Hittite Journal of Science and Engineering, 2022, 9 (1) 09–18
ISSN NUMBER: 2148–4171
DOI: 10.17350/HJSE19030000250

Diagnosing Diabetes with Machine Learning Techniques

Omer Faruk Akmese
Hitit University, Department of Computer Engineering, Corum, Turkey

ABSTRACT

The rate of diabetes is rapidly increasing worldwide. Early detection of diabetes can help prevent or delay the onset of diabetes by initiating lifestyle changes and taking appropriate preventive measures. Prediabetes and type 2 diabetes have proved to be early detection problems. There is a need for easy, rapid, and accurate diagnostic tools for the early diagnosis of diabetes in this context. Machine learning algorithms can help diagnose diseases early. Numerous studies are being conducted to improve the speed, performance, reliability, and accuracy of diagnosing with these methods for a particular disease. This study aims to predict whether a patient has diabetes based on diagnostic measurements in a dataset from the National Institute of Diabetes and Digestive and Kidney Diseases. Eight different variables belonging to the patients were selected as the input variable, and it was estimated whether the patient had diabetes or not. Of the 768 records examined, 500 (65.1%) were healthy, and 268 (34.9%) had diabetes. Ten different machine learning algorithms have been applied to predict diabetic status. The most successful method was the Random Forest algorithm with 90.1% accuracy. Accuracy percentages of other algorithms are also between 89% and 81%. This study describes a highly accurate machine learning prediction tool for finding patients with diabetes. The model identified in the study may be helpful for early diabetes diagnosis.

Keywords:
Diabetes; Diagnosis; Machine learning; Classification; Prediction

Article History:
Received: 2021/09/12
Accepted: 2022/01/10
Online: 2022/03/30

***Correspondence to**: Ömer Faruk AKMEŞE, E-mail: ofarukakmese@hitit.edu.tr; Phone: +90 364 227 4533; Fax: +90 364 227 4535*

INTRODUCTION

Diabetes is a lifelong illness that occurs when the pancreas secretory gland does not produce sufficient hormone insulin levels or when the hormone insulin it produces cannot be used effectively. The patient cannot efficiently use the glucose that has been taken from the foods they have eaten, and as a result, the blood sugar level increases [1]. The chronic hyperglycemia of diabetes is associated with critical complications such as long-term damage, dysfunction and failure of the eyes, kidneys, nerves, heart and blood vessels and other different organs [2]. Failure to regularly intervene to maintain blood glucose levels at normal levels can cause many problems. Early diagnosis of diabetes is extremely important for effective treatments. However, some patients are unaware of their condition until complications occur [3]. There are three types of diabetes: type I diabetes, type II diabetes, and gestational diabetes [4]. Insulin-dependent Diabetes Mellitus (IDDM), which requires the injection of insulin to the patient as a result of the human body's inability to produce enough insulin, is classified as type I. Type II, which occurs when body cells cannot use insulin properly, is known as Non-Insulin Dependent Diabetes Mellitus (NIDDM). Type III Gestational Diabetes occurs with increased blood sugar in pregnant women where diabetes is not detected earlier [5]. Type 2 diabetes accounts for about 90% of cases of diabetes. The most common type 2 diabetes is considered a "silent disease," and disease indications may not be noticed for many years [6]. Obesity is considered the main cause of Type 2 diabetes in people genetically predisposed to the disease [7]. Early diagnosis and lifestyle changes or medical interventions can help prevent type 2 diabetes in many high-risk individuals [8]–[10]. For this reason, early diagnosis of diabetes is a crucial step to taking the necessary precautions.

Along with the increase in the world population, the number of patients with diabetes is increasing sig-

nificantly. The main causes of increased diabetes are malnutrition, overweight, ageing, ease of transportation, inactivity, introducing computers into everyday life, the Internet, smartphones, tablets, and constant stress in business life. In the world, 450 million people are fighting against diabetes [11]. In Turkey, there are over 10 million diabetics. Diabetic patients in Turkey is almost two times the world average in terms of population. Turkey is the country with the fastest increase in diabetes in Europe. According to data from 2015, in Turkey, it is stated that one out of every six people is fighting against diabetes [11]. A large amount of money is spent on treating diabetes mellitus, seen in a wide variety. In addition, the treatment process requires severe care and work.

Data mining techniques are preferred in all areas, as computers' data processing and computing capacities increase rapidly. Health is one of the areas that makes the most use of its support in the diagnosis and treatment process. Medical diagnosis is a complex and important process requiring actual patient data, accurate medical literature knowledge, and clinical experience. The medical diagnosis process is much more complicated than the identification processes in other sectors because it includes many unexpected situations. Clinical decisions are often made based on physicians' perceptions and experiences [12]. However, patients may not always express their complaints correctly. The rapid increase in the amount of data also causes difficulties in decision making.

Data mining and machine learning techniques are widely used in diabetes studies [5], [13]–[22]. The number of studies that predict disease diagnosis with machine learning is increasing. Researchers for the diagnosis of diabetes conduct many studies. Sowjanya et al. [23]: developed an android-based application to raise awareness of diabetes. In application, machine learning techniques have been used to predict diabetes among users. The system also provides information about diabetes and some suggestions about the disease. Orabi et al. [24] designed a system for estimating diabetes. While the proposed system achieved high accuracy using the decision tree algorithm, the results were satisfactory. Nongyao et al. [25], in their study "Comparison of Classifiers for the Risk of Diabetes Prediction Diabetes Mellitus," applied an algorithm that classifies the risk of diabetes mellitus. The model is designed with Decision Tree, Artificial Neural Networks, Logistic Regression, Random Forest, and Naive Bayes algorithms. Findings show that the best performance of the disease risk classification is the Random Forest algorithm. According to Humar et al. [26] proposed a hybrid Neural Network System with Artificial Neural Network and Fuzzy Neural Network with 79.16% accuracy for Diabetes diagnosis. Mohammed et al. [27] proposed SVR, a hybrid method with the NSGA-II method, for diabetes detection and achieved 86.13% accuracy. Mujumdar and Vaidehi [5] achieved 98.8% accuracy using the AdaBoost classifier in their study. Faruque et al. [19], with the C4.5 decision tree, achieved an accuracy of 73.5% to predict diabetes. On the other hand, Sonar and Jaya Malini [19] achieved an accuracy of 85% with the Decision Tree in their research. Zou et al. [28] showed that the accuracy of diabetes prediction could be 80.8% with random forest. In their study, Kaur and Kumari [21] achieved the best 89% accuracy for diabetes prediction with the SVM-linear model. Acar et al. achieved a performance of 87.06% with the LS-SVM method in their study in which they presented the estimation of diabetes mellitus with biometric measurements [29].

