Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2023) 31: 722 – 738

doi:10.55730/1300-0632.4013

A practical framework for early detection of diabetes using ensemble machine learning models

Qusay Saihood, Emrullah Sonuç*
Department of Computer Engineering, Karabuk University, Karabük, Turkey

Received: 27.02.2023 • **Accepted/Published Online:** 25.05.2023 • **Final Version:** 28.07.2023

Abstract: The diagnosis of diabetes, a prevalent global health condition, is crucial for preventing severe complications. In recent years, there has been a growing effort to develop intelligent diagnostic systems for diabetes utilizing machine learning (ML) algorithms. Despite these efforts, achieving high accuracy rates using such systems remains a significant challenge. Recent advancements in ensemble ML methods offer promising opportunities for early detection of diabetes, as they are known to be faster and more cost-effective than traditional approaches. Therefore, this study proposes a practical framework for diagnosing diabetes that involves three stages. The data preprocessing stage encompasses several crucial tasks, including handling missing values, identifying outliers, balancing the data, normalizing the data, and selecting relevant features. Subsequently, the hyperparameters of the ML algorithms are fine-tuned using grid search to improve their performance. In the final stage, the framework employs ensemble techniques such as bagging, boosting, and stacking to combine multiple ML algorithms and further enhance their predictive capability. Pima Indians Diabetes Database open-access dataset was used to test the performance of the proposed models. The experimental results of this framework indicate the superiority of ensemble methods in diagnosing diabetes compared to individual ML models. The stacking method achieved the best accuracy among the ensemble methods, with the stacked random forest (RF) and support vector machine (SVM) model attaining an accuracy of 97.50%. Among the bagging methods, the RF model yielded the highest accuracy, while among the boosting methods, eXtreme Gradient Boosting (XGB) model achieved the highest accuracy rates of 97.20% and 97.10%, respectively. Moreover, our proposed framework outperforms other ML models as confirmed by the comparison. The study has demonstrated that ensemble methods are crucial for accurate diabetes diagnosis, enabling early detection through efficient preprocessing and calibrated models.

Key words: Machine learning, ensemble learning, diabetes diagnosis, classification

1. Introduction

Diabetes is a chronic condition characterized by insufficient insulin production in the pancreas (type 1 diabetes, or T1D) or ineffective use of insulin in the body (type 2 diabetes, or T2D) [1]. Insulin, a hormone responsible for regulating blood sugar levels, is essential in maintaining proper health. When diabetes goes uncontrolled, it leads to hyperglycemia, which over time can cause significant damage to various body systems, particularly the nerves and blood vessels [2]. The World Health Organization (WHO) reports that diabetes is one of the leading causes of death worldwide. Furthermore, 422 million people, primarily in low- and middle-income countries,

*Correspondence: esonuc@karabuk.edu.tr

are living with the disease. As a result, diabetes is responsible for approximately 1.5 million deaths per year worldwide.[1]

In recent years, a significant increase has been observed in diabetics. Over the past several decades, the number of cases and the prevalence of diabetes have consistently risen. This condition can result in serious complications such as heart disease, hypertension, and stroke [3]. Early diagnosis and prediction of diabetes are critical to mitigating and avoiding complications of this disease [4]. Diagnosis of diabetes is often delayed due to a lack of attention to personal health and regular visits to healthcare facilities. This may lead to severe complications, even the loss of the patient's life. To address this issue, patients' health records and clinical data can be utilized to predict and diagnose diabetes in its early stages [5]. An intelligent system can aid in the prediction of diabetes by analyzing and extracting insights from clinical data.

The field of medicine has benefited greatly from the advancements in artificial intelligence (AI) and machine learning (ML) techniques [6]. ML helps to develop intelligent algorithms to perform tasks based on patterns and inference rather than explicit instructions. These algorithms are able to analyze a large datasets and uncover hidden patterns [7]. Hence, researchers have designed various ML models for the diagnosis of diseases [8]. Most of the traditional ML algorithms have been performed in the area, such as decision tree (DT), support vector machine (SVM), K-nearest neighbors (KNN) [9], various deep learning techniques and artificial neural networks (ANNs) models [10, 11]. However, no single technique has been deemed universally suitable for disease diagnosis. This presents a challenge for researchers to design innovative and competitive models for the diagnosis of diabetes. One of the contributing factors is the lack of attention paid to data preprocessing, which can significantly improve diagnostic results, but the effectiveness of ML refinements is highly dependent on the quality of data used in the training phase [12]. For these reasons, the accuracy of the diagnosis remains a major challenge for researchers.

Ensemble modeling refers to the technique of generating multiple models with diverse algorithms or training datasets for predicting an outcome. The predictions of each base model are then combined to produce a final prediction for the unseen data. The primary objective of ensemble models is to decrease the generalization error in prediction [13]. Ensemble methods are preferred over single models for two main reasons. Firstly, an ensemble model can enhance the performance of the model and produce superior predictions compared to a single model. Secondly, it also strengthens the robustness of the model by decreasing the dispersion of predictions and performance. There are only a few ensemble models proposed for diabetes diagnosis in the literature [14, 15]. Additionally, while preprocessing is recognized as a crucial step in improving the performance of ML models [16, 17], it appears that many studies in this field do not devote sufficient attention to this aspect. Thus, the significance and effectiveness of preprocessing techniques in this context have not been fully explored. This study proposes a three-stage ensemble learning approach to improve the accuracy of a diabetes diagnosis. The main contributions and limitations can be listed as follows:

- The three stages involved in the study are data preprocessing, hyperparameter tuning using the grid search method, and combining individual ML algorithms through ensemble learning models (bagging, boosting, and stacking).

- The experimental results of this framework indicate the superiority of ensemble methods in diagnosing diabetes compared to individual ML models.

[1]WHO (2023). Diabetes [online]. Website https://www.who.int/health-topics/diabetes [accessed 20 February 2023].

- The study highlights the importance of combining individual ML models through ensemble methods for accurate diabetes diagnosis, which could effectively detect the condition in its early stages.
- The superiority of the proposed model is confirmed through k-fold cross validation and a comparative analysis of its performance against other ML models reported in existing literature.
- The generalizability of the study findings to other datasets may be limited due to the exclusive use of the Pima Indians Diabetes Database (PIDD) open-access dataset.

The rest of the paper is organized as follows. Section 2 provides a brief overview of the relevant literature in the field. Section 3 provides a comprehensive description of the proposed method, including its various stages. Section 4 presents and evaluates the results obtained, with a comprehensive discussion of the findings. Finally, Section 5 concludes the study and suggests directions for future works.

