Original Article

MEDICINE PALLIATIVE CARE

DOI: 10.47582/jompac.1307319

J Med Palliat Care. 2023;4(4):270-276

Optimized machine learning based predictive diagnosis approach for diabetes mellitus

Erkan Akkur[1], Fuat Türk[2]

[1]Turkish Medicines and Medical Devices Agency, Ankara, Turkey
[2]Department of Computer Engineering, Faculty of Engineering and Architecture, Kırıkkale University, Kırıkkale, Turkey

Cite this article as: Akkur E, Türk F. Optimized machine learning based predictive diagnosis approach for Diabetes mellitus. *J Med Palliat Care*. 2023;4(4):270-276.

Received: 30.05.2023 ◆ **Accepted**: 11.07.2023 ◆ **Published**: 30.08.2023

ABSTRACT

Aims: Diabetes mellitus is a metabolic disease caused by elevated blood sugar. If this disease is not diagnosed on time, it has the potential to pose a risk to other organs and tissues. Machine learning algorithms have started to preferred day by day in the detection of this disease, as in many other diseases. This study suggests a diabetes prediction approach incorporating optimized machine learning (ML) algorithms.

Methods: The framework presented in this study starts with the application of different data pre-processing processes. Random forest (RF), support vector machine (SVM), K-nearest neighbor (K-NN) and decision tree (DT) algorithms are used for classification. Grid search is utilized for hyperparameter optimization of algorithms. Different performance evaluation measures are used to find the algorithm that best predicts diabetes. PIMA Indian dataset (PID) is chosen for testing the experiments. In addition, it is investigated to what extent the attributes in the data set affect the result using Shapley additive explanations (SHAP) analysis.

Results: As a result of the experiments, the RF algorithm achieved the highest success rate with 89.06%, 84.33%, 84.33%, 84.33% and 0.88% accuracy, precision, sensitivity, F1-score and AUC scores. As a result of the SHAP analysis, it is found that the "Insulin", "Age" and "Glucose" attributes contributed the most to the prediction model in identifying patients with diabetes.

Conclusion: The hyperparameter optimized RF approach proposed in the framework of the study provided a good result in the prediction and diagnosis of diabetes mellitus when compared with similar studies in the literature. As a result, an expert system can be designed to detect diabetes early in real time using the proposed method.

Keywords: Machine learning, diabetes mellitus, data preprocessing, grid search, random forest

INTRODUCTION

Diabetes mellitus (DM) is one of the diseases that threaten public health at significant rates. Insufficient or no insulin production in the body for any reason or insensitivity of body tissues to insulin causes diabetes.[1] Symptoms such as dry mouth, nocturia, polyuria, polydipsia, loss of appetite, blurred vision, weight loss, itching, recurrent fungal infections are frequently seen in diabetic patients.[2] DM can cause many problems in a person's health due to the effects it creates. Common problems caused by diabetes include heart diseases, vascular diseases, vision loss, kidney failure, and nervous system diseases.[3] According to published statistics, there are more than 500 million diabetics worldwide as of 2021. It is predicted that the incidence of DM will reach over 600 million in 2030 and over 700 million in 2045. In 2021, DM caused over 6.5 million deaths in 2021.[4] Therefore, early diagnosis of diabetes is essential procedure in terms of reducing the incidence of diabetes and reducing the problems that diabetes can cause.[5]

In order to diagnose diabetes in the early period, it is done by examinations by health professionals and examining blood samples taken from patients in a laboratory environment.[1] However, due to the fact that diabetes is a disease that progresses without showing many symptoms, it may not be clearly diagnosed at times. The artificial intelligence (AI), which is successfully used in the diagnosis and prediction of many diseases such as cancer, heart, skin, genetic and neurological disorders, can be used in the prediction of diabetes.[6] In addition to learning from known data, AI is a structure that includes analytical algorithms and allows computers to perform many complex operations. In addition to learning from known data, AI is a structure that includes analytical algorithms and allows computers to perform many complex operations. The working areas of AI are shown in **Figure 1**. Machine learning (ML) is a field of AI that helps a computer learn with the data it uses, increases the

