RESEARCH ARTICLE | MAY 21 2026

Diabetes prediction approach using machine learning and a novel dataset FREE

Maalim A. Aljabery ✉; Zainab N. Nemer; Adalla M. Chyaid

AIP Conf. Proc. 3408, 040162 (2026)
https://doi.org/10.1063/5.0329621

Diabetes Prediction Approach Using Machine Learning and a Novel Dataset

Maalim A. Aljabery[1, a)], Zainab N. Nemer[1, b)], Adalla M. Chyaid[1, c)]

[1]*Computer Science Dept., College of Computer Science and Information Technology, University of Basrah, Basrah, Iraq*

[a)]*Corresponding author: maalim.aljabery@uobasrah.edu.iq*
[b)]*zainab.nemer@uobasrah.edu.iq.*
[c)]*adala.gyad@uobasrah.edu.iq.*

Abstract. Diabetes is a chronic case with major health risks, making early diagnosis critical for appropriate intervention and effective management. Although current diagnostic techniques have limitations in correctness and consistency, assure the requirement for more functional predictive tools. This study handles this issue by applying Machine Learning (ML) techniques to predict diabetes, utilizing diabetes_prediction_dataset.csv, which is publicly obtainable from Kaggle. This dataset covers diverse medical and demographic attributes along with the patient's diabetes condition (positive or negative). Our research goal is to evaluate the predictive performance of four ML techniques: Stochastic Gradient Descent (SGD), Support Vector Machine (SVM), Random Forest (RF), and Constant classifier, a simple baseline method. The SGD, SVM, and RF models reached optimal predictive achievement, with Area Under the Curve (AUC), classification accuracy, precision, recall, and F1 score 1.0. This research demonstrated its superior ability to classify diabetic as well as non-diabetic patients accurately. As a comparison, the Constant classifier achieved a classification accuracy of 0.915 while presenting a limited discriminative ability, as reflected in an AUC of 0.500, a recall of 0.915, a precision of 0.837, and an F1 score of 0.874. All these results emphasize the effectiveness of advanced ML techniques, and one of the most effective in accurately predicting diabetes in the real world is SGD. This study highlights the potential of ML in healthcare applications.

Keywords: diabetes, machine learning, stochastic gradient descent, support vector machine, random forest, constant classifier, classification accuracy

INTRODUCTION

Diabetes is an escalating health concern. Although accurate diagnostic methods are available for identifying diabetes, predictive tools to assess the likelihood of a healthy individual developing diabetes in the future are less common. To combat diabetes effectively, it is essential to develop strategies that can aid healthcare organizations in predicting the risk of a non-diabetic patient becoming diabetic shortly [1]. Failure to identify high-risk groups for diabetes can lead to severe health complications, increasing the burden of the disease. Predictive tools, therefore, can potentially reduce substantial medical costs and constraints for diabetic patients while enhancing timely interventions and treatments.

Currently, the commercial application of predictive tools for diabetes diagnosis through web and mobile applications is limited. To address this gap, we employ classical data mining techniques to build predictive models integrated into a user-friendly platform that includes both a web service and a mobile application. This solution offers accessibility and interactivity, combining model expertise within an intuitive interface. Our work utilizes tabular data from a public domain dataset, which includes a substantial representation of individuals with diabetes or diabetes-related conditions. Our predictive approach aims to serve as an auxiliary diagnostic tool, potentially assisting in early medical intervention. We developed this prototype using classification techniques, specifically SGD, SVM, RF, and a Constant classifier by implementing a website and a mobile app using Android and the Ionic framework.

The International Scientific Conference of Engineering Sciences and Advances Technologies
AIP Conf. Proc. 3408, 040162-1–040162-13; https://doi.org/10.1063/5.0329621

The research is organized into several sections to provide a clear study overview. It begins with an Introduction, which includes the background, objectives, and significance of the research topic. This is followed by a Literature Review, discussing existing studies and highlighting gaps addressed by our research. The Methodology section outlines the dataset, classification models, and performance metrics used for evaluation. In the Results section, we present the findings, focusing on model performance. The Results Discussion analyzes these findings and addresses study limitations. Finally, the Conclusion summarizes the key outcomes and suggests future research directions, with a comprehensive list of References at the end.

LITERATURE REVIEW

Diabetes is a prevalent disease worldwide, consuming a significant portion of healthcare resources. Many individuals exhibit prediabetic symptoms but may be unaware of them. If these early warning signs are not addressed, they can progress to more severe conditions, such as cardiovascular disease and hypertension [2]. Several statistical methods, including various ML and linear regression techniques, have been applied in the analysis of diabetes. However, few data-driven models have been developed to analyze patterns leading to prediabetes, and limited research has focused on risk prediction and the long-term outcomes of prediabetes [3].

In this research, we conducted survival analysis using four ML techniques, Stochastic Gradient Descent (SGD), Support Vector Machine (SVM), Random Forest (RF), and a Constant classifier, to examine factors associated with the development of prediabetes worldwide. We also analyzed competing risk factors related to prediabetes progression. Rather than censoring diabetic patients, we focused on the onset of prediabetes as the primary event of interest.

Our results identified several key factors that increase the risk of prediabetes among individuals with diabetes. Notably, high intake of sugar-sweetened beverages and prolonged sitting emerged as significant modifiable factors associated with elevated risk of prediabetes. Conversely, having a supportive environment for physical activity, walking at least 10 minutes, and engaging in some forms of exercise significantly reduces the risk of developing prediabetes.

Previous studies have developed AI algorithms for the early diagnosis of diabetes based on data from participants who later developed the disease. However, few long-term studies have focused on identifying key factors for prediabetes development using large, novel datasets, particularly with a specific emphasis on prediabetes analysis.