Clinical records of the National Institute of Diabetes and Digestive and Kidney Diseases were used in this study. Vincent Sigillito of Johns Hopkins University is the database donor. These data include age, pregnancies, glucose, blood pressure, skin thickness, insulin, body mass index, and diabetes lineage function variables associated with diabetes in women with suspected diabetes. This study aimed to create an effective predictive model with high sensitivity and selectivity to better identify patients at risk for diabetes based on patient measurements. In the study, patients with diabetes were tried to be determined with ten machine learning algorithms. These are, in order of accuracy: Random Forest, Gradient Boosting, XGB, LGBM, Decision Tree, AdaBoost, Support Vector Machine, Logistic Regression, kNN, Naive Bayes algorithms. Accuracy percentages and experimental performances of all algorithms were compared. Compared to the previous ones, this study is a more comprehensive study that includes many algorithms used in diabetes diagnosis, aiming to compare their performance and find the best among them.

EXPLORATORY DATA ANALYSIS

The approach to analyzing datasets using visual methods to summarise their key features and seeing what the data can say beyond the task of modelling or hypothesis testing is often called Exploratory Data Analysis (EDA) [30]. EDA aims to perform initial investigations on data before formal modelling and graphical representations and visualizations to discover patterns, review assumptions, and test hypotheses. Data visualizations contain explanatory and comparative charts to effectively illustrate abstract and concrete ideas. Summary information about key features and hidden trends in data can help identify problems, and their resolution can improve accuracy in diagnosing diabetes.

Understanding and Visualizing Data

In this study, the diabetes dataset in Kaggle was used to model and test the proposed method [31]. The selected

Table 1. Dataset description

Name	Type	Description	Role
Outcome	Categorical	0(no diabetes) / 1(diabetes)	Target
Pregnancies	Numerical	Number of times pregnant	Input
Glucose	Numerical	Plasma glucose concentration is an oral glucose tolerance test.	Input
BloodPressure	Numerical	Diastolic blood pressure (mm Hg)	Input
SkinThickness	Numerical	Triceps skin fold thickness (mm)	Input
Insulin	Numerical	2-Hour serum insulin (mu U/ml)	Input
BMI	Numerical	Body mass index	Input
DiabetesPedigree Function	Numerical	Diabetes pedigree function	Input
Age	Numerical	Age (years)	Input

Categorical: Data that can be grouped and cannot be expressed numerically. Numerical: Data expressed as numbers, not letters or words that cannot be grouped. Target: Estimated output variable Input: Attribute, predictor, feature

dataset is part of a larger dataset maintained by the National Institutes of Diabetes and Digestive and Kidney Diseases. This data set has been used by many researchers in predictive analyses [14], [17], [21], [22], [28]. The dataset consists of data used for diabetes research on women of Pima Indian heritage, aged 21 and over, living in Phoenix, the 5th largest city of the State of Arizona in the USA. The data set consists of 768 observations and eight independent numerical variables. The target variable is specified as "result"; 1 indicates positive diabetes test result, 0 indicates negative. The name of the data, the data type definition, and its role are shown in Table 1.

Table 2 shows summary statistics, including measures of central tendency such as mean and median and measures of distribution such as standard deviation, which are useful in providing a quick and simple description of the dataset and its characteristics. Pregnancies appear in a realistic range from 0 to 17. Some other attributes in the data (Glucose, BloodPressure, SkinThickness, Insulin, BMI) include the value 0, which is not possible in practice. In this case, the impossible 0 values need to be corrected. All impossible values were corrected by replacing them with mean values at the pre-processing stage. The 'DiabetesPedigreeFunction' is a function that scores the probability of diabetes based on family history, with a realistic range of 0.08 to 2.42. Age has a realistic range from 21 to 81. The Outcome, in the target variable, 0 represents healthy people, and 1 represents those with diabetes.

Table 2. Basic descriptive statistics

	Count	Mean	Std	Min	25%	50%	75%	Max
Pregnancies	768	3.85	3.37	0	1	3	6	17
Glucose	768	120.89	31.98	0	99	117	140.25	199
BloodPressure	768	69.10	19.36	0	62	72	80	122
SkinThickness	768	20.54	15.95	0	0	23	32	99
Insulin	768	79.79	115.24	0	0	30.5	127.25	846
BMI	768	31.99	7.88	0	27	32.	36.60	67.1
DiabetesPedigree Function	768	0.47	0.33	0.078	0	0.37	0.62	2.42
Age	768	33.24	11.76	21	24	29	41	81
Outcome	768	0.34	0.47	0	0	0	1	1

Count: Number of values in the dataset. Mean: The average of values. Std: The standard deviation of values. Min: The smallest value. 25%: The value at the 25% percentile. 50%: The value at the 50% percentile. 75%: The value at the 75% percentile. Max: The largest value.

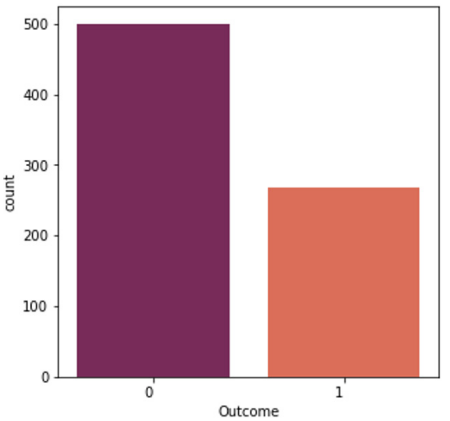

Figure 1. Count of diabetes Outcome.

