



A hybrid super ensemble learning model for the early-stage prediction of diabetes risk

Ayşe Dođru¹ · Selim Buyrukođlu² · Murat Arı¹

Received: 31 July 2022 / Accepted: 22 December 2022 / Published online: 5 January 2023
© International Federation for Medical and Biological Engineering 2023

Abstract

Diabetes mellitus has become a rapidly growing chronic health problem worldwide. There has been a noticeable increase in diabetes cases in the last two decades. Recent advances in ensemble machine learning methods play an important role in the early detection of diabetes mellitus. These methods are both faster and less costly than traditional methods. This study aims to propose a new super ensemble learning model to enable an early diagnosis of diabetes mellitus. Super learner is a cross-validation-based approach that makes better predictions by combining prediction results of more than one machine learning algorithm. The proposed super learner model was created with four base-learners (logistic regression, decision tree, random forest, gradient boosting) and a meta learner (support vector machines) as a result of a case study. Three different dataset were used to measure the robustness of the proposed model. Chi-square was determined as an optimal feature selection technique from five different techniques, and also hyper-parameter settings were made with GridSearch. Finally, the proposed new super learner model achieved to obtain the best accuracy results in the detection of Diabetes mellitus compared to the base-learners for the early-stage diabetes risk prediction (99.6%), PIMA (92%), and diabetes 130-US hospitals (98%) dataset, respectively. This study revealed that super learner algorithms can be effectively used in the detection of diabetes mellitus. Also, obtaining of the high and convincing statistical scores shows the robustness of the proposed super learner model.

Keywords Super learner · Ensemble · Machine learning · Chi-square · Diabetes

1 Introduction

Diabetes is a serious health illness (metabolic) resulting in higher than normal levels of sugar in the human body. Diabetes disease directly caused 6.7 million deaths in 2021 according to World Health Organisation (WHO). Additionally, the amount of time spent on diabetes in health institutions has increased by around 316% in the last 15 years [1]. Although it is not a contagious disease, it is appeared as an epidemic disease due to the increase in the prevalence of diabetes [2]. Diabetes occurs when the insulin hormone-producing by the pancreas is insufficient or not generated in the body. The mission of the insulin hormone is to carry glucose in the blood to cells,

tissues, and organs. However, if the glucose level in the blood rises abnormally, it leads to hyperglycemia when the insulin hormone cannot accomplish this mission [3].

The main short-term symptoms of hyperglycemia are sudden weight loss, excessive thirst (polydipsia), frequent urination (polyuria), weakness, overeating (polyphagia), blurred vision, genital thrush, and itching [4]. In the long-term, it causes serious diseases such as kidney failure (nephropathy), peripheral neuropathy, cardiovascular, gastrointestinal, foot ulcers, and loss of vision (retinopathy) [5]. Diabetes has four clinical types, including type one, type two, gestational diabetes, and prediabetes. Type one is usually seen in people under 30 years old. This is a type of autoimmune illness generated by the demolition of beta cells, which are involved in producing insulin in the body. Type two diabetes is the most common diabetes in the world. It happens when the insulin hormone produced by the pancreas is not sufficient or the body creates resistance against insulin. Gestational diabetes occurs during pregnancy and its symptoms disappear when the pregnancy ends. However, monitoring the mother and the fetus during pregnancy is extremely important for their health. There is a

✉ Selim Buyrukođlu
sbuyrukoglu@karatekin.edu.tr;
selimbuyrukoglu@hotmail.com

¹ Department of Electrical and Electronics Engineering,
Çankırı Karatekin University, 18100 Çankırı, Turkey

² Department of Computer Engineering, Çankırı Karatekin
University, 18100 Çankırı, Turkey

high risk of type two diabetes disease in the future. In prediabetes diabetes, the level of blood glucose stands above regular.

The classical methods may not always be sufficient in the diagnosis of diabetes. The popularity and effectiveness of artificial intelligence techniques play an essential role in health science along with technological improvement. Especially, machine learning algorithms have become more preferred because they enable to obtain robust, reliable, and convincing results in a short time compared to the classical methods.

Machine learning has occurred as an encouraging tool to support decision-making processes in different fields, including health, safety, sports, environment, and biomedicine [6–8]. In literature, there are a variety of studies have been carried out on the detection of diabetes mellitus. A study [9] was proposed to predict diabetes by employing decision tree (DT), random forest (RF), and artificial neural networks (ANNs). The PIDD and Luzhou diabetes dataset were used to get information about the employed machine learning algorithms. Principal component analysis (PCA), maximum suitability (mRMR), and minimum redundancy were applied as feature selection techniques. The structure of ANN consists of one input, output, and ten hidden layers. The machine learning algorithms provided the best classification scores when using the features obtained through the mRMR feature selection technique. The highest accuracy score was achieved with the machine learning algorithm DT at 76.13% in the utilization of the Luzhou dataset. On the other hand, RF reached to provide the highest precision score (77.21%) in the use of the PIDD dataset. In a different study, Shamreen Ahamed and Sumeet Arya [10] proposed a study to predict diabetes using the PIDD diabetes dataset. Seven machine learning algorithms, including rf, dt, logistic regression (lr), gradient Boosting (GB), XGBoost (XGB), extra decision trees (ET), and light gradient boosting machine (LGBM) were employed in the estimated process. The results show that the most accurate estimation performance was achieved with the LGBM algorithm with a value of 95.2%. Furthermore, Emon et al. [11] used a dataset which is created according to the survey results obtained from Sylhet Diabetes Hospital for diabetes disease risk estimation. A different study was conducted with 11 different machine learning algorithms. Accuracy, precision, F-1 score, AUC, and recall statistical scores were measured. As a result of this study, the best accuracy performance was obtained with RF (98% accuracy). The accuracy results of other employed machine learning algorithms used in the study are as follows: support vector machines (SVM) (96%), AdaBoost (AB) (94%), DT (93%), LR (92%), MLP (92%), KNN (91%), Gaussian process (88%), Naive Bayes (NB) (84%), boosting (92%). Also, in a different study, a new hybrid feature selection approach is combined with a machine learning approach to classify diabetes mellitus [12]. In the study, correlation heatmap and sequential forward selection (SFS) feature selection approaches were employed to obtain the optimal feature subset sequentially. Then, support vector machines (SVM), random

forest (RF), and artificial neural network (ANN) algorithms are created based on the features obtained as a result of the hybrid feature selection approach. In the end, the best accuracy score is obtained by employing a hybrid feature selection approach with ANN (99.1%).

Ensemble learning algorithms were also employed in a different study for the detection of Diabetes Mellitus [13]. AB, XGB, LGBM, Soft Voting, and RF ensemble learning models were used to predict diabetes disease. The PIDD dataset is split into 70:30 training and test sets. This study purposes to measure the efficiency of the ensemble learning algorithms with and without data preprocessing using tenfold cross-validation. According to the results of this study, the best accuracy score (94%) was obtained through the LGBM machine learning algorithm after data preprocessing. A different study was proposed by Yadav and Pal [14] used on the detection of diabetes mellitus by applying rule-based machine learning algorithms (DT, Jrip, OneR). A dataset from the UCI machine learning repository was utilized in this study. The OneR algorithm creates a rule for each class and tries to choose the best rule with fewer errors. The Jrip algorithm calculates the size of the class and aims to reduce the errors in the dataset. Missing values were deleted from the dataset in the data pre-processing. Nine features have been chosen based on the chi-square feature selection technique. These features are serum insulin, plasma glucose, triceps skin fold thickness, class, age, body mass index, times pregnant, diastolic blood pressure, and diabetes pedigree function. In this study, the incremental ensemble learning model (DT, Jrip, OneR) and the bagging ensemble learning model (DT, Jrip, OneR) were applied in the detection of diabetes. According to the results of the experiment, the best accuracy score (98%) was obtained by employing the bagging ensemble learning model.