RELATED WORKS

The summaries of previous studies utilizing ML methods are presented below:

- Jaggi, et.al. [4], in this study, the researchers used supervised ML algorithms like SVM, Naive Bayes (NB) classifier, and LightGBM to train on the actual data of 520 patients suffering from diabetes and the researchers' potential diabetic patients age within the range of 16 to 90. Through comparative analysis of recognition and classification accuracy, the performance of SVM was the best.
- Xue, et.al. [5], this study found that the SVM achieved the highest accuracy depending on evaluations using a confusion matrix. However, it should be continually updated with additional datasets to improve reliability. Data Mining (DM) algorithms and ML techniques have significantly contributed to medical diagnostics, offering valuable support to clinicians in assessing disease status.
- Gündoğdu, Serdar [6], this research introduced a method based on a combination of multiple linear regression (MLR), RF, and XGBoost (XG) to diagnose diabetes from questionnaire data. MLR-RF algorithm is applied to select the features, and XG is applied for classification. The dataset used is the diabetic data from Sylhet, Bangladesh. It includes 520 instances, 320 diabetics, and 200 control instances. The classifier's performance is measured concerning ACC, precision, F1 score, recall (SEN, sensitivity), and AUC. The results show that the proposed system achieves an AUC of 99.3%, an accuracy of 99.2%, and a prediction time reached 0.04825 seconds.
- Bhat, et.al. [7], analyze a clinical dataset collected from a doctor in the Indian district of Bandipora between April 2021 and February 2022. Machine Learning Algorithms MLA are increasingly significant in healthcare because of their predictive capabilities which help enhance disease prediction and reduce costs. The study presents a methodology for diabetes risk prediction in North Kashmir using six MLAs: RF, Multi-Layer Perceptron (MLP), SVM, Gradient Boosting (GB), Decision Tree (DT), and Logistic Regression (LR). Among these, RF achieved the highest accuracy at 98%, followed by MLP at 90.99% and SVM at 92%.

- Viswanatha, et.al. [8], the study effectively identifies diabetes prevalence and enhances prediction accuracy. This study presents a predictive model to determine the likelihood of diabetes development based on diagnostic measures from the PIMA Indians Diabetes dataset and one from Vanderbilt related to rural African Americans in Virginia. Logistic regression serves as the primary algorithm, implemented in Python IDEs. Feature selection is performed using two methods, and aggregation techniques like Maximum Voting enhance model performance. The study achieved accuracies of approximately 78% for dataset 1 and 93% for dataset 2. Additionally, factors such as data preprocessing and cross-validation are identified as crucial for improving model accuracy and runtime.

RESEARCH OBJECTIVES

Diabetes prediction using gradient-boosting-based ML models has garnered significant research interest, fueled by recent advancements in big data and data analytics. The availability of various sensors, such as multichannel Bluetooth and Z-wave smart sensors for personal health monitoring, has led to the generation of large volumes of real-time, multivariate clinical and temporal data [9]. Despite extensive research on diabetes prediction and associated risk factors, there has been limited scholarly discussion on predicting diabetes using a combination of meteorological variables and risk factors derived from blood test results [10]. The primary goal of this study is to predict diabetes using ML Models and to provide insights into disease progression. We also aim to forecast future progression rates based on changes in hemoglobin, glucose, red blood cells, and cholesterol levels.

The specific objectives of this research are as follows:

- Develop an ML model to predict diabetes trends using hemoglobin, glucose, red blood cells, and cholesterol blood test results.
- Predict the progression from normal to abnormal diabetes levels using an ML model.
- Identify the most effective features and ML model for predicting future diabetes probabilities.
- Forecast future disease progression rates based on increases or decreases in hemoglobin, glucose, red blood cells, and cholesterol levels.
- Provides a recommendation for managing and preventing diabetes progression based on the results, which could be valuable for AI-driven health platforms.

BACKGROUND AND SIGNIFICANCE

This research proposes types of neural networks combined with a voting algorithm to predict diabetes based on a diabetes dataset. Given that different error metrics can yield varying results, four error metrics were used. The data was divided into training and testing sets with tenfold cross-validation to train the models, through which unique attributes were identified that enhanced model performance rather than relying on local model attributes. Experimental results show that prediction accuracy for the testing data is sufficiently high, indicating that combining unique attributes, deep learning, and a voting method produces reliable results. This approach shows promise for use in screening tools or assessing an individual's potential risk of developing diabetes.

This study also investigates key risk factors including personal, pregnancy-related, blood, and urine test metrics for diabetes prediction. Based on these factors, prediction models (including SGD, SVM, RF, and a Constant classifier focusing on the majority class) were developed to identify individuals at risk of diabetes. The research includes two versions of the dataset, differing in size: The smaller dataset provides only 10 out of 25 local factors and includes personal information. Model evaluation results and significant findings for both dataset types are presented. ***Figure 1*** illustrates a block diagram of the workflow for diabetes prediction analysis using ML algorithms.

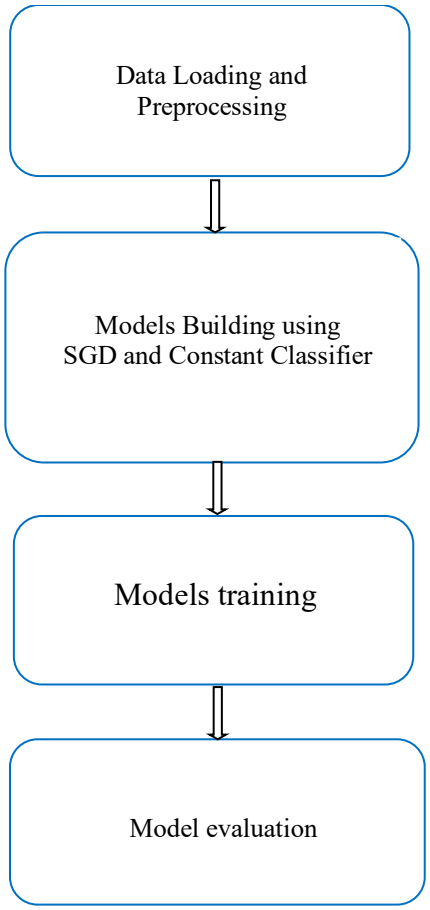


FIGURE 1. Block Diagram of Workflow for Diabetes Prediction Analysis Using ML Algorithms

Healthcare systems are increasingly utilizing ML techniques to improve the early detection and management of chronic diseases, like diabetes. This status comes from the ability of the body to produce sufficient insulin or effectively use it. So, if the patient is left untreated, diabetes becomes progressively damages vital organs such as the heart, liver, kidneys, eyes, and brain [11].

To predict diabetes when analyzing fundamental health indicators like insulin and blood glucose levels, Body Mass Index (BMI), and Hemoglobin, researchers have benefited from ML and Deep Learning (DL) techniques. Within this concept, advanced predictive techniques, which depend on stacked ensembles' methods, have demonstrated the perspective of enhancing diagnostic accuracy as well as supporting clinical decision-making [11].

