www.dergipark.gov.tr
ISSN:2148-3736
El-Cezerî Journal of Science and Engineering
Vol: 10, No: 2, 2023 (409-419)
DOI: 10.31202/ecjse.1242207

ECJSE

Research Paper

The Diagnosis of Diabetes Mellitus with Boosting Methods

Hilal KOÇAK[1a], Gürcan ÇETIN[1b]

[1]Information Systems Engineering Department, Muğla Sıtkı Koçman University, Muğla, Turkey
[b]gcetin@mu.edu.tr

Received: 25.01.2023 **Accepted:** 09.05.2023

Abstract: In addition to the damage, it can cause to various organs, diabetes mellitus (DM) also increases a person's risk of developing other serious health conditions. These can include heart disease, stroke, and nerve damage. Furthermore, DM is a leading cause of blindness and kidney failure. However, with proper management and treatment, many of the complications of DM can be prevented or delayed. Thus, early detection and treatment of DM are crucial. With the advancement of machine learning technology, new opportunities have emerged in the field of medicine. Many disease detection research relies on machine learning techniques, with a particular emphasis on boosting algorithms. Boosting algorithms are used to improve the accuracy of predictions made by other weak models such as decision trees. Using knowledge discovery methods, boosting algorithms are examined and compared on a diabetes dataset in this study. The performance of the boosting algorithms is evaluated by generating ROC curves and comparing average accuracy values. When the study's results were evaluated in terms of precision, Gradient Boosting, AdaBoost, CatBoost, LightGBM, and XGBoost algorithms gives success rates of %85, %83, %88, %86, and %87, respectively.

Keywords: Diabetes mellitus, boosting algorithms, machine learning

1. Introduction

Diabetes is a life-long disease with an increasing number of sufferers. According to the International Diabetes Federation (IDF), 1 in 11 people live with diabetes. Also, the last edition of the IDF diabetes atlas shows that 463 million adults live with diabetes worldwide. Unfortunately, it is predicted that the number of adults with diabetes will reach 578 million by 2030 and 700 million by 2045 [1].

In diabetic patients, the hormone responsible for converting foods into energy and adjusting blood sugar is not secreted by the pancreas. This hormone called insulin is responsible for bringing blood sugar to body cells. The inability to use insulin effectively causes glucose levels to rise too high in the blood, which is called hypoglycaemia. In patients with insufficient insulin hormones, symptoms such as hunger, restlessness, sweating, and fainting begin to appear. The level of blood sugar should be adjusted with treatments. Diagnosis of diabetes is made according to plasma glucose values in the morning blood sample after 9 hours of night fasting [2]. The patient's fasting blood glucose value of 126 mg/dl or above means that she/he has diabetes. High blood sugar for long periods can cause various dysfunctions and damage, especially in blood vessels, nerves, heart, and kidneys [3]. Diabetes also is one of the leading causes of blindness and visual impairment in adults. Every year, many patients die of heart attack and stroke due to complications related to diabetes. Considering all the threats, diabetes has become a serious health problem worldwide [4]. Therefore, as with many diseases, it is very important to diagnose diabetes early and to take measures for complications [5].
Recently, the increase in medical data and the easier storage of them has made the use of decision-support systems widespread in the detection of diseases [6]. The purpose of these systems is to assist the experts by making the diagnosis of the disease quickly and with high accuracy [7]. This situation

How to cite this article
Koçak H., Çetin G., "The Diagnosis of Diabetes Mellitus with Boosting Methods," *El-Cezerî Journal of Science and Engineering*, 2023, 10 (2), 409-419.
ORCID: [a]0000-0003-2602-8557; [b]0000-0003-3186-2781

facilitated the analysis of huge amounts of data and led to significant developments in terms of detecting, treating, or taking precautions for diseases [8].