The visual representation of quantitative data for communication and analysis is called data visualization [32]. With the increase in data types and diversity, there is a need for more analysis and presentation types that reveal the relationships between variables and summarise complex data with simple and easy-to-understand visuals [33]–[35]. Visualization of data is often performed without a model or hypothesis testing. Analysts can quickly and easily identify patterns, trends, and outliers from charts and charts [35]. The following visualizations were made in the data set to summarise complex data with simple and understandable visuals, reveal the relationships between variables, and identify patterns, trends, and outliers. Of the 768 clinical records analyzed according to Fig. 1, 500 (65.1%) were healthy, and 268 (34.9%) had diabetes.

Heatmaps can be used to cross-examine multivariate data, show variance between variables, show whether any variables are similar to each other, and detect whether there is a correlation between variables. Fig. 2 shows the heatmap

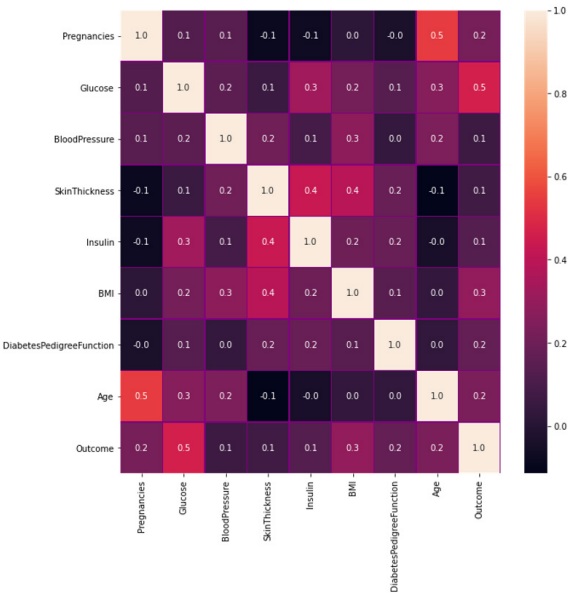

Figure 2. Heatmap

of the attributes in the data set. Accordingly, almost all attributes have weak linear correlations. This indicates that most variables are more likely to have nonlinear relationships.

To obtain accurate predictions in classification, mostly focused on the relationship between diabetes features in the dataset and the target feature resulting from diabetes. For example, when Glucose and BMI increase by 1 unit according to the heatmap, the positive Outcome of diabetes increases by 0.5 units and 0.3 units, respectively. The graph shows cases where diabetes patients generally have a higher number of Glucose, BMI, Age, Pregnancies and DiabetesPedigreeFunction. Glucose is the best indicator of diabetes outcome in this situation. It is seen that with strongly correlated features, the target class can be predicted more easily, and more meaningful results can be drawn.

MATERIALS AND METHODS

In the study, exploratory data analysis was performed to understand the data better. The dataset was preprocessed to provide reliable and acceptable results for estimating diabetes classifications. The modelling process was analyzed and evaluated with various classifier models, accuracy, precision, recall, and F1 score performance metrics. Data analysis in the research was carried out using the Python (3.8) programming language.

Pre-processing

Before analysis, pre-processing consists of steps to transform the raw data into a clean and organized dataset. Databases can have many quality control issues. Preprocessing aims to evaluate and improve data quality to allow reliable analysis [36]. Preprocessing refers to the transformations applied to the data before analysis. In this process, raw data is converted into an understandable dataset. Various techniques such as min-max, variance, deviation, standardization, mean scaling and elimination of missing values in the data set were applied in the preprocessing process [37]. In addition, outliers were also removed from the dataset.

In the pre-processing stage, it was observed that some records of the attributes in the data set contain the '0' value, which is not possible, some records contain outliers, and some records have missing values. Data containing missing, outlier and not possible value '0' were replaced with mean values. Thus, a data set without other noises with impossible values was obtained.

Feature Engineering

Feature engineering is the task of improving predictive modelling performance by transforming the feature space in a dataset. [38]. They are methods that enable to extraction of new features for machine learning models from raw data. Thus, better results can be obtained in terms of model performance. Feature engineering either changes the form of the variables in the data set or generates new and different variables in a machine learning process. Feature engineering was used in the study to improve the analyzability of the dataset further. In addition to generating new variables, the form of the variables has also been changed with the One-Hot Encoding and Robust Scaler methods. In this study, while nine variables were included in the original data set, the number of variables resulted in 17 with feature extraction.

Table 3 contains a cross-section showing the first five records of the data set. In addition to the new variables produced by feature extraction, it is seen that the form of the variables in the data set has changed. As such, the data is ready for analysis.

The architecture of the proposed model

The data set was estimated by algorithms of Random Forest, Gradient Boosting, XGB, LGBM, Decision Tree, AdaBoost, Support Vector Machine, Logistic Regression, kNN and Naive Bayes algorithms. The reason for choosing these algorithms was that they are widely used in the literature and give relatively better results in the existing data set.

Random Forest

Random forest algorithm is based on combining Decision trees and Bagging methods and falls under Ensemble methods [39]. Random forest is a flexible machine learning method used for regression or classification problems. In its simplest form, a random forest combines a large number of generated decision trees to obtain a more accurate prediction.

Table 3. Samples of the dataset

	0	1	2	3	4
Pregnancies	0.600	-0.400	1.000	-0.400	-0.600
Glucose	0.765	-0.790	1.630	-0.691	0.494
BloodPressure	0.000	-0.375	-0.500	-0.375	-2.000
SkinThickness	1.000	0.143	0.571	-0.714	1.000
Insulin	1.000	0.000	1.000	-0.127	0.978
BMI	0.170	-0.599	-0.962	-0.434	1.214
DiabetesPedigreeFunction	0.230	-0.019	0.271	-0.186	0.749
Age	1.235	0.118	0.176	-0.471	0.235
Outcome	1.000	0.000	1.000	0.000	1.000
New_Glucose_Class_Prediabetes	1.000	0.000	1.000	0.000	0.000
New_BMI_Range_Healty	0.000	0.000	1.000	0.000	0.000
New_BMI_Range_Overweight	0.000	1.000	0.000	1.000	0.000
New_BMI_Range_Obese	1.000	0.000	0.000	0.000	1.000
New_BloodPressure_HS1	0.000	0.000	0.000	0.000	0.000
New_BloodPressure_HS2	0.000	0.000	0.000	0.000	0.000
New_SkinThickness_1	0.000	0.000	0.000	0.000	0.000
NewInsulinScore_1	0.000	1.000	0.000	1.000	0.000

Gradient Boosting

Gradient Boosting is a powerful machine learning technique used for regression or classification problems, one of the ensemble methods. Each tree is grown using information from previously grown trees. The basic idea is to minimize the error and determine the target outputs for the next model. This technique relies on the progress of subsequent forecasts by learning from previous forecast errors [40].