Two studies have utilized hybrid feature-chosen techniques to make efficient diabetes predictions, like Combining Correlation Matrices (CCM) and Sequential Forward Selection (SFS). Techniques such as Random Forests (RF), Support Vector Machines (SVM), and Artificial Neural Networks (ANN) were used to examine the effectiveness of these characteristics and features. Metrics related to accuracy, AUC, precision, F1 score, and recall are commonly recorded [12].

In this research, our group applied SGD, SVM, RF, and Constant classifiers, focusing on these techniques metrics to evaluate the reliability and efficiency of diabetes predictions.

METHODOLOGY

Dataset Source

The dataset of diabetes disease prediction is a combination of medical data and demographic information along with the diabetes patients' status (even if it is positive or negative). These data include attributes like gender, age, body mass index (BMI), heart disease, hypertension, HbA1c level, blood glucose level, and smoking history [13].

It can be used to construct ML models to predict diabetes cases in patients related to their demographic information and medical history. This dataset can help all healthcare professionals identify patients at risk of developing diabetes and personalize treatment programs for them. Furthermore, researchers can use this dataset to examine relationships among diverse medical and demographic features and the probability of developing diabetes. The *diabetes_prediction_dataset.csv* file contains medical data and patients' demographic information related to their diabetes status (positive or negative).

It contains diverse attributes that can be used to build ML models to predict the risk of diabetes in patients, taking into account their medical history, demographics, and other details. Table 1, illustrates the description of these attributes.

TABLE 1. Diabetes Prediction Dataset: A Comprehensive Dataset for Predicting Diabetes with Medical and Demographic Data

Feature name	Description
gender	It refers to an individual's biological sex, which can influence their susceptibility to diabetes. The categories include male, female, and other.
age (0.08 - 80)	It is a significant factor, as diabetes is more frequently diagnosed in older adults. The age ranges from 0 to 80 years.
hypertension (0 - 1)	It is a medical condition characterized by persistently elevated blood pressure in the arteries. In the dataset, it is represented by a binary value: 0 indicates the absence of hypertension, while 1 signifies its presence.
Heart_disease (0 - 1)	It is another medical condition linked to a higher risk of developing diabetes. In the dataset, it is represented by a binary value: 0 indicates the absence of heart disease, while 1 signifies its presence.
Smoking_history	It is also recognized as a risk factor for diabetes and can worsen the complications associated with the condition. In our dataset, there are five categories: not current, former, no information, current, and never.
BMI (10 - 95.7)	Body Mass Index (BMI) is a measure of body fat calculated from an individual's weight and height. Elevated BMI values are associated with an increased risk of diabetes. In the dataset, BMI ranges from 10.16 to 71.55. A BMI of less than 18.5 is classified as underweight, 18.5 to 24.9 is considered normal, 25 to 29.9 is categorized as overweight, and a BMI of 30 or higher is classified as obese.
HbA1c_level (3.5- 9)	Hemoglobin A1c (HbA1c) level reflects an individual's average blood sugar levels over the previous 2 to 3 months. Elevated HbA1c levels are associated with a higher risk of developing diabetes, with levels exceeding 6.5% typically indicating the presence of diabetes.
blood_ glucose_ level (80 - 300)	Blood glucose level indicates the concentration of glucose in the bloodstream at any given moment. Elevated blood glucose levels are a significant indicator of diabetes.
diabetes (0 - 1)	It is the target variable being predicted, where a value of 1 signifies the presence of diabetes and a value of 0 indicates its absence.
Label	Count
Female59%	No Info36%
Male41%	never35%
Other (18)0%	Other (29089)29%

Figures (2-8), show the dataset summary in the form of a data card that includes an overview of the dataset's purpose, feature summary, data range, and dataset size information on feature type and missing values. This figure allows for a quick high-level understanding of the dataset.

