




RESEARCH ARTICLE

EARLY-STAGE DIABETES RISK PREDICTION USING MACHINE LEARNING
TECHNIQUES BASED ON ENSEMBLE APPROACH

Tuğba PALABAŞ^{1,*}

¹ Biomedical Engineering, Faculty of Engineering, Zonguldak Bulent Ecevit University, Zonguldak, Turkey

tugba.palabas@gmail.com -  [0000-0002-6985-6494](https://orcid.org/0000-0002-6985-6494)

Abstract

Diabetes Mellitus which is considered as one of the deadliest is a common, chronic disease. It also causes the emergence of many diseases, especially neuropathy, nephropathy, and retinopathy. In this context, initiating a rapid treatment process is very important by accurately evaluating the symptoms and making an early diagnosis of the disease. This study aims to provide an effective model that can determine the risk of diabetes at an early stage with the best accuracy. For this purpose, classification algorithms frequently used in diabetes risk prediction are supported by ensemble approaches. Firstly, the performance of Naive Bayes (NB), Trees-J48, k Nearest Neighbor (kNN), and Sequential Minimal Optimization (SMO) classifiers are analyzed separately by using a dataset of 520 samples collected with direct questionnaires from Sylhet Diabetes Hospital patients in Sylhet, Bangladesh. Then, the effects of Adaboost, Bagging and Random Sub-Space (RSS) algorithms on classifier success are investigated and it is shown that the J48 classifier based on Adaboost approach has the best accuracy. Finally, the Wrapper Subset Eval (WSE) feature extraction algorithm is applied to reduce the estimation cost and increase classification success. Thus, the best accuracy (99%) is achieved using reduced data set with proposed classifier method.

Keywords

Diabetes,
Classification,
Ensemble Approach,
Feature Extraction

Time Scale of Article

Received :29 June 2023
Accepted : 17 July 2024
Online date :30 July 2024

1. INTRODUCTION

The body uses carbohydrates, proteins and fats within foods as an energy source. While these nutrients, which are broken down into small particles to be absorbed, are digesting, a simple sugar called glucose is released. Glucose is an important nutrient source for all organs, especially the mammalian brain. Thus, glucose must be taken into the cell to be used the energy. It is the insulin hormone located behind the stomach and released from the pancreatic gland, which allows glucose to enter the cell and be stored as glycogen [1-2]. Diabetes Mellitus is a disease caused by excessively high levels of glucose as a result of the pancreas not producing enough insulin for the body, or the insulin produced by the pancreas not being used effectively by the body [3]. Thus, the glucose that passes from food into the blood cannot be used and the sugar level in the blood increases. This situation causes damage to many tissues and organs in the long term. Moreover, it triggers many diseases such as heart diseases, kidney problems and blindness.

According to the 2021 International Diabetes Federation (IDF) data, around 537 million adults aged 20-79 years live with diabetes worldwide in 2021 and this number is expected to increase to 643 in 2030 and 784 million in 2045. According to the Federation's report, one in 11 adults has diabetes and one in every two adults is unaware they have diabetes [4-5]. Early diagnosis of the disease is extremely

*Corresponding Author: tugba.palabas@gmail.com

important to prevent or slow the progress of advanced complications by allowing rapid intervention [6]. However, diagnosis is a very complex stage as it requires the evaluation of many factors together [7-9]. Machine learning algorithms, such as Naive Bayes (NB), Decision Trees (DT), k Nearest Neighbor (kNN) and Support Vector Machine (SVM) are frequently performed in healthcare systems to predict various disorders such as Diabetes Mellitus and they provide a great contribution to progress of the diagnosis process rapidly and accurately [10]. In addition, automatic taking of the patient's history and computer-aided decision-making provides a significant benefit in the physician's ability to initiate an effective and rapid treatment process, as possible problems are detected in early.