2. Related work

In the field of medical diagnosis, researchers have demonstrated significant interest in the use of ML algorithms. One of the popular subjects in this domain is the diagnosis of diabetes, which has received considerable attention in the literature. Table 1 provides a summary of the ML models that have been proposed for the prediction of diabetes. Despite the progress that has been made, it remains a challenge to develop more accurate prediction methods.

The efficacy of different ML algorithms (ANN, SVM, KNN, DT, naive Bayes (NB), linear regression, random forest (RF), and adaptive boosting (AdaBoost)) were evaluated for the diagnosis of diabetes [12]. The data preprocessing steps included the selection of significant features from the PIDD dataset using the Pearson correlation as feature selection method. The data was divided into testing and training sets using k-fold cross validation and the train/test splitting method, and the ML algorithms were then evaluated using the testing set. ANN demonstrated the highest accuracy, while the other techniques achieved competitive results. In a similar study, ANN was again found to be the most effective technique, surpassing both the RF and K-mean [18]. ANN provided the best accuracy and was superior to both RF and K-mean. To address the overfitting problem in ANNs, dropout method was applied to the ANN model, which resulted in a significant improvement in accuracy [19]. DT, SVM, and NB were investigated for early diagnosis of diabetes and found that the NB performed best, with an accuracy of 76.30% [20]. To improve the accuracy of diabetes prediction, logistic regression (LR) was also suggested [21]. In another study [22], the effectiveness of a set of rarely ML algorithms (Reduces Error Pruning Tree, KStar, oneR, PART, sequential minimal optimization, and BayesNet) were examined, using 10-fold cross validation for training and testing. A well-tuned MLP (multilayer perceptron) was suggested for early prediction of diabetes, and compared its performance with a set of other ML algorithms which are K-mean, fuzzy c-means clustering (FCM), ANN, and convolutional neural network (CNN). The MLP demonstrated the best performance for early diabetes prediction among all comparative methods [23].

Another effort to improve the accuracy of diabetes diagnosis was data preprocessing process which involved the selection of the most significant samples and the detection of outliers using the radial basis function (RBF) [24]. The performance of a set of ML algorithms was then compared, and the MLP network with optimal stochastic gradient descent (SGD) outperformed the others. In order to achieve a more precise diagnosis, a voting method was utilized to combine individual ML algorithms, including DT, NB, SVM, RF, and ANN [25]. The proposed system was evaluated on the PIDD dataset, and the ensemble learning model that

combined DT and ANN models demonstrated the highest accuracy. In another study, researchers conducted a study to assess the effectiveness of LR and ensemble methods for the diagnosis of diabetes, both with and without feature selection methods [26]. Performance experiments were performed on two datasets, PIDD and Vanderbilt, and the ensemble methods were built by combining LR, NB, SVM, KNN, and DT using stacking and voting methods. The results showed that the voting method achieved the highest accuracy, outperforming the other techniques when comparing the performance on both datasets. The popularity of ensemble learning models in disease diagnosis has been growing in recent years, as a result of their demonstrated success. Five ensemble learning models (RF, AdaBoost, gradient boosting (GB), eXtreme Gradient Boosting (XGB), and a voting model) were investigated for diabetes diagnosis [27]. Before applying the ensemble learning models, the PIDD dataset was balanced, as well as the hyperparameter tuning of the models using the grid search method. The proposed ensemble learning models were tested on the PIDD dataset, with the XGB model exhibiting the best performance. A stacking model consisting of a SVM as a metalearner and four different ML models as base-learners, was also employed during the testing phase. It is worth highlighting that the stacking model, e.g., light gradient boosting machine (LightGBM), demonstrated superior performance scores across multiple diabetic datasets, including the PIDD dataset [14, 15].

Table 1. Summary of the recent studies for prediction of diabetes.

Ref.	Year	Data balancing	Feature slection	ML algorithms	Best method (accuracy)
[19]	2018	-	-	DL.	DL (88.41%)
[20]	2018	-	-	NB, SVM, DT.	NB (76.30%)
[25]	2018	-	✓	DT, NB, SVM, RF, ANN, Voting.	Voting (94.50%)
[22]	2019	-	-	REPTree, KStar, oneR, PART, SMO, BayesNet.	REPTree (74.48%)
[18]	2019	-	✓	ANN, RF, K-mean.	ANN (75.70%)
[24]	2020	-	-	NB, SVM, RF, DT, RFE, MLP-SGD, RFE + MLP-SGD.	MLP-SGD (96.90%)
[21]	2021	-	-	LR.	LR (75.32%)
[26]	2021	-	✓	LR, NB, SVM, KNN, DT, Stacking, Voting.	Voting (77.83%)
[27]	2021	✓	-	RF, AdaBoost, GB, XGB, Voting.	XGB (80.00%)
[12]	2021	-	✓	ANN, SVM, KNN, DT, NB, LR, RF, AdaBoost.	ANN (88.60%)
[23]	2022	-	-	K-mean, FCM, ANN, CNN, MLP.	MLP (86.08%)
[15]	2023	-	-	KNN, NB, LightGBM, Adaboost, RF.	Stacking (90.76%)
[14]	2023	-	✓	LR, DT, RF, GB, SVM.	Stacking (92.00%)

3. The proposed framework

This study comprises of several sequential stages. In the first stage, the data was prepared effectively by implementing the data preprocessing including filling the missing values, outlier detection, data balancing, data normalization, and feature selection. Filling the missing values helps to ensure that no information is lost during the modeling process. Handling outliers can improve the performance of the model by reducing the impact of

extreme values. Data balancing ensures that the model is not biased towards a particular class, which can help in building a more robust and accurate model. Data normalization can bring all features to the same scale and reduce the effect of variables with a wide range of values. Feature selection can help to identify the most relevant variables and reduce the complexity of the model, which can lead to improved performance and a more straightforward interpretation of the results. Combining these steps in an ensemble method can result in a more accurate and robust model that is capable of handling complex datasets with missing values, outliers, and imbalanced classes. In the second stage, the hyperparameters of the ML algorithms were tuned using the grid search. Later, the ensemble methods were built by combining individual well-tuned ML models. In the final stage, the performance of the models were evaluated using common performance evaluation measures: accuracy, precision, recall, and F1-score. A general overview of the proposed framework was depicted in Figure 1.