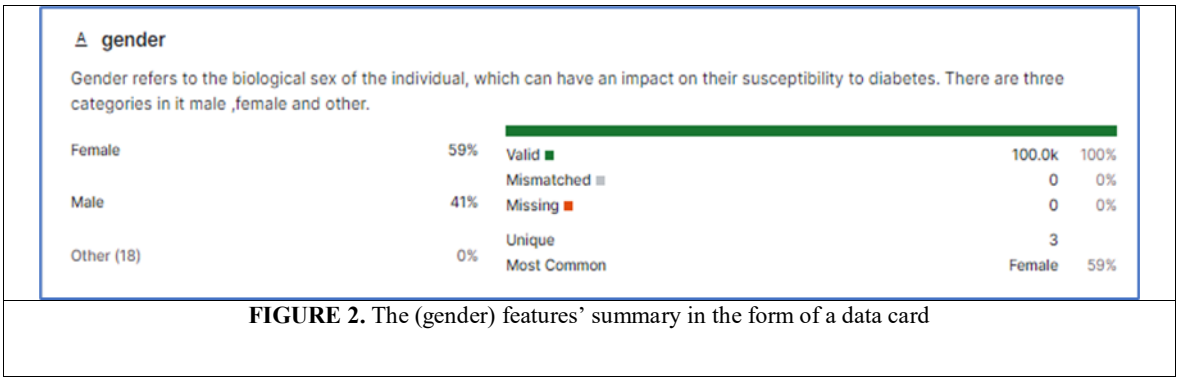


FIGURE 2. The (gender) features' summary in the form of a data card

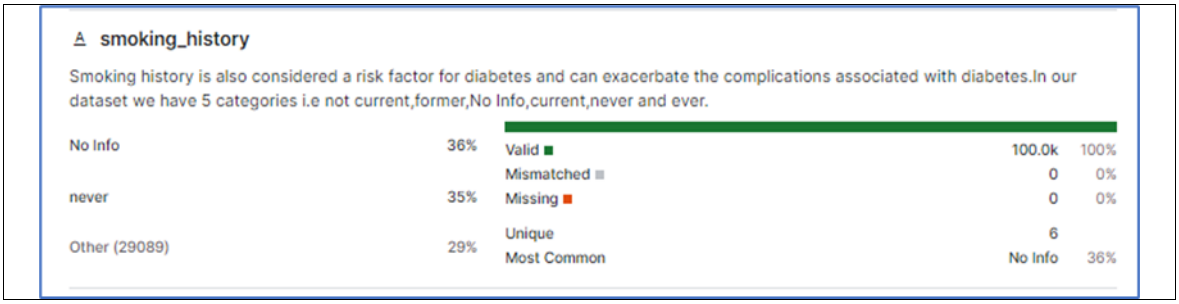


FIGURE 3. The (smoking history) features' summary in the form of a data card

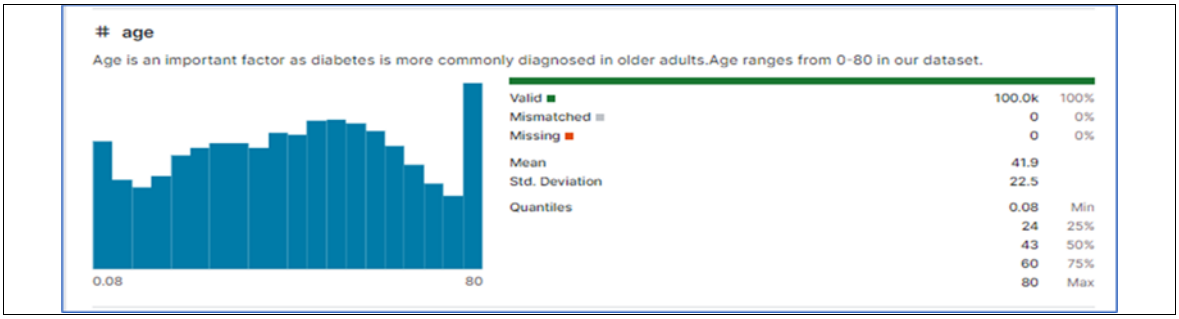


FIGURE 4. The (age) features' summary in the form of a data card

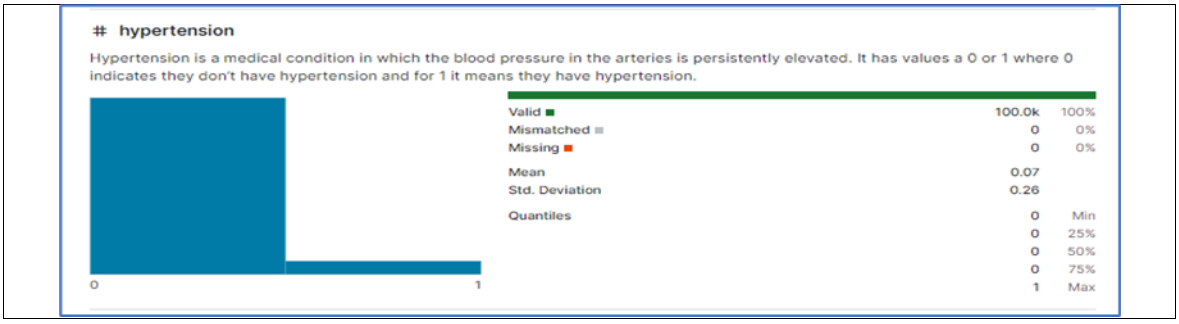


FIGURE 5. The (hypertension) features' summary in the form of a data card

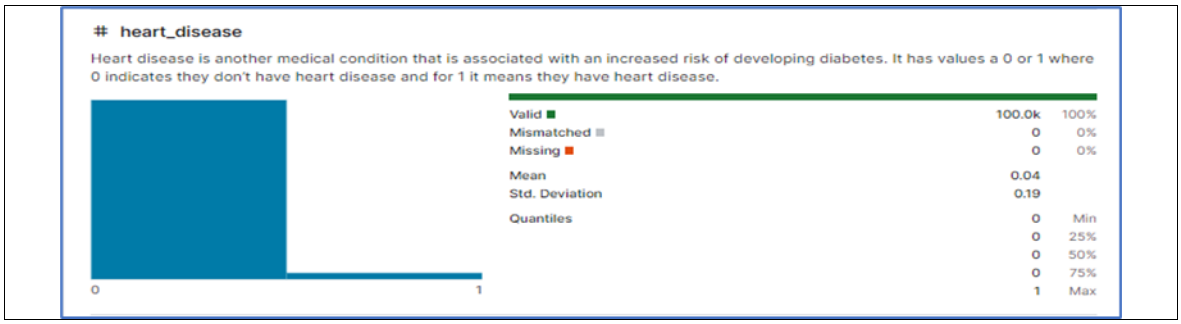


FIGURE 6. The (heart_disease) features' summary in the form of a data card

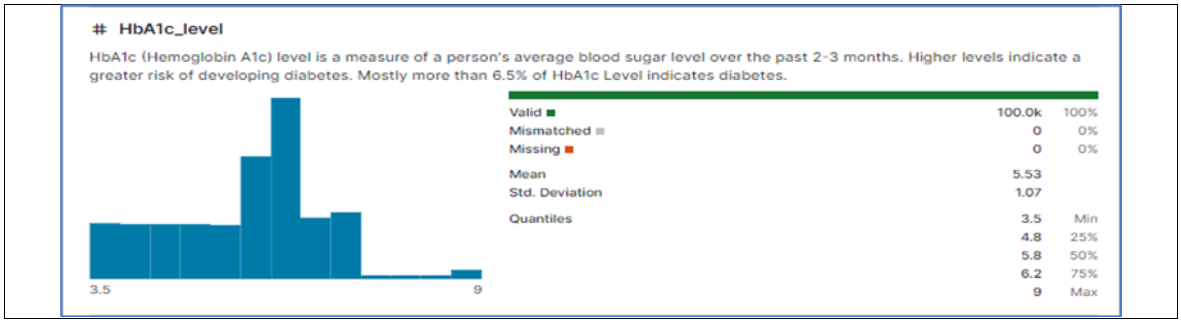


FIGURE 7. The (HbA1c_level) features' summary in the form of a data card

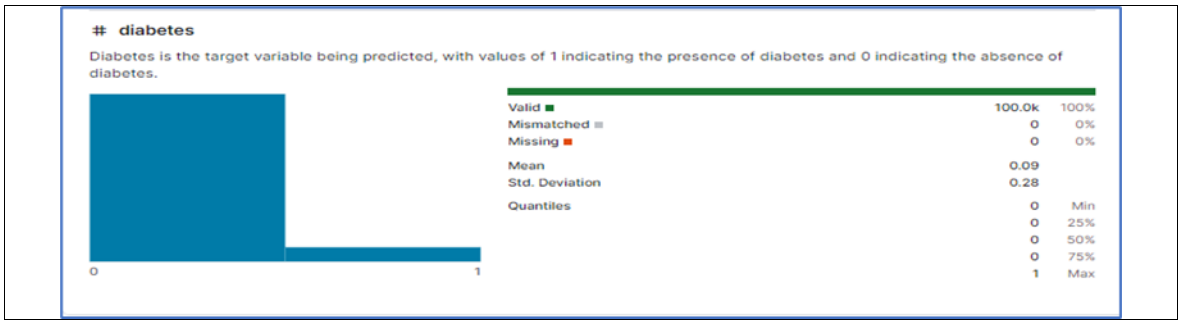


FIGURE 8. The (diabetes) features' summary in the form of a data card

Data Sample Selection

The details of the dataset selected are as follows:

- Sampling type: Random sample with 25 data instances, with replacement, stratified (if possible), deterministic.
- Input: 100000 instances.
- Sample: 25 instances.
- Remaining: 99975 instances.

The condition of the data sample selected is illustrated in Figure 9.

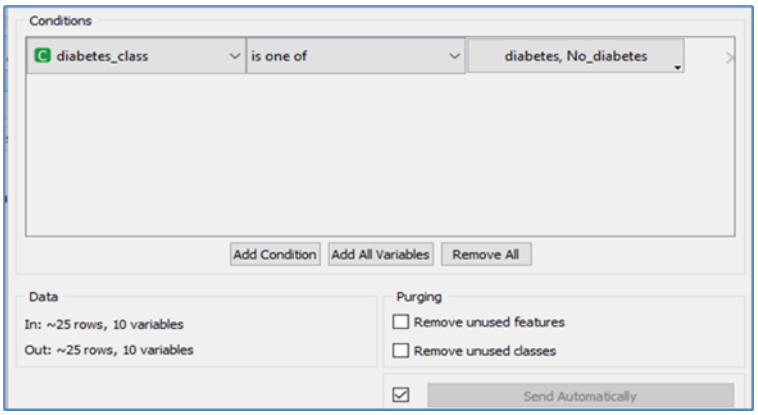


FIGURE 9. Data sample selected

Stochastic Gradient Descent (SGD)

In a basic supervised learning framework, each instance is represented as a pair consisting of an arbitrary input and a scalar output. A loss function is selected to quantify the cost of the predictions compared to the actual output, and a function parameterized by a weight vector is chosen. Stochastic Gradient Descent (SGD) updates the parameters based on each training instance, introducing randomness into the process based on the instances randomly selected during each iteration [14].

In a deployed system, this stochastic algorithm can process instances in a single pass since it does not require retaining information about which instances were handled in previous iterations. Consequently, it is generally much faster and can also be utilized for online learning. SGD performs frequent updates, which leads to high variance and causes significant fluctuations in the objective function. However, this variability enables SGD to explore new, potentially more optimal local minima. Despite this advantage, SGD often overshoots, making it difficult to reach the exact minimum without appropriate adjustments to the learning rate [15].

Support Vector Machine (SVM)

Vapnik first introduced SVMs. It depends on the binary classification, so it isolates the vector training set into two classes (a_1, b_1), (a_2, b_2),. ., (a_m, b_m), where $a_i \in R^d$ which indicates the vector of feature space of d-dimensional, and $b_i \in \{-1, +1\}$ is a label of the class [16].