Machine learning is a field that stands out with the use of the features of artificial intelligence such as reasoning, planning, correlation inference, and knowledge evaluation with statistical methods. Results from many recent studies have shown that machine learning algorithms have grateful power in classifying and diagnosing various diseases [9]. When state-of-the-art methods of machine learning for automatic diagnosis of diabetes are examined; Qu et al. compared the success of neural network, random forest, and decision tree algorithms in diabetes detection. Although there was not much difference between them, they obtained 77% and 80% accuracy rates from the random forest algorithm in the two different datasets used [10]. Kaur et al. found the accuracy rates from 5 classification models they applied on the Pima Indian Diabetes dataset to be 0.89 for linear kernel SVM, 0.84 for radial basis function SVM model, 0.88 for kNN algorithm, 0.86 for ANN and 0.83 for MDR based model [11]. Besides this study with the same dataset, Rawat et al. applied Bagging and Adaboost techniques to diagnose diabetes mellitus and they predicted the disease with 81.77% and 79.69% accuracy rates respectively [12]. Chen et al. realized LogitBoost and AdaBoost algorithms to detect diabetes mellitus and they achieved diagnosis with above 90% accuracy of classification [13]. Xu et al. bring forward a type II diabetes prediction model by using the random forest algorithm and they achieved an 85% accuracy rate [14]. Kalaiselvi et al. analyzed Naive Bayes and K-means algorithms and they proposed Adaptive Neuro Fuzzy Inference System with reaching around %80 accuracy [15]. Pangaribuan et al. have shown they can predict diabetes mellitus with a satisfactory accuracy rate by implementing an ANN-based Extreme Machine Learning algorithm [16].

It may not be sufficient to use only one algorithm for the detection of a disease, because each algorithm has different advantages and disadvantages. Therefore, it is important to use more than one algorithm for the detection of diabetes [17]. On the other hand, algorithms like voting, bagging, and boosting that classify diseases using ensemble models produce better results. These algorithms combine weak learners, which are learning algorithms with poor performance, to create strong learners. Therefore, in this study, the performances of boosting algorithms, one of the ensemble models, were evaluated by applying them to the diabetes dataset. Gradient Boosting, AdaBoost, CatBoost, LightGBM, and XGBoost algorithms have been chosen for the study, and their results have been compared using ROC curves and average accuracy values. As a result of the research, the performance value ranged between 0.83-0.88. Furthermore, the CatBoost algorithm had the highest success rate of 0.88.

The article continues with methods in section 2, where the features of boosting algorithms are examined, and part 4 which is Exploratory Data Analysis, where the implementation of the application is explained. Finally, it ends with a comparison and discussion of the created models.

2. Experimental Methods

2.1. Boosting Algorithms

Ensemble models generalize and diversify the results by combining decisions from multiple models to create the best classifier [18]. Many models have been invented to build ensemble methods like Boosting [19] algorithms for various applications. Boosting Algorithms are methods of converting weak classifiers to strong classifiers across iterations. Inferences are made from the collection of trees obtained by giving different weights to the dataset. From observations with equal weight at the beginning, the weights of incorrectly classified observations are increased while correctly classified examples lose weight. A new model is created with updated weights, predictions are made, and the cycle continues as long as the model improves. Each new model emerges by correcting the error of the previous model. The final model is derived from the weighted average of all previous models.

2.1.1 Adaptive Boosting (AdaBoost)

It is a boosting algorithm that focuses on weighting and training where multiple sequential models are created, each correcting the errors in the next model [20]. AdaBoost puts weight on misclassified training data and the next model tries to accurately predict these values. When constructing the next model, more weight is given to incorrectly predicted data points. Weights are determined using the error value. For example, the higher the error, the higher the weight assigned to the observation. To summarize AdaBoosts' differences from the random forest are combining many weak learners, some stumps have more say than others and each stump (subtree) is made by taking the previous stump's mistake into account.

2.1.2 Gradient Boosting

It is the boosting algorithm used for both regression and classification problems. In the Gradient Boosting approach, new trees are created by making use of the weighted estimates of previous trees [21]. As long as the model continues to evolve, new trees are created by looking at the prediction errors of previous trees. The difference between the predicted value produced for any problem and the target value is tried to be reduced at each iteration. The disadvantage of the algorithm is that the combined trees can cause overfitting [22].