XGB

The XGB (eXtreme Gradient Boosting) algorithm, which was brought to the literature by Chen and Guestrin, is an effective algorithm that has been frequently used in supervised machine learning applications such as regression and classification [41]. It is an efficient optimization of the gradient boosting technique [42]. It uses an approach to find the best decision tree model to achieve higher speed and better performance. XGBoost is one of the most successful machine learning algorithms [43].

LGBM

It is a machine learning algorithm with a decision tree approach that emerged late 2017 [44]. The algorithm released by Microsoft has the advantages of low memory consumption and providing high accuracy. LGBM is an optimized ensemble learning algorithm based on the Gradient Boosted Decision Tree. This model uses a histogram-based algorithm on high-dimensional data to speed up the computation time and avoid overloading the prediction system [45], [46].

Decision Tree

Decision Trees are a type of supervised machine learning where data is continuously divided according to a certain parameter. A decision tree is a method used to divide a data set into smaller clusters by applying rules. In other words, it is based on the principle of dividing large amounts of data into smaller data groups [47]. The decision tree contains the concepts of nodes and leaves. Nodes represent where data is divided, and leaves represent decisions. The tree structure used is easy to understand and interpret as it can be visualized.

AdaBoost

AdaBoost is a machine learning approach that combines many relatively weak and erroneous rules to create an accurate prediction rule. Freund and Schapire's AdaBoost algorithm was the first practical boosting algorithm and remained one of the most widely used. [48]. The AdaBoost algorithm produces strong classifiers with weak classifiers. In each cycle, the weights are adjusted, and a committee of weak classifiers is formed. While the weights of the training samples incorrectly classified by the existing weak classifier are increased, the weights of the correctly classified training samples are decreased. The AdaBoost algorithm has good performance due to generating expanding diversity [49].

Support Vector Machine

SVM (Support Vector Machine) is one of the machine learning methods used for classification. The high accuracy of SVM, which is widely used in classification problems, has made this method widespread. This method differs because the number of operations and algorithm complexity is low. SVM is divided into two as linear or nonlinear [50]. The algorithm proposed by Vapnik in 1963 was a linear classifier model [51]. However, in 1992 Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik proposed a way to construct nonlinear classifiers [52]. In linear cases, there is a separation of classes with the help of a decision function obtained from the training data. The line that divides the data set into two is called the decision line. The main purpose of this model is to determine the hyperplane that will best separate the classes from each other. In the case of a nonlinear dataset, the kernel method is used because SVMs cannot draw a linear hyperplane. The kernel method greatly increases the classification accuracy for nonlinear data.

Logistic Regression

Logistic regression is a statistical method used to analyze a data set that determines an outcome and has independent variables. Although it is called regression, there is a classification here. Recently, logistic regression has come to the forefront and has become an intensively used method due to its ease of use and interpretation of numerical data. It is generally used in medicine, biology, and economics [53].

kNN

kNN (K Nearest Neighbor) is one of the machine learning methods used for classification. The k value determines the number of elements in the classification. It searches for the nearest neighbours in the dataset while estimating. Euclidean, Manhattan, Minkowski and Hamming functions can be used in distance calculations [40].

Naive Bayes

The Naive Bayes algorithm is a supervised learning algorithm based on Bayes' theorem, used in solving classification problems. Naive Bayes classification assumes that the variables are independent of the classes. It is a probabilistic classifier; that is, it makes predictions based on the probability of an object. The Naive Bayes Classifier is a simple and effective classification algorithm that can make fast predictions that help build fast machine learning models. The traditional Naive Bayes classifier is

still widely used as a popular learning algorithm for data mining applications due to its simplicity [54]–[56].

Table 4 lists the parameters used in ten machine learning techniques.

In the study, calculations were made using hold-out and cross-validation methods. Similar results were obtained in both methods. Since the data set is large enough, the hold-out method, which is faster and requires less computational power, was preferred. Some of the data (75%) were used in training the model and some of it (25%) in testing. The proposed method is shown in Fig. 1.

Table 4. Parameters in algorithms

No	Algorithm	Parameters
1	Random Forest	{'criterion': 'gini', 'n_estimators': 100, 'max_depth': 10}
2	Gradient Boosting	{'criterion': 'friedman_mse', 'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 100, 'subsample': 1.0 }
3	XGB	{'booster': 'gbtree', 'learning_rate': 0.3, 'max_depth': 6, 'n_estimators': 100}
4	LGBM	{'boosting_type': 'gbdt', 'learning_rate': 0.1, 'max_depth': -1, 'n_estimators': 100, 'subsample': 1.0}
5	Decision Tree	{'criterion': 'gini', 'max_depth': 10, 'min_samples_leaf': 1, 'min_samples_split': 2, }
6	AdaBoost	{'learning_rate': 1, 'n_estimators': 100, }
7	SVM	{'C': 1.0, 'cache_size': 200, 'coef0': 0.0, 'kernel': 'rbf', 'max_iter': -1}
8	Logistic Regression	{'C': 1.0, 'max_iter': 1000, 'tol': 0.0001}
9	kNN	{'leaf_size': 30, 'metric': 'minkowski', 'n_neighbors': 5}
10	Naive Bayes	{'var_smoothing': 1e-09}

