Turkish Journal of Electrical Engineering and Computer Sciences

Volume 31 | Number 1 Article 8

1-1-2023

Early diagnosis of pancreatic cancer by machine learning methods using urine biomarker combinations

İREM ACER

FIRAT ORHAN BULUCU

SEMRA İÇER

FATMA LATİFOĞLU

Follow this and additional works at: https://journals.tubitak.gov.tr/elektrik

Part of the Computer Engineering Commons, Computer Sciences Commons, and the Electrical and Computer Engineering Commons

Recommended Citation

ACER, İREM; BULUCU, FIRAT ORHAN; İÇER, SEMRA; and LATİFOĞLU, FATMA (2023) "Early diagnosis of pancreatic cancer by machine learning methods using urine biomarker combinations," *Turkish Journal of Electrical Engineering and Computer Sciences*: Vol. 31: No. 1, Article 8. https://doi.org/10.55730/1300-0632.3974

Available at: https://journals.tubitak.gov.tr/elektrik/vol31/iss1/8

This Article is brought to you for free and open access by TÜBİTAK Academic Journals. It has been accepted for inclusion in Turkish Journal of Electrical Engineering and Computer Sciences by an authorized editor of TÜBİTAK Academic Journals. For more information, please contact academic.publications@tubitak.gov.tr.

Turkish Journal of Electrical Engineering & Computer Sciences

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

Research Article

Turk J Elec Eng & Comp Sci
(2023) 31: 112 – 125
© TÜBİTAK
doi:10.55730/1300-0632.3974

Early diagnosis of pancreatic cancer by machine learning methods using urine biomarker combinations

İrem ACER[1,3], **Fırat ORHANBULUCU**[2,3], **Semra İÇER**[3,*], **Fatma LATİFOĞLU**[3]

[1]Department of Biomedical Device Technology, Kütahya Dumlupınar University, Kütahya, Turkey
[2]Department of Biomedical Engineering, İnönü University, Malatya, Turkey
[3]Department of Biomedical Engineering, Erciyes University, Kayseri, Turkey

Received: 28.04.2022 • **Accepted/Published Online:** 25.10.2022 • **Final Version:** 19.01.2023

Abstract: The most common type of pancreatic cancer is pancreatic ductal adenocarcinoma (PDAC), which accounts for the vast majority of pancreatic cancers. The five-year survival rate for PDAC due to late diagnosis is 9%. Early diagnosed PDAC patients survive longer than patients diagnosed at a more advanced stage. Biomarkers can play an essential role in the early detection of PDAC to assist the health professional. Machine learning and deep learning methods are used with biomarkers obtained in recent studies for diagnostic purposes. In order to increase the survival rates of PDAC patients, early diagnosis of the disease with a noninvasive test is a critical need. Our study offers a promising approach for the early detection of PDAC with noninvasive urinary biomarkers and carbohydrate antigen 19-9 (CA19-9). The Kaggle Urinary Biomarkers for Pancreatic Cancer (2020) open-access dataset consisting of 590 participants was used in this study. Seven machine learning classifiers (support vector machine (SVM), naive Bayes (NB), k-nearest neighbors (kNN), random forest (RF), light gradient boosting machine (LightGBM), AdaBoost, and gradient boosting classifier (GBC)) to detect PDAC disease classifier were used. Binary and multiple classification processes were carried out. Data was validated in our study using 5–10-fold crossvalidation. This study aimed to determine the best machine learning model by analyzing the performance of machine learning models in determining the classes of healthy controls, pancreatic disorders, and patients with PDAC. It is a remarkable finding that ensemble learning models were more successful in all our groups. The most successful classification method in classifying healthy controls and patients with PDAC was CV-10, while the GBC (92.99%) model was (AUC = 0.9761). The most successful classification method in classifying patients with pancreatic disorders and PDAC was CV-10, while the LightGBM (86.37%) model was (AUC = 0.9348). In the classification of healthy controls, pancreatic disorders, and patients with PDAC, the most successful classification method was CV-5, while the GBC (72.91%) model was (AUC = 0.8733).