Corresponding Author: Fuat Türk, fturk@kku.edu.tr

performance of systems, and uses mathematical models for these operations. These algorithms are critical for building a data model that predicts future states from known data. ML can be grouped supervised learning algorithms such as classification and regression process, unsupervised learning algorithms such as cluster analysis and reinforcement learning algorithms such as decision making. Neural network is type of ML algorithm created by modeling neural networks in the brain of living things. Deep learning is a kind of machine learning algorithm model that automatically creates the hierarchy of the presented data by using multi-layered neural networks as a model.[7] ML contributes to the interpretation of health data, which is especially difficult to learn and cannot be analyzed by traditional statistical methods. While building ML algorithms, different hyperparameter sets should be tested and appropriate hyperparameters should be selected. The only way to determine them is through multiple experiments on models by selecting a set of hyperparameters. This process is called hyperparameter optimization. Choosing the right hyperparameters directly affects the performance of ML algorithms. Considering that there are tens of hyperparameters and tens of values that these hyperparameters can take for a ML algorithm, it is clear how difficult it would be to try all combinations one by one and choose the best combination. For ML algorithms, it is useful to use the hyperparameter optimization method to determine the best hyperparameters. Hyperparameter optimization is the process of finding the most appropriate hyperparameter combination according to the success metric determined for a ML algorithm.[8] Therefore, hyperparameter optimization is an extremely useful process for building a successful model.

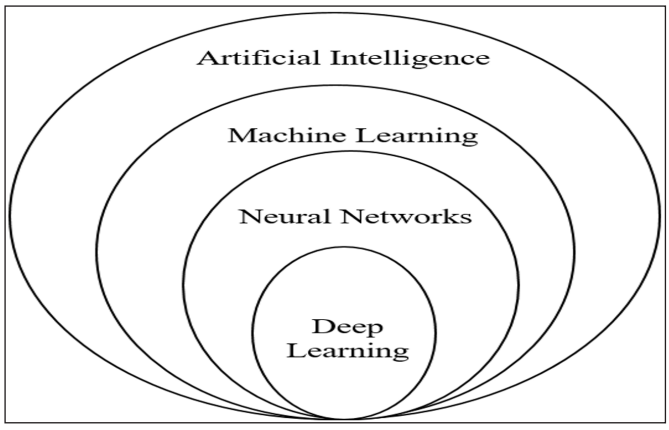

Figure 1. The working areas of AI

In this study, using a data set, an effective prediction model is proposed to determine whether a patient has diabetes with different ML algorithms. Within this scope, this study proposes an effective predictive diagnostic model for diabetes using four different supervised ML models predictive diagnosis approach for diabetes with the four ML algorithms optimized by grid search (GS) hyperparameter optimization.

The ML algorithms used in the study are as follows: Decision Tree (DT), Support Vector Machine (SVM), K-Nearest Neighbor (K-NN) and Random Forest (RF).

In this study, the diagnostic performances of ML in diabetes are compared with each other to create an effective predictive model and the most successful method is tried to be determined.

Different performances criteria are used to determine the ML method that best detects diabetes.

Shapley additive explanations (SHAP) values are utilized to show the effect of attributes on model success.

LITERATURE SURVEY

The studies related to the prediction of DM are investigated, it is seen that many classification models have been proposed by using ML algorithms. One of the most common open datasets used for the prediction DM is the dataset named "Pima Indians Diabetes" (PID). One of the studies in which this dataset was included Birjais et al.[9] has conducted. Gradient boosting (GB), logistic regression (LR), and naive Bayes (NB) classifiers were preferred to predict whether a person is diabetic. GB algorithm with 86% accuracy achieved the best result. Tigga et al.[10] preferred LR, K-NN, SVM, NB, DT and RF algorithms. The 10-fold cross validation results are 77%, 74.2%, 77%, 75.6%, 74.9% and 77.4%, respectively. Singh and Singh[11] utilized a stacked ensemble approach for prediction of DM. The proposed model achieved 83.8% of accuracy. Lyngdoh et al.[12] utilized K-NN, NB, SVM, DT and RF algorithms for prediction of DM. K-NN algorithm achieved 76% of accuracy. Kumari et al.[13] applied different ML algorithms for prediction process. Soft voting classifier (SVC) achieved of 79.08% accuracy Chang et al.[14] suggested ML-based model for prediction of DM. RF algorithm achieved 82.26% of accuracy compared to other ML algorithms. Yakut[15] utilized RF, Extra Tree Classifier and Gaussian Process Classifier for prediction of DM. RF achieved 81.71% of accuracy.