Many studies have been performed about early diagnosis by using data mining and statistical analysis methods that consider the general complaints of the patient and the prominent symptoms of the disease. For instance, Khafaga et al. [11] analyzes the diabetes dataset with NB, Logistic Regression (LR), and Random Forest (RF) Algorithms. Authors obtain the best accuracy using RF method on this dataset after applying 10-Fold Cross Validation and Percentage Split evaluation techniques. Similarly, Islam et al. [12] proposed a methodology using three ensemble techniques, AdaBoost, Bagging, and RF for estimation of the early diabetes risk. To test the success of the classification algorithms, the same diabetes dataset in the UCI machine learning repository is used and it is shown that RF algorithm provides maximal accuracy, precision, recall, and F-measure. Then Laila et al. [10] present an integrative approach that combines classification algorithms with association rules to improve prediction accuracy. Namely, they present a method by using Local Outlier Factor, Balanced Bagging Classifier, and association rules for early-stage prediction of diabetes. As a result, they obtain the prediction accuracy (97.36%). Pima Indians Diabetes dataset is another dataset which is widely used in much research. [Sisodia D and Sisodia DS [13] compare the performance of NB, SVM and DT classifiers in early diagnosis of diabetes according to various parameters such as Recall, F-Measure, Accuracy and ROC using this dataset in the UCI database. It is shown that the NB method achieves the highest correct classification success (76.30%). Naz and Ahuja [14] present a study comparing data mining classification techniques as Artificial Neural Network (ANN), NB, DT and Deep Learning (DL) and the accuracy is obtained by these functional classifiers within the range of 90–98%. Peker et al. [15] present a study by using diabetes data obtained from Köycegiz and Dalaman State Hospitals of Mugla in Turkey and comparing different algorithms as RF, Feed Forward Neural Network (FFNN), DT, kNN and SVM.

In this study, firstly, Early-Stage Diabetes Risk Prediction dataset in the UCI machine learning repository is analyzed with NB, Trees-J48, kNN and sequential minimal optimization (SMO) algorithm. Then, the effects of ensemble algorithms on the performance of these classifier are examined in detail by using three different methods (Adaboost [16], Bagging [17], Random Sub-Spaces (RSS) [18]). The success of classifiers and the effect of ensemble algorithms on classification performance are compared for 4 different criteria (Accuracy, Kappa, Root Mean Squared Error (RMSE) and Area Under of Curve (AUC)). In addition, to reduce the cost in the collection of data related to the detection of diabetes risk and increase the detection speed, the importance ratios of the features are determined by using the Wrapper Subset Eval (WSE) algorithm When the attributes are removed from the data set according to these rates, classification success is increased. Thus, the features are reduced in the preprocessing step, allowing the training and testing steps to be executed faster. A higher success is achieved in a shorter time.

2. MATERIAL AND METHOD

"The first sentence of Section 2 should be changed as "The Early-Stage Diabetes Risk Prediction dataset" in "UC Irvine Machine Learning Repository database" is analyzed in this study. This dataset has been collected using direct questionnaires from the patients of Sylhet Diabetes Hospital of Sylhet province in Bangladesh and approved by a consultant. It contains the sign and symptom data of newly diabetic or would be diabetic patient [19]. The content of the training data consists of 520 examples of 17 attributes shown in Table 1.

All results are obtained through with the “Waikato Environment for Knowledge Analysis (WEKA)” machine learning software.

Table 1. For the Early-Stage Diabetes Risk Prediction dataset, the attributes, which indicative the risk factors, and value type of attributes.

	Attribute	Value Type
1	Age (Age)	Numeric
2	Gender (Male/Female)	Nominal
3	Polyuria (Yes/No)	Nominal
4	Polydipsia (Yes/No)	Nominal
5	Sudden weight loss (Yes/No)	Nominal
6	Weakness (Yes/No)	Nominal
7	Polyphagia (Yes/No)	Nominal
8	Genital thrush (Yes/No)	Nominal
9	Visual blurring (Yes/No)	Nominal
10	Itching (Yes/No)	Nominal
11	Irritability (Yes/No)	Nominal
12	Delayed healing (Yes/No)	Nominal
13	Partial paresis (Yes/No)	Nominal
14	Muscle stiffness (Yes/No)	Nominal
15	Alopecia (Yes/No)	Nominal
16	Obesity (Yes/No)	Nominal
17	Class (Positive/Negative)	Nominal

2.1. Feature Extraction

First, 17 risk factors expressed in Table 1 are considered as features. A feature vector with 520x17 is obtained by using 520 positive (class1) and negative (class2) records with diabetes or symptoms in the 20-65 age range. The performance of classifiers and the effects of ensemble algorithms on classification success are investigated by using this vector. Then, feature extraction is performed using WSE–Greedy Stepwise (GS) algorithm among 17 features and the importance ratio of the features is determined as shown in Table 2. The performance criteria are evaluated again for the same methods ignoring 6th, 13th, 16th risk factors.

2.2. Classification

In this study, 4 classification algorithms, namely NB, Trees-J48, kNN, and SMO are used, and the classification performances of these algorithms are evaluated and presented comparatively. To improve the success of classification, Adaboost, Bagging, RSS algorithms, which are ensemble learning methods, are used together with the classifiers.