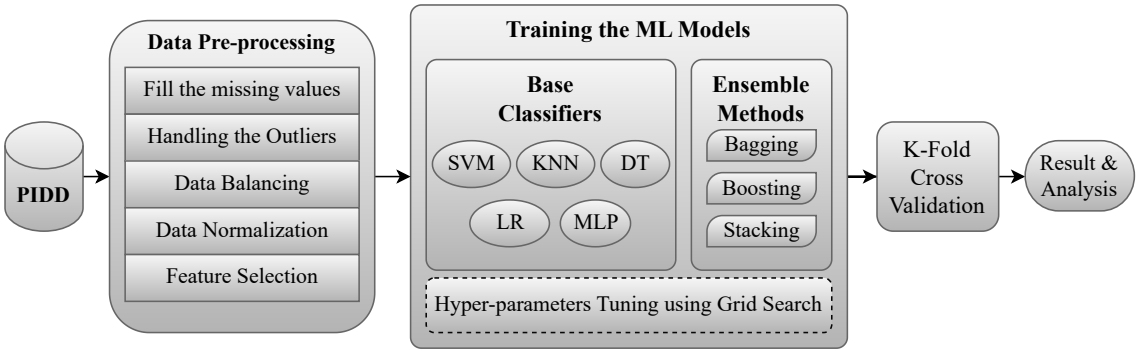

Figure 1. The proposed framework.

PIDD dataset which can be downloaded from UCI machine learning repository[2] was used for experimentation. Like many medical datasets, the dataset used in this study contains missing values and a disparity in its distribution. The dataset comprises 768 records of patients, with 268 being diabetic and 500 nondiabetic. It contains 8 features and a target variable called "outcome", which are illustrated in Table 2.

Table 2. The features of the Pima Indians Diabetes Database dataset.

#	Feature name	Data type
1	Age	Continuous
2	Pregnancies	Discrete
3	Glucose level	Continuous
4	Blood pressure	Continuous
5	Skin thickness	Continuous
6	Insulin level	Continuous
7	BMI	Continuous
8	Diabetes pedigree function	Continuous
9	Outcome	Discrete

3.1. Data preprocessing

The preprocessing of the diabetes data in this study was performed to address the inconsistencies and noise in the raw data. This is a critical step for ML algorithms as the quality of the input data has a significant impact

[2]https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

on the efficiency and accuracy of the model [28]. Therefore, the preprocessing stage is crucial for achieving high performance in building a model.

3.1.1. Filling the missing values

Five columns in PIDD dataset were noted to contain missing values (glucose, insulin, blood pressure, BMI, and skin thickness) that must be filled in. To handle these missing values, a mean imputation approach was employed in this study, where the mean value of each column was calculated based on the target class to which the missing value belonged (see Table 3).

Table 3. Average values calculated for missing values of the relevant feature.

Target (class)	Glucose	Insulin	Blood pressure	BMI	Skin thickness
0 (Nondiabetic)	107	102.5	70	30.1	27
1 (Diabetic)	140	169.5	74.5	34.3	32

3.1.2. Outliers detection

The interquartile range (IQR) is used to detect outliers. The IQR, which measures the distance between the first quartile (Q1) and the third quartile (Q3), is a widely accepted measure of central tendency and provides insight into the distribution of the data. By determining the range that encompasses the middle half of the data, the IQR is a useful tool for identifying potential outliers. The following steps were followed to detect and process outliers:

- Determining the first quartile (Q1).
- Determining the third quartile (Q3).
- Calculating the quartile range (IQR) by subtracting Q1 from Q3.
- The lower and upper limits of the normal data range were established using (1) and (2), respectively.

$$Lower\ whisker\ =\ Q1\ -\ (1.5 \times IQR) \tag{1}$$

$$Upper\ whisker\ =\ Q3\ +\ (1.5 \times IQR) \tag{2}$$

3.1.3. Data balancing

The categories in the PIDD dataset are imbalanced, with the diabetic category having a sample size of 268, and the nondiabetic category having a sample size of 500. To address this imbalance, the synthetic minority oversampling technique (SMOTE) [29] was employed to increase the size of the smaller category to be equivalent to the larger category. Following the implementation of the SMOTE, the distribution of each category was adjusted to achieve a state of equilibrium among them.

3.1.4. Data normalization

The min-max normalization method was applied to the numerical column values in the PIDD dataset to achieve a common scaling between 0 and 1 while maintaining the value ranges. This normalization was performed through Equation (3) [30]:

$$Y = (a - min(a)) \ / \ (max(a) - min(a)), \tag{3}$$

where Y represents the normalized value, a represents the original value, $min(a)$ represents minimum value, and $max(a)$ represents the maximum value.

3.1.5. Feature selection

The selection of relevant and influential features in a dataset plays a significant role in enhancing the performance of ML algorithms while also reducing computational requirements [31]. Feature selection methods reduce the dimensionality of the input data by identifying the most important features. There exist several methods for building feature selection methods, including filter methods, wrapper methods, and embedded methods [32]. To determine the most important features of PIDD dataset, the recursive feature elimination (RFE) was applied. RFE is one of the most popular encapsulation methods, due to its ease of configuration and effectiveness in selecting important features. The RFE operates by utilizing an ML algorithm as an estimator to select the features that are most relevant to the target [33]. It starts with all the features at the beginning and removes one by one from the least important features until the best subset of features is achieved. Consequently, four features were selected: age, glucose level, skin thickness, and insulin level.

3.2. ML algorithms

After completing preprocessing, which involved data preparation and feature selection, the ML algorithms for diagnosing diabetes were implemented. The following paragraphs provide a brief overview of each ML algorithm utilized.

DT is a well-known and effective supervised learning method that can be applied to both classification and regression problems. The structure of the DT is represented in the form of a tree, with the highest node being the root and each internal node representing a test on a feature. Each internal node represents a test on a feature, each branch indicates the result of the test, and each leaf node indicates a class label [34]. To diagnose diabetes, the CART (classification and regression tree) approach was employed, which generates binary trees and uses the entropy as a method for feature selection in classification problem

SVM is a powerful supervised ML model that is frequently utilized for classification problems. It classifies data by constructing a hyperplane or line that separates two classes within a dataset. The optimal hyperplane is determined by calculating the distance between the points of the two different classes and identifying the points closest to each class, referred to as support vectors. The hyperplane that offers the greatest margin between the hyperplane and the support vectors is then chosen [20].