Moreover, a study Saxena et al. [15] focuses on the detection of diabetes using the super learner learning model. In this study, the dataset was divided into 75:25 training and test sets. tenfold cross-validation was applied. The mean reduction method of the Extra Trees classifier was utilized for feature selection. As a result of this technique, 12 features (polyphagia, age, itching, visual blurring, delayed healing, alopecia, irritability, partial paresis, sudden weight loss, gender, polyuria, and polydipsia) were chosen in the early-stage diabetes risk prediction dataset, and 4 features were chosen (glucose, age, BMI, diabetes pad function) in the PIMA dataset. Ten base learners were chosen at Level-0 (LR, LDA, KNN, SVM, DT, ADA, GB, NB, RF, ET) of the super learner model. GB was adopted as a meta learner in level 1. The performance results obtained as a result of the experiments are 97% for the early-stage diabetes risk prediction dataset and 86% for the PIMA dataset, subsequently.

To the best of the author's knowledge, super or stacked ensemble models have not been commonly used in classifying diabetes mellitus. Thus, this study aims to determine an

appropriate super learner model for the detection of diabetes. To the best of the author's knowledge, super-ensemble models have not been commonly used in classifying diabetes mellitus. Thus, this study aims to determine an appropriate super-learner model for the detection of diabetes. A case study was carried out in the creation of the proposed super learner model. In this sense, four different super ensemble models were created using a variety of different machine learning algorithms. Single-based machine learning algorithms were used in the level 0 of the first super learner model. Then, bagging ensemble learning algorithms were employed in the level 0 of the second super learner model while boosting ensemble learning algorithms were used in the level 0 of the third super learner model. Finally, the fourth super learner model was created with the combination of single-base, bagging, and boosting algorithms. Additionally, a different case study was also carried out to determine the optimal meta-learner (level 1) algorithm. kNN, ANN, AdaBoost, and SVM algorithms were used in the determination of the most beneficial meta-learner algorithm for the proposed super learner. In the end, the best performance was obtained through the fourth super ensemble model (level 0: LR, KNN, NB, GB; level 1: SVM). We found out the novelty of this study by performing both level 0 and level 1 case studies. Furthermore, this work aims to address the following investigation questions based on the purpose of this study:

1. Which ensemble learning algorithm enables obtaining a more efficient statistical score in the prediction of diabetes?
2. Can the super learner algorithm provide a better statistical score in the detection of diabetes than the other machine learning algorithms (single-based, bagging, and boosting)?
3. Can the super learner algorithm provide robust statistical scores for different datasets related to diabetes?

The essay has been organized in the following way. Firstly an overview of diabetes followed by details of literature. Section 2 provides information about the proposed super learner algorithm. Section 3 presents the findings of the research, and ultimately, in Sect. 4, the conclusion and future work are mentioned.

2 Proposed super learner approach

In this section, the proposed approach used in this study is explained in detail. Figure 1 illustrates the structure of this approach which consists of the following processes: dataset, data preparation, split-validation dataset, super learner model, and assessment metrics. Also Google Colaboratory (Google Colab) Framework, a cloud service based on Jupiter

Notebooks, was preferred as the development environment [16]. It allows writing and executing code for free, recording and sharing forecasts and analysis. In order to use the super learner algorithm efficiently and effectively, the ML-Ensemble (mlens) Python library is used.

2.1 Dataset

This study was conducted with the early-stage diabetes risk prediction dataset. It was prepared by the doctor's approval of the results of the direct questionnaires made to diabetes patients at Sylhet Diabetes Hospital in Sylhet, Bangladesh. This dataset is publicly available from the UCI Machine Learning Repository [17]. Detailed information on the features in the dataset is shown in Table 1. The dataset consists of 16 attributes and one class of information in line with the information obtained from a total of 520 patients. This dataset consists of 320 with diabetes and 200 without diabetes. There are 328 male and 192 female patient records in the dataset. The rate of male patients with diabetes is 45% and the rate of female patients is 90%.

Two more datasets are also used in this study namely PIMA and diabetes 2. The reason to use the datasets is to show the robustness of the proposed approach. Detailed information about the datasets is given in Sect. 3.2.

2.2 Feature selection

Various feature selection techniques were used to obtain the most effective features including gain ratio, variance reduction, information gain, gini index, and chi-squared. Different features were selected based on these feature selection techniques. The highest accuracy score is achieved using the nine selected features based on chi-square. These features are polyuria, polydipsia, sudden weight loss, partial paresis, gender, irritability, polyphagia, alopecia, and age. Due to the best accuracy score is obtained with the features obtained through the chi-square, the rest of this section gives the working principle of it. In statistics, the chi-square test is used to test two independent events. It is a preferred method when the target variable is categorical in classification problems in feature selection methods [18]. The chi-square formula is given in Eq. (1).

$$\chi^2 = \sum \frac{(O - E)^2}{E} \quad (1)$$

where X refers to, O stands for observed value. and E stands for expected value. To use the chi-square test in Python, the "SelectKBest" and "chi2" methods in the sklearn library have been implemented. Chi-square is a non-parametric test to measure the degree of association used when the variables are independent and categorical. For each attribute, a hypothesis test is performed and the hypotheses are set up as follows:

H_0 (Null hypothesis) : (polyuria, polydipsia, sudden weight loss, partial paresis, gender, irritability, polyphagia, alopecia, and age) has no role to play in diagnosing diabetes.

H_1 (alternate hypothesis) : (polyuria, polydipsia, sudden weight loss, partial paresis, gender, irritability, polyphagia, alopecia, and age) has role to play in diagnosing diabetes.

For each attribute a contingency table is comprised. In Table 2, shows an example contingency table showing the relationship of gender with diabetes in the early-stage diabetes risk estimation dataset.

$$\text{Expected value} = \frac{\text{Row Total} * \text{Column Total}}{\text{Grand Total}} = \frac{(A+B)(B+D)}{(A+B+C+D)} \quad (2)$$

Expected values are obtained by applying the Expected (2) formula. Then, using the Expected and Observed values, the chi-squared formula is applied. The calculated version of the chi-square formula is presented in Table 3. After

calculating the chi-square score, the selected attributes are those with the maximum chi-square score.

2.3 Methods of splitting and validation of the dataset

In this study, the nested cross-validation method was used to improve the accuracy of the model. Initially, the dataset was divided into the 70:30 training and testing datasets based on the hold-out technique. Then, the k-fold cross-validation technique was applied to the training set. Generally, the k value is chosen between 3 and 10 [19]. Assigning a larger value may increase the cost when calculating. In this study, tenfold cross-validation was performed on the training dataset.

