Data Partitioning (Splitting) Our data set has one thousand records for partitioning the data. We used 70% to train the data and 30% to test the data.

6. **Model testing and experimental results.** We use the two supervised machine learning techniques, random forest, and decision trees, and evaluate the model using precision (Specificity), recall (Sensitivity), F- measure score, and accuracy. We compare the previous experimental results results of two previous papers shown in table 3 which these two authors used decision tree and random forest classifiers.

	Gender	AGE	Urea	Cr	HbA1c	Chol	TG	HDL	LDL	VLDL	BMI	CLASS
0	0	-0.401144	-0.033558	-0.843755	-1.367086	-0.552264	-1.271422	1.726423	-1.171805	-1.142773	-1.136588	0
1	1	-3.130017	-0.162272	0.034128	-1.367086	-0.999066	-0.793842	-0.100319	-0.473465	-0.948622	-1.341523	0
2	0	-0.401144	-0.033558	-0.843755	-1.367086	-0.552264	-1.271422	1.726423	-1.171805	-1.142773	-1.136588	0
3	0	-0.401144	-0.033558	-0.843755	-1.367086	-0.552264	-1.271422	1.726423	-1.171805	-1.142773	-1.136588	0
4	1	-2.334096	1.511012	-0.843755	-1.367086	0.073258	-1.175906	-1.013689	-0.573228	-1.336923	-1.751395	0
...	...	...	...	...	...	...	...	...	...	...	...	...
995	1	1.986619	1.768441	1.954496	-0.519989	2.396625	-0.507294	0.204138	-0.772754	-0.948622	0.093026	2
996	1	-2.561502	-1.127628	-0.075608	1.617923	-0.641625	-0.029715	-1.318146	-0.174176	1.769492	1.568563	2
997	1	-2.675205	1.511012	1.076613	-0.641002	-0.641625	-1.080390	0.204138	-0.174176	1.769492	-0.439806	2
998	1	-1.765581	0.674370	-0.130475	-0.641002	0.430699	-0.220747	1.421966	0.324638	1.769492	2.244851	2
999	1	0.053668	0.159514	0.308466	-0.560326	-0.909705	-0.507294	-0.100319	0.424401	-0.754471	0.707833	2

1000 rows × 12 columns

FIGURE 7. Data after Scaling by Standardization

TABLE 3. The Previous Authors' Experimental Results

No	*AuthorsName*	*Model*	*Precision*	*Recall*	*F1Score*	*Accuracy*
1	(Alhussan et al., 2023)	DT	90.9%	75.33%	77.9%	73%
		RF	95.30%	96.80%	96.10%	95.19%
2	(Mehedi Hassan et al., 2022)	DT	90.9%	84.2%	90.9%	88.5%
		RF	96.03%	98.40%	97.20%	98.03%

We have implemented the statistical measure (Pearson's correlation coefficient), calculated for each input variable with the target. Depending on the correlation of the features with the target and the feature importance scores obtained from the built models.

we have selected a subset of our features, including the three risk factors (Gender, Age, and BMI), to try building a very simple model with no laboratory result variables. These three features, especially BMI and Age, still significantly correlate with the target. The correlation coefficients of the three variables concerning the target variable are (Gender is 0.1), (Age is 0.44), and (BMI is 0.58), We used these three features to build our decision tree model. respectively. figure 8 show the decision tree confusion matrix and features before and after data-balancing also with selecting the three risk factors, also figure 9 shows the random forest using cross validation. we use cross validation and select features from the datasets using random forest classifier shown in figure 9 and the value of the predicted classes.

Our result denominates that our new proposed algorithm performs very well. The decision tree-based classifier with data balancing gives the highest accuracy (99.3%) compared to other built models.

After Balancing by oversampling the minority class, we have built the second decision tree model It shows an accuracy of also (99.3%). Next, We built the third decision tree model using a feature subset of the three risk factors (Gender, Age, and BMI), which