2.2.1. Naïve Bayes

Naive Bayes classifier that is a statistical classification method based on Bayes theorem finds the class to which the samples belong, assuming that the attributes are independent of each other. To do so, it determines the conditional probability $P(A|B)$ of event A for given event B. Namely, the case $P(C_i|Y)$ for the class C_i that maximizes the probability is calculated as below [20]:

$$P(C_i|Y) = \frac{P(Y|C_i)P(C_i)}{P(Y)} \quad (1)$$

$P(C_i|Y)$ is the posterior probability of C_i given Y. $P(Y|C_i)$ is the conditional probability of Y given C_i . $P(C_i)$ is the prior probability of i_{th} C class. $P(Y)$ is the prior probability for Y [21].

2.2.2. Trees-J48

In decision tree methods, the training set is recursively divided into subsets from root to leaves considering certain criteria. The leaf level of the tree represents the class labels. The most important problem in decision trees is that determining the root node and the order of branching to the leaves. For this purpose, the entropy measurement shown in Eq.2 is frequently used. Accordingly, the division criterion from root node to leaves is determined by calculating the entropy-dependent information gain [22-24].

$$H(X) = - \sum_{i=1}^n (p_i) \log_2(p_i) \tag{2}$$

Here, X is an attribute, p_i is each element with i_{th} position of each X element. A small entropy value indicates that the uncertainty and indecision about the result is small. So, a class or attribute has a small entropy is selected for another step [25].

2.2.3. k-Nearest Neighbor (kNN)

The kNN classifier is an instance-based method. Accordingly, the distance (d) of the relevant sample to j_{th} sample in the training set is calculated to determine the class of i_{th} x sample in p-dimensional space. The class has the most samples, which has the smallest d value among the k samples, is determined as the target class. The Euclidean Distance method shown in Eq.3 is frequently used in distance measurement [26-27].

$$d(i, j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \dots + |x_{ip} - x_{jp}|^2} \tag{3}$$

3. RESULTS

In the study, for the purpose of predicting the risk of early diabetes, the values recorded for 520 samples are examined in terms of 17 features that may be associated with the disease. In the dataset, positive and negative case of diabetes or symptoms are labeled as 'class1' and 'class2'.

Table 2. Evaluation of classification success of NB, J48, kNN and SMO algorithms and comparison of performance criteria of classifiers based on ensemble approaches.

Ens	Classifier	Accuracy (%)	Kappa	RMSE	AUC
None	NB	88	0.742	0.316	0.95
	J48	95	0.887	0.222	0.96
	kNN	98	0.951	0.149	0.98
	SMO	93	0.854	0.263	0.92
Adaboost	NB	88	0.745	0.296	0.96
	J48	98	0.963	0.129	0.99
	kNN	98	0.951	0.149	0.98
	SMO	93	0.849	0.229	0.98
Bagging	NB	87	0.738	0.316	0.95
	J48	97	0.927	0.171	0.99
	kNN	97	0.939	0.152	0.99
	SMO	93	0.841	0.244	0.96
RSS	NB	88	0.738	0.311	0.94
	J48	96	0.907	0.196	0.99
	kNN	98	0.959	0.158	0.99
	SMO	89	0.754	0.279	0.96

In classification step, NB, J48, kNN and SMO classifier performances are investigated using k (15) fold cross-validation method. In this context, accuracy rate, kappa coefficient, RMSE and AUC performance

measures are obtained. Then, the effect of ensemble algorithms (Ens), Adaboost, Bagging, RSS, to the success of classifier is evaluated separately using these performance parameters as shown in Table 2.

Here, accuracy represents the rate at which the class of each sample is correctly labeled and is calculated as in Equation 4.

$$Accuracy = \frac{tp+tn}{tp+tn+fp+fn} \quad (4)$$

In the equation, f_p , f_n , t_p , and t_n are the number of false positives, false negatives, true positives and true negatives, respectively.

When ensemble algorithms are not used, the best estimates of diabetes risk are obtained using kNN (97.69%). However, the accuracy rate for J48 increases significantly when classification performances are supported by ensemble algorithms. In particular, the J48 classification based on the Adaboost algorithm provides the maximum accuracy rate among all methods.