2.1.3 LightGBM

LightGBM is a histogram-based algorithm that reduces the computational cost by making discrete binary variables with continuous values. LightGBM is a histogram-based algorithm. Since the training time of decision trees is directly proportional to the calculation and therefore the number of splits, LightGBM both shortens the training time and reduces resource usage [23].

In decision trees, two strategies are used: the depth-oriented strategy, in which the balance of the tree is maintained while the tree grows, and the leaf-oriented strategy in which the division process continues to reduce loss. With this, LightGBM has less error rate and learns faster from XGBoost [24]. However, because the leaf-oriented growth approach causes the model to be prone to over-learning when the data is small, the algorithm is more suitable for use in big data [25]. LightGBM also uses two novel techniques different from other algorithms. These are Gradient Based One Way Sampling (GOSS) and Exclusive Feature Bundling (EFB) which deals with the number of data samples and variables.

Traditional gradient boosting scans all data samples to calculate the information gain for each variable, while GOSS uses only the key data, thus reducing the number of data without much impacting the distribution of the data. Besides, EFB aims to create stronger features by combining variables and thus reduce the number of features.

In summary, EFB combines variables to reduce dimensionality, while GOSS reduces data size to compute knowledge acquisition by neglecting less important data. With these two functions, LightGBM has more efficiency and scalability [26]. To prevent overlearning in LightGBM; the number of leaves, the maximum depth, the minimum number of data in the leaf, the number of data to be used in each iteration, and the learning rate can be optimized.

2.1.4 CatBoost

The boosting algorithm, which is a combination of the words "Category" and "Boosting" and emerged with the development of GBM, was developed by Yandex. It works with high learning speed on both categorical and numerical data. One hot encoding method [27], which is one of the common methods to deal with categorical variables, can sometimes cause too many new features to be added for high cardinality features. This algorithm has a different method that can encode categorical data and thus speed up data preprocessing by executing various statistical combinations on both categorical and numerical features [28]. Also, CatBoost builds symmetrical trees, achieving a high prediction rate that prevents overfitting. CatBoost runs on the GPU as long as the appropriate graphics card is used which provides an advantage for problems with long learning time by taking memory copies for unexpected situations such as sudden shutdown of the computer [29]. The basic parameters that can be optimized in the CatBoost algorithm are the maximum length of the tree, the learning rate, the number of trees to be created, categorical features, and the parameters used to prevent overlearning.

2.1.5 XGBoost

It is a scalable machine-learning method developed by Tianqi Chen and Carlos Guestrin. The fact that the XGBoost algorithm gives state-of-the-art results in many competitions on Kaggle has shown that it can be applied to a wide range of problems [30]. XGBoost is an algorithm created by strengthening Gradient Boosting using fewer software and hardware resources to give better results in a shorter time. It works well in more complex problems with large datasets. It has been shown in many studies that it has higher predictive power and higher operating speed compared to other boosting algorithms. XGBoost uses the following novel techniques [31].

- *Coping With Missing Data Effectively:* One of the biggest challenges in data mining is missing data in the dataset. To extract meaningful information from the dataset, the missing data must be filled in or removed from the dataset. XGBoost is good at finding patterns in datasets with sparse data, as error values (residuals) are also calculated on rows with missing values after the prediction.
- *Regularization for Overfitting:* XGBoost has exclusively regularization methods like L1 and L2 [31].
- *Triggling with the Depth Priority Approach:* XGBoost creates the splits up to the specified max_depth value and then begins to prune the tree backward and removes any splits beyond which there are no positive gains.
- *Cache Awareness and Out Of Core Computing:* XGBoost computes the similarity score and output value in the cache. Since the cache memory can be used at the maximum level, fast calculations can be made.