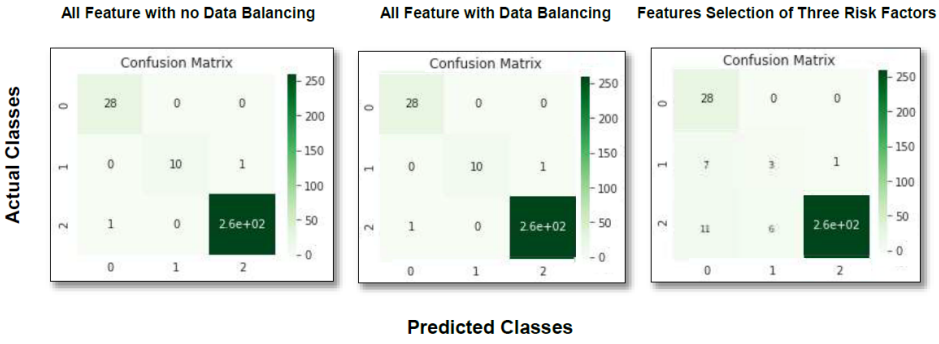


FIGURE 8. The Classification Model (Decision Tree) Results – Confusion Matrix

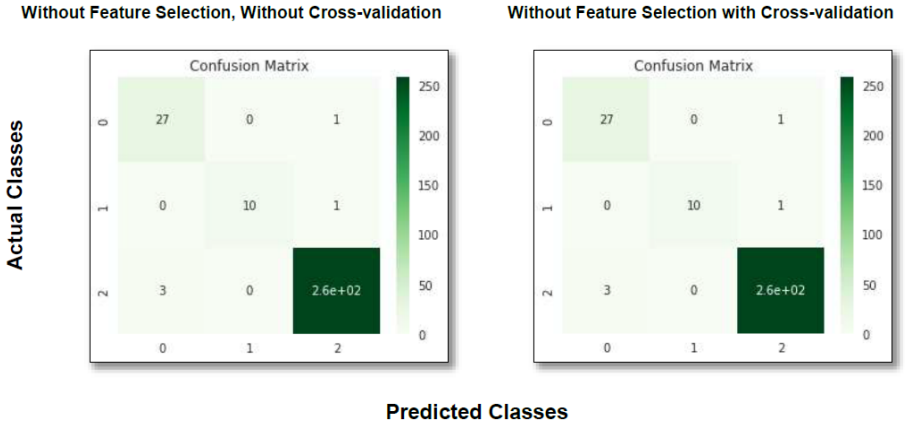


FIGURE 9. The Classification Model (Random Forest) Results-Confusion Matrix

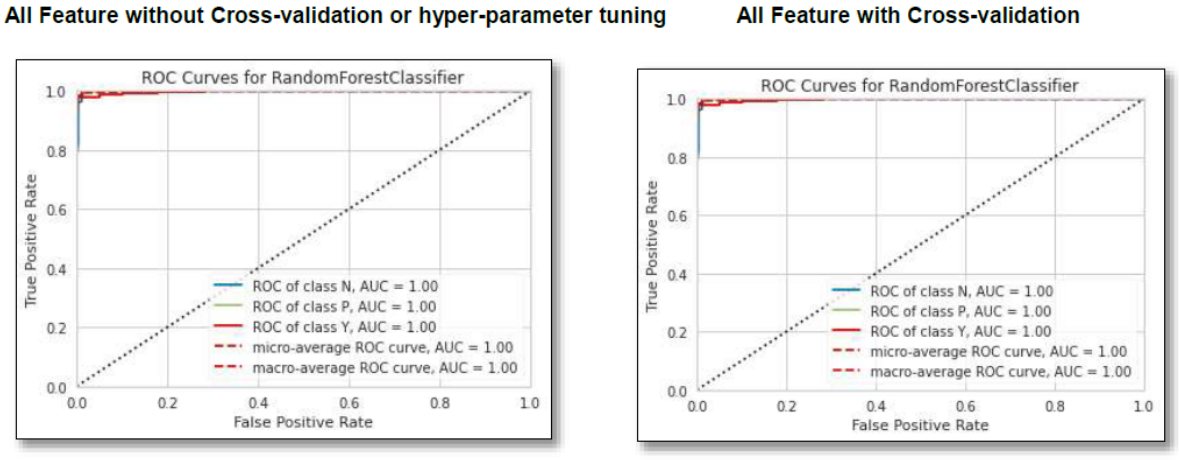
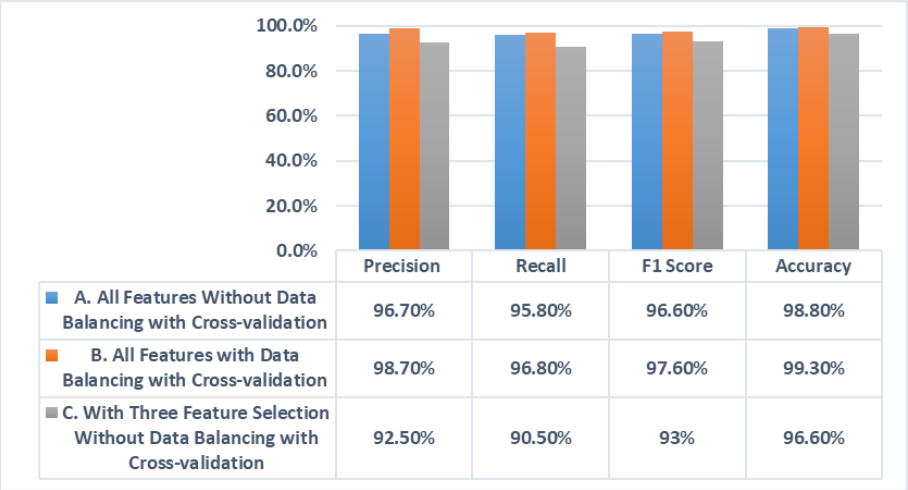