SVM is an ML model widely used for classification and optimization tasks. To apply SVM, the first step is to collect labeled training data, where each item's category is pre-defined. The next step is to select features, choosing characteristics that best differentiate between categories. During training, the SVM algorithm identifies a dividing line between the two classes that maximizes the margin, or distance, between this line and the closest training points from each class, creating a clear boundary for classification [17].

Random Forest (RF)

They are an ensemble of tree predictors, where each tree is built based on results from an independently sampled random vector, with the same classification applied across all trees in the forest [18].

Constant Classifier

A classifier is called a "constant" if it assigns every example to the same class. A training dataset is described as "(linearly) indiscriminate" if a constant classifier minimizes the misclassification rate compared to all linear classifiers on that dataset. General sufficient conditions guarantee a positive probability of obtaining an indiscriminate dataset. Conversely, additional general sufficient conditions are provided to ensure that the probability of obtaining an indiscriminate dataset is zero [19]. In classification tasks, the training sample (data set D) consists of an ordered list of pairs (x_1, y_1) …. (x_n, y_n), where each pair (x_i, y_i) represents an "example." Here, xi is the "predictor vector" or

"input" for the i- th instance, and y_i is the associated class. The number of examples, n is called the "sample size," and k denotes the "dimension" of each predictor vector. In binary classification problems (common in learning theory), the label y_i is either +1 or -1, representing "positive" and "negative" classes, respectively [20].

For simplicity, suppose k >1. Note that it is possible for K≤ n, meaning there may be as many predictors as examples. When k ≥ n, fitting a linear classifier often requires regularization. In linear classification, the aim is to find an affine function as in equation 1:

$$f(x) = b + w.x \tag{1}$$

In most cases, correctly matches the class y_i for each xi in D, where $w \in R^k$ and $b \in R$. The symbol "·" represents the inner product in R^k, and sgn(t) denotes the sign of t, with sgn (0)=0 and for sgn(t)=t/|t| for t≠0 [20]. The "misclassification rate" of *f* on dataset *D* refers to the proportion of examples (x_i,y_i) for which the classifier disagrees. In this context, the conditions under which a constant classifier (with w=0) is the best-performing linear classifier, achieving the lowest misclassification rate on *D* compared to all other linear classifiers. This situation arises when dataset *D* possesses the property that a constant classifier minimizes the misclassification rate [21].