LR is a supervised learning classification model that fits data to a logistic curve to forecast the likelihood of an occurrence using a statistical model. This model is utilized to estimate discrete values, such as 0/1 or yes/no, based on a set of independent variables. The model fits the data to a logistic curve, representing the probabilities in favor of a specific event. LR is a multivariable method that endeavors to establish a functional relationship between two or more predictor variables and a single dependent variable [21]. In the present study, Binary LR is used to predict the membership of two categorical outcomes, namely, “diabetic” and “nondiabetic”.

KNN is one of the simplest ML algorithms that can be applied to both classification and regression problems. KNN assumes that convergent objects are the same. In other words, similar things are close to each other. To classify an instance, KNN computes the distances between the instance and all the instances in the training set. Then, it selects the "k" nearest neighbors based on the computed distances and determines the class label of the instance by taking a majority vote among the k nearest neighbors [35]. In this study, the Minkowski distance was used to measure the distances between instances, and multiple values of k were tried to determine the optimal value for k.

MLP is a subtype of feed-forward ANNs designed to address nonlinear problems. The MLP structure consists of three layers, the first layer is the input layer, the middle layer is represented by one or more hidden layers, and the last layer is the output layer [7]. The input layer inputs the data to the first hidden layer, which consists of neurons. Within each neuron, the sum of the product of the input vectors multiplied by the initial weights with the added bias value is calculated. The activation function is then applied to the sum to obtain the output of the neuron. This output is then propagated forward through the rest of the hidden layers, until the final output is obtained through the output layer. The error between the actual and predicted values is calculated, and the weights are updated using the backpropagation method, which minimizes the error rate [36].

3.3. Tuning the hyperparameters of ML algorithms

ML algorithms have a set of hyperparameters that are used to control the learning process. Optimal tuning of hyperparameters helps greatly to improve the performance and stability of ML algorithms [37]. While manual experimentation can be used to set hyperparameters, this process can be time-consuming. To address this issue, in the present study, grid search is applied to determine the optimal hyperparameters for the ML models. A set of hyperparameters and their values are fed to the grid search. The model is then trained with different combinations of hyperparameter values to determine the optimal set [38]. Table 4 shows the parameter configurations for each classifier.

3.4. Ensemble methods

Ensemble methods, which incorporate a combination of multiple ML algorithms, offer a solution to the challenge of improving the generalizability and robustness of individual ML model. These methods are often seen to mitigate the variance and overfitting issues that are commonly encountered when relying solely on a single technique. The construction of ensemble methods typically involves one of three main approaches: bagging, boosting, or stacking methods [39].

3.4.1. Bagging methods

Bootstrap aggregation, commonly referred to as bagging, is a widely used ensemble method that leverages the parallel combination of multiple instances of the same model to enhance model performance and stability. The bagging method operates in two phases: bootstrapping and aggregation. During the bootstrapping phase, base classifiers are trained on subsets of the dataset and produce independent predictions for each classifier. The aggregation stage then compiles the predictions of these classifiers and employs voting to determine the final prediction [28, 40]. The benefit of using bagging methods is to combine weak individual classifiers to form a strong classifier that is more robust than weak classifiers. Some of the most well-known examples of bagging methods include the RF model and extra tree model [39].

Table 4. The hyperparameters and input values for the classifiers.

Algorithm	Hyperparameters and their possible values	Selected value
DT	Max depth = range(1,20)	8
	Criterion = [‘gini’, ‘entropy’]	entropy
	Splitter = [‘best’, ‘random’]	best
	Max features = [‘auto’, ‘sqrt’, ‘log2’]	auto
SVM	Degree = range (1,10)	5
	Gamma = [‘scale’, ‘auto’]	scale
	Kernel = [‘linear’, ‘poly’, ‘rbf’, ‘sigmoid’]	rbf
LR	Penalty = [‘l1’, ‘l2’, ‘elasticnet’, ‘none’]	none
	Solver = [‘lbfgs’, ‘liblinear’, ‘newton-cg’, ‘newton-cholesky’, ‘sag’, ‘saga’]	lbfgs
	Max iteration = [100, 1000, 10000]	10000
KNN	N neighbors = range(1,20)	4
	Weights = [‘uniform’, ‘distance’]	distance
	Algorithm = [‘auto’, ‘ball_tree’, ‘kd_tree’, ‘brute’]	ball_tree
	Leaf size = range(20,50)	40
	Metric = [‘minkowski’, ‘euclidean’]	minkowski
	P = range(1,4)	2
MLP	Hidden layer sizes = range(100,400)	300
	Activation =[‘identity’, ‘logistic’, ‘tanh’, ‘relu’]	relu
	Solver = [‘lbfgs’, ‘sgd’, ‘adam’]	lbfgs
	Learning rate = [‘constant’, ‘invscaling’, ‘adaptive’]	adaptive

RF model is a widely used ensemble method that is based on the bagging method. In this model, several classification and regression trees (CARTs) are generated, where each tree is trained on a subset of the original dataset. The predictions made by all the decision trees within the forest are aggregated and the final classification decision is made based on a majority vote of the trees [28, 41].

Similar to the RF model, extra tree (ET) model is also based on the compilation of a set of CARTs. Despite this similarity, differences can be observed in the training methodology utilized by the two models. Whereas the RF model trains its individual CARTs on a variety of random subsamples of the dataset, the ET model employs the original dataset to train its CARTs [41]. In the present study, a bagging model was built of each individual ML algorithm (DT, SVM, LR, KNN, and MLP), in addition to the usage of both the RF and ET models.

3.4.2. Boosting methods

The boosting method is a sequential ensemble learning model that aims to reduce variance and improve prediction accuracy by combining weak classifiers. The fundamental concept behind boosting methods is to identify models that complement each other, such that each subsequent model aims to rectify the prediction errors made by the previous model. The final classification result is determined through voting, as with bagging methods [42].

Adaptive boosting (AB), also known as AdaBoost trains a primary learning model on the dataset, followed by additional models on the same dataset. AB is referred to as an adaptive method as subsequent weak learners are modified to give weight to cases that were misclassified by earlier classifiers [27].

The gradient boosting (GB) model is a type of boosting method that combines a set of weak classifiers, usually CARTs. It is constructed in a stage-wise fashion, with the difference from other methods being that it optimizes a differentiable arbitrary loss function [27, 41].

XGBoost is short for eXtreme Gradient Boosting (XGB), which is an improvement of the GB model. It is distinguished from the GB model by its ability to sample both rows and columns, as opposed to just columns in the GB model. Furthermore, XGB has the advantage of consuming fewer computational resources and achieving superior results in a shorter period [27, 41].