XGBoost builds decision trees in all possible scenarios to maximize the learnings score for each variable. Such algorithms are called "Greedy Algorithms". This process can take a very long time in large datasets. Instead of examining each value in the data, XGBoost divides the data into quantiles and works according to those parts. Also, it uses the weighted quantile sketch algorithm to find the most accurate split points.

2.2. Model Evaluation

2.2.1. Confusion Matrix

It is used to evaluate the performance of the models. The confusion matrix given in Table 1 shows how accurately the generated model can predict the actual values [32].

Table 1. Confusion Matrix

	Predicted Negatives	Predicted Positives
Negatives	True Negative(TN)	False Positive(FP)
Positives	False Negative(FN)	True Positive(TP)

2.2.2 Receiver Operating Characteristic Curve (ROC)

ROC is the graphic that is used to evaluate the performance of the algorithm used, as well as to reveal its visualization and accomplishment [33]. In the ROC curve, there is False Positive Rate on the X-axis and True Positive Rate on the Y-axis. In the graphic, while the true positive rate refers to sensitivity, the false positive rate is equal to 1-specificity.

In Equation (1) Sensitivity, also called Recall, is the proportion of the actual positive cases predicted as positive.

$$Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{1}$$

In Equation (2), Specificity is the proportion of actual negative cases predicted as negative.

$$1 - Specificity = 1 - \frac{True\ Negative}{True\ Negative + False\ Positive} \tag{2}$$

In the Fig.1 the higher the level under the curve shows the higher discrimination performance between classes.

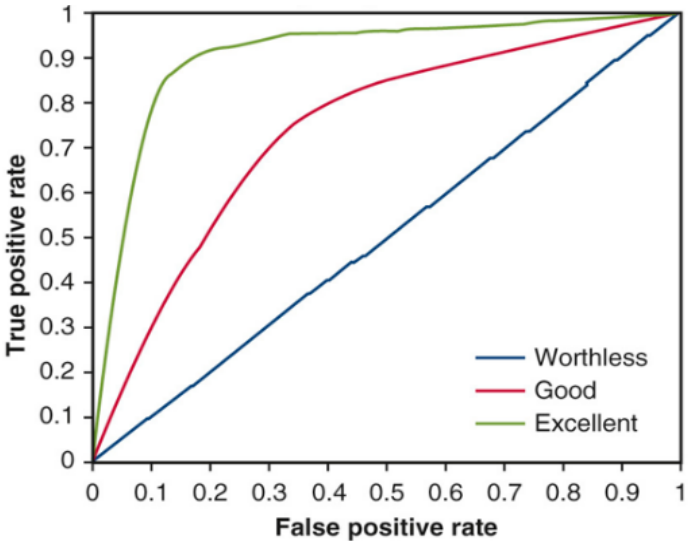

Figure 1. ROC of Three Representative Models [34]

2.3. Exploratory Data Analysis

2.3.1. Data Preprocessing

The dataset includes two class labels that are diabetes and nondiabetes. Fig. 2 shows that the dataset is unbalanced because the number of non-patients is more concentrated in the dataset. This dataset consists of two parts: healthy people and diabetes. It contains information on 500 healthy and 268 patients diagnosed with diabetes. The dataset consists of 9 attributes which are times of pregnancy, glucose in the blood, Diastolic blood pressure (mm Hg), triceps skin fold thickness (mm), 2-h serum insulin (mu U/ml), diabetes pedigree function, plasma glucose concentration 2 hours in an oral glucose tolerance test, age, and class (0 or 1). The Diabetes Pedigree Function refers to the genetic factor of diabetes [17].