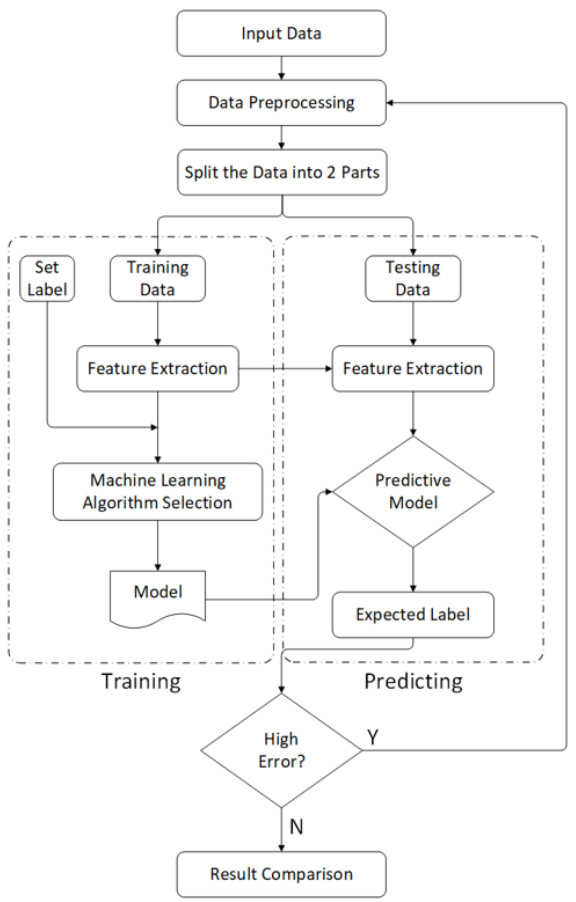

Figure 3. The proposed model

According to the architecture shown in Fig. 3, the collected data are subjected to pre-processing by the researchers. The fact that the quality of the data greatly influences the prediction result means that pre-processing plays an important role in the model [24].

Measurement

The following metrics were used to measure classification performance [57]. (TP: true positives, TN: true negatives, FN: false negatives, FP: false positives.)

Accuracy: Expresses the total accuracy rate.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} = \frac{TP + TN}{P + N}$$

AUC: The AUC value measures the accuracy of a diagnostic test. It is calculated according to the area under the ROC curve.

Precision: It expresses the ratio of correctly detected Positive classes to all positives.

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{TP}{P'}$$

Recall: It expresses the ratio of correctly detected Positive classes to true positives.

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{TP}{P}$$

F1-score: The F1-score is the harmonic mean of the sensitivity and precision.

$$\text{F1-score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

RESULT AND DISCUSSION

Table 5 shows the results of the ten machine learning methods used in the study. Accordingly, the Random Forest algorithm is the most accurate conclusion.

Table 4 shows the Accuracy percentage, F1, Precision and Recall values of each model. Also, the most successful method according to the percentage of accuracy is found as Random Forest. In the algorithms selected according to Table 6, the complexity matrix is seen for the Random Forest method. This method gives the most accurate result with 90.1%.

According to Table 6, the model correctly predicted those with diabetes to be 90.1%. The rate of healthy people defined as ill (Type I error) was 11.8%, while the proportion of patients diagnosed as healthy (Type II error) was 9.2%.

Table 5. Accuracy values of models using data set

Method	Accuracy (%)	F1	Precision	Recall
Random Forest	90.1	0.90	0.90	0.90
Gradient Boosting	89.5	0.89	0.90	0.90
XGB	88.5	0.88	0.88	0.89
LGBM	88.5	0.88	0.88	0.89
Decision Tree	86.4	0.87	0.87	0.86
AdaBoost	86.4	0.86	0.86	0.86
Support Vector Machine	86.4	0.86	0.86	0.86
Logistic Regression	85.9	0.86	0.86	0.86
kNN	85.4	0.85	0.85	0.85
Naive Bayes	81.7	0.82	0.83	0.82

Accuracy: Total accuracy rate. F1-score: The F1-score is the harmonic mean of the sensitivity and precision. Precision: Ratio of correctly detected Positive classes to all positives. Recall: Ratio of correctly detected Positive classes to true positives.

Table 6. Results of Random Forest

Accuracy:90.1%	True 1	True 0	Total	Class Precision	
Pred. 1	45 (TP) Correct Decision	6 (FP) Type I error	51 (P')	88.2%	90% (weighted avg)
Pred. 0	13 (FN) Type II error	128 (TN) Correct Decision	141 (N')	90.8%	
Total	58 (P)	134 (N)	192		
Class Recall	77.5%	95.5%			
	90% (weighted avg)				

TP: true positives, TN: true negatives, FN: false negatives, FP: false positives. Precision: Ratio of correctly detected Positive classes to all positives. Recall: Ratio of correctly detected Positive classes to true positives.

The Receiver Operating Characteristic (ROC) curve is expressed as the ratio of sensitivity to specificity. ROC can also be expressed as the fraction of true positives, false positives. The ROC curve provides the opportunity to compare different tests and diagnostic activities of different practitioners, monitor practitioners' development, determine the test's discrimination power, and monitor the quality of the laboratory results [58]. A diagnostic test is useful to the extent that it distinguishes patients well from their health. In the case where the diagnostic test has no separation characteristics, the value of the area under the ROC curve is 0.50. In the case of an excellent diagnostic test, this value should be 1. The test should have a value between these two. The area under the ROC curve is called AUC (Area Under Curve). The larger the AUC value, the better the diagnostic test can discriminate it.