FIGURE 10. The Classification Models (Random Forest) -ROC/AUC

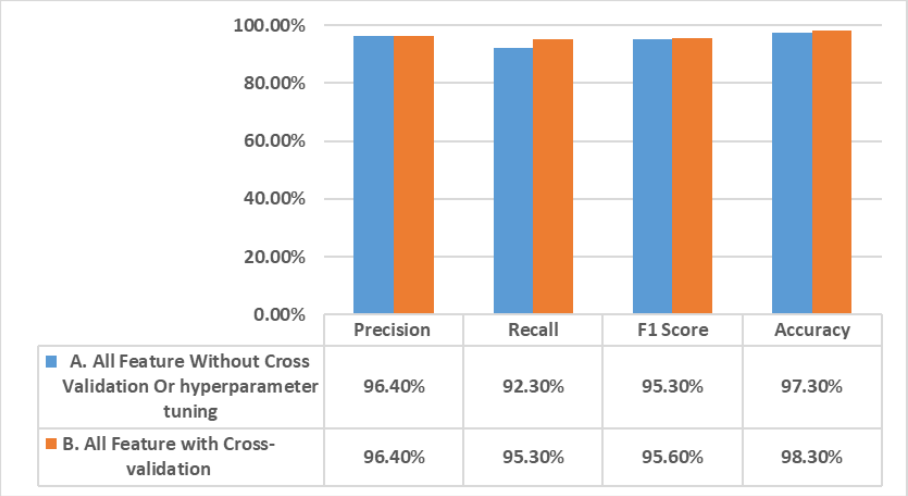
has a accuracy score (96.6%). We fine-tuned the three decision tree models using 10-fold cross-validation techniques to determine the models' unbiased estimates and compare their performances.

The results of classification models with and without cross validation shown in figure 10. The performance of proposed method is re ported in Table 4 with figure 11 for decision tree and figure 12 for random Forest.



	Precision	Recall	F1 Score	Accuracy
A. All Features Without Data Balancing with Cross-validation	96.70%	95.80%	96.60%	98.80%
B. All Features with Data Balancing with Cross-validation	98.70%	96.80%	97.60%	99.30%
C. With Three Feature Selection Without Data Balancing with Cross-validation	92.50%	90.50%	93%	96.60%

FIGURE 11. Decision Tree Proposed Method Experimental Results



	Precision	Recall	F1 Score	Accuracy
A. All Feature Without Cross Validation Or hyperparameter tuning	96.40%	92.30%	95.30%	97.30%
B. All Feature with Cross-validation	96.40%	95.30%	95.60%	98.30%

FIGURE 12. Random forest Proposed Method Experimental Results

TABLE 4. Proposed Method Experimental Results

Model Name	**Classification**	**Precision**	**Recall**	**F1 Score**	**Accuracy**
Decision Tree	A. All Features Without Data Balancing with Cross-validation	96.7%	95.8%	96.6%	98.8%
	B. All Features with Data Balancing with Cross-validation	98.7%	96.8%	97.6%	99.3%
	C. With Three Feature Selection Without Data Balancing with Cross-validation	92.5%	90.5%	93%	96.6%
Random Forest	A. All Features Without Cross-validation or Hyper-parameter Tuning	96.4%	92.3%	95.3%	97.3%
	B. All Features with Cross-validation	97.4%	95.3%	95.6%	98.3%

7. **Conclusions.** In this study, machine learning techniques were employed to classify and predict three types of diabetes. We conducted a comparative analysis of supervised learning methods, focusing on two techniques—decision tree and random forest—evaluating their performance with various hyperparameters and selection criteria. Our algorithm utilized three key feature selections to assess the accuracy results.