METHODS

All procedures were carried out in accordance with the ethical rules and the principles of the Declaration of Helsinki. An open data set was used in this study. Thus, ethics committee approval was not obtained.

The suggested methodology for diagnosing diabetes is presented in this section. The flowchart of suggested predictive model for diabetes mellitus is shown in the Figure 2. The suggested model is starting with diabetes

dataset acquisition. Then, various data pre-processing process are performed. After data pre-processing stage, ML algorithms process is starting. In the next stage, the hyperparameters of ML algorithms is automatically selected by using GS approach. Then, using performance measure metrics, the ML algorithms are compared and the best classifier is determined for prediction of diabetes. At the last stage, the effect of the features on the result of the model with the best prediction rate is determined by SHAP analysis. All the experiments were conducted using a Jupyter Notebook (6.3.0)[16] environment running Phyton (3.8.8)[17] on Windows 10. A personal computer used to run the simulation created in the study is equipped with an Intel Core i7 8750 CPU processor and 16 GB of memory. The simulation libraries are used as follows: Data pre-processing, data splitting, ML modelling, evaluation and plotting (sklearn, PyOD, Numpy, pandas, SciPy, matplotlib, skicitplot and seaborn).[18]

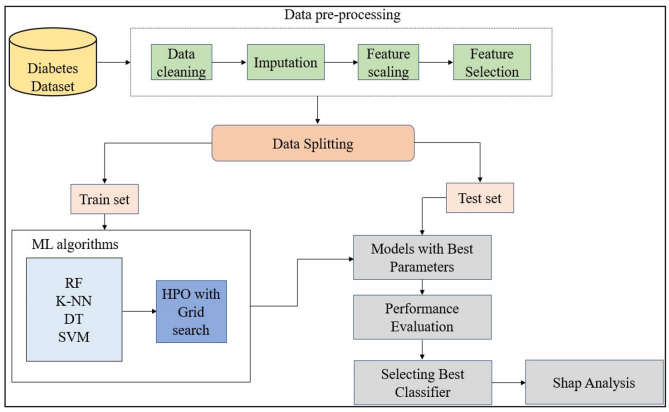

Figure 2. The flowchart of suggested predictive model for diabetes mellitus

Dataset

The approach developed within the scope of the study for the detection of DM is studied on the PIMA INDIANS (PIMA) data set. This dataset was originally created by the National Institute of Diabetes and Digestive and Kidney Diseases. It is open and accessible from the University of California, Irvine UCI AI repository. There are a total of 768 samples in the dataset, 500 from the non-diabetic class and 268 from the diabetic class. This dataset comprises eight independent and one output attributes. The output attribute has two classes, where '0' represents non diabetic and '1' represents diabetic.[20]

Data Pre-processing

Data preprocessing is the first stage to make the diabetes dataset raw data available for the prediction process. The data pre-processing stage consists of data cleaning, imputation, feature scaling and feature selection. In the data cleaning phase, operations such as cleaning the outliers detected in the data set, removing or completing the missing data are performed. These operations reduce the noise on the data. **Table 1** presents the statistical characteristics of the attributes in the data set. When the dataset was examined, it was seen that there were no missing values in the dataset and some attributes had zero values. In general, the Glucose, Insulin, BMI and blood pressure range can never start from zero values. Therefore, imputation process is required to fill the missing values. Imputation is a technique to replace missing data with some substitute values to preserve most of the data/information in the dataset.[21] Values with missing data in the data set were filled using the mean and median values of the features. In the next step, the feature scaling process is applied to the data sets whose missing data are completed. Feature scaling is one of the essential issues in preprocessing process before fitting it into the ML algorithms. This process can make a weak ML algorithm a better one. In this study, Min-max scaling technique is utilized for feature scaling. The principle this technique is illustrated in **Equation 1**. In this method, the largest and smallest values in a data set are taken into account. All other data are normalized to these values.[22]

Table 1. The statistical characteristics of the attributes in the PIMA INDIANS Diabetes Dataset