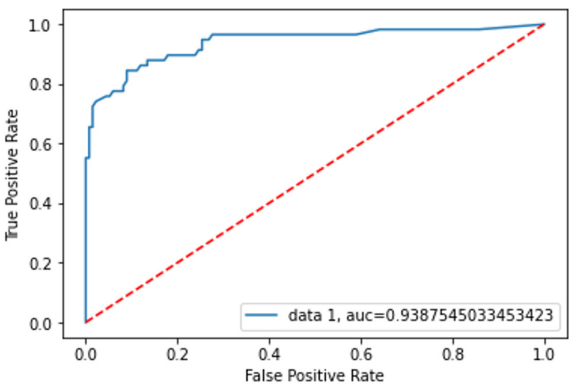

Figure 4. Receiver Operating Characteristic Curve (ROC AUC)

Table 7. Classification accuracies of other classifiers in the literature

	Authors	Year	Method	Best Accuracy
1	Mujumdar and Vaidehi [5]	2019	Logistic Regression	96%
2	Iyer et. al. (2015) [13]	2015	Naive Bayes	79.5%
3	Hasan et. al. [14]	2020	Ensembling AB+XB	95%
4	Meng et. al. [15]	2013	Decision Tree	78%
5	Lai et. al. [16]	2019	GBM	84.7%
6	Sarwar et. al. [17]	2018	SVM and KNN	77%
7	Faruque, et. al. [19]	2019	C4.5 Decision Tree	73.5%
8	Sonar and JayaMalini [20]	2019	Decision Tree	85%
9	Wei et. al. [21]	2018	Deep Neural Network	77.8%
10	Kaur and Kumari [22]	2020	Linear Kernel SVM	89%
11	Nai-arun and Moungmai [25]	2015	Random Forest	85.5%
12	Zangooei et. al. [27]	2014	SVR using NSGA-II	86.1%
13	Zou et. al. [28]	2018	Random Forest	80.8%
14	Acar et. al. [29]	2011	LS-SVM classifier	87.06%
15	Akmeşe (This study)	2022	Random Forest	90.1%

Fig. 4 shows the area under the curve and ROC curve (AUC) for the Random Forest algorithm. The area under the ROC curve is close to 1. Accordingly, it can be said that the analysis is close to perfect distinctiveness.

In Table 7, it is possible to see some studies on diabetes prediction in the literature. Differences in prediction percentages are due to the data set, the algorithms used, and the methodological differences in the studies.

Researchers have made comparative analyses using different machine learning algorithms to evaluate their predictive performance and select the most efficient ones in many studies. In studies reviewed, it is seen that the prediction accuracy is over 80%. However, factors such as the size of the data set and the number of features can significantly affect the algorithm's performance. For this reason, an algorithm with the best performance in a dataset may have a lower prediction accuracy in different data sets [59]. Because diabetes is a disease that can cause many complications, how to predict exactly this disease using Machine learning is worth studying. Early prediction of such diseases can be controlled and save human life. The type of diabetes cannot be predicted from the data set. Therefore, future work may predict the type of diabetes and increase the percentage of accuracy.

CONCLUSION

In this study, machine learning algorithms were used to predict the diagnosis of diabetes. All 768 patients studied were female, and the mean age was also 33 (21-81 years). Of the individuals in the data set, 268 (34.9%) were patients, and 500 (65.1%) were healthy individuals. Ten different machine learning algorithms have been applied to predict diabetic status. The estimation accuracy is considered the most important factor in the study. It has been determined that the Random Forest algorithm achieves the best success rate with a 90.1% correct prediction rate. Accuracy percentages of other algorithms are also

between 81% and 89%. New patient data can be added to train the model in future studies. It is thought that a better result can be obtained for estimation as the number of data increases.

CONFLICT OF INTEREST

Authors approve that to the best of their knowledge, there is not any conflict of interest or common interest with an institution/organization or a person that may affect the review process of the paper.