3.4.3. Stacking methods

Stacking is a type of ensemble method that has been demonstrated to be effective in improving the performance of individual ML algorithms. This method involves combining heterogeneous weak learners through the use of a metaclassifier. The stacking method consists of two primary levels, namely Level 0 and Level 1, or the metaclassifier level. At Level 0, multiple base ML algorithms are trained on the training dataset. The predictions of the base methods from Level 0 are then passed to the upper layer (Level 1), where one of the base ML algorithms is designated as the metaclassifier, which often utilizes a LR. In Level 1, the metaclassifier is trained on both the predictions from Level 0 and the training dataset to produce the final predictions.

For this work, various models including DT, SVM, LR, KNN, MLP, and RF, are constructed by combining the various ML algorithms used at Level 0, while the LR is employed as the metaclassifier at Level 1.

3.5. Performance metrics

The performance of the ML algorithms was evaluated using four quality metrics, namely: accuracy, precision, recall, and F1-score. In this analysis, samples that contain diabetes were considered positive and were represented by the value "1", while healthy samples were considered negative and represented by the value "0". Equations (4) through (7) present the mathematical formulations of the performance metrics. The following definitions were used to compute the performance metrics:

- **True positive (TP)**: Individuals diagnosed with diabetes who have diabetes.
- **True negative (TN)**: Individuals diagnosed as healthy who do not have diabetes.
- **False positive (FP)**: Individuals diagnosed with diabetes who are healthy.
- **False negative (FN)**: Individuals diagnosed as healthy who have diabetes.

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \tag{4}$$

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{7}$$

4. Results and discussion

The Python programming language was utilized to perform classification procedures.[3] All ML algorithms were implemented using the scikit-learn library, except for the XGB, which is sourced from the XGBoost library. The models were trained by utilizing feature selection methods to identify crucial features and subsequently assessed using 5-fold cross validation on a 20% sample of test data. The results achieved in diagnosing diabetes through the utilization of the proposed models are analyzed and discussed in multiple sections.

4.1. Experimental results of ML algorithms

Five well-tuned individual ML algorithms were applied to diagnose diabetes. Among the all algorithms, DT achieved the highest level of accuracy, with a rate of 95.50%, along with a precision of 94.97%, recall of 95.79%, and F1-score of 95.34%. Additionally, the MLP achieved promising results that are relatively close to those of the DT, with an accuracy rate of 94.90%, precision of 92.13%, recall of 97.70%, and F1-score of 94.87%. According to the results, SVM ranks third with an accuracy of 92.00%, followed by KNN with an accuracy of 91.40%, and LR with the lowest accuracy of 85.80%. Table 5 presents the results of the performance criteria for the individual ML algorithms employed for diagnosing diabetes.

Table 5. Classification results of ML algorithms.

Algorithm	Accuracy	Precision	Recall	F1-score
DT	0.955	0.950	0.958	0.953
SVM	0.920	0.871	0.986	0.925
LR	0.858	0.829	0.903	0.864
KNN	0.914	0.878	0.962	0.918
MLP	0.949	0.921	0.978	0.949

4.2. Experimental results of ensemble methods

In order to enhance the accuracy of diabetes diagnosis, the bagging technique was employed. In this approach, distinct bagged models were constructed for each of the individual algorithms utilized, as well as incorporating both RF and ET models. The results, as presented in Table 6, indicate that the performance of the RF model, the bagged DT model, and the ET model was highly competitive, achieving the highest, second highest, and third highest levels of accuracy, with 97.20%, 97.10%, and 96.10% accuracy, respectively. Additionally, the bagged MLP model demonstrated a slight improvement over the individual MLP model. However, the bagged KNN model and the bagged LR model did not exhibit any significant improvement.

Three of the most widely used boosting models (AB, GB, and XGB) were constructed for the purpose of diagnosing diabetes. The results of the performance evaluations, as displayed in Table 7, reveal that the XGB model achieved the highest level of accuracy at 97.10%, outperforming both the AB and GB models. The GB model achieved the second highest accuracy rate of 96.90%, while the AB model recorded the lowest accuracy of 96.60%.

For stacking methods, two-stage performance experiments were conducted. In the first stage, many stacked models were built by combining different ML (DT, SVM, LR, KNN, and MLP) algorithms. Stacked models that combine DT with various other algorithms performed better than other models that do not contain

[3]https://github.com/qusayallahibi95/Ensemble-ML-models-for-diabetes-diagnosis

a DT. As illustrated in Figure 2, the stacked DT+SVM model achieved the highest accuracy with a rate of 97.20%. The stacked DT+KNN+SVM+LR model and the stacked DT+LR models followed closely, achieving the second- and third-highest accuracy rates of 96.80% and 96.70%, respectively.

Table 6. Classification results of bagging models.

Bagging models	Accuracy	Precision	Recall	F1-score
RF	0.972	0.958	0.988	0.972
ET	0.961	0.934	0.994	0.963
Bagged DT	0.971	0.960	0.984	0.972
Bagged SVM	0.925	0.875	0.992	0.930
Bagged LR	0.845	0.829	0.872	0.849
Bagged KNN	0.912	0.870	0.970	0.917
Bagged MLP	0.950	0.922	0.984	0.952

Table 7. Classification results of boosting models.

Boosting models	Accuracy	Precision	Recall	F1-score
AB	0.966	0.955	0.978	0.966
GB	0.969	0.950	0.990	0.970
XGB	0.971	0.957	0.986	0.971

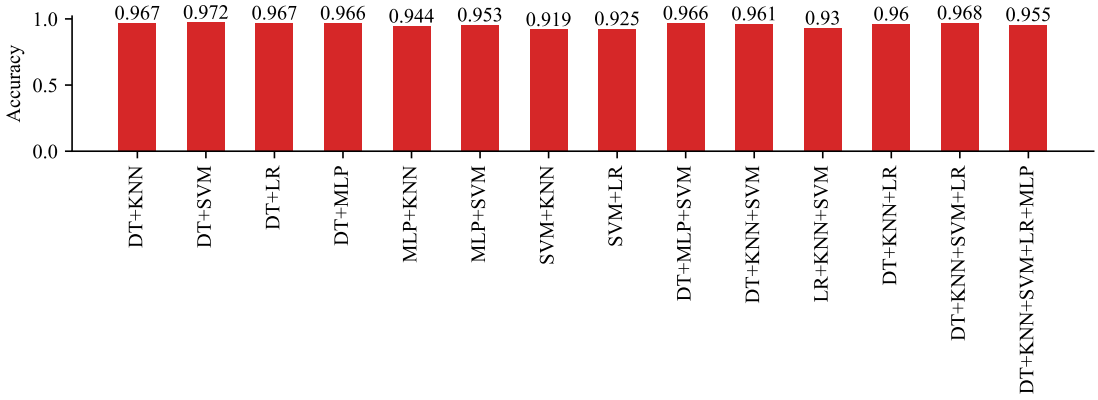

Figure 2. Accuracy performances of the various stacked models.