Attributes	Mean	Std.	Min	Max	Zero values
Pregnancies	3.84	3.37	0	17	0
Glucose	120.89	31.97	0	199	5
Blood pressure	69.1	19.35	0	122	35
Skin thickness	20.53	15.95	0	99	227
Insulin	79.8	115.24	0	846	374
Body mass index	32	7.88	0.07	67.1	11
Diabetes pedigree function	0.47	0.33	21	2.42	0
Age	33.24	11.76	0	81	0

$$X = \frac{x - \min(x)}{\max(x) - \min(x)}$$

The last step is selection of relevant feature. This process aims to reduce the number of attributes when building a predictive model.[19] In this study, the statistical correlations are used to identify critical features that could contribute significantly to ML algorithm and to achieve optimum performance. Correlation is utilized to measure the strength of the relationship between two attributes that are required in real life. Thus, it can predict the value of a variable with the help of other attributes associated with it. It is a type of bivariate statistics. The correlation matrix is defined as a table of all bivariate or zero-order correlations between and among the attributes in the dataset.[23] A correlation heatmap plot for feature selection is presented in **Figure 3**. It depicts shows correlations between and among all relevant features.

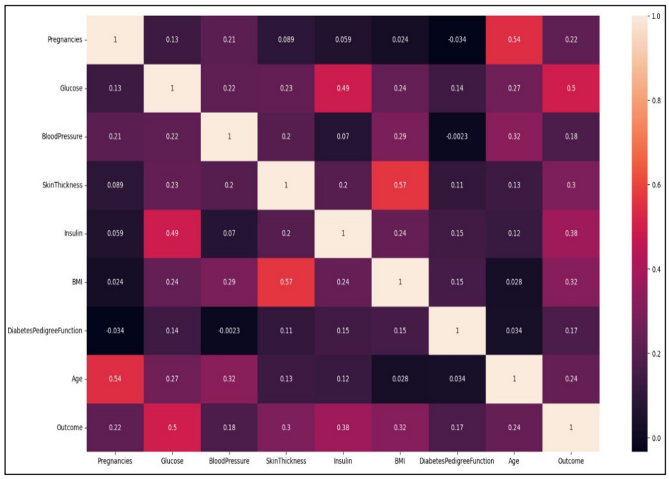

Figure 3. The correlation heatmap for feature selection

From the **Figure 4**, it can see a glimpse that the magnitude of correlation between "Outcome" output attribute and the independent attributes. The correlation coefficients are shown between output attribute and the independent attributes are shown in **Figure 5**. It can be clearly that the BP and DPF correlation sizes are less than "0.2" that is a low correlation with the outcome. Therefore, blood pressure (BP) and diabetes pedigree function (DPF) are eliminated from the primary diabetes dataset. Finally, after the feature selection process, the pregnancies, glucose, skin thickness (ST), insulin, body mass index (BMI) and age attributes are determined as the most relevant.

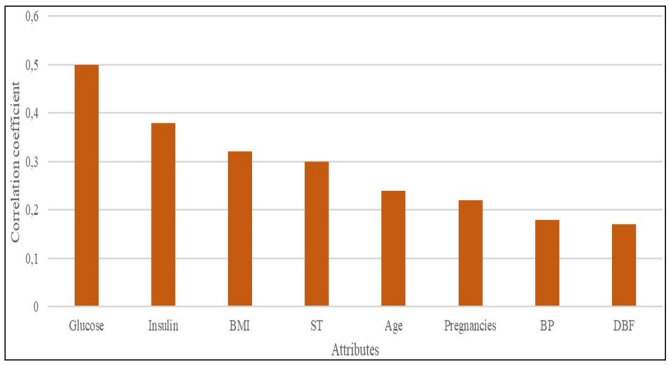

Figure 4. The correlation coefficients are shown between output attribute and the independent attributes