References

1. K. G. M. M. Alberti, P. Zimmet, and J. Shaw, "International Diabetes Federation: A consensus on Type 2 diabetes prevention," Diabet. Med., vol. 24, no. 5, pp. 451–463, 2007, doi: 10.1111/j.1464–5491.2007.02157.x.
2. D. O. F. Diabetes, "Diagnosis and classification of diabetes mellitus," Diabetes Care, vol. 33, no. SUPPL. 1, 2010, doi: 10.2337/dc10–S062.
3. M. Franciosi et al., "Use of the Diabetes Risk Score for Opportunistic Screening of Undiagnosed Diabetes and Impaired Glucose Tolerance: The IGLOO (Impaired Glucose Tolerance and Long–Term Outcomes Observational) study," Diabetes Care, vol. 28, no. 5, pp. 1187–1194, May 2005, doi: 10.2337/diacare.28.5.1187.
4. Z. Tao, A. Shi, and J. Zhao, "Epidemiological Perspectives of Diabetes," Cell Biochem. Biophys., vol. 73, no. 1, pp. 181–185, Sep. 2015, doi: 10.1007/S12013–015–0598–4.
5. A. Mujumdar and V. Vaidehi, "Diabetes Prediction using Machine Learning Algorithms," Procedia Comput. Sci., vol. 165, pp. 292–299, 2019, doi: 10.1016/j.procs.2020.01.047.
6. P. Hossain, B. Kawar, and M. El Nahas, "Obesity and Diabetes in the Developing World — A Growing Challenge," N. Engl. J. Med., vol. 356, no. 3, pp. 213–215, 2007, doi: 10.1056/nejmp068177.
7. F. Mercaldo, V. Nardone, and A. Santone, "Diabetes Mellitus Affected Patients Classification and Diagnosis through Machine Learning Techniques," Procedia Comput. Sci., vol. 112, pp. 2519–2528, 2017, doi: 10.1016/j.procs.2017.08.193.
8. J. Tuomilehto et al., "Prevention of Type 2 Diabetes Mellitus by Changes in Lifestyle among Subjects with Impaired Glucose Tolerance," New England Journal of Medicine, vol. 344, no. 18. pp. 1343–1350, 2001, doi: 10.1056/nejm200105033441801.
9. J. L. Chiasson, R. G. Josse, R. Gomis, M. Hanefeld, A. Karasik, and M. Laakso, "Acarbose for prevention of type 2 diabetes mellitus: the STOP–NIDDM randomized trial," Lancet, vol. 359, no. 9323, pp. 2072–2077, Jun. 2002, doi: 10.1016/S0140–6736(02)08905–5.
10. A. Ramachandran, C. Snehalatha, S. Mary, B. Mukesh, A. D. Bhaskar, and V. Vijay, "The Indian Diabetes Prevention Programme shows that lifestyle modification and metformin prevent type 2 diabetes in Asian Indian subjects with impaired glucose tolerance (IDPP–1)," Diabetologia, vol. 49, no. 2, pp. 289–297, 2006, doi: 10.1007/s00125–005–0097–z.
11. T. Diyabet, V. Başkanı, and P. M. Temel, "DİYABET ORANI 10 YILDA YÜZDE 100 ARTTI," pp. 10–12, 2017.
12. L. Parthiban and R. Subramanian, "Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm," Int. J. Biol. Med. Sci., vol. 3, no. 3, pp. 157–160, 2008.
13. A. Iyer, J. S, and R. Sumbaly, "Diagnosis of Diabetes Using Classification Mining Techniques," Int. J. Data Min. Knowl. Manag. Process, vol. 5, no. 1, pp. 01–14, 2015, doi: 10.5121/ijdkp.2015.5101.
14. M. K. Hasan, M. A. Alam, D. Das, E. Hossain, and M. Hasan, "Diabetes prediction using ensembling of different machine learning classifiers," IEEE Access, vol. 8, pp. 76516–76531, 2020, doi: 10.1109/ACCESS.2020.2989857.
15. X. H. Meng, Y. X. Huang, D. P. Rao, Q. Zhang, and Q. Liu, "Comparison of three data mining models for predicting diabetes or prediabetes by risk factors," Kaohsiung J. Med. Sci., vol. 29, no. 2, pp. 93–99, 2013, doi: 10.1016/j.kjms.2012.08.016.
16. H. Lai, H. Huang, K. Keshavjee, A. Guergachi, and X. Gao, "Predictive models for diabetes mellitus using machine learning techniques," BMC Endocr. Disord., vol. 19, no. 1, pp. 1–9, 2019, doi: 10.1186/s12902–019–0436–6.
17. M. A. Sarwar, N. Kamal, W. Hamid, and M. A. Shah, "Prediction of diabetes using machine learning algorithms in healthcare," ICAC 2018 – 2018 24th IEEE Int. Conf. Autom. Comput. Improv. Product. through Autom. Comput., no. September, pp. 6–7, 2018, doi: 10.23919/IConAC.2018.8748992.
18. A. U. Haq et al., "Intelligent machine learning approach for effective recognition of diabetes in e–healthcare using clinical data," Sensors (Switzerland), vol. 20, no. 9, 2020, doi: 10.3390/s20092649.
19. M. F. Faruque, Asaduzzaman, and I. H. Sarker, "Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus," 2nd Int. Conf. Electr. Comput. Commun. Eng. ECCE 2019, pp. 7–9, 2019, doi: 10.1109/ECACE.2019.8679365.
20. P. Sonar and K. Jaya Malini, "Diabetes prediction using different machine learning approaches," Proc. 3rd Int. Conf. Comput. Methodol. Commun. ICCMC 2019, no. Iccmc, pp. 367–371, 2019, doi: 10.1109/ICCMC.2019.8819841.
21. S. Wei, X. Zhao, and C. Miao, "A comprehensive exploration to the machine learning techniques for diabetes identification," IEEE World Forum Internet Things, WF–IoT 2018 – Proc., vol. 2018–Janua, pp. 291–295, 2018, doi: 10.1109/WF–IoT.2018.8355130.
22. H. Kaur and V. Kumari, "Predictive modelling and analytics for diabetes using a machine learning approach," Appl. Comput. Informatics, 2019, doi: 10.1016/j.aci.2018.12.004.
23. K. Sowjanya, A. Singhal, and C. Choudhary, "MobDBTest: A machine learning based system for predicting diabetes risk using mobile devices," Souvenir 2015 IEEE Int. Adv. Comput. Conf. IACC 2015, pp. 397–402, 2015, doi: 10.1109/IADCC.2015.7154738.
24. K. M. Orabi, Y. M. Kamal, and T. M. Rabah, "Early predictive system for diabetes mellitus disease," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2016, vol. 9728, pp. 420–427, doi: 10.1007/978–3–319–41561–1_31.
25. N. Nai–Arun and R. Moungmai, "Comparison of Classifiers for the Risk of Diabetes Prediction," Procedia Comput. Sci., vol.