2.4 Super learner model

As seen in Fig. 2, the super learner model consists of single-based and meta learner algorithms. This section initially

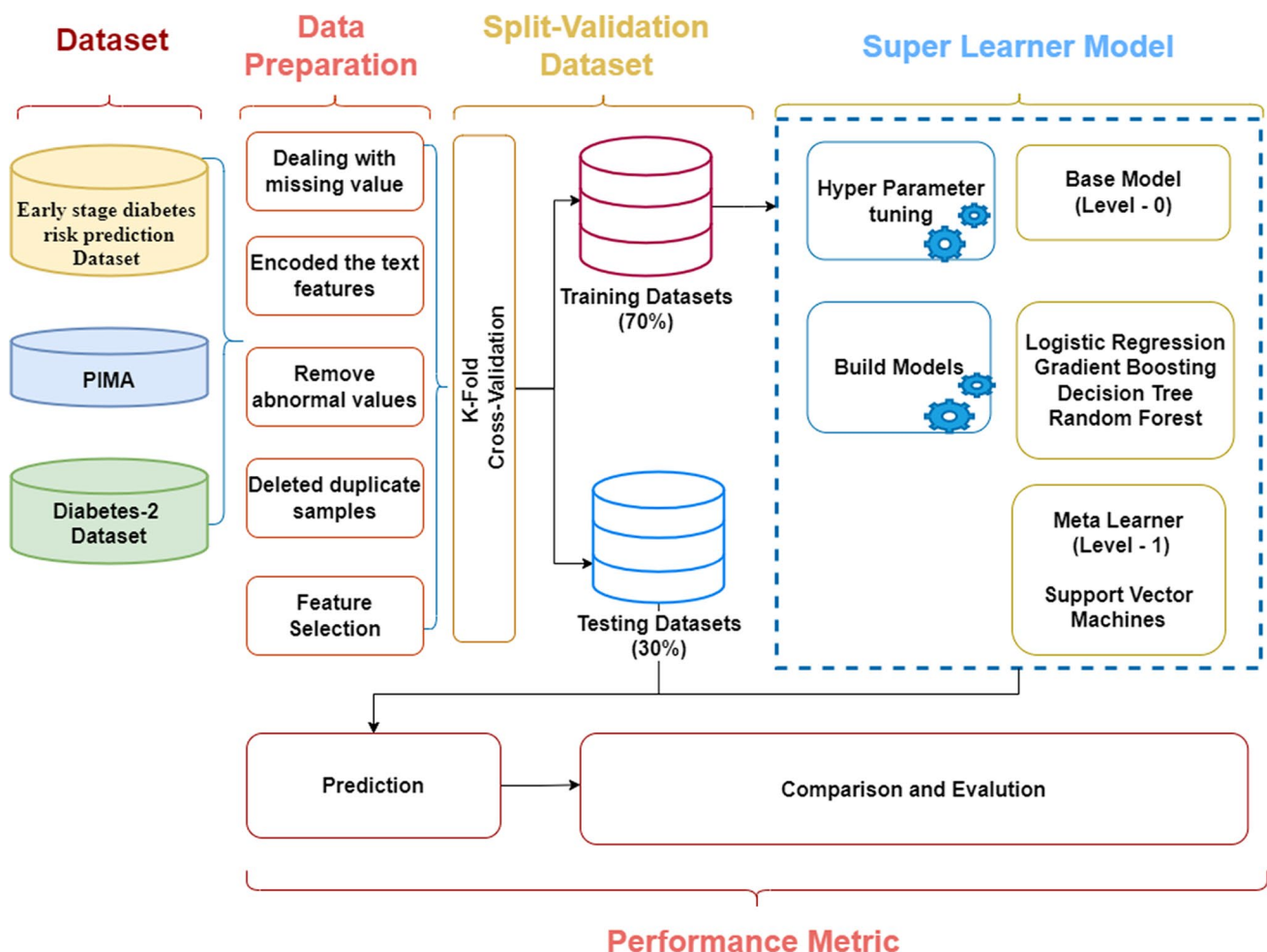


Fig. 1 The architecture of the diabetes estimation approach

gives information about the working principle of the super learner. Super learner has two levels including base learner (level 0) and meta learner (level 1). Then, information is given about the employed algorithms in level 0 and level 1.

2.4.1 Super learner model

The super learner algorithm is a kind of ensemble learning algorithm. Super learner is a cross-validation-based approach to make better predictions by combining prediction results of more than one machine learning algorithm [20]. Combining more than one machine learning algorithm makes the super learner more preferable than the single-based algorithms [11, 14, 21, 22]. The super learner ensemble learning model consists of level 0 and level 1. Level 0 consists of single-based machine learning algorithms. These machine learning algorithms are trained using the training dataset. Later meta-learner is trained using cross-validation (CV) on the outputs of the base classifiers. In estimating cases, CV is considered fundamental to the super learner for exploring the performance of the trained algorithm [23]. In the training set, each candidate is trained, and then the outcome is obtained in validation set for each observation. The level 1 (meta learner) combines the prediction results from level 0. Figure 2 shows the data flow diagram adapted from [17] to the our recommended super learner model.

The working principle of the super learner model is given below:

- To train single base models, the original dataset is split into V blocks.
- Each single base model is trained with one of these blocks.
- Parameter settings are made for each model to achieve the best result.
- Base models are predicted with k-fold cross-validation. All estimates out of k-fold are stored. Each base model has fit the total training dataset therefore, they can use the model later to make estimation in new instances not seen throughout training.
- Finally, the predictions of the base models are given as input parameters to the super learner model. The meta-learner is trained with out-of-fold predictions of other models. Thus, it is trained on out-of-fold predictions from each model and learns to combine the best predictions from multiple models. In fact, it does both jobs simultaneously.

2.4.2 Base learner machine learning algorithms

Working principles of the used single-based algorithms in level 0 are presented in this Section. Four different algorithms were used in level 0 which are decision tree (DT), logistic regression (LR), random forest (RF), and gradient boosting (GB). In the determination process of the appropriate algorithms in level 0, a combination of diverse machine learning algorithms (KNN, NB, AB, ET) was tried to obtain the best statistical scores. In the end, these four machine learning algorithms (DT, LR, RF, GB) enabled us to provide the best accuracy score compared to the others. Rest of this section gives information about the selected single-based algorithms.

Logistic regression is a predictive statistical analysis algorithm used in datasets where the target variable is of categorical type. It uses the sigmoid function to cost function. A threshold value is defined to be able to classify the dataset. In this sense, if a classification value is below the threshold value, it is labeled as one class. Contrary to this, it is labeled as another class [24].

Gradient boosting is an ensemble learning algorithm used to solve both classification and regression problems. Stumps are added to the model sequentially. The gradient descent function is applied to reduce the error that occurs during this process. The gradient boosting algorithm aims to obtain strong learners from weak learners [25].

Table 1 Early-stage diabetes risk prediction dataset explanation

Number	Feature name	Explanation
1	Age	1. 20–35, 2. 36–45, 3. 46–55, 4. 56–65, 5. Above 65
2	Gender	1. Male, 2. Female
3	Polyuria	1. Yes, 2. No
4	Polydipsia	1. Yes, 2. No
5	Sudden weight loss	1. Yes, 2. No
6	Weakness	1. Yes, 2. No
7	Polyphagia	1. Yes, 2. No
8	Genital thrush	1. Yes, 2. No
9	Visual blurring	1. Yes, 2. No
10	Itching	1. Yes, 2. No
11	Irritability	1. Yes, 2. No
12	Delayed healing	1. Yes, 2. No
13	Partial paresis	1. Yes, 2. No
14	Muscle stiffness	1. Yes, 2. No
15	Alopecia	1. Yes, 2. No
16	Obesity	1. Yes, 2. No
17	Class (target)	1. Positive, 2. Negative

Table 2 Contingency table showing the relationship of gender with diabetes in the early-stage diabetes risk estimation dataset

Gender	Non diabetes	Diabetes	Total
Female	A = 19	B = 173	192
Male	C = 181	D = 147	328
Total	200	320	520

A decision tree is one of the bagging ensemble learning algorithms. It consists of a tree-like structure that starts with a root node, splits into sub-nodes/branches according to possible outcomes, and ends with a leaf node [26, 27].