Performance Evaluation

In the field of classification problems, assessing model performance is essential for understanding its effectiveness and reliability. Various metrics provide insights into how well a model distinguishes between different classes. Among these, the Area Under the Curve (AUC) serves as a vital tool for assessing the model's discriminative ability across various threshold settings. Additionally, accuracy, precision, recall, and F1 score are fundamental metrics that help gauge the model's performance in correctly identifying positive instances and minimizing false predictions. We comprehensively evaluate our models by utilizing these metrics, ensuring the findings are robust, interpretable, and applicable in real-world scenarios. This evaluation is essential for refining models, improving predictive accuracy, and ultimately achieving more reliable outcomes in research studies [22], The metrics are:

Accuracy

Accuracy is a metric of the measure's overall classification technique effectiveness. Accuracy is defined as the attribution of exactly classified instances to the whole number of instances within the dataset. It is calculated by implementing the following formula [17]:

$$\text{Accuracy} = \frac{\text{True Positives (TP)} + \text{True Negatives (TN)}}{\text{Total Instances}} \tag{2}$$

Precision

Precision evaluates the value of the positive predictions made by the technique. It is the attribution of TP predictions to all positive predictions made (for both TP and false positives (FP). It is calculated by implementing the following formula [17]:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

Recall

Recall or sensitivity, also known as TP rate, measures the technique's capability to identify all pertinent instances and the attribution of TP predictions to all number of existent positive instances. It is calculated by implementing the following formula [17]:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{False Negatives (FN)}} \tag{4}$$

F1 Score

F1 Score is an efficiency mean of recall and precision; it provides an equivalence between these two metrics. It is particularly beneficial in scenarios where an uneven distribution of the classes is found. The F1 Score is calculated by implementing the following formula [17]:

$$\text{F1 Score} = \frac{2\ (\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})} \tag{5}$$

This structured explanation gives a clear overview of each metric, its significance, and the corresponding mathematical formulations, without directly borrowing from the original text.

AUC

The AUC curve is a fundamental metric used to evaluate the performance of classification models across different threshold settings. The area under the ROC curve (AUC) quantifies the model's ability to differentiate between classes [23].

A higher AUC value indicates a better capacity of the model to distinguish between the positive and negative classes. For instance, when the AUC is close to 1, it signifies that the model can accurately classify instances of disease versus no disease, suggesting strong predictive performance [16].

Where:

- True positive refers to the instances correctly classified as belonging to the positive class. True Negative represents the instances correctly classified as belonging to the negative class.
- False Positive occurs when the model incorrectly classifies a negative instance as positive.
- False Negative happens when a positive instance is incorrectly classified as negative [18].

RESULTS AND DISCUSSION

Figures 10 and 11 show two different tables summarizing the datasets used in the study. The first table (Data Table) contains a large dataset with 100,000 instances and 9 features, with no missing values. It also includes a discrete class with 2 values and no meta-attributes. This indicates that the original dataset is extensive, providing a sufficient number of samples for accurate and reliable analysis. In contrast, the second table (Sample Data Table) has only 25 instances but the same features and class properties as the original dataset. This smaller dataset represents a sample extracted from the larger one, potentially used for preliminary testing or to observe model performance on a limited subset before applying the whole dataset, as shown in ***Figure 12***.

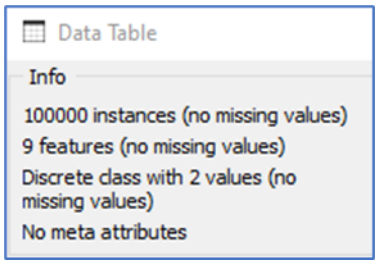


FIGURE 10. Data table details

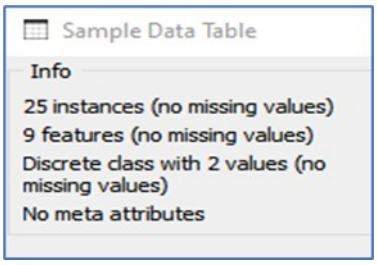
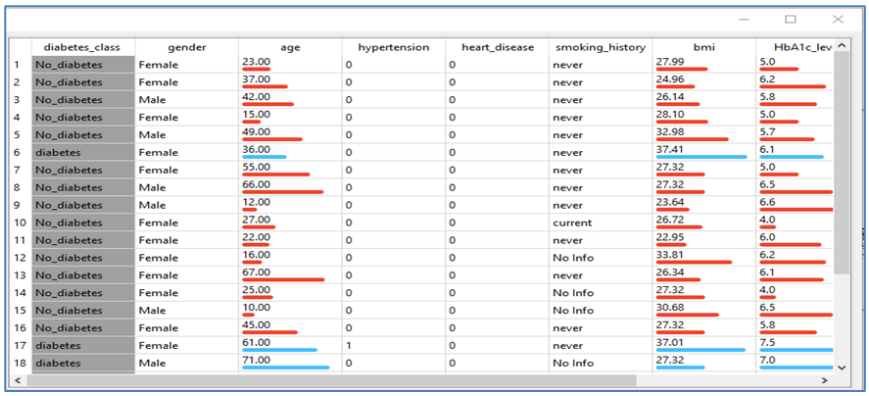


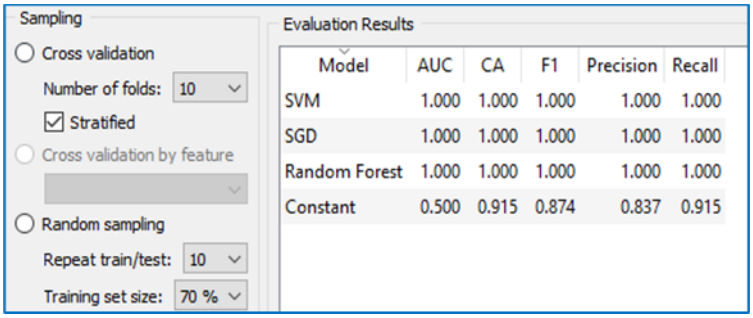
FIGURE 11. Sample of the data table

The results of this study underscore the promising potential of ML techniques, specifically SGD, SVM, and RF for accurately predicting diabetes. The impressive performance of these models achieving an AUC, Classification Accuracy, F1-score, Precision, and Recall all at 1.0 illustrates their capability to distinguish diabetic from non-diabetic patients with high reliability, as evaluation results are shown in Figure 13. This outcome suggests that advanced algorithms like SGD, SVM, and RF can offer critical support in healthcare diagnostics by providing highly accurate predictive insights, which could be essential in early diabetes detection and management.