69, pp. 132–142, 2015, doi: 10.1016/j.procs.2015.10.014.
26. H. Kahramanli and N. Allahverdi, "Design of a hybrid system for the diabetes and heart diseases," Expert Syst. Appl., vol. 35, no. 1–2, pp. 82–89, 2008, doi: 10.1016/j.eswa.2007.06.004.
27. M. H. Zangooei, J. Habibi, and R. Alizadehsani, "Disease Diagnosis with a hybrid method SVR using NSGA–II," Neurocomputing, vol. 136, pp. 14–29, 2014, doi: 10.1016/j.neucom.2014.01.042.
28. Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Predicting Diabetes Mellitus With Machine Learning Techniques," Front. Genet., vol. 9, no. November, pp. 1–10, 2018, doi: 10.3389/fgene.2018.00515.
29. V. . ACAR, E , ÖZERDEM, M , AKPOLAT, "Forecasting Diabetes Mellitus with Biometric Measurements.," Int. Arch. Med. Res., vol. 1, no. 1, pp. 28–42, 2011.
30. J. Tukey, "Exploratory data analysis," 1977, Accessed: Sep. 08, 2021. [Online]. Available: http://theta.edu.pl/wp–content/uploads/2012/10/exploratorydataanalysis_tukey.pdf.
31. R. S. Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, "Pima Indians Diabetes Database," Https://Www.Kaggle.Com/Uciml/Pima–Indians–Diabetes–Database, 2016. https://www.kaggle.com/uciml/pima–indians–diabetes–database (accessed Aug. 01, 2021).
32. "Tufte: The visual display of quantitative information – Google Akademik." https://scholar.google.com/scholar_lookup?title=The Visual Display of Quantitative Information&publication_year=2001&author=E. Tufte (accessed Sep. 08, 2021).
33. S. Lavalle, E. Lesser, R. Shockley, M. S. Hopkins, and N. Kruschwitz, "Big Data , Analytics and the Path From Insights to Value Big Data , Analytics and the Path From Insights to Value," no. 52205, 2011.
34. R. Agrawal, A. Kadadi, X. Dai, and F. Andres, "Challenges and opportunities with big data visualization," 7th Int. ACM Conf. Manag. Comput. Collect. Intell. Digit. Ecosyst. MEDES 2015, pp. 169–173, Oct. 2015, doi: 10.1145/2857218.2857256.
35. S. Nestorov, B. Jukić, N. Jukić, A. Sharma, and S. Rossi, "Generating insights through data preparation, visualization, and analysis: Framework for combining clustering and data visualization techniques for low–cardinality sequential data," Decis. Support Syst., vol. 125, no. March, p. 113119, 2019, doi: 10.1016/j.dss.2019.113119.
36. C. M. Salgado, C. Azevedo, H. Proença, and S. M. Vieira, Setting the Stage: Rationale Behind and Challenges to Health Data Analysis. 2016.
37. S. B. Kotsiantis and D. Kanellopoulos, "Data pre–processing for supervised leaning," Int. J. ;, vol. 1, no. 2, pp. 1–7, 2006, doi: 10.1080/02331931003692557.
38. F. Nargesian, H. Samulowitz, U. Khurana, E. B. Khalil, and D. Turaga, "Learning feature engineering for classification," IJCAI Int. Jt. Conf. Artif. Intell., vol. 0, no. August, pp. 2529–2535, 2017, doi: 10.24963/ijcai.2017/352.
39. L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.
40. Ö. F. AKMEŞE, "Karın Ağrısı ile Acil Servise Başvuran Hastalarda Akut Apandisit Tanısı için Makine Öğrenmesi Yaklaşımlarının Kullanımı," Kırıkkale University, 2020.
41. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., vol. 13–17–August–2016, pp. 785–794, Aug. 2016, doi: 10.1145/2939672.2939785.
42. J. Friedman, "Greedy Function Approximation : A Gradient Boosting Machine Author (s): Jerome H . Friedman Source : The Annals of Statistics , Vol . 29 , No . 5 (Oct ., 2001), pp . 1189–1232 Published by : Institute of Mathematical Statistics Stable URL : http://www," Ann. Stat., vol. 29, no. 5, pp. 1189–1232, 2001.
43. W. Zhao, J. Li, J. Zhao, D. Zhao, J. Lu, and X. Wang, "XGB model: Research on evaporation duct height prediction based on XGBoost algorithm," Radioengineering, vol. 29, no. 1, pp. 81–93, 2020, doi: 10.13164/re.2020.0081.
44. G. Ke et al., "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," Adv. Neural Inf. Process. Syst., vol. 30, 2017, Accessed: Nov. 28, 2021. [Online]. Available: https://github.com/Microsoft/LightGBM.
45. W. Cai, R. Wei, L. Xu, and X. Ding, "A method for modelling greenhouse temperature using gradient boost decision tree," Inf. Process. Agric., Sep. 2021, doi: 10.1016/J.INPA.2021.08.004.
46. M. Massaoudi, S. S. Refaat, I. Chihi, M. Trabelsi, F. S. Oueslati, and H. Abu–Rub, "A novel stacked generalization ensemble–based hybrid LGBM–XGB–MLP model for Short–Term Load Forecasting," Energy, vol. 214, p. 118874, Jan. 2021, doi: 10.1016/J.ENERGY.2020.118874.
47. K. S. Albayrak A., "VERİ MADENCİLİĞİ: KARAR AĞACI ALGORİTMALARI VE İMKB VERiLERİ ÜZERİNE BİR UYGULAMA * DATA MINING: DECISION TREE ALGORITHMS AND AN APPLICATION ON ISE DATA," no. May, 2014.
48. R. E. Schapire, "Explaining AdaBoost," Empir. Inference Festschrift Honor Vladimir N. Vapnik, pp. 37–52, Jan. 2013, doi: 10.1007/978–3–642–41136–6_5.
49. T. K. An and M. H. Kim, "A new Diverse AdaBoost classifier," Proc. – Int. Conf. Artif. Intell. Comput. Intell. AICI 2010, vol. 1, pp. 359–363, 2010, doi: 10.1109/AICI.2010.82.
50. V. Vapnik, The Nature of Statistical Learning Theory. Springer science & business media, 2013.
51. AIZERMAN and M. A., "Theoretical Foundations of the Potential Function Method in Pattern Recognition Learning," Autom. Remote Control, vol. 25, pp. 821–837, 1964, Accessed: Nov. 27, 2021. [Online]. Available: https://ci.nii.ac.jp/naid/10021200712.
52. Boser Berhard E., G. I. M., and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in In Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory, 1992, pp. 144–152.
53. E. Ürük, "İstatistiksel Uygulamalarda Lojistik Regresyon Analizi," Marmara University, 2007.
54. D. Soria, J. M. Garibaldi, F. Ambrogi, E. M. Biganzoli, and I. O. Ellis, "A 'non–parametric' version of the naive Bayes classifier," Knowledge–Based Syst., vol. 24, no. 6, pp. 775–784, Aug. 2011, doi: 10.1016/J.KNOSYS.2011.02.014.
55. "Naive Bayes Classifier in Machine Learning – Javatpoint." https://www.javatpoint.com/machine–learning–naive–bayes–classifier (accessed Nov. 29, 2021).
56. M. Hall, "A Decision Tree–Based Attribute Weighting Filter for Naive Bayes," Res. Dev. Intell. Syst. XXIII – Proc. AI 2006, 26th SGAI Int. Conf. Innov. Tech. Appl. Artif. Intell., pp. 59–70, Dec. 2006, doi: 10.1007/978–1–84628–663–6_5.
57. Gorunescu Florin, Data Mining: Concepts, Models and Techniques. Berlin: Springer Science & Business Media, 2011.

58. A. Dirican, "Tanı Testi Performanslarının Değerlendirilmesi ve Kıyaslanması," Cerrahpaşa Tip Dergisi, vol. 32, no. 1, pp. 25-30, 2001.
59. I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine Learning and Data Mining Methods in Diabetes Research," Comput. Struct. Biotechnol. J., vol. 15, pp. 104–116, 2017, doi: 10.1016/j.csbj.2016.12.005.