Random forest is a bagging ensemble learning algorithm created by combining multiple decision trees [28]. Each decision tree is trained independently and produces a prediction. If it is a classification problem, the highest vote, and if it is a regression problem, the average value is taken and the class labeling is done [29].

2.4.3 Meta learner machine learning algorithm

Support vector machines (SVM) is a supervised machine learning algorithm often used in classification problems. A decision

Table 3 The grand total of the chi-square values of gender in the early-stage diabetes risk estimation dataset

Gender	Observed	Expected	O-E	χ^2
Female diabetes	173	118.15	54.84	25.45
Female not diabetes	19	73.84	54.84	40.73
Male diabetes	147	201.84	54.84	14.90
Male not diabetes	181	126.15	54.84	23.84
Total	520	520		104.94

boundary line is drawn to be able to group the data in the dataset correctly in n-dimensional space. This decision boundary line is called the hyperplane. The purpose of SVM is to determine the maximum distance between remaining members of two classes. Kernel and c coefficients from SVM parameters are one of them. Kernels are transformed using the kernel function to linearly separate data that is not linearly separated [30].

2.4.4 Hyperparameter tuning

There are some techniques used to adjust the parameter settings most appropriately during the model setup of machine learning algorithms. For the hyperparameter settings in this dataset, the most appropriate parameter selection was made by using GridSearch method. In Table 4, hyperparameter settings of the employed machine learning algorithms in the early-stage diabetes risk prediction dataset are given. It is noted that these algorithms were used in the creation of the proposed super learner approach. SVM was used as a meta learner while the rest of the algorithms were used as a base learner.

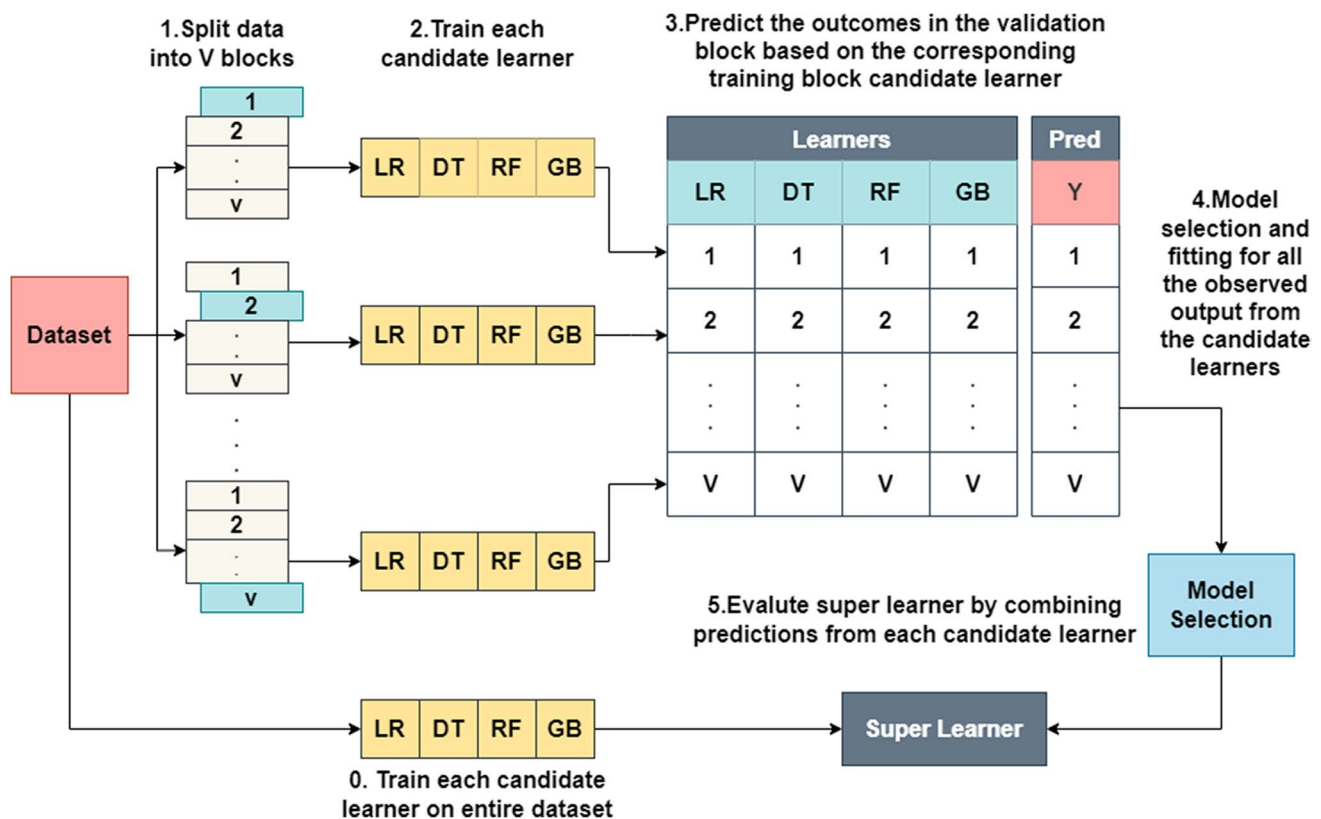


Fig. 2 Super learner data flow [20]

Table 4 Hyperparameter tuning of early-stage diabetes risk prediction dataset

Model	Parameter	Range	Best
LR	C	[0.001, 0.01, 0.1, 1, 10, 100, 1000]	1
	Penalty	L1, l2	L2
RF	n_estimators	10, 20, 50, 100, 200, 300, 400, 500, 1000	100
	criterion	'gini', 'entropy'	Entropy
GB	learning_rate	0.001, 0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0	0.2
	criterion	'friedman_mse', 'mse', 'mae'	friedman_mse
	min_samples_leaf	[1, 6]	4
DT	n_estimators	10, 20, 50, 100, 200, 300, 400, 500, 1000	400
	criterion	'gini', 'entropy'	Gini
	max_depth	[1, 10]	9
	min_samples_split	[1, 10]	2
	min_samples_leaf	[1, 5]	1
SVM	Kernel	'linear', 'poly'	Linear
	C	[0.001, 0.01, 0.1, 1, 10, 100, 1000]	100

2.5 Evaluation metrics

In this study, F1-Score, recall, precision, accuracy, and ROC curve evaluation metrics were used to compare ensemble learning models and single-based algorithms. There are 4 parameters we need to know before moving on to the calculation values. The formulas for the evaluation metrics are given in Eqs. (3–8).