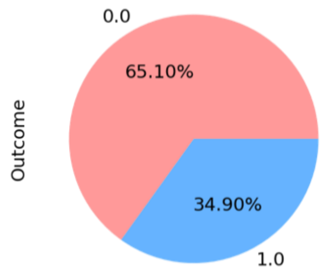

Figure 2. Class Variables in the Dataset

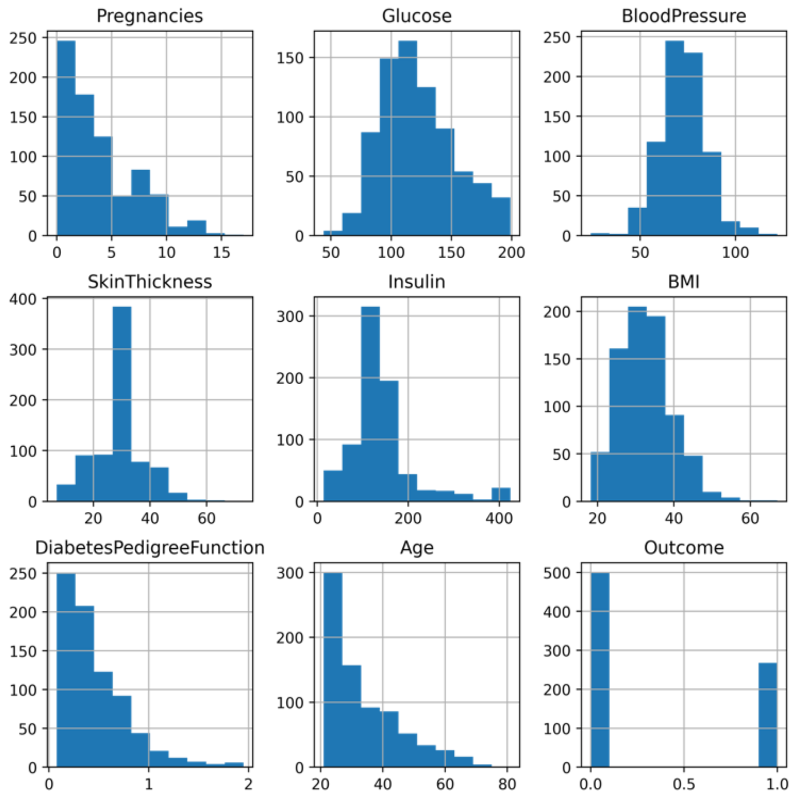

Figure 3. Distributions of Data by Columns

Fig. 3 shows in what range and density the data are distributed according to the features. When the data were examined, rows containing 0 in columns such as Blood Pressure, Skin Thickness, Glucose, Insulin, and BMI were detected. Since these values can't be 0 in a living person, these values are filled with the median values of the relevant column according to the class variable. Besides this, since some rows contain outlier data in columns such as Skin Thickness Diabetes Pedigree Function and Insulin, they are updated according to the min and max values in the relevant field in the dataset according to the Interquartile Range method.

3. Results

3.1. Experimental Results for AdaBoost

As seen in Table 2 when the AdaBoost algorithm was applied to the diabetes dataset, it was correctly estimated that 148 out of 161 patients were sick in the test dataset that the model did not know. The remaining 13 patients were misclassified. The accuracy rate is 83%.

Table 2. Confusion Matrix for AdaBoost results

	Predicted 0	Predicted 1
Actual 0	148	9
Actual 1	13	61

3.2. Experimental Results for Gradient Boosting

As seen in Table 3 when the Gradient Boosting Algorithm was applied to the diabetes dataset, it was correctly estimated that 146 out of 153 patients were sick in the test dataset that the model did not know. The remaining 7 patients were misclassified. The accuracy rate is 85%.

Table 3. Confusion Matrix for Gradient Boosting results

	Predicted 0	Predicted 1
Actual 0	146	11
Actual 1	7	67

3.3. Experimental Results for CatBoost

As seen in Table 4 when the Gradient Boosting Algorithm was applied to the diabetes dataset, it was correctly estimated that 149 out of 160 patients were sick in the test dataset that the model did not know. The remaining 11 patients were misclassified. The accuracy rate is 86%.

Table 4. Confusion Matrix for CatBoost results

	Predicted 0	Predicted 1
Actual 0	149	8
Actual 1	11	63

3.4. Experimental Results for XGBoost

As seen in Table 5 when the XGBoost Algorithm was applied to the diabetes dataset, it was correctly estimated that 145 out of 156 patients were sick in the test dataset that the model did not know. The remaining 11 patients were misclassified. The accuracy rate is 87%.