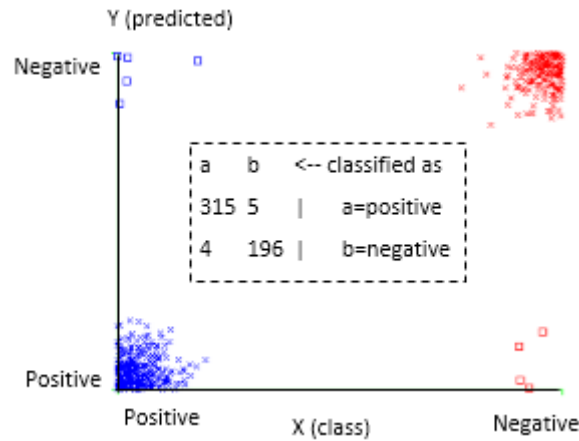


Figure 1. Confusion matrix and distribution of correctly and incorrectly classified samples explaining the classification accuracy of J48 algorithm based on Adaboost ensemble approach.

As shown in Figure 1, the classes of 315 positive and 196 negative samples are correctly labeled. On the other hand, only 5 positive and 4 negative samples are labeled in incorrect class. As a result, 98.26 % correct classification success is achieved.

Similarly, the highest value of the Kappa coefficient (κ) is also obtained by the J48 classifier based on the Adaboost approach. Kappa coefficient is mainly used to calculate the classification accuracy. To calculate κ value, two different probabilities, $Pr(a)$ and $Pr(e)$, are used as follows [29].

$$\kappa = \frac{Pr(a)+Pr(e)}{1-Pr(e)} \quad (5)$$

Here, $Pr(a)$ is the actual observed agreement, while $Pr(e)$ is the probability of this fit occurring by chance. "1" indicates a perfect fit, and "0" indicates a poor fit.

Third performance criteria RMSE is calculated using the square root of mean squares error (MSE). It measures the mean size of errors and deals with deviations from the true value. So, the lower of RMSE is the better prediction and "0" indicates a perfect fit [30]. As seen in the Table 2, all ensemble algorithms decrease the RMSE value of the J48 classifier and Adaboost has ensured that the best RMSE is obtained. Lastly, AUC criteria are examined for all classification algorithms in the table. AUC is the area of the two-dimensional measure under the ROC (Receiver Operating Characteristic) curve and its value ranges

are [0, 1]. Accordingly, if a model prediction is 100% wrong, AUC is "0"; a model predictions are 100% accurate, AUC is "1" [31-33].

More specifically, in ROC curve, the true positive rate (sensitivity is calculated by Eq. 6) is plotted as a function of the false positive rate (specificity is calculated by Eq. 7) for different cut-off points [34-36].

$$Sensitivity = \frac{tp}{tp+fn} \tag{6}$$

$$Specificity = \frac{tn}{tn+fp} \tag{7}$$

Thus, each point on the curve correspond to a sensitivity/specificity pair. If 100% sensitivity and specificity value is obtained, the ROC curve, which is very close to the upper left corner (the area under the curve is larger), shows a perfect separation without overlapping in the two distributions [37].

Accordingly, Figure 2 (a-b-c-d) shows the ROC curves for NB, J48, kNN and SMO classifiers, respectively. As can be seen from these figures, the most obvious distinction between the classes and the left-justified graph is provided in panel (c) by the kNN method. Then, the best success is achieved with J48 in (b), NB in (a), and SMO in (d), respectively.

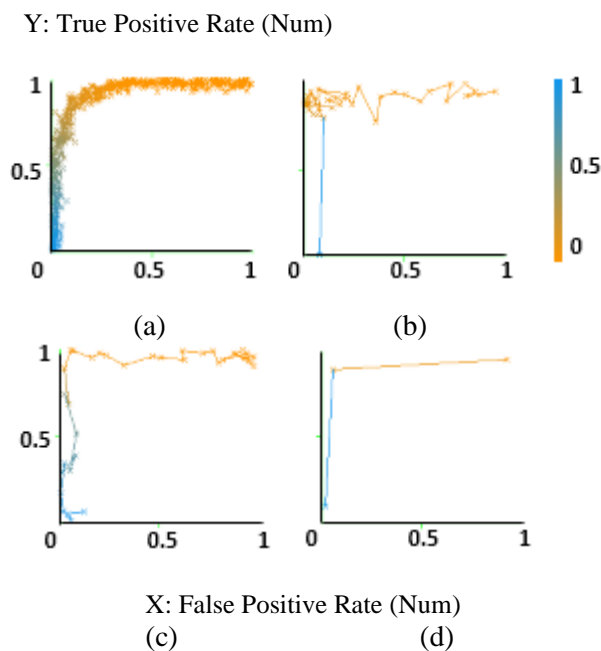


Figure 2. ROC curves include sensitivity/specificity pair representation which expresses the classification success of four different classification methods (a) NB, (b) J48, (c) kNN, (d) SMO.