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{FP} + \text{TN} + \text{FN})} \quad (3)$$

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \quad (4)$$

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \quad (5)$$

$$\text{Specificity} = \frac{\text{TN}}{(\text{TN} + \text{FP})} \quad (6)$$

$$\text{F1-Score} = \frac{(2 * \text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (7)$$

$$\text{ROC} = \frac{\text{TPR}}{\text{FPR}}, \text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \quad (8)$$

where TP (true positive) is the number of patients with true positive diabetes in the model; TN (true negative) is the number of patients with true negative diabetes in the model; FP (false positive) is the model number of patients with false-positive diabetes and FN (false negative) refers to the number of patients with false-negative diabetes of the

model. In addition, AUC represents the area under the ROC Curve and takes a value between zero and 1 [31]. AUC score is also used to support the investigation questions (please see Table 5, 6, 7, 8, 9 and Table 10).

3 Results and discussion

3.1 Evaluation scores of the proposed super learner for early-stage diabetes risk prediction dataset

The machine learning algorithm accuracy performances are presented in Fig. 3 and the ROC curves results are presented in Fig. 4. The proposed super learner (SL) algorithm provided the best accuracy (99.6%) and ROC curve performances compared to other machine learning algorithms. In contrast to this, LR and SVM algorithms have the worst accuracy scores (93%) and ROC curve scores.

Bagging ensemble learning algorithms (RF, DT) achieved to provide the second-best accuracy

Table 5 Machine learning algorithms for F1-score, precision, recall, and AUC performance evaluation reports

Model	F1-score	Precision	Recall	AUC
LR	0.95	0.94	0.95	0.96
DT	0.93	0.98	0.89	0.98
RF	0.94	0.99	0.90	0.98
GB	0.98	0.99	0.97	0.97
SVM	0.95	0.94	0.95	0.97
SL	0.99	0.99	0.99	0.99

Table 6 Studies with early-stage diabetes risk prediction dataset

Author	Year	Model	Machine learning algorithms	Best model	Accuracy
Sadhu and Jadli [32]	2021	Single base	LR, NB, RF, MLP, DT, KNN, SVM	RF	0.987
Alpan and Ilgi [33]	2020	Single base	LR, NB, RF, RT, DT, KNN, SVM	KNN	0.98
Xue et al. [34]	2020	Single base	SVM, NB, LightGBM	SVM	0.98
Ilyas Özer [35]	2020	Single base	LSTM	LSTM	0.986

Table 7 Selected features from PIMA and diabetes 2 datasets employing the chi-square feature selection technique

Data Set	Selected features
PIMA	Insulin, glucose, age, BMI, pregnancies
Diabetes 2	race, gender, age, admission_type_id, discharge_disposition_id, admission_source_id, time_in_hospital, num_lab_procedures, num_procedures, num_medications, number_inpatient, number_diagnoses, max_glu_serum, A1Cresult, metformin, repaglinide, nateglinide, glimepiride, glipizide, glyburide, pioglitazone, rosiglitazone, acarbose, insulin, glyburide-metformin, change, diabetesMed

Table 8 Hyper-parameter tuning of different machine learning algorithms for PIMA and diabetes 2 dataset

Model	Parameter	Range	Best	
			PIMA	Diabetes 2
LR	C	[0.001, 0.01, 0.1, 1, 10, 100, 1000]	100	0.1
	Penalty	L1, l2	L2	L2
RF	n_estimators	10, 20, 50, 100, 200, 300, 400, 500, 1000	100	150
	criterion	'gini','entropy'	Entropy	Entropy
GB	learning_rate	0.001,0.01,0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0	0.2	0.8
	criterion	'friedman_mse','mse','mae'	friedman_mse	friedman_mse
	min_samples_leaf	[1,6]	4	4
	n_estimators	10, 20, 50, 100, 200, 300, 400, 500, 1000	400	1000
DT	criterion	'gini','entropy'	Gini	Gini
	max_depth	[1,10]	3	9
	min_samples_split	[1,10]	2	2
	min_samples_leaf	[1,5]	1	4
SVM	Kernel	'linear','poly'	Linear	Linear
	C	[0.001, 0.01, 0.1, 1, 10, 100, 1000]	100	100

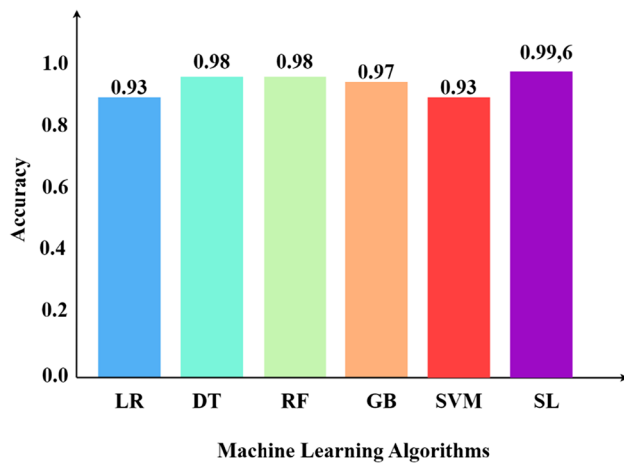
Table 9 Accuracy performance comparison using two different dataset

Data Set	LR	RF	DT	GBM	SVM	SL
PIMA	0.82	0.83	0.83	0.86	0.82	0.92
Diabetes 2	0.58	0.93	0.62	0.91	0.93	0.98

Table 10 F1-Score, precision, and recall performance comparison using two different dataset

Dataset	Model	F1-Score	Precision	Recall	Auc
PIMA	LR	0.70	0.84	0.59	0.785
	DT	0.76	0.75	0.78	0.795
	RF	0.72	0.85	0.63	0.832
	GB	0.79	0.81	0.78	0.82
	SVM	0.68	0.81	0.59	0.775
	SL	0.91	0.90	0.92	0.921
Diabetes 2	LR	0.56	0.59	0.54	0.58
	DT	0.87	0.83	0.93	0.86
	RF	0.93	0.90	0.97	0.931
	GB	0.92	0.88	0.96	0.678
	SVM	0.56	0.59	0.53	0.58
	SL	0.98	0.98	0.98	0.978

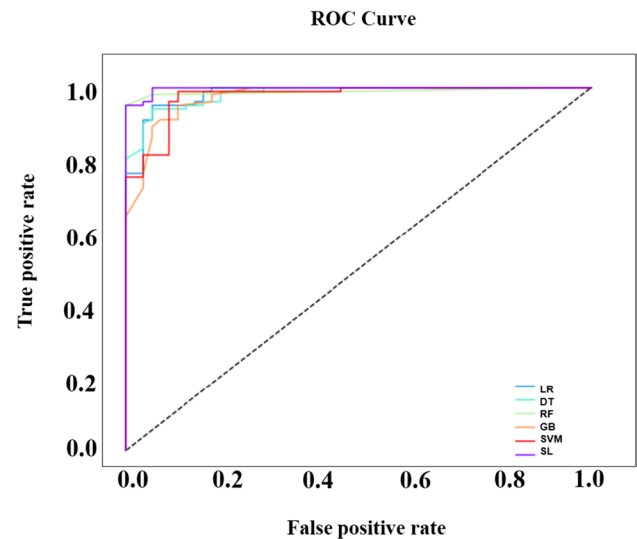
performance with 98% accuracy. Then, the gradient

**Fig. 3** Accuracy performances of the employed machine learning algorithms

boosting ensemble learning algorithm provided 97% accuracy. It reveals that bagging ensemble learning algorithms can be considered better than boosting ensemble (GB) algorithms in the use of the early-stage diabetes risk prediction dataset.

The F1-score, precision, and recall performances of the employed machine learning algorithms are presented in Table 5. Super learner (SL) algorithm has better statistical performance compared to the other machine learning algorithms.