	diabetes_class	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_lev
1	No_diabetes	Female	23.00	0	0	never	27.99	5.0
2	No_diabetes	Female	37.00	0	0	never	24.96	6.2
3	No_diabetes	Male	42.00	0	0	never	26.14	5.8
4	No_diabetes	Female	15.00	0	0	never	28.10	5.0
5	No_diabetes	Male	49.00	0	0	never	32.98	5.7
6	diabetes	Female	36.00	0	0	never	37.41	6.1
7	No_diabetes	Female	55.00	0	0	never	27.32	5.0
8	No_diabetes	Male	66.00	0	0	never	27.32	6.5
9	No_diabetes	Male	12.00	0	0	never	23.64	6.6
10	No_diabetes	Female	27.00	0	0	current	26.72	4.0
11	No_diabetes	Female	22.00	0	0	never	22.95	6.0
12	No_diabetes	Female	16.00	0	0	No Info	33.81	6.2
13	No_diabetes	Female	67.00	0	0	never	26.34	6.1
14	No_diabetes	Female	25.00	0	0	No Info	27.32	4.0
15	No_diabetes	Male	10.00	0	0	No Info	30.68	6.5
16	No_diabetes	Female	45.00	0	0	never	27.32	5.8
17	diabetes	Female	61.00	1	0	never	37.01	7.5
18	diabetes	Male	71.00	0	0	No Info	27.32	7.0

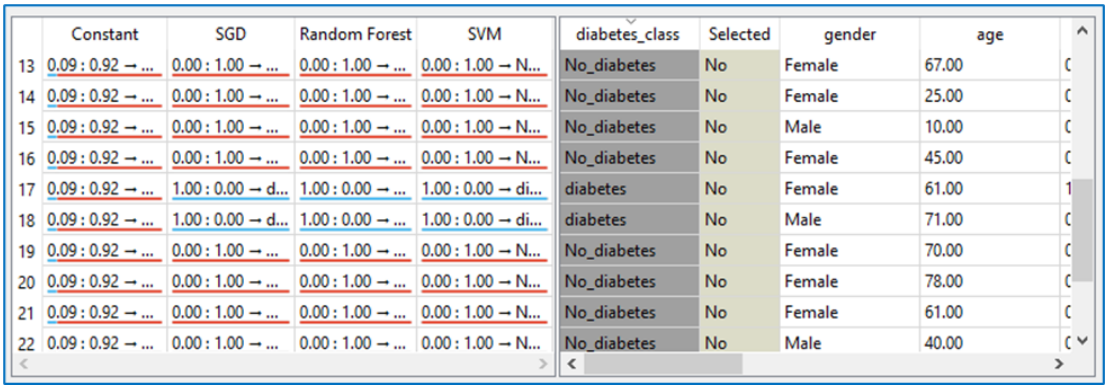
FIGURE 12. Application sample of dataset details with predictions highlighted (Blue: diabetes, Red: No_diabetes)



Model	AUC	CA	F1	Precision	Recall
SVM	1.000	1.000	1.000	1.000	1.000
SGD	1.000	1.000	1.000	1.000	1.000
Random Forest	1.000	1.000	1.000	1.000	1.000
Constant	0.500	0.915	0.874	0.837	0.915

FIGURE 13. Evaluation Results

In contrast, the Constant classifier, though demonstrating a reasonable classification accuracy of 0.915, failed to discriminate between diabetic and non-diabetic patients effectively, as shown by its AUC of 0.500. Effective, as indicated by its AUC of 0.500. The limited performance of the Constant classifier reinforces its usefulness as a baseline method rather than a reliable predictor. The low F1 score, Precision, and Recall values indicate an inability to capture complex patterns in the dataset successfully identified by other models, especially when it comes to distinguishing true positive cases. This highlights the importance of using advanced algorithms in predictive models for healthcare. The results are consistent with other findings in healthcare ML applications, where models such as SGD, SVM, and RF often outperform simpler classifiers because they improve decision boundaries on complex datasets. However, it is important to acknowledge that the use of a highly curated and balanced dataset may have contributed to these models' outstanding metrics. *Real-world healthcare data may introduce additional complexities, such as missing values and class imbalance*, that could affect model performance. Future research could address these aspects by incorporating larger, more diverse datasets to evaluate the robustness and generalizability of these models in different populations. In conclusion, our findings demonstrate that ML, particularly through algorithms like SGD, SVM, and RF, can enhance predictive accuracy in diabetes diagnostics compared with other techniques, as shown in ***Figure 14***.



	Constant	SGD	Random Forest	SVM	diabetes_class	Selected	gender	age
13	0.09 : 0.92 → ...	0.00 : 1.00 → ...	0.00 : 1.00 → ...	0.00 : 1.00 → N...	No_diabetes	No	Female	67.00
14	0.09 : 0.92 → ...	0.00 : 1.00 → ...	0.00 : 1.00 → ...	0.00 : 1.00 → N...	No_diabetes	No	Female	25.00
15	0.09 : 0.92 → ...	0.00 : 1.00 → ...	0.00 : 1.00 → ...	0.00 : 1.00 → N...	No_diabetes	No	Male	10.00
16	0.09 : 0.92 → ...	0.00 : 1.00 → ...	0.00 : 1.00 → ...	0.00 : 1.00 → N...	No_diabetes	No	Female	45.00
17	0.09 : 0.92 → ...	1.00 : 0.00 → d...	1.00 : 0.00 → ...	1.00 : 0.00 → di...	diabetes	No	Female	61.00
18	0.09 : 0.92 → ...	1.00 : 0.00 → d...	1.00 : 0.00 → ...	1.00 : 0.00 → di...	diabetes	No	Male	71.00
19	0.09 : 0.92 → ...	0.00 : 1.00 → ...	0.00 : 1.00 → ...	0.00 : 1.00 → N...	No_diabetes	No	Female	70.00
20	0.09 : 0.92 → ...	0.00 : 1.00 → ...	0.00 : 1.00 → ...	0.00 : 1.00 → N...	No_diabetes	No	Female	78.00
21	0.09 : 0.92 → ...	0.00 : 1.00 → ...	0.00 : 1.00 → ...	0.00 : 1.00 → N...	No_diabetes	No	Female	61.00
22	0.09 : 0.92 → ...	0.00 : 1.00 → ...	0.00 : 1.00 → ...	0.00 : 1.00 → N...	No_diabetes	No	Male	40.00

FIGURE 14. Application sample of compared prediction of the two models (Blue: diabetes, Red: No_diabetes)

By leveraging these techniques, healthcare systems can move closer to more reliable, data-driven decision-making processes, which are crucial for the early detection and timely intervention of chronic conditions like diabetes.