In Figure 3, classifier performances based on Adaboost algorithm are evaluated. Here, the distinction between classes can be seen quite clearly in panel (b). In addition, the curve leans almost entirely to the left. In (c) and (d), the kNN and SMO methods have similar distribution and steepness. In (a), NB is the graph where the curve is furthest to the left side.

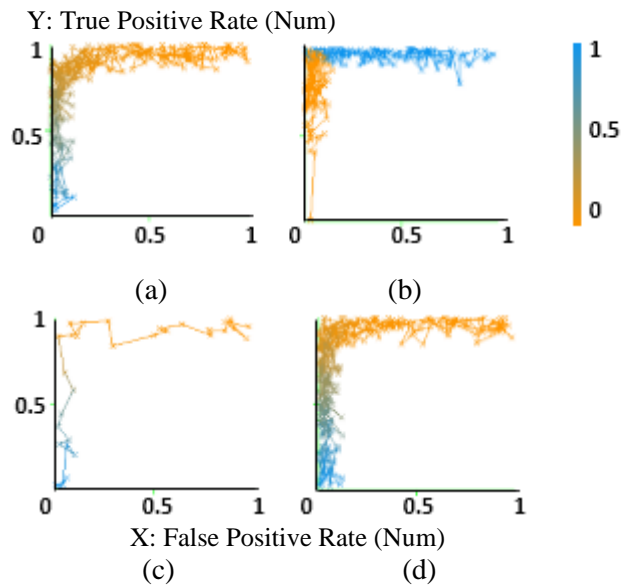


Figure 3. ROC curves of classifiers based on Adaboost ensemble algorithms (a) NB, (b) J48, (c) kNN, (d) SMO.

In Figure 4, the classifier performances based on the Bagging algorithm show that in panel (b) and (c), J48 and kNN classifier achieve the highest performance with left-justified ROC curves and a similar separation between classes. Then, SMO in (d) and NB in (a) with the worst performance are shown.

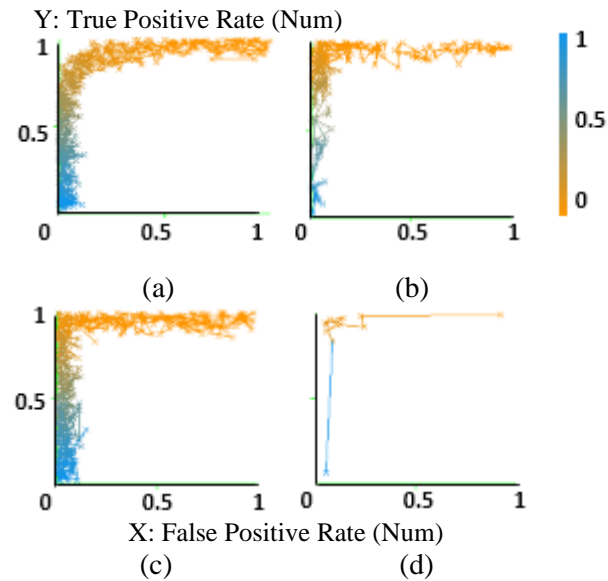


Figure 4. ROC curves of classifiers based on Bagging ensemble algorithms (a) NB, (b) J48, (c) kNN, (d) SMO.

The ROC curves obtained by the RSS algorithm in Figure 5 are quite like the Bagging algorithm, so the performance order is obtained as (b)=(c)>(d)>(a).

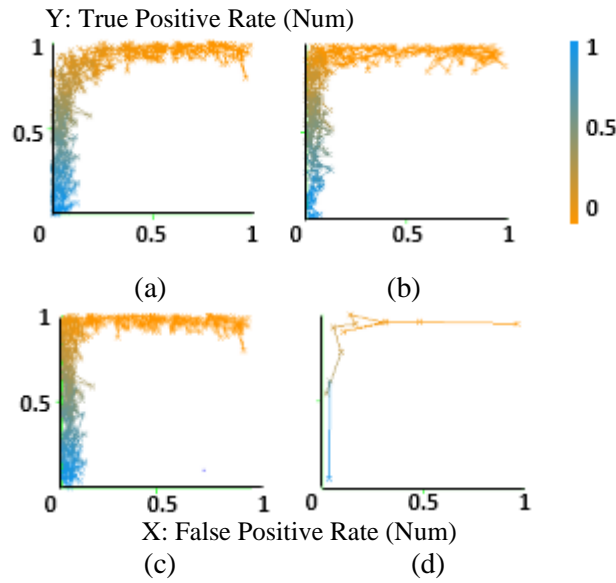


Figure 5. ROC curves of classifiers based on Bagging ensemble algorithms (a) NB, (b) J48, (c) kNN, (d) SMO.