They have been given the studies conducted in the literature using the early-stage diabetes risk prediction dataset with single base machine learning algorithms in Table 6. The best accuracy score is obtained by employing the RF model [32]. The lowest accuracy score is provided using kNN by Alpan and Ilgi [33] and SVM by Xue

**Fig. 4** ROC Curve Performances of the employed machine learning algorithms

et al. [34]. It can be inferred from Table 6 that the bagging ensemble learning algorithm (RF) provides better performances than the other models presented in Table 6.

3.2 The efficiency of the proposed super learner approach based on different dataset

The performance of the proposed super learner algorithm is evaluated using two different datasets, namely PIMA and diabetes 130-US hospitals.

- **Pima Indians Diabetes dataset (PIMA):** Pima Indians Diabetes dataset (PIMA) is obtained from the UCI Machine Learning Repository. As with considerably medical data, there are missing values and distribution disparity. In addition, the dataset involves 768 patient records, 200 of whom are diabetic and 500 are not. It further includes 8 features (pregnancy, glucose, blood pressure, skin thickness, skin thickness, BMI, diabetes pedigree function, and age) and a target variable (outcome).
- **Diabetes 130-US hospitals dataset:** Diabetes 130-US hospitals for the years 1999–2008 dataset was received from the UCI Machine Learning Repository. It involves 100,000 patient records collected from 130 hospitals in the USA between 1999 and 2008. There are 70,000 diabetic patients in this dataset. Diabetes in 130-US hospitals for the years 1999–2008 dataset is named diabetes 2 in this study. It consists of 50 features. On the other hand, the chi-square feature selection technique is used in the feature selection process to obtain the most effective features from these two datasets. The selected features are given in Table 7.

Table 11 Studies with PIMA and diabetes 2 datasets

Author	Year	Dataset	Model	Machine learning algorithms	Best model	Accuracy
Kumari et al. [36]	2021	PIMA	Voting	S0: LR, NB,RF S1:Soft Voting (LR, NB,RF)	SL	0.79
Akula et al. [37]	2019	PIMA	Weighted average	S0: KNN, SVM, DT, RF, GB, MLP, NB S1: SVM, RF, and GB	SL	0.89
Lai et al. [38]	2019	PIMA	Bagging	RF, GB	RF	0.85
Birjais et al. [39]	2019	PIMA	Boosting	LR, NB, GB	GB	0.86
Zou et al. [9]	2018	PIMA	Single base	RF, DT, ANN	RF	0.77
Hammoudeh et al. [40]	2018	Diabetes 2	Single base	CNN	CNN	0,95
Saxena et al. [15]	2021	D1:Early-stage diabetes risk prediction D2:PIMA	Super learner	S0: LR, LDA, KNN, SVM, DT, ADA, GB, NB, RF, ET S1: GB	SL	D1: 0.97 D2:0.86
Kabir and Ludwig [21]	2019	PIMA	Super learner	S0:GB-RF-DNN, S1:GLM	SL	0.88
Our model	2022	D1: early-stage diabetes risk prediction D2:PIMA D3: diabetes 2	Super learner	S0: LR, DT, RF, GB S1:SVM	SL	D1:0.996 D2:0.92 D3:0.98

To measure the efficiency and robustness of the proposed super learner approach, the employed algorithms for the early-stage diabetes risk prediction dataset were also employed for these two datasets (please see Table 9 for the accuracy results). For the hyper-parameter settings in the datasets, the best suitable parameter selection was made by employing GridSearch and RandomizedSearch techniques. Table 8 illustrates the hyper-parameter settings of the employed machine learning algorithms in the PIMA and diabetes 2 datasets values led to the optimal performance.

3.2.1 Evaluation results of the super learner approach for PIMA and diabetes 2

Existing studies in the literature generally use single base machine learning algorithms to achieve convincing performances. However, the determination of an optimal machine learning model is not an easy task. Also, obtaining of convincing and similar results using different datasets shows the robustness of the model. Consequently, the proposed super learner ensemble model is applied to PIMA and diabetes 2 datasets to show the robustness of the model. Data preprocessing steps were applied to both datasets. The chi-square test was applied as the feature selection technique. Their datasets are separated into test (30%) and training sets (70%) using the hold-out approach.

Tables 9 and 10 provide the evaluation scores of the super learner approach and the used machine learning algorithms in the creation of the super learner approach. As seen in Table 9, the super learner approach achieved

to provide convincing accuracy scores for both PIMA (0.92) and diabetes 2 (0.98).

In Table 10, it is shown that the super learner approach has the better F-measure, recall, and precision scores compared to the other employed algorithms. The super learner approach has also better statistical scores in the use of diabetes 2 than the use of PIMA.

3.2.2 Comparison of previous ensemble studies with current super learner approach

This section compares the accuracy performance of the studies on diabetes risk assessment with the proposed super learner approach. As illustrated in Table 11, three datasets (PIMA, diabetes 2, and early-stage diabetes risk prediction) were preferred by the authors.

As shown in Table 11, different machine learning models have been employed to classify Diabetes patients using the PIMA dataset. These models are voting stacked learner, weighted average stacked learner, bagging, boosting, super learner, and single base models. The best accuracy (89%) performance is achieved by applying the super ensemble learner model [33]. A different super learner ensemble learning model also achieved to provide an 88% accuracy score [20]. GLM algorithm was used as meta learner in the study proposed by [20] while the SVM, RF, and GB were separately used as meta-models in the study proposed by [33]. In this sense, SVM, RF, and GB algorithms can be considered more effective algorithms than GLM when they are used as a meta-model in the super learner. Kabir and Ludwig [21] in the super learner

Table 12 Characteristics of the datasets

Data set	Number of features	Sample size	Output rate	
			Rate (%)	Values (number of diabetes /total number of patients)
Early-stage diabetes risk	9	520	61	320/520
PIMA	5	768	35	268/768
Diabetes 2	27	101,766	69	70,230/101766

models suggested that three machine learning algorithms were employed as the base learner at Level 0. However, this model could be created by employing more machine learning algorithms. As the meta learner, only the GLM machine learning algorithm was applied, but alternatively machine learning algorithms could be used to obtain an appropriate meta-model. Also, in that study, no information was given about whether the feature selection method was applied to the dataset.

On the other hand, the voting stacked ensemble learning model proposed by [36] is not provided a convincing statistical score (79%) for the PIMA dataset. It means that the selection of stacked ensemble learner model type is important. It should be noted that voting, weighted average, blending, stacking, and super learner models are included as stacking ensemble models. In our study, a type of super learner ensemble learning algorithm is proposed as presented in Sect. 2. If we check the our experimental results, it can be said that the application of the feature selection technique (chi-square) improves the performance in the prediction of diabetes mellitus. the proposed super learner ensemble model achieved to increase 0.03 accuracy in the detection of diabetes mellitus compared to the proposed weighted average stacked model. [37].

In literature, only one study has focused on analyzing the diabetes 2 dataset by applying a single base algorithm (CNN) to the best of our knowledge. This study provided a convincing accuracy score (95%) [40]. If the CNN algorithm was used in any stacked ensemble type as a base learner, this model could provide better classification scores. On the other hand, the proposed super learner ensemble model provides a better score (98%) than the CNN.