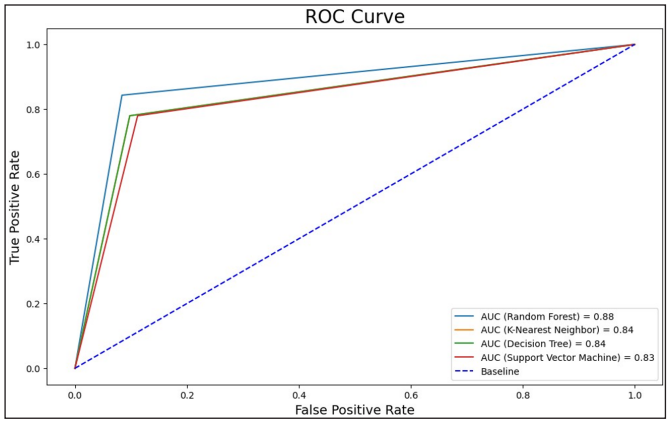

Figure 5. AUC-ROC Curve

Splitting the Dataset

Data splitting divides existing data into two parts. The first is used as training data to build a predictive model while the other is used as test data to evaluate the performance. In this study, 10-fold stratified cross validation (SCV) method is utilized as data splitting approach. SCV is an enhanced version of the k-fold CV. In k-fold CV, the dataset is divided into k subsets of equal size and samples are randomly selected for each subset or layer. Each subset is used for testing in turn and the rest is used for the training set. The model is evaluated k times model which each subset is utilized once as the test. Nevertheless, each subset is stratified so they contain approximately the same proportion of class labels as the original dataset during using SCV. Thus, the variance between predictions is diminished and the mean estimate error becomes more reliable.[24]

Machine Learning Algorithms

Because of the preprocessing step and the training/test sets partition, we processed it to fit the ML algorithms. Therefore, this section discusses the multiple supervised learning algorithms selected in this study to classify individuals with and without diabetes. Four classification algorithms used for predictions: RO, DT, K-NN and SVM. Grid search approach is utilized to automatically select the best hyperparameter for the ML algorithms. DT builds a structure a decision tree using the attributes in the dataset. This algorithm is utilized to predict classes based on the values of the attributes in the dataset. The tree starts from a root node and branches according to the characteristic of the dataset. Each branch contains a link between a property and possible values for that property. Each leaf node is associated with a class or output value. The K-NN algorithm looks at the K number of data points closest to that data point to estimate the class or value of a data point. SVM aims to find a function that can parse the training data using class labels. Since class labels are often called "yes-no" or "positive-negative", the most appropriate separator between these expressions, namely the hyperplane. In other words, SVM two decisions maximizes the distance between the boundary and the optimal separator aims to find the hyperplane. RF is formed by combining many DT. Each tree is applied to a subset of the dataset and its outputs are combined to make classification. The algorithm creates decision trees by choosing a random subset of the variables in the dataset. This helps prevent overfitting and increases generalization. Also, each tree has a different structure due to the selection of different subsets, resulting in greater diversity.[25]

Grid Search

Grid search approach is the one of most used hyperparameter optimization methods. It is a technique used in hyperparameter tuning to find the values that give the best performance in a given model. In this technique, the model is trained with all combinations of parameters given by the user, and it is an important process as the best parameters found affect the performance of the whole model. However, overfitting may occur during the optimization process. The overfitting problem can be reduced by applying the CV method. The CV technique trains a model with a dataset and tests it with various datasets. To determine the best combination of learning, Grid search with CV (GSCV) is utilized. Then, the set of parameter combinations with the highest accuracy is selected for each algorithm. After the selection of the best parameter set, the estimation process of the data begins. Using the k-fold CV technique, the dataset splits into training and testing part. 10-fold CV method is utilized in order to ten different sets of training and test. 10-fold CV method is utilized for each dataset to determine the average of diabetes prediction. With using the grid search and CV model together, hyperparameter optimization is obtained as a result of various experiments.[26]

Shapley Additive Explanations (SHAP) Analysis

Selecting the right algorithm with the right data has positive effect on the result in ML applications. ML algorithms takes the inputs and produce an output. Although the performance of the outputs can be measured by various techniques, the results are closed to make interpretation. In short, a black-box processed the data and produce an output. Explainable artificial intelligence methods can be applied to understand what is inside this black box mechanism. SHAP is one of these methods thar allows the ML model to be interpreted as a black box and not to be a black box. This method utilizes a game theory to identify ML algorithms. In this theory, the extent to which each player influences the game can be measured. In ML algorithms, it is possible to measure how much each attribute affects the result. In classification models using the SHAP technique, it can also be observed to what extent the features affect the result according to the classification type.[27]

Performance Evaluation Metrics

Evaluating the success of the models created in ML problems means determining the prediction success of the models. The confusion matrix is used to evaluate the relationship between the actual output values and the predicted values obtained after applying the models. From the confusion matrix, different performance metrics can be produced. The accuracy metric expresses the overall success of the model. Accuracy value, true number of classified samples divided by the total number of samples is obtained with. Precision, predicted positively how many of the values are actually positive metric showing diabetes disease. Also, the precision metric demonstrates the classifier's ability to eliminate false positives. Recall is expressed as a measure of ability to predict TP. F1-score is utilized to express the balance between precision and recall. One of the success evaluation criteria of classification models is ROC (Region of Curve). It explains how good the algorithm is at predicting. The area under the ROC Curve (AUC) can be considered as a summary of model ability, in other words, model performance.[28]

RESULTS

The hyperparameters of each algorithm, their search ranges and best combinations of hyperparameters of algorithms are shown in Table 2.