The decision tree model, when utilizing all available variables, yielded the best performance in terms of classification accuracy, precision, recall, and F1-score. However, the decision tree model, when limited to just three risk factor variables, also demonstrated strong performance. This simplified model is particularly useful for identifying individuals at high risk during initial screenings. It is well-suited for public health initiatives aimed at the general population, as it helps reduce screening costs while effectively targeting those most likely to benefit from intervention.

In the future, Using big data to understand the epidemiology of diabetes and to track treatment outcomes across populations could lead to more effective public health strategies and tailored therapies for individuals.

REFERENCES

[1] B. S. Ahamed, M. S. Arya, and A. O. Nancy V, "Prediction of Type-2 Diabetes Mellitus Disease Using Machine Learning Classifiers and Techniques," Frontiers in Computer Science, vol. 4. Frontiers Media S.A., May 10, 2022. doi: 10.3389/fcomp.2022.835242.

[2] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Predicting Diabetes Mellitus with Machine Learning Techniques," Front Genet, vol. 9, Nov. 2018, doi: 10.3389/fgene.2018.00515.

[3] N. A. Farooqui, R., and A. Tyagi, "Prediction Model for Diabetes Mellitus Using Machine Learning Techniques," International Journal of Computer Sciences and Engineering, vol. 6, no. 3, Mar. 2018, pp. 292–296, doi: 10.26438/ijcse/v6i3.292296.

[4] J. Davies, "A Comparative Analysis of Destructive Methods and Non-Destructive Methods with Machine Learning and Deep Learning Approaches for Rice Leaf Disease Identification," Journal of Information Hiding and Multimedia Signal Processing, vol. 15, no. 2, 2024, pp. 87–97.

[5] T. Nguyen Tu, "A new operator of enhancing dark images and application in binarizing dark images," Journal of Information Hiding and Multimedia Signal Processing, vol. 15, no. 1, 2024, pp. 1–9.

[6] M. Alghamdi, M. Al-Mallah, S. Keteyian, C. Brawner, J. Ehrman, and S. Sakr, "Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford Exercise Testing (FIT) project," PLoS One, vol. 12, no. 7, Jul. 2017, doi: 10.1371/journal.pone.0179805.

[7] T. Abbas et al., "Predicting long-term type 2 diabetes with support vector machine using oral glucose tolerance test," PLoS One, vol. 14, no. 12, Dec. 2019, doi: 10.1371/journal.pone.0219636.

[8] M. Minarul, I. Raju, S. Sarker, and M. M. Islam, "Chronic Kidney Disease Prediction using Ensemble Machine Learning," Journal of Information Hiding and Multimedia Signal Processing , vol. 14, 2023, no. 1, pp. 1–9.

[9] Hussain and S. Naaz., "Prediction of diabetes mellitus: Comparative study of various machine learning models," in Advances in Intelligent Systems and Computing, Springer, 2021, pp. 103–115. doi 10.1007/978-981-15-5148-2-10.

[10] Vukelić I, Šuša B, Klobučar S, Buljević S, Liberati Pršo A-M, Belančić A, Rahelić D, Detel D. Exosome-Derived microRNAs: Bridging the Gap Between Obesity and Type 2 Diabetes in Diagnosis and Treatment. Diabetology, vol. 5, no. 7, 2024;pp:706-724. https://doi.org/10.3390/diabetology5070052

[11] F. Alaa Khaleel and A. M. Al-Bakry, "Diagnosis of diabetes using machine learning algorithms," Mater Today Proc, Jul. 2021, doi: 10.1016/j.matpr.2021.07.196.

[12] S. Malik, S. Harous, and H. El-Sayed, "Comparative analysis of machine learning algorithms for early prediction of diabetes mellitus in women," in Lecture Notes in Networks and Systems, Springer, 2021, pp. 95–106. doi: 10.1007/978-3-030-58861-8-7.