CONCLUSION

This study examined the effectiveness of ML techniques for diabetes prediction using a dataset obtained from Kaggle containing medical and demographic information of patients along with their diabetes status. The comparison among SGD, SVM, RF, and a Constant classifier provided insightful results. The SGD, SVM, and RF models performed exceptionally well, achieving perfect scores across all evaluation metrics (AUC = 1.0, CA = 1.0, F1 = 1.0, Precision = 1.0, and Recall = 1.0), demonstrating their ability to accurately predict diabetes status. In contrast, the Constant classifier showing a respectable Classification Accuracy (CA = 0.915), had poor discriminatory power with an AUC of 0.500, and reflected its limited utility for nuanced predictions. These results highlight the importance of employing advanced ML techniques for complex classification problems such as diabetes prediction. The high accuracy of the SGD, SVM, and RF models suggests that they can be effectively used in medical decision-making processes to support early diagnosis and treatment. Future work can build on this study by incorporating additional features or techniques for further refining prediction accuracy and model interpretability, potentially improving clinical outcomes for diabetic patients.

ACKNOWLEDGMENTS

We extend our sincere gratitude to the staff of Basra General Teaching Hospital, Chronic Diseases Department, Diabetes Patients Division, for their invaluable assistance in testing the application on a patient group to determine diabetes prediction accuracy and for granting us full access to their data.

REFERENCES

1. N. P. Tigga and S. Garg, “Prediction of type 2 diabetes using machine learning classification methods,” Procedia Comput. Sci., vol. 167, no. 2019, pp. 706–716, 2020.
2. A. Site, J. Nurmi, and e. S. Lohan, “machine-learning-based diabetes prediction using multisensor data,” IEEE Sens. J., vol. 23, no. 22, pp. 28370–28377, 2023.
3. I. Tasin, t. U. Nabil, s. Islam, and r. Khan, “diabetes prediction using machine learning and explainable AI techniques,” Healthc. Technol. Lett., Vol. 10, no. 1–2, pp. 1–10, 2023.
4. A. K. Jaggi, a. Sharma, n. Sharma, r. Singh, and P.S. Chakraborty, “diabetes prediction using machine learning,” lect. Notes networks syst., Vol. 185, No. 09, pp. 383–392, 2021.
5. J. Xue, f. Min, and f. Ma, “research on diabetes prediction method based on machine learning,” j. Phys. Conf. Ser., vol. 1684, no. 1, 2020.
6. S. Gündoğdu, “Efficient prediction of early-stage diabetes using XGBoost classifier with random forest feature selection technique,” Multimed. Tools appl., vol. 82, no. 22, pp. 34163–34181, 2023.
7. S. S. Bhat, v. Selvam, g. A. Ansari, m. D. Ansari, and m. H. Rahman, “Prevalence and early prediction of diabetes using machine learning in North Kashmir: a case study of District Bandipora,” Comput. Intell. Neurosci., vol. 2022, 2022.
8. V. V, r. A.c, d. Murthy, and. T., “diabetes prediction using Machine learning approach,” ssrn electron. J., vol. 10, no. 8, pp. 75–82, 2023.
9. P. Zhang et al., “erratum to ‘global healthcare expenditure on diabetes for 2010 and 2030’ [diabetes res. Clin. Pract. 87 (1) (2010) 293-301],” diabetes res. Clin. Pract., vol. 92, no. 2, p. 301, 2011.
10. K. Wu, “Optimizing diabetes prediction with machine learning: model comparisons and insights,” J. Sci. Technol., vol. 5, no. 4, pp. 41–51, 2024.
11. M. A. Rahim, m. A. Hossain, m. N. Hossain, j. Shin, and k. S. Yun, “stacked ensemble-based type-2 diabetes prediction using machine learning techniques,” ann. Emerg. Technol. Comput., vol. 7, no. 1, pp. 30–39, 2023.
12. S. Buyrukoğlu and A. Akbaş, “Machine learning based early prediction of type 2 diabetes: a new hybrid feature selection approach using correlation matrix with heatmap and SFS,” Balk. J. Electr. Comput. Eng., vol. 10, no. 2, pp. 110–117, 2022.
13. A. Murphy and d. Wang, “stochastic gradient descent,” radiopaedia.org, no. February 2018.

14. O. Osho, “An overview: stochastic gradient descent classifier, linear discriminant analysis, deep learning and naive bayes classifier approaches to network intrusion detection,” vol. 10, no. 04, pp. 294–308, 2021.
15. Maalim. A. Aljabery and s. Kurnaz, “applying data mining techniques to predict hearing aid type for audiology patients,” j. Inf. Sci. Eng., vol. 36, no. 2, pp. 205–215, 2020.
16. A. Sabir, H. A. Ali, and Maalim A. Aljabery, “ChatGPT tweets sentiment analysis using machine learning and data classification,” Inform., vol. 48, no. 7, pp. 103–112, 2024.
17. S. Kurnaz and Maalim. A. Aljabery, “Predict the type of hearing aid of audiology patients using data mining techniques,” ACM Int. Conf. Proceedings ser., pp. 2–6, 2018.
18. A. J. Der van Veen and A. Paulraj, “An analytical constant modulus algorithm,” ieee trans. Signal process., vol. 44, no. 5, pp. 1136–1155, 1996.
19. C. R. Johnson, P. Schniter, t. J. Endres, j. D. Behm, d. R. Brown and R. A. Casas, “blind equalization using the constant modulus criterion: a review,” Proc. IEEE, vol. 86, no. 10, pp. 1927–1949, 1998.
20. V. Zarzoso and P. Comon, “optimal step-size constant modulus algorithm,” Electr. Eng., no. 33, 2004.
21. H. El massari, z. Sabouri, s. Mhammedi, and n. Gherabi, “diabetes prediction using machine learning algorithms and ontology,” J. Ict stand., vol. 10, no. 2, pp. 319–338, 2022.
22. A. Dutta et al., “Early prediction of diabetes using an ensemble of machine learning models,” Int. J. Environ. Res. Public health, vol. 19, no. 19, pp. 1–25, 2022.