In the second stage, stacked models were built by combining the RF model with various individual ML algorithms to enhance performance. These combinations produced superior results compared to the stacked models in the first stage, with the stacked RF+SVM model achieving the highest accuracy of 97.50%. The stacked RF+SVM+LR+DT model ranked second with an accuracy of 97.40%, while the stacked RF+LR and stacked RF+SVM+LR models ranked third with an accuracy rate of 97.30%. The accuracy of the stacked RF+ML models is shown in Figure 3.

In order to demonstrate the efficacy of the proposed models for the purpose of diabetes diagnosis, a comparative analysis was performed utilizing existing studies, as presented in Table 8. The results show that

the accuracy of diabetes diagnosis is more precise than other approaches [12, 27]. Among the comparative methods, the highest accuracy was achieved using voting methods with an accuracy of 94.50% [25], whereas the stacked RF+SVM model in this study achieved an accuracy of 97.50%. Additionally, the proposed XGB model outperformed the XGB models used in another study [27]. The results of this study also revealed that the proposed stacked models were more accurate in diagnosing diabetes compared to the existing stacking models [26]. Figure 4 presents the confusion matrices of the ensemble models developed using different combinations of individual classifiers. These matrices were used to evaluate the performance of the models based on true positives, true negatives, false positives, and false negatives. The results presented in the confusion matrices provide insight into the accuracy of the models in predicting outcomes for different classes. According to our evaluation criteria, the confusion matrix obtained from the stacking model demonstrated the highest accuracy in predicting different class outcomes, with a significant number of true positives and true negatives and a relatively lower number of false positives and false negatives. These findings suggest that the stacking model, with the given ensemble classifier combination, exhibited the best overall performance. In contrast, the bagging model exhibited slightly lower accuracy than the stacking model, but still showed a good balance between true positives and true negatives with fewer false negatives compared to the boosting model. Based on our results, we recommend the stacking model with the given individual classifier combination if the aim is to achieve a good balance between accuracy and efficiency.

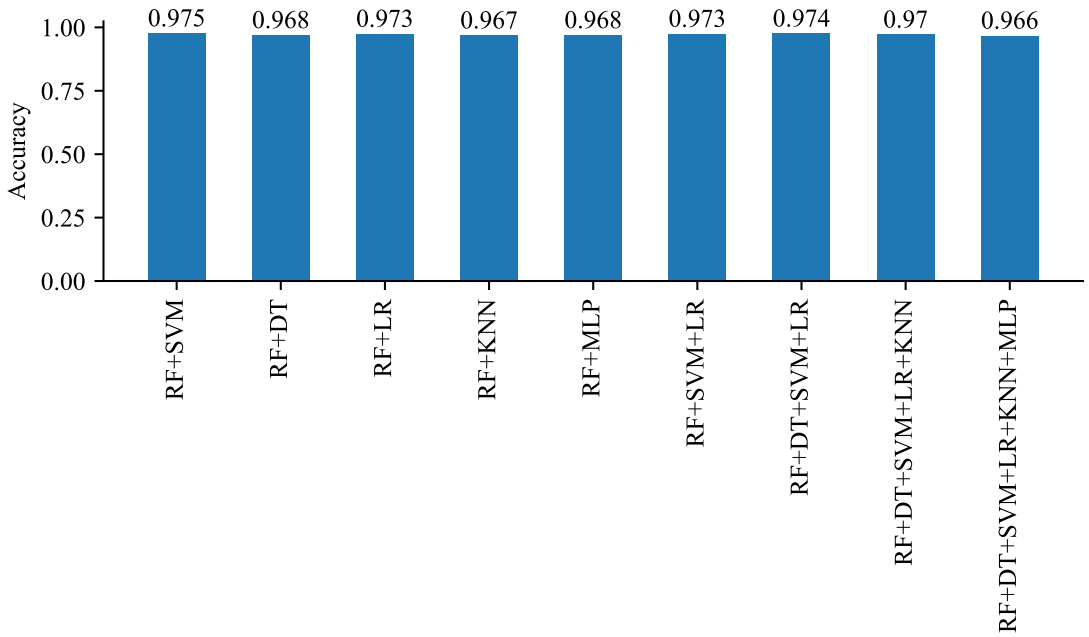

Figure 3. Accuracy performances of the RF-based stacked models.

As a result, ensemble learning models, when combined with data preprocessing techniques, have been shown to yield better predictions for diabetes diagnosis than individual ML models. Data preprocessing can identify and correct errors, reduce noise and outliers, and normalize data distributions. By using ensemble models that combine the predictions of several individual models trained on preprocessed data, the accuracy of the final prediction can be further improved. Ensemble models also mitigate the impact of bias and variability

in the data, resulting in more robust and reliable predictions. The findings of this study demonstrate that the prediction accuracy of a learning model is not only dependent on the performance of the model, but can also be improved through efficient data preprocessing, feature selection, and hyperparameter tuning. Furthermore, the study highlights the potential of ensemble methods in combining weak classifiers to build strong models.

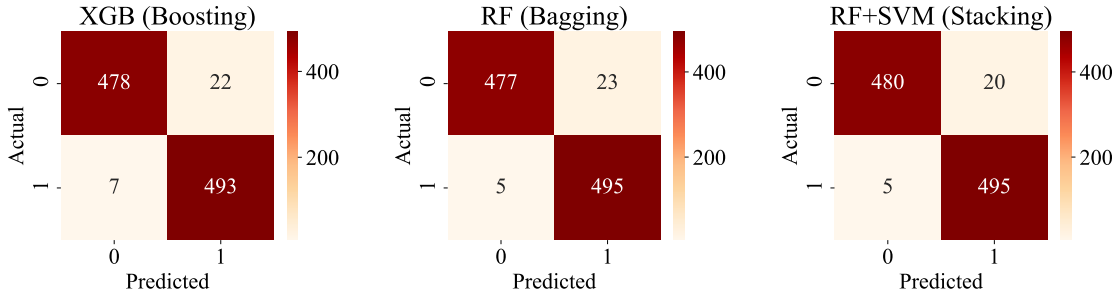

Figure 4. Confusion matrices for ensemble models.