Table 2. The hyperparameters of each machine learning algorithms, their search ranges and best combinations of hyperparameters of machine learning algorithms

Algorithm	Hyperparameters	Search Range	Best Parameter
RF	"N_estimators"	[100, 200, 500,1000]	500
	"Max_features"	[3-7]	7
	"Min_samples_split"	[2-30]	5
	"Max_depth"	[3, 5, 8]	5
	"Min_samples_leaf"	[2-10]	5
K-NN	"N_neighbors"	[1-31]	5
DT	"Max_feature"	["auto", "sqrt", "log2"]	"log2"
	"Ccp_alpha"	[0.1, 0.01, 0.001]	0,01
	"Max_depth"	[5-9]	6
	"Criterion"	"entropy", "gini"	"entropy"
SVM	"C"	[1, 10, 100, 1000]	10
	"Gamma"	[1, 0.1, 0.01, 0.001, 0.0001]	0,1
	"Kernel"	["rbf", "linear"]	"rbf"

The prediction performances of used ML algorithms in DM diagnosis Table 3.

Table 3. The prediction performance of machine learning algorithms for diabetes mellitus

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
RF	89.06	84.33	84.33	84.33
K-NN	85.94	81.01	77.99	79.47
DT	85.42	79.54	78.36	78.95
SVM	85.02	78.87	77.99	78.42

AUC-ROC is one of the performance metrics used to evaluate the success of ML algorithms that explains how much algorithm distinguish between classes. Figure 5 illustrates the AUC-ROC curve for the ML algorithms.

DISCUSSION

Because diabetes is a worldwide public health threat, many researchers have motivated to develop ML applications to automate diabetes diagnosis as much as possible. In this study, different ML algorithms are used on a data set to detect diabetes. 10-fold SCV is used to for train-test split process. Hyperparameters of ML algorithms were determined by grid search 10-fold CV method. Data preparation, data preprocessing, creation of ML algorithms, statistical analyzes were performed using the Python program and its libraries. This study aims to compare the performance four different ML algorithms for diabetes prediction. The results obtained in terms of performance evaluation criteria of ML algorithms are presented in **Table 3** and **Figure 5**. In the first algorithm of this experiment is RF. In the RF algorithm, the result scored by applying GSCV method of best parameter the performance metrics of accuracy, precision, recall, F1-score and AUC-ROC is 89.06%, 84.33%, 84.33%, 84.33% and 0.88 respectively. The results obtained with the best parameters as a result of the hyperparameter optimization process of the K-NN used as the second algorithm is 85.94% of accuracy, 81.01% precision, 77.99% of recall, 79.47% of F1-score and 0.84 of AUC-ROC. The third ML algorithm used for diabetes prediction in the study is DT that shows 85.42%, 79.54%, 78.36%, 78.95% and 0.84 scores for performance evaluation metrics, respectively. The final ML algorithm SVM tested in this experiment achieves 85.02% accuracy, 78.87% precision, 77.99% recall, 78.42% F1-score, and 0.83 AUC score. When comparing the classification rates of ML, RF shows the best predictive ability to predict diabetes.

Clinical decisions that can be taken with ML algorithms in the field of health are important for the lives of patients. It is useful to have both accurate and interpretable prediction models in these applications. Therefore, to interpret the model, the SHAP technique can be used to discover the importance of each feature in determining the predicted output. The model can be interpreted based on the SHAP values that explain the contribution of each feature to the prediction. **Figure 6** shows the feature importance for the prediction of DM, and the contribution of each attribute to the performance of the algorithm. This figure depicts the SHAP values for the RF algorithm with the highest prediction rate.

According to this graph, "Insulin", "Age", and "Glucose" attributes contributed the most to the prediction model in identifying patients with diabetes. It is seen that the least contributing attributes are "Pregnancies", "BloodPressure" and "BMI".