A super learner ensemble model has been proposed for the diabetes risk assessment by Saxena et al. [15]. This model is applied to the early-stage diabetes risk prediction and PIMA datasets. The proposed model provided 97% (for Early-stage diabetes risk prediction) and 88% (for PIMA). Ten machine learning algorithms are used as base learners. However, the use of multiple training machine learning algorithms as base-learner significantly increases the cost and complexity of computation. Therefore, the optimum number of selected base learners may play an important role. On the other hand, in the selection of meta learners, only the

GBM machine learning algorithm was used in level 1. A variety of machine learning algorithms could be used to determine the optimal meta-learner. Thus, the model could provide better statistical performance compared to the current version of it. In the study, the extra tree was used as a feature selection technique. This feature selection technique is one of the embedded methods. To create the optimal diabetes risk assessment model, different feature selection techniques (filter and wrapper) could be used in terms of the determination of an appropriate feature selection technique. On the other hand, the proposed super learner ensemble model in this study achieved to provide 2.6% better accuracy performance in the use of the early-stage diabetes risk prediction dataset compared to the study proposed by Saxena et al. [15]. Additionally, this approach again achieves to provide better (6%) performance in the uses of the PIMA dataset.

As given in Table 11, the proposed super learner model (our model) was applied to three different datasets. According to the obtained results, the proposed super learner achieve to provide the best accuracy scores for each dataset (Early-Stage diabetes risk prediction: 99.6%, PIMA: 92%, diabetes 2: 98%) compared to the models presented in Table 11. Also, obtaining of the convincing and high accuracy scores using these three datasets through the proposed super learner algorithm shows its robustness. Robustness can be defined as a term widely used in machine learning. It concerns with the rate which varies model performances in the use of input (new samples) versus training data [41, 42]. If the model robust, the rate should not deviate significantly. Some aspects of robustness can be listed as follows: number of features, sample size, output rate. Table 12 shows some characteristics of the datasets.

As can be seen in Table 12, there are differences between the number of features and sample size in the datasets. In particular, the difference between the diabetes 2 dataset and the other datasets is large. Additionally, the differences about the output rate of PIMA dataset and the other datasets are around 35%. At the end, the employed super ensemble learning approach can be considered robust based on the information presented in Table 12.

Different feature selection techniques were performed to make better the performances of the proposed super learner. The best result was acquired through the chi-square feature selection technique. Therefore, the chi-square can be regarded as an effectual feature selection technique in the diabetes risk prediction tasks. In other words, the filter-based feature selection method is more effective than other methods (embedded and wrapper) based on the obtained results.

4 Conclusion and future directions

In this study, a new super learner ensemble model was proposed to detect Diabetes mellitus. Two different case studies were carried out to determine the optimal feature selection technique and structure of

the super learner. Initially, five different feature selection techniques were used to obtain the most effective features in the detection of mellitus in the first case study. The feature selection techniques are gain ratio, variance reduction, information gain, gini index, and chi-squared. The best statistical score was achieved using the selected nine features based on the chi-square and hyperparameter settings with the GridSearch method as a result of the first case study. Then, nested cross-validation was used to improve the statistical performances of the proposed super learner model. As highlighted in Sect. 2, the structure of a super learner model consists of two levels, namely level 0 (base-learner) and level 1 (meta-learner). Thus, the second case study was carried out to obtain the optimal structure of the super learner. In this sense, the best performance was obtained using four base-learners (LR, DT, RF, and GB) and a meta-learner (SVM) as a result of the second case study.

Three different datasets were used to measure the robustness of the proposed super learner model. The datasets are early-stage diabetes risk prediction, PIMA, and diabetes 2. The obtained accuracy performances according to test results of the super learner model are as follows; Early-stage diabetes risk prediction of 99.6%, PIMA of 92%, and diabetes 2 of 98%. Overall, we can conclude that the super learner ensemble model is better than single base, bagging, and boosting machine learning algorithms in the classification of diabetes mellitus.

In future investigations, it might be possible to use a variety of deep learning algorithms as base learners in a super learner model. Thus, the efficiency of the machine and deep learning algorithms can be compared in the detection of diabetes mellitus. Also, the proposed super learner approach can be applied in treatment and recovery techniques for health problems that will require long-term treatment such as other chronic diseases and cancer.

Acknowledgements The authors gratefully acknowledge partial support of the Faculties of Engineering at Çankırı Karatekin University.

Author contribution Publicly available data were used in this study. Conceptualization, formal analysis, methodology, and writing – original draft was performed by Ayşe Doğru, Selim Buyrukoğlu, and Murat Ari. Resources, software, supervision, writing – review, and editing were organized by Ayşe Doğru, Selim Buyrukoğlu, and Murat Ari.

Data availability The data that support the findings of this study are available on request from the corresponding author.

Declarations

Conflict of interest The authors declare no competing interests.

References

1. Federation ID (2021) “IDF Diabetes Atlas | Tenth Edition,” International Diabetes Federatio 2021. [Online]. Available: <https://diabetesatlas.org/>. [Accessed: 03-Nov-2021]
2. Sreedharan J et al (2015) Incidence of type 2 diabetes mellitus among Emirati residents in Ajman, United Arab Emirates. *Korean J Fam Med* 36(5):253–257. <https://doi.org/10.4082/KJFM.2015.36.5.253>
3. Ki R (2007) Diabetes treatment—bridging the divide. *N Engl J Med* 356(15):1499–1501. <https://doi.org/10.1056/NEJMP078030>
4. Gavin JR et al (2003) Report of the expert committee on the diagnosis and classification of diabetes mellitus. *Diabetes Care* 26(SUPPL. 1):s5–s20. <https://doi.org/10.2337/diacare.26.2007.s5>
5. Horton WB, Barrett EJ (2021) Microvascular Dysfunction in diabetes mellitus and cardiometabolic disease. *Endocr Rev* 42(1):29–55. <https://doi.org/10.1210/ENDREV/BNAA025>
6. Buyrukoğlu S, Yılmaz Y, Topalcengiz Z (2022) “Correlation value determined to increase Salmonella prediction success of deep neural network for agricultural waters,” *Environ Monit Assess* 2022 1945 vol. 194, no. 5, pp. 1–12 April 2022. <https://doi.org/10.1007/S10661-022-10050-7>
7. Savaş T, Savaş S (2021) “Tekdüzen Kaynak Bulucu Yoluyla Kimlik Avı Tespiti için Makine Öğrenmesi Algoritmalarının Özellik Tabanlı Performans Karşılaştırması,” *Politeknik Derg* pp. 1–1, Dec. 2021. <https://doi.org/10.2339/POLITEKNİK.1035286>
8. Buyrukoğlu S, Serkan S (2022) Stacked-based ensemble machine learning model for positioning footballer. *Arab J Sci Eng* 2022:1–13. <https://doi.org/10.1007/S13369-022-06857-8>
9. Zou Q, Qu K, Luo Y, Yin D, Ju Y, Tang H (2018) “Predicting Diabetes mellitus with machine learning techniques,” *Front Genet* vol. 9, Nov. 2018. <https://doi.org/10.3389/fgene.2018.00515>
10. Shamreen Ahamed B, Sumeet Arya M (2021) “Prediction of type-2 diabetes using the LGBM classifier methods and techniques,” May 2021
11. Emon MU, Keya MS, Kaiser MS, Islam MA, Tanha T, Zulfiker MS (2021) Primary stage of diabetes prediction using machine learning approaches. *Proceed - Int Conference Artificial Intel Smart Syst ICAIS 2021*:364–367. <https://doi.org/10.1109/ICAIS50930.2021.9395968>
12. Buyrukoğlu S, Akbaş A (2022) Machine Learning based early prediction of type 2 diabetes: a new hybrid feature selection approach using correlation matrix with heatmap and SFS. *Balk. J Electr Comput Eng* 10(2):110–117. <https://doi.org/10.17694/BAJECE.973129>
13. Taz NH, Islam A, Mahmud I (2021) “A comparative analysis of ensemble based machine learning techniques for diabetes identification,” 2021, pp. 1–6. <https://doi.org/10.1109/icres t51555.2021.9331036>
14. Yadav DC, Pal S (2021) An experimental study of diversity of diabetes disease features by bagging and boosting ensemble method with rule based machine learning classifier algorithms. *SN Comput Sci* 2(1):50. <https://doi.org/10.1007/s42979-020-00446-y>
15. Saxena S, Mohapatra D, Padhee S, Sahoo G K (2021) Machine learning algorithms for diabetes detection: a comparative evaluation of performance of algorithms. *Evol. Intell* 1:1–17. <https://doi.org/10.1007/S12065-021-00685-9>
16. Google Colab, “Google Colab,” Getting Started - Introduction 2020. [Online]. Available: <https://research.google.com/colaboratory/faq.html>. [Accessed: 01-Nov-2022]
17. Islam MMF, Ferdousi R, Rahman S, Bushra HY (2020) “UCI machine learning repository: early stage diabetes risk prediction dataset. Data Set,” [Online]. Available: <https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset>. [Accessed: 14-Oct-2021].
18. Urdan TC (2010) “Statistics in plain english,” *Stat Plain English* p. 211, 2010, <https://doi.org/10.4324/9781410612816>
19. Yadav S, Shukla S (2016) “Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification,” *Proc - 6th Int Adv Comput Conf IACC 2016*, pp. 78–83. <https://doi.org/10.1109/IACC.2016.25>