Table 5. Confusion Matrix for XGBoost results

	Predicted 0	Predicted 1
Actual 0	145	12
Actual 1	11	63

3.5. Experimental Results for LightGBM

As seen in Table 6 when the LightGBM Algorithm was applied to the diabetes dataset, it was correctly estimated that 146 out of 168 patients were sick in the test dataset that the model did not know.

Table 6. Confusion Matrix for LightGBM results

	Predicted 0	Predicted 1
Actual 0	146	11
Actual 1	12	62

The remaining 12 patients were misclassified. The accuracy rate is 86%.

4. Comparison of Techniques

Considering the accuracy rates, all boosting algorithms applied gave satisfactory results. It has been observed that the CatBoost algorithm gives a better result with a small difference compared to the others since there is a high cardinality categorical feature in the diabetes dataset.

Table 7. Accuracy Rates of Algorithms

Boosting Algorithms	Accuracy Rates
Gradient Boosting	0.85
AdaBoost	0.83
CatBoost	0.88
LightGBM	0.86
XGBoost	0.87

As seen in Fig 4, the ability of all algorithms to separate the classes is gratifying and close to each other.

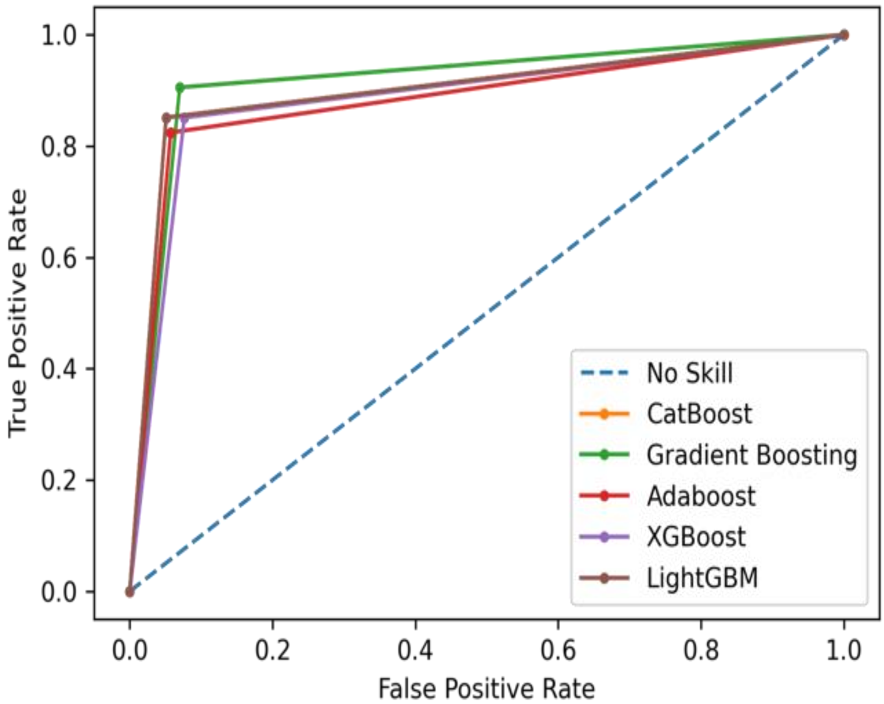

Figure 4. ROC Curves of Boosting Algorithms

5. Conclusions

Diabetes is a very oldest disease in the world. And the use of boosting algorithms in disease classification certainly gives pleasant results. In diabetes disease, deviation of blood sugar from normal can cause various complications. Specifically in the advanced stages of the disease, it causes much more insurmountable health problems. In this study, the prediction of diabetes disease, which is very important for early diagnosis and treatment, was made using boosting algorithms. In this context, the working performances of Gradient Boosting, XGBoost, LightGBM, Adaptive Boosting, and CatBoost algorithms on the diabetes dataset were evaluated. Confusion matrices and roc curves of the algorithms are examined. It has been proven that the CatBoost algorithm has the highest performance rate, with results that are close to each other. To improve the performance of diabetes detection, various machine learning techniques, including voting and bagging, will be used in upcoming studies. Additionally, it aims to explore the extent to which artificially increasing the number of data through the use of techniques like oversampling and SMOTE can improve performance.