Table 8. Comparison with other studies.

Reference	ML model	Accuracy (%)
[12]	ANN	88.60
[14]	Stacking	92.00
[15]	Stacking	90.76
[18]	ANN	75.70
[19]	DL	88.41
[20]	NB	76.30
[21]	LR	75.32
[22]	REPTree	74.48
[23]	MLP	86.08
[24]	MLP-SGD	96.90
[25]	Voting	94.50
[26]	Voting	77.83
[27]	XGB	80.00
This study	XGB (boosting)	97.10
This study	RF (bagging)	97.20
This study	RF+SVM (stacking)	97.50

5. Conclusion

In the field of medical diagnosis, it is of utmost importance to develop a system that is capable of effectively and accurately diagnosing diseases, particularly in the early stages to prevent further health complications. This study presents a framework for the diagnosis of diabetes that has been developed with the objective of achieving high accuracy in the diagnosis. The proposed framework encompasses three key stages: data preprocessing, hyperparameter tuning, and the application of ensemble methods.

Data preprocessing involves filling missing values, detecting outliers, balancing the data, normalizing the data, and selecting the most relevant features. The second stage entails the tuning of the hyperparameters of the

ML algorithms using grid search method. Finally, in the third stage, the framework leverages ensemble methods such as bagging, boosting, and stacking to combine individual ML algorithms and enhance their performance. Ensemble methods effectively reduce variance and address the problem of overfitting, resulting in a robust and high-performing model.

The proposed framework, incorporating effective data preprocessing and well-tuned ensemble learning models, has demonstrated promising results in the diagnosis of diabetes. In particular, stacking methods outperformed the other ensemble methods, with the stacked RF+SVM model achieving the highest accuracy of 97.50%. Among the bagging methods, the RF model achieved the highest accuracy of 97.20%, while the XGB model among the boosting methods demonstrated the best result with an accuracy of 97.10%. The experimental studies on PIDD have demonstrated that all ensemble models outperformed individual algorithms. The superior performance of ensemble models can be attributed to several factors. Boosting effectively identifies and reduces errors in each iteration, resulting in improved model performance. Bagging is particularly effective in decision trees, enhancing the accuracy and robustness of the models. Additionally, a comprehensive grid search with different hyperparameters is employed to fine-tune individual models used in stacked ensemble models, leading to improved overall model performance. Future work should focus on training models on larger datasets to further improve performance and explore the integration of deep learning techniques in ensemble methods.

References

[1] Jaiswal V, Negi A, Pal T. A review on current advances in machine learning based diabetes prediction. Primary Care Diabetes. 2021; 15 (3): 435-43. https://doi.org/10.1016/j.pcd.2021.02.005

[2] American Diabetes Association. Diagnosis and classification of diabetes mellitus. Diabetes care. 2014; 37(Supplement_1): S81-90. https://doi.org/10.2337/dc14-S081

[3] Tun NN, Arunagirinathan G, Munshi SK, Pappachan JM. Diabetes mellitus and stroke: a clinical update. World journal of diabetes. 2017; 8(6): 235. https://doi.org/10.4239/wjd.v8.i6.235

[4] American Diabetes Association Professional Practice Committee. 13. Older Adults: Standards of Medical Care in Diabetes—2022. Diabetes Care. 2022; 45(Supplement_1): S195–207. https://doi.org/10.2337/dc22-S013

[5] Khan FA, Zeb K, Al-Rakhami M, Derhab A, Bukhari SA. Detection and prediction of diabetes using data mining: a comprehensive review. IEEE Access. 2021; 9: 43711-35. https://doi.org/10.1109/ACCESS.2021.3059343

[6] Vetter N. The promise of artificial intelligence: a review of the opportunities and challenges of artificial intelligence in healthcare and clinical trials in skeletal dysplasia: a paradigm for treating rare diseases. British Medical Bulletin. 2021; 139(1): 1-3. https://doi.org/10.1093/bmb/ldab022

[7] Saihood Q, Sonuç E. The Efficiency of Classification Techniques in Predicting Anemia Among Children: A Comparative Study. In: Emerging Technology Trends in Internet of Things and Computing. Cham: Springer International Publishing; 2022. p. 167–81. https://doi.org/10.1007/978-3-030-97255-4_12

[8] Sharma T, Shah M. A comprehensive review of machine learning techniques on diabetes detection. Visual Computing for Industry, Biomedicine, and Art. 2021; 4: 1-6. https://doi.org/10.1186/s42492-021-00097-7

[9] Tuppad A, Patil SD. Machine learning for diabetes clinical decision support: a review. Advances in Computational Intelligence. 2022; 2 (2): 22. https://doi.org/10.1007/s43674-022-00034-y

[10] Frimpong EA, Oluwasanmi A, Baagyere EY, Zhiguang Q. A feedforward artificial neural network model for classification and detection of type 2 diabetes. In Journal of Physics: Conference Series 2021 (Vol. 1734, No. 1, p. 012026). IOP Publishing. https://doi.org/10.1088/1742-6596/1734/1/012026

[11] Bani-Salameh H, Alkhatib SM, Abdalla M, Al-Hami MT, Banat R et al. Prediction of diabetes and hypertension using multi-layer perceptron neural networks. International Journal of Modeling, Simulation, and Scientific Computing. 2021; 12 (02): 2150012. https://doi.org/10.1142/S1793962321500124

[12] Khanam JJ, Foo SY. A comparison of machine learning algorithms for diabetes prediction. ICT Express. 2021; 7 (4): 432-9. https://doi.org/10.1016/j.icte.2021.02.004

[13] Kotu V, Deshpande B. Data mining process. Predictive analytics and data mining. 2015: 17-36. https://doi.org/10.1016/B978-0-12-801460-8.00002-1

[14] Doğru A, Buyrukoğlu S, Arı M. A hybrid super ensemble learning model for the early-stage prediction of diabetes risk. Medical & Biological Engineering & Computing. 2023: 1-3. https://doi.org/10.1007/s11517-022-02749-z

[15] Sai MJ, Chettri P, Panigrahi R, Garg A, Bhoi AK et al. An Ensemble of Light Gradient Boosting Machine and Adaptive Boosting for Prediction of Type-2 Diabetes. International Journal of Computational Intelligence Systems. 2023; 16 (1): 14. https://doi.org/10.1007/s44196-023-00184-y