[13] N. P. Miriyala, R. L. Kottapalli, G. P. Miriyala, G. Lorenzini, C. Ganteda, and V. A. Bhogapurapu, "Diagnostic Analysis of Diabetes Mellitus Using Machine Learning Approach," Revue d'Intelligence Artificielle, vol. 36, no. 3, pp. 347–352, Jun. 2022, doi: 10.18280/ria.360301.

[14] O. Daanouni, B. Cherradi, and A. Tmiri, "Type 2 Diabetes Mellitus Prediction Model Based on Machine Learning Approach," 2020, pp. 454–469. doi: 10.1007/978-3-030-37629-1-33.

[15] R. Saxena, S. K. Sharma, M. Gupta, and G. C. Sampada, "A Novel Approach for Feature Selection and Classification of Diabetes Mellitus: Machine Learning Methods," Comput Intell Neurosci, vol. 2022, 2022, doi: 10.1155/2022/3820360.

[16] S. Padhy, S. Dash, S. Routray, S. Ahmad, J. Nazeer, and A. Alam, "IoT-Based Hybrid Ensemble Machine Learning Model for Efficient Diabetes Mellitus Prediction," Comput Intell Neurosci, vol. 2022, 2022, doi: 10.1155/2022/2389636.

[17] Kaur and V. Kumari, "Predictive modeling and analytics for diabetes using a machine learning approach," Applied Computing and Informatics, vol. 18, no. 1–2, pp. 90–100, Jan. 2022, doi: 10.1016/j.aci.2018.12.004.

[18] M. S. Islam, M. K. Qaraqe, H. T. Abbas, M. Erraguntla, and M. Abdul-Ghani, "The prediction of diabetes development: A machine learning framework," in Middle East Conference on Biomedical Engineering, MECBME, IEEE Computer Society, Oct. 2020. doi: 10.1109/MECBME47393.2020.9292043.

[19] N. F. Shafiq, "Exploring the Benefits of Feature Selection Based on Bat Algorithm and Deep Learning in Brain Cancer Diagnosis," *Al-Esraa University College Journal for Engineering Sciences*, vol. 6, no. 9, pp. 68–79, Dec. 2024, doi: 10.70080/2790-7732.1005.

[20] M. Maniruzzaman et al., "Accurate Diabetes Risk Stratification Using Machine Learning: Role of Missing Value and Outliers," J Med Syst, vol. 42, no. 5, May 2018, doi: 10.1007/s10916-018-0940-7.

[21] S. Kumari, D. Kumar, and M. Mittal, "An ensemble approach for classification and prediction of diabetes mellitus using a soft voting classifier," International Journal of Cognitive Computing in Engineering, vol. 2, Jun. 2021, pp. 40–46, doi: 10.1016/j.ijcce.2021.01.001.

[22] "Rashid, Ahlam , 'Diabetes Dataset,' Mendeley Data, V1, 2020 doi: 10.17632/wj9rwkp9c2.1".

[23] R. Kowsar and A. Mansouri, "Multi-level analysis reveals the association between diabetes, body mass index, and HbA1c in an Iraqi population," Sci Rep, vol. 12, no. 1, Dec. 2022, doi: 10.1038/s41598-022-25813-y.

[24] M. R. Rajput and S. S. Khedgikar, "Diabetes prediction and analysis using medical attributes: A Machine learning approach", doi: 10.37896/JXAT14.01/314405.

[25] S. S. Qasim, R. J. Khanjar, J. N. Hasoon, B. A. Abad, A. H. Fadil, and Shajan. M. Alsowaidi, "An Intelligent System Using Deep Learning for Healthcare Monitoring in Light of the COVID-19 and Future Pandemics Based on IoT," *Al-Esraa University College Journal for Engineering Sciences*, vol. 6, no. 9, pp. 48–67, Dec. 2024, doi: 10.70080/2790-7732.1004.

[26] Jian Y, Pasquier M, Sagahyroon A, Aloul F. A Machine Learning Approach to Predicting Diabetes Complications. Healthcare (Basel).vol. 9, no. 12, Dec 2021, doi: 10.3390/healthcare9121712.

[27] M. A. Aboelazm, "Design and implementation of analog modulated signals radio monitoring receiver based on SDR technology," Journal of Information Hiding and Multimedia Signal Processing, vol. 13, no. 1, 2022, pp. 64–77.

[28] P.-H. Dinh et al., "A novel based on parameter optimization for enhancing images," Journal of Information Hiding and Multimedia Signal Processing , vol. 13, no. 1, 2022, pp. 98–105.