20. Van Der Laan MJ, Polley EC, Hubbard AE (2007) “Super learner,” *Stat Appl Genet Mol Biol* 6. 1. <https://doi.org/10.2202/1544-6115.1309>
21. Kabir MF, Ludwig SA (2019) Enhancing the performance of classification using super learning. *Data-Enabled Discov Appl* 3(1):1–13. <https://doi.org/10.1007/s41688-019-0030-0>
22. Perveen S, Shahbaz M, Guergachi A, Keshavjee K (2016) Performance analysis of data mining classification techniques to predict diabetes. *Procedia Comput Sci* 82:115–121. <https://doi.org/10.1016/j.procs.2016.04.016>
23. Phillips RC, van der Laan MJ, Lee H, Gruber S (2022) “Practical considerations for specifying a super learner,” *arXiv*, p. [arXiv: 2204.06139](https://arxiv.org/abs/2204.06139), Apr. 2022
24. Hosmer DW, Lemeshow S, Sturdivant RX (2013) “Applied logistic regression: third edition,” *Appl Logist Regres Third Ed.*, pp. 1–510, <https://doi.org/10.1002/9781118548387>
25. Mason L, Baxter J, Bartlett P, Frean M (2000) “Boosting algorithms as gradient descent,” *Adv Neural Inf Process Syst* pp. 512–518
26. Quilan JR (1988) Decision trees and multi-valued attributes I *Machine intelligence* 11
27. Rokach L, Maimon O (2006) “Decision trees,” in *Data Mining and Knowledge Discovery Handbook*, Springer-Verlag 165–192
28. Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32. <https://doi.org/10.1023/A:1010933404324>
29. Liaw A, Wiener M (2002) “Classification and regression by randomForest,”
30. Cortes C, Vapnik V (1995) Support-vector networks. *Mach Learn* 20(3):273–297. <https://doi.org/10.1007/bf00994018>
31. Fawcett T (2006) An introduction to ROC analysis. *Pattern Recognit Lett* 27(8):861–874. <https://doi.org/10.1016/j.patrec.2005.10.010>
32. Sadhu A, Jadli A (2021) “This work is licensed under a Creative Commons Attribution 4.0 International License Early-Stage Diabetes Risk Prediction: A Comparative Analysis of Classification Algorithms,” *Int Adv Res J Sci Eng Technol* 2 (8). <https://doi.org/10.17148/IARJSET.2021.8228>
33. Alpan K, İlgı GS (2020) “classification of diabetes dataset with data mining techniques by using weka approach,” *4th Int Symp Multidiscip Stud Innov Technol ISMSIT 2020 - Proc.* <https://doi.org/10.1109/ISMSIT50672.2020.9254720>.
34. Xue J, Min F, Ma F (2020) “Research on diabetes prediction method based on machine learning,” *J Phys Conf Ser Pap • OPEN ACCESS J Phys Conf Ser* (1684) 12062. <https://doi.org/10.1088/1742-6596/1684/1/012062>
35. Ozer I (2020) “Uzun Kısa Dönem Bellek Ağlarını Kullanarak Erken Aşama Diyabet Tahmini early-stage diabetes prediction using long short-term memory networks. *Müh Bil ve Araş Derg* 2(2):50–57
36. Kumari S, Kumar D, Mittal M (2021) An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. *Int J Cogn Comput Eng* 2:40–46. <https://doi.org/10.1016/j.ijcce.2021.01.001>
37. Akula R, Nguyen N, Garibay I (2019) “Supervised machine learning based ensemble model for accurate prediction of type 2 diabetes,” in *Conference Proceedings - IEEE SOUTHEASTCON*, vol. 2019-April, <https://doi.org/10.1109/SoutheastCon42311.2019.9020358>
38. Lai H, Huang H, Keshavjee K, Guergachi A, Gao X (2019) “Predictive models for diabetes mellitus using machine learning techniques,” *BMC Endocr Disord* 19 1. <https://doi.org/10.1186/s12902-019-0436-6>
39. Birjais R, Mourya AK, Chauhan R, Kaur H (2019) Prediction and diagnosis of future diabetes risk: a machine learning approach. *SN Appl Sci* 1(9):1–8. <https://doi.org/10.1007/s42452-019-1117-9>
40. Hammoudeh A, Al-Naymat G, Ghannam I, Obied N (2018) Predicting hospital readmission among diabetics using deep learning. *Procedia Comput Sci* 141:484–489. <https://doi.org/10.1016/j.procs.2018.10.138>
41. Steinhart J (2018) “Robust learning: information theory and algorithms A Dissertation Submitted To The Department Of Computer Science And The Committee On Graduate Studies Of Stanford University In Partial Fulfillment Of The Requirements For The Degree Of Doctor Of Philosophy,”
42. Li JZ (2018) “Principled approaches to robust machine learning and beyond,”

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.



Ayşe Doğru received B.Sc. degree from Trakya University, Turkey, in 2010 and M.Sc. degree from Çankırı Karatekin University, Turkey, in 2022. Her research interests include Data Science, Machine Learning, and Deep Learning.



Selim Buyrukoğlu received Ph.D. degree from Loughborough University, UK in 2018. He became an Asst. Prof. at Çankırı Karatekin University. His research directions include machine learning and deep learning methods.



Murat An received Ph.D. degree from Gazi University, Turkey, in 2004. He is currently working as Professor at Çankırı Karatekin University. His areas of interest are Embedded Systems, Artificial Intelligence, and Deep Learning.