[16] Yağ İ, Altan A. Artificial Intelligence-Based Robust Hybrid Algorithm Design and Implementation for Real-Time Detection of Plant Diseases in Agricultural Environments. Biology. 2022; 11 (12): 1732. https://doi.org/10.3390/biology11121732

[17] Sezer A, Altan A. Detection of solder paste defects with an optimization-based deep learning model using image processing techniques. Soldering & Surface Mount Technology. 2021; 33 (5): 291-8. https://doi.org/10.1108/SSMT-04-2021-0013

[18] Alam TM, Iqbal MA, Ali Y, Wahab A, Ijaz S et al. A model for early prediction of diabetes. Informatics in Medicine Unlocked. 2019; 16: 100204. https://doi.org/10.1016/j.imu.2019.100204

[19] Ashiquzzaman A, Tushar AK, Islam MR, Shon D, Im K et al. Reduction of overfitting in diabetes prediction using deep learning neural network. In IT Convergence and Security 2017: pp. 35-43. Springer Singapore. https://doi.org/10.1007/978-981-10-6451-7_5

[20] Sisodia D, Sisodia DS. Prediction of diabetes using classification algorithms. Procedia computer science. 2018; 132: 1578-85. https://doi.org/10.1016/j.procs.2018.05.122

[21] Tigga NP, Garg S. Predicting type 2 diabetes using logistic regression. In Proceedings of the Fourth International Conference on Microelectronics, Computing and Communication Systems: MCCS 2019 2021 (pp. 491-500). Springer Singapore. https://doi.org/10.1007/978-981-15-5546-6_42

[22] Larabi-Marie-Sainte S, Aburahmah L, Almohaini R, Saba T. Current techniques for diabetes prediction: review and case study. Applied Sciences. 2019; 9 (21): 4604. https://doi.org/10.3390/app9214604

[23] Sivasankari SS, Surendiran J, Yuvaraj N, Ramkumar M, Ravi CN et al. Classification of Diabetes using Multilayer Perceptron. In2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE) 2022 ; pp. 1-5. IEEE. https://doi.org/10.1109/ICDCECE53908.2022.9793085

[24] Ranjeeth S, Kandimalla VA. Predicting Diabetes Using Outlier Detection and Multilayer Perceptron with Optimal Stochastic Gradient Descent. In 2020 IEEE India Council International Subsections Conference (INDISCON) 2020 ; pp. 51-56. IEEE. https://doi.org/10.1109/INDISCON50162.2020.00023

[25] Kaur P, Sharma N, Singh A, Gill B. CI-DPF: A cloud IoT based framework for diabetes prediction. In 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON) 2018 : pp. 654-660. IEEE. https://doi.org/10.1109/IEMCON.2018.8614775

[26] Rajendra P, Latifi S. Prediction of diabetes using logistic regression and ensemble techniques. Computer Methods and Programs in Biomedicine Update. 2021; 1: 100032. https://doi.org/10.1016/j.cmpbup.2021.100032

[27] Alluri RP, Hemavathy R. Diabetes Prediction Using Ensemble Techniques. International Journal of Applied Engineering Research. 2021; 16 (5): 410-5.

[28] Han J, Pei J, Tong H. Data mining: concepts and techniques. Morgan kaufmann; 2022.

[29] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research. 2002; 16: 321-57. https://doi.org/10.1613/jair.953

[30] Singh D, Singh B. Investigating the impact of data normalization on classification performance. Applied Soft Computing. 2020; 97: 105524. https://doi.org/10.1016/j.asoc.2019.105524

[31] Sharma NV, Yadav NS. An optimal intrusion detection system using recursive feature elimination and ensemble of classifiers. Microprocessors and Microsystems. 2021; 85: 104293. https://doi.org/10.1016/j.micpro.2021.104293

[32] Khaire UM, Dhanalakshmi R. Stability of feature selection algorithm: A review. Journal of King Saud University-Computer and Information Sciences. 2022; 34 (4): 1060-73. https://doi.org/10.1016/j.jksuci.2019.06.012

[33] Senan EM, Al-Adhaileh MH, Alsaade FW, Aldhyani TH, Alqarni AA et al. Diagnosis of chronic kidney disease using effective classification algorithms and recursive feature elimination techniques. Journal of Healthcare Engineering. 2021. https://doi.org/10.1155/2021/1004767

[34] Ghiasi MM, Zendehboudi S, Mohsenipour AA. Decision tree-based diagnosis of coronary artery disease: CART model. Computer methods and programs in biomedicine. 2020; 192: 105400. https://doi.org/10.1016/j.cmpb.2020.105400

[35] Xing W, Bei Y. Medical health big data classification based on KNN classification algorithm. IEEE Access. 2019; 8: 28808-19. https://doi.org/10.1109/ACCESS.2019.2955754

[36] Fath AH, Madanifar F, Abbasi M. Implementation of multilayer perceptron (MLP) and radial basis function (RBF) neural networks to predict solution gas-oil ratio of crude oil systems. Petroleum. 2020; 6 (1): 80-91. https://doi.org/10.1016/j.petlm.2018.12.002

[37] Weerts HJ, Mueller AC, Vanschoren J. Importance of tuning hyperparameters of machine learning algorithms. arXiv preprint arXiv:2007.07588. 2020. https://doi.org/10.48550/arXiv.2007.07588

[38] Sumathi B. Grid search tuning of hyperparameters in random forest classifier for customer feedback sentiment prediction. International Journal of Advanced Computer Science and Applications. 2020; 11 (9). https://doi.org/10.14569/IJACSA.2020.0110920

[39] Shorewala V. Early detection of coronary heart disease using ensemble techniques. Informatics in Medicine Unlocked. 2021; 26: 100655. https://doi.org/10.1016/j.imu.2021.100655

[40] Richman R, Wüthrich MV. Nagging predictors. Risks. 2020; 8 (3): 83. https://doi.org/10.3390/risks8030083

[41] Al Bataineh A, Manacek S. MLP-PSO hybrid algorithm for heart disease prediction. Journal of Personalized Medicine. 2022; 12 (8): 1208. https://doi.org/10.3390/jpm12081208

[42] Yaman MA, Rattay F, Subasi A. Comparison of bagging and boosting ensemble machine learning methods for face recognition. Procedia Computer Science. 2021; 194: 202-9. https://doi.org/10.1016/j.procs.2021.10.074