As a result, both the AUC values obtained in Table 2 and the ROC curves shown in Figure 2-3-4-5 indicate that ensemble algorithms increase the classifier performance. In other words, as seen in the table, the performance of classifiers based on ensemble algorithms in general resulted in AUC values close to “1”. J48 classifier provides the maximum AUC value for all ensemble algorithms. On the other hand, ROC curves very close to the upper left corner are obtained with a perfectly discriminating distribution as seen in panels (b) of the four figures presented. In addition, it becomes very difficult to perform the machine learning task or gain insight into the data in the case of very large datasets. Because the complexity of the model and the time required to train the model also increase by the reason of increasing data size. Moreover, inaccurate, or less reliable results may be obtained. In this context, feature selection algorithms provide to obtain a better classification performance in a shorter time by removing some features that are unrelated or less important to the dependent variable from the data set.

Table 3. The significant rates of attributes which are obtained with the WSE algorithm.

	Attribute	Significant Rate (%)
1	Age (Age)	100
2	Gender (Male/Female)	100
3	Polyuria (Yes/No)	100
4	Polydipsia (Yes/No)	100
5	Sudden weight loss (Yes/No)	67
6	Weakness (Yes/No)	20
7	Polyphagia (Yes/No)	40
8	Genital thrush (Yes/No)	47
9	Visual blurring (Yes/No)	40
10	Itching (Yes/No)	53
11	Irritability (Yes/No)	47
12	Delayed healing (Yes/No)	100
13	Partial paresis (Yes/No)	13
14	Muscle stiffness (Yes/No)	40
15	Alopecia (Yes/No)	87
16	Obesity (Yes/No)	0

For instance, when "wrapper feature selection algorithm (Wrapper Subset Eval-WSE) following the greedy search approach based on bi-directional elimination (Stepwise Selection)" is used, it is

determined that "obesity" as an uncorrelated feature (0%) in the original data set consisting of 17 attributes belonging to each individual. In addition, "weakness" and "partial paresis" are less correlated features (20% and 13%) as seen in Table 3. When these features are removed from the dataset and the same algorithms are run again, the effect of WSE on the correct classification performance is clearly seen.

Table 4. The effect of WSE algorithm to classifiers 'performance criteria.

Ens	Classifier	Accuracy (%)	Kappa	RMSE	AUC
Adaboost	NB	89	0.775	0.288	0.96
	J48	99	0.967	0.120	1.00
	kNN	98	0.947	0.255	0.98
	SMO	92	0.834	0.228	0.98

As can be seen in Table 4, an overall increase in performance measures is observed for the NB, J48 and kNN classifiers. More specifically, the accuracy of J48 classifier increases to 99% by using Adaboost ensemble algorithm with 15-fold cross validation. In addition, improvement is observed in other performance criteria as well.

Table 5. The comparison of the studies' performance on Early-stage diabetes risk prediction dataset published by UCI.

Method	input number	Techniques	Accuracy
[11]	12	kNN	97.36%
[12]	16	RF+10-Fold Cross Validation+ Percentage Split Evaluation	99.00%
[38]	16	Adaboost and Bagging+NBTree	98.65%
[39]	16	Multi-Layer Perceptron+Improved Crow Search Algorithm	
		*one hidden layer	97.69%
		*two hidden layer	96.92%
proposed	13	WSE+Adaboost +J48	99.00%

In [11], 97.36% classification success is achieved by using the kNN classifier with 12 input features. To increase the success in [12], the RF classifier is supported by the Percentage Split Evaluation method and 99% accurate classification is provided with 16 inputs. In [38], 98.65% success is obtained by using 16 inputs with Adaboost, Bagging, NBTree algorithms. [39] uses Multi-Layer Perceptron and Improved Crow Search Algorithm. The method has 97.69% and 96.92% success for one and two hidden layers for 16 inputs. On the other hand, the proposed method, not only saves time and memory as it uses less input, but also achieves a very high success rate of 99%.