Authors' Contributions

HK and GC designed the structure. In addition, HK and GC have realized boosting experiments, theoretical calculations for the data set and wrote the article. Both authors read and approved the final draft.

Competing Interests

The authors declare that they have no competing interests.

References

[1]. International Diabetes Federation, "Home", https://idf.org/, (Accessed Mar. 13, 2023).

[2]. I. Iancu, M. Mota, and E. Iancu, "Method for the analyzing of blood glucose dynamics in diabetes mellitus patients," *2008 IEEE International Conference on Automation, Quality and Testing, Robotics*, Cluj-Napoca, Romania, 2008, pp. 60-65.

[3]. A. Krasteva, V. Panov, A. Krasteva, A. Kisselova, and Z. Krastev, "Oral cavity and systemic diseases—Diabetes Mellitus," *Biotechnology & Biotechnological Equipment*. vol. 25, pp. 2183-2186, 2011.

[4]. A. Bener, E. J. Kim, F. Mutlu, A. Eliyan, H. Delghan, E. Nofal, L. Shalabi, N. Wadi, "Burden of diabetes mellitus attributable to demographic levels in Qatar: an emerging public health problem," *Diabetology Metabolic Syndrome*. vol. 8, no. 4, pp. 216-20, 2014.

[5]. N. Wang, and G. Kang, "A monitoring system for type 2 diabetes mellitus," *IEEE 14th International Conference on e-Health Networking, Applications and Services (Healthcom)*, 2012, pp. 62-67.

[6]. B. Robson, S. Boray, "Data-mining to build a knowledge representation store for clinical decision support. Studies on curation and validation based on machine performance in multiple choice medical licensing examinations," *Computers in Biology and Medicine*. vol. 73, pp.71-93, 2016.

[7]. V. Vijayan, A. Ravikumar. "Study of Data Mining Algorithms for Prediction and Diagnosis of Diabetes Mellitus," *International Journal of Computer Applications*, vol. 95, pp. 12-16, 2014.

[8]. O. S. Lupse, M. Crisan-Vida, L. Stoicu-Tivadar, E. Bernard, "Supporting diagnosis and treatment in medical care based on big data processing", *Studies in Health Technology and Informatics*, vol. 197, pp. 65-9, 2014.

[9]. U. Neumann, N. Genze, D. Heider, "EFS: an ensemble feature selection tool implemented as R-package and web-application," *BioData Mining*, vol. 10, no. 1, 2017.

[10]. Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, H. Tang, "Predicting Diabetes Mellitus With Machine Learning Techniques," *Frontiers in Genetics*. vol. 9, 2018.

[11]. H. Kaur, V. Kumari, "Predictive Modelling and Analytics for Diabetes using a Machine Learning Approach," *Applied Computing and Informatics*, vol. 18, no. 1/2, pp. 90-100, 2018.