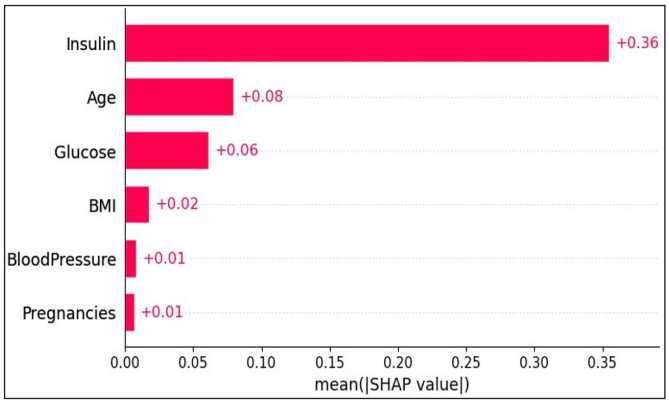

Figure 6. SHAP values of the RF algorithm

The comparison of the model proposed in this study with similar studies in the literature with the same data set are presented in **Table 4**. When similar studies in the literature are examined, it is seen that the models proposed for the prediction of DM reach accuracy rates between 75-90%. The GSCV-RF model proposed in this study also has a good prediction rate as a result of the comparison.

Table 4. Comparison of the suggested model with the similar studies in the literature

References	Year	Methods	Accuracy (%)
Birjais et al.[9]	2019	GB	86
Sing & Sing[11]	2020	Stacked ensemble approach	83.8
Lyngdogh et al.[12]	2021	K-NN	76
Kumari et al[13]	2021	SVC	79.08
Chang et al.[14]	2022	RF	82.26
Yakut[15]	2023	RF	81.77
Our proposed		GSCV+RF	89.06

The model suggested in this study was tested on an open access diabetes dataset. This situation is the most important limitation of our study. It is planned to predict the diabetes disease that may develop over time with high sensitivity and accuracy by using the clinical data, genetic data, past hospital visit data and current patient findings in the open access diabetes dataset that we used as a prototype in our study.

CONCLUSION

ML algorithms can enable the diagnosis of diseases by using datasets obtained in the field of health. In this study, an approach using ML algorithms for the diagnosis of diabetes, which is an important health problem worldwide, is proposed. The classification process is conducted by using four different ML algorithms on a diabetes dataset which is widely used in the literature. Since the data set used does not have a balanced bit structure, a series of data preprocessing steps were applied. The RF, K-NN, DT and SVM algorithms used in the study contain many hyperparameters. Choosing

the right combination of hyperparameters increases the success rate of ML algorithms. For this reason, GSCV method is used to select the most suitable hyperparameters of ML algorithms. The classification rates of the algorithms are evaluated with different performance criteria. As a result of the comparisons, the GSCV-RF model achieves the highest classification rate, with 89.06%, 84.33%, 84.33%, 84.33% and 0.88 accuracy, precision, sensitivity, F1-score and AUC-ROC. In this study, unlike the studies in the literature, the extent to which the attributes in the data set affect the result is investigated using SHAP analysis. The order of importance of the qualities that have an impact on the success of the model has been revealed in terms of interpreting this established model from the perspective of a healthcare professional. As a result of this analysis, it can be concluded that "Insulin", "Glucose", "Age" parameters have significant place in the diagnosis of diabetes. In this study, an ML estimator tool is presented to identify diabetic and non-diabetic individuals with high accuracy. It is thought that hospitals or diabetes prevention programs can benefit from the suggested approach.

ETHICAL DECLARATIONS

Ethics Committee Approval: Not applicable. An opensource dataset was utilized for the study.

Referee Evaluation Process: Externally peer-reviewed.

Conflict of Interest Statement: The authors have no conflicts of interest to declare.

Financial Disclosure: The authors declared that this study has received no financial support.

Author Contributions: All of the authors declare that they have all participated in the design, execution, and analysis of the paper, and that they have approved the final version.