4. DISCUSSION

Diabetes Mellitus, which is at the forefront of the diseases of the age, is a type of disease that is very common all over the world. Moreover, long-term high blood sugar due to diabetes can cause permanent damage to the whole body, especially the cardiovascular system, kidneys, or eyes. For this reason, early diagnosis of the disease and initiation of the treatment process are also vital for the prevention of other diseases. In this paper, it is aimed that a prediction system is modeled for detection of early-stage diabetes. To do so, firstly, the performances of NB, J48, kNN, SMO classifiers are compared for four

performance criteria which are Accuracy, Kappa, RMSE, AUC. 98% accuracy rate is achieved with the kNN algorithm. Then, classifier performances are examined separately based on Adaboost, Bagging and RSS ensemble approximations. The results show that the highest classification success of 98% is obtained when the J48 classifier is used together with the Adaboost algorithm. The other performance criteria Kappa, RMSE and AUC values are 0.963, 0.129 and 0.99, respectively. Finally, the WSE feature extraction algorithm is applied to the data set and the irrelevant or least relevant "obesity" (0%), "weakness" (20%) and "partial paresis" (13%) attributes are removed from the diabetes dataset. Thus, the classification success of the J48 algorithm based on the Adaboost ensemble approach increases to 99% when 13 inputs are used. In addition, Kappa, RMSE and AUC values are 0.967, 0.120 and 1.00, respectively.

Thus, it has been shown that the Adaboost ensemble algorithm can contribute to obtaining effective results in the training of different artificial intelligence models to be defined in diabetes risk prediction in the future.

CONFLICT OF INTEREST

The author(s) stated that there are no conflicts of interest regarding the publication of this article.

CRedit AUTHOR STATEMENT

Tuğba Palabaş: Investigation, Formal and computational analysis, Writing - original draft, Visualization, Conceptualization, Supervision.

REFERENCES

- [1] Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. How cells obtain energy from food. In *Molecular Biology of the Cell*. 4th edition. Garland Science, 2002.
- [2] Mergenthaler P, Lindauer U, Dienel GA, Meisel A. Sugar for the brain: the role of glucose in physiological and pathological brain function. *Trends in neurosciences*, 36(10), 587-597, 2013.
- [3] Brutsaert EF. *Diabetes mellitus (DM)*. Merck Manual, 2020.
- [4] International Diabet Federation, "IDF Diabetes Atlas". <https://diabetesatlas.org/>(16.05.2023).
- [5] Sağlık Bakanlığı, "Kronik Hastalıklar". <https://www.saglik.gov.tr/yazdir?2DE933CD45A7AD200096270A9E25E935> (16.05.2023).
- [6] Marshall SM, Flyvbjerg A. Prevention and early detection of vascular complications of diabetes. *Bmj*, 333(7566), 475-480, 2006.
- [7] Sümbül H, Yüzer AH. Development of diagnostic device for COPD: a MEMS based approach. *Int J Comput Sci Network Secur*. 2017;17 (7):196–203.
- [8] Sümbül H, Yüzer AH. Estimating the value of the volume from acceleration on the diaphragm movements during breathing. *J Eng Sci Technol*. 2018;13(5):1205–1221.