[12]. V. Rawat, S. Suryakant, “A Classification System for Diabetic Patients with Machine Learning Techniques,” *International Journal of Mathematical, Engineering and Management Sciences*. vol. 4. pp. 729-744, 2019.
[13]. P. Chen, C. Pan, “Diabetes classification model based on boosting algorithms,” *BMC Bioinformatics*, vol. 19, 2018.
[14]. W. Xu, J. Zhang, Q. Zhang, and X. Wei, "Risk prediction of type II diabetes based on random forest model," *Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*, 2017, pp. 382-386.
[15]. C. Kalaiselvi, G. Nasira, “A New Approach for Diagnosis of Diabetes and Prediction of Cancer Using ANFIS,” *Proceedings - World Congress on Computing and Communication Technologies*, 2014, pp. 188-190.
[16]. J. Pangaribuan, S. Suharjito, “Diagnosis of Diabetes Mellitus Using Extreme Learning Machine,” *International Conference on Information Technology Systems and Innovation, ICITSI 2014 - Proceedings.* 2014.
[17]. J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes, “Using the ADAP learning algorithm to forecast the onset of diabetes mellitus,” *Proceedings of the Annual Symposium on Computer Applications in Medical Care*, 1988, pp. 261-265.
[18]. X. Ying, “Ensemble Learning”, 2014
[19]. P. Bartlett, Y. Freund, W. S. Lee, R. E. Schapire, “Boosting the margin: A new explanation for the effectiveness of voting methods,” *The Annals of Statistics*, vol. 26, no. 5, pp. 1651-1686, 1998.
[20]. P. Harrington “Machine Learning in Action,” Manning, Publications, 2012.
[21]. J. H. Friedman, “Stochastic gradient boosting,” *Computational Statistics & Data Analysis,* vol. 38, pp. 367- 378, 2002.
[22]. S. Soshnikov, C. Lee, V. Vlassov, M. Gaidar, and S. Vladimirov, "A Comparison of Some Predictive Models for Modeling Abortion Rate in Russia," *14th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, 2013, pp. 115-120.
[23]. S. Xiaosong, Y. Cheng, D. Xue, “Classification Algorithm of Urban Point Cloud Data based on LightGBM,” *IOP Conference Series: Materials Science and Engineering*. vol. 631. no.5, 2019.
[24]. Y. Wang, T. Wang, “Application of Improved LightGBM Model in Blood Glucose Prediction,” *Applied Sciences*, vol. 10, no. 9, 2020.
[25]. X. Ma, J. Sha, D. Wang, Y. Yu, Q. Yang, X. Niu, “Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning,” *Electronic Commerce Research and Applications*, vol. 31, pp. 24-39, 2018.
[26]. G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T. Liu, “LightGBM: A Highly Efficient Gradient Boosting Decision Tree,” *31st Conference on Neural Information Processing Systems*, 2017.
[27]. J. Friedman, T. Hastie, and R. Tibshirani, “Additive logistic regression: a statistical view of boosting,” *The annals of statistics*, vol. 28, no. 2, pp. 337-407, 2000.
[28]. H. Alshari, A. Saleh, A. Odabas, “CPU Performansı için Gradyan Artırıcı Karar Ağacı Algoritmalarının Karşılaştırılması,” *Erciyes Üniversitesi Fen Bilimleri Enstitüsü Fen Bilimleri Dergisi*, vol. 37, no. 1, pp. 157-168, 2021.

[29]. J. Hancock, T. Khoshgoftaar, “CatBoost for Big Data: an Interdisciplinary Review”, *Journal of Big Data*, 2020.

[30]. R. E. Schapire, Y. Freund, P. Bartlett, W. S. Lee, “Boosting the margin: A new explanation for the effectiveness of voting methods,” *Machine Learning: Proceedings of 14th Int. Conference*, 1997, pp. 322-330.

[31]. J. Ma, Z. Yu, Q. Yuanhao, J. Xu, Y. Cao, “Application of the XGBoost Machine Learning Method in PM2.5 Prediction: A Case Study of Shanghai,” *Aerosol and Air Quality Research*, vol. 20. no. 1, 2020.

[32]. S. Visa, B. Ramsay, A. Ralescu, E. Knaap. “Confusion Matrix-based Feature Selection,” *22nd Midwest Artificial Intelligence and Cognitive Science Conference*, 2011.

[33]. T. Fawcett, “Introduction to ROC analysis,” *Pattern Recognition Letters*, vol. 27, pp. 861-874, 2006.

[34]. V. Ferraris, “Should We Rely on ROC Curves? – From Submarines to Medical Tests, the Answer Is a Definite Maybe,” *The Journal of Thoracic and Cardiovascular Surgery*, vol. 157, 2018.