REFERENCES

1. American Diabetes Association. Diagnosis and classification of diabetes mellitus. *Diabetes Care.* 2011;37 (Suppl_1):62-S69.
2. Priya G, Kalra S, Dasgupta A, Grewal E, Diabetes insipidus: a pragmatic approach to management. *Cureus.* 2021;13(1):e12498- e12498.
3. Prabhakar PK, Pathophysiology of secondary complications of diabetes mellitus. *Pathophysiology*. 2016;9(1):32-36.
4. Sun H, Saeedi P, Karuranga S, et al. IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes Res Clin Pract.* 2022;183:109119.
5. Sönmez A, Ozdoğan O, Arıcı M, et al. Diyabette kardiyovasküler ve renal komplikasyonların önlenmesi, tanısı ve tedavisi için Endokrinoloji Kardiyoloji Nefroloji (ENKARNE) Uzlaşı Raporu. *Turk J Endocrinol Metab.* 2021;25(4):392-411.
6. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nat Med.* 2022;28(1):31-38.
7. Ghaffar Nia N, Kaplanoglu E, Nasab A. Evaluation of artificial intelligence techniques in disease diagnosis and prediction. *Discover Artificial Intelligence.* 2023;3(1):5. doi:10.1007/s44163-023-00049-5
8. Ali YA, Awwad EM, Al-Razgan M, Maarouf A, Hyperparameter search for machine learning algorithms for optimizing the computational complexity. *Processes.* 2023;11(2):349.
9. Birjais R, Mourya AK, Chauhan R, Kaur H, Prediction and diagnosis of future diabetes risk: A machine learning approach. *SN Appl Sci.* 2019;9(1):1-8.
10. Tigga, NP, Garg S. Prediction of type 2 diabetes using machine learning classification methods. *Procedia Comput Sci.* 2020;167: 706-716.
11. Singh, N, Singh P. Stacking-based multi-objective evolutionary ensemble framework for prediction of diabetes mellitus. *Biocybern Biomed Eng.* 2020;40(1):1-22.
12. Lyngdoh AC, Choudhury NA, Moulik S. Diabetes disease prediction using machine learning algorithms. 2020 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES), Langkawi Island, Malaysia. 2021:517-521.
13. Kumari S, Kumar D, Mittal M, An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. *Int J Cog Comp in Eng.* 2021;2:40-46
14. Chang V, Ganatra MA, Hall K, Golightly L, Xu QA. An assessment of machine learning models and algorithms for early prediction and diagnosis of diabetes using health indicators. *Healthcare Analytics.* 2022;2(1):100118.
15. Yakut Ö. Diabetes prediction using colab notebook-based machine learning methods. *IJCESEN.* 2023;9(1):36-41.
16. Kluyver T, Ragan-Kelley B, Pérez F, et al. Jupyter Notebooks—A publishing format for reproducible computational workflows. In Positioning and Power in Academic, Players, Agents and Agendas; IOS Press: Amsterdam, The Netherlands. 2016;pp. 87-90.
17. The Python Library Reference, Release 3.8.8, Python Software Foundation. Available online: https://www.python.org/downloads/release/python-388/ (accessed on 10 May 2023).
18. Kumar VH. Python libraries, development frameworks and algorithms for machine learning applications. *IJERT.* 2018;7(4):2278-0181.
19. Pima Indians Diabetes Database | Kaggle, https://www.kaggle.com/datasets/uciml/pima-indiansdiabetes-database/ Accessed 09 May. 2023.
20. Joshi, AP, Patel BV, Data preprocessing: The techniques for preparing clean and quality data for data analytics process. Orient. *J Comput Sci Technol.* 2021;13(0203):78-81.
21. Ahsan MM, Mahmud MP, Saha PK, Gupta KD, Siddique Z. Effect of data scaling methods on machine learning algorithms and model performance. *Technologies.* 2021;9(3):52.
22. Venkatesh B, Anuradha J, A review of feature selection and its methods. *Cybern Inform Tech (CIT).* 2019;19(1):3-26.
23. Jamaluddin NSA, Kadir SA, Abdullah A, Alias SN, Learning strategy and higher order thinking skills of students in accounting studies:Correlation and regression analysis. *Univers J Educ.* 2020;8(3C):85-90.
24. Prusty S, Patnaik S, Dash SK. SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer. *Front Nanosci.* 2022;4:972421.
25. Ibrahim I, Abdulazeez A, The role of machine learning algorithms for diagnosing diseases. *J App Sci Techol Trends.* 2021;2(01):10-19.
26. Belete DM, Huchaiah MD, Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results. *Int J Comput Appl.* 2022;44(9):875-886.
27. Nohara Y, Matsumoto K, Soejima H, Nakashima N, Explanation of machine learning models using shapley additive explanation and application for real data in hospital. *Comput Methods Programs Biomed.* 2022;214:106584.
28. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res.* 2011;12:2825-2830.