- [9] Sümbül H, Yüzer AH. Measuring of diaphragm movements by using iMEMS acceleration sensor. In 2015 9th International Conference on Electrical and Electronics Engineering (ELECO) EEE; 2015: 166–170.
- [10] Laila UE, Mahboob K, Khan AW, Khan F, Taekeun W. An ensemble approach to predict early-stage diabetes risk using machine learning: An empirical study. *Sensors* 2022; 22(14), p 5247.
- [11] Khafaga DS, Alharbi AH, Mohamed I, Hosny KM. An Integrated Classification and Association Rule Technique for Early-Stage Diabetes Risk Prediction. In *Healthcare* 2022; 10(10), p 2070.
- [12] Islam MM, Ferdousi R, Rahman S, Bushra HY. Likelihood prediction of diabetes at early stage using data mining techniques. In *Computer vision and machine intelligence in medical image analysis* (pp. 113-125). Springer, Singapore, 2020.
- [13] Sisodia D, Sisodia DS. Prediction of diabetes using classification algorithms. *Procedia computer science* 2018; 132, p 1578-1585.
- [14] Naz H, Ahuja S. Deep learning approach for diabetes prediction using PIMA Indian dataset. *Journal of Diabetes & Metabolic Disorders* 2020; 19, p 391-403.
- [15] Peker M, Özkaraca O, Şaşar A. Use of orange data mining toolbox for data analysis in clinical decision making: The diagnosis of diabetes disease. In *Expert System Techniques in Biomedical Science Practice* 2018; p 143-167.
- [16] Kalaycı TE. Comparison of machine learning techniques for classification of phishing web sites. *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi* 2018; 24(5), p 870-878.
- [17] Aytuğ O, Korukoğlu S. A review of literature on the use of machine learning. *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi* 2016; 22(2), p 111-122.
- [18] Özdemir A, Aytuğ O, Ergene VÇ. Machine learning and ensemble learning based method using online employee assessments to identify and analyze job satisfaction factors. *Avrupa Bilim ve Teknoloji Dergisi* 2022; 40, p 19-28.
- [19] UCI Machine Learning Repository, “Early-stage diabetes risk prediction dataset”. <https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset> (16.05.2023).
- [20] Tsymbal A, Puuronen S, Patterson DW. Ensemble feature selection with the simple Bayesian classification. *Information fusion* 2003; 4(2), p 87-100.
- [21] Banchhor C, Srinivasu N. Integrating Cuckoo search-Grey wolf optimization and Correlative Naive Bayes classifier with Map Reduce model for big data classification. *Data & Knowledge Engineering* 2020; 127, p 101788.
- [22] Altaş D, Gülpınar V. A Comparison of Classification Performances Of The Decision Trees and The Artificial Neural Networks: European Union, *Trakya Üniversitesi Sosyal Bilimler Dergisi* 2012; 14(1) p 1-22.
- [23] Kavzoğlu T, Çölkesen İ. Classification of Satellite Images Using Decision Trees: Kocaeli Case. *Harita Teknolojileri Elektronik Dergisi* 2010; 2(1), p 36-45.

- [24] Sangeorzan L. Effectiveness analysis of ZeroR and J48 classifiers using WEKA toolkit. *Bulletin of the Transilvania University of Brasov. Series III: Mathematics and Computer Science* 2019; p 481-486.
- [25] Chen CH. A novel multi-criteria decision-making model for building material supplier selection based on entropy-AHP weighted TOPSIS. *Entropy* 2020; 22(2), p 259.
- [26] Hemmatian F, Sohrabi MK. A survey on classification techniques for opinion mining and sentiment analysis. *Artificial intelligence review* 2019; 52(3), p 1495-1545.
- [27] Alharbi Y, Alferaidi A, Yadav K, Dhiman G, Kautish S. Denial-of-service attack detection over IPv6 network based on KNN algorithm. *Wireless Communications and Mobile Computing* 2021; p 1-6.
- [28] Platt JC. Fast training of support vector machines using sequential minimal optimization, advances in kernel methods. *Support vector learning* 1999; p 185-208.
- [29] McHugh ML. Interrater reliability: the kappa statistic. *Biochemia medica* 2012; 22(3), p 276-282.
- [30] Alghamdi A S, Polat K, Alghoson A, Alshdadi AA, Abd El-Latif AA. A novel blood pressure estimation method based on the classification of oscillometric waveforms using machine-learning methods. *Applied Acoustics* 2020; 164, p 107279.
- [31] Pepe MS, Cai T, Longton G. Combining predictors for classification using the area under the receiver operating characteristic curve. *Biometrics* 2006; 62(1), p 221-229.
- [32] Hajian-Tilaki K. Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. *Caspian journal of internal medicine* 2013; 4(2), p 627.
- [33] Kemalbay G, Alkış BN. Prediction of stock index movement direction with multiple logistic regression and k-nearest neighbors algorithm. *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi* 2020; 27(4), p 556-569.
- [34] Janssens ACJ, Martens FK. Reflection on modern methods: Revisiting the area under the ROC Curve. *International journal of epidemiology* 2020; 49(4), p 1397-1403.
- [35] Ruisánchez I, Jiménez-Carvelo AM, Callao MP. ROC curves for the optimization of one-class model parameters. A case study: Authenticating extra virgin olive oil from a Catalan protected designation of origin. *Talanta* 2021; 222, p 121564.
- [36] Cihan P, Kalipsiz O, Gökçe E. Computer-aided diagnosis in neonatal lambs. *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi* 2020; 26(2), p 385-391.
- [37] Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical chemistry* 1993; 39(4), p 561-577.
- [38] Taser PY. Application of bagging and boosting approaches using decision tree-based algorithms in diabetes risk prediction. *Proceedings* 2021; 74(1), p 6.
- [39] Wijayaningrum VN, Saragih TH, Putriwijaya NN. Optimal multi-layer perceptron parameters for early stage diabetes risk prediction. In *IOP Conference Series: Materials Science and Engineering* 2021; 1073(1), p 012070.