Key words: Pancreatic cancer, urine biomarker, machine learning, ensemble learning, classification

1. Introduction

The importance of cancer is increasing day by day due to aging and population growth worldwide [1]. The most common type of pancreatic cancer is PDAC, which accounts for the vast majority of pancreatic cancers [2]. The mortality rate is high because it is usually diagnosed when the disease is advanced. The five-year survival rate for PDAC due to late diagnosis is 9% [3]. PDAC is currently the third leading cause of cancer-related death in the United States [4]. In the study by Rahib et al., it is estimated that pancreatic cancer will be the second most lethal type of cancer worldwide by 2030 [5].

[*]Correspondence: ksemra@erciyes.edu.tr

This work is licensed under a Creative Commons Attribution 4.0 International License.

Recent studies have focused on identifying specific biomarkers for PDAC [6–8]. Biomarkers can play an essential role in the early detection of PDAC to assist the health professional. CA19-9 is currently the most routinely used serum biomarker for PDAC. Nevertheless, it does not have the necessary accuracy to detect PDAC by itself [9]. Biomarkers that could potentially improve the earlier detection of PDAC in combination or alone with CA19-9 of the biomarkers examined in studies were mentioned [6, 10]. Advances in data analysis give good results in classifying cancer types using artificial intelligence. Artificial intelligence applications in pancreatic cancer can improve early diagnosis, treatment, and survival rates [11]. Diagnosis of pancreatic cancer has mostly been made through personal health records and imaging methods [12–14]. Machine learning and deep learning methods are used with biomarkers obtained in the recent studies for diagnostic purposes. Almeida et al. created an artificial neural network model to predict PDAC based on gene expression. They classified network tumor samples with an f1 score of 0.83 for normal samples and 0.88 for PDAC samples [11]. Honda et al. correctly diagnosed 97.2% of cancer patients in the training cohort and 94.4% of healthy controls using the SVM model based on four plasma proteins. In the test set, it was detected 91% correctly [15]. Hsieh et al. developed a prediction model for pancreatic cancer in patients with type 2 diabetes using logistic regression and artificial neural network models. As a result of their studies, they stated that the logistic regression (LR), (AUC = 0.727) model was more successful than the artificial neural network (AUC = 0.605) model [16]. In the study of Khatri et al., gene-based PDAC classifier performance using a support vector machine showed a sensitivity (Sens.) of 0.88 and a specificity (Spec.) of 0.80 for blood datasets, a sensitivity of about 0.94 and a specificity of 0.97 for tissue datasets [17]. Radon et al. analyzed urine samples and identified three biomarkers for early detection of radon PDAC [8]. They then examined a new biomarker with additional urine samples collected [18]. They stated that substitution of REG1B (Sens. and Spec. >85%) biomarker instead of REG1A biomarker (LYVE-1+REG1A+TFF1 Sens. = 76.9%, Spec. = 89.8%) increased the detection performance of resectable PDAC. Their study applied logistic regression to CA19-9 and its combinations (LYVE1, REG1, TFF1, creatinine, and age) by separating training and test data in a 1:1 ratio. The success of urinary biomarkers in the study was confirmed with many samples.

In order to increase the survival rates of PDAC patients, early diagnosis of the disease with a noninvasive test is a critical need. Our study offers a promising approach for the early detection of PDAC with noninvasive urinary biomarkers and CA19-9. To the best of our knowledge, no comparative classification studies have been conducted using various machine learning models with urine biomarkers for PDAC detection. The use of machine learning methods in PDAC patients can increase early diagnosis and survival. This study aimed to accurately diagnose PDAC patients using urine biomarkers and various machine learning methods. The main outputs of this study can be listed as follows:

(1) To detect PDAC disease, binary, and multiple classifications were made with seven machine learning models.

(2) The classification model was designed using SVM, NB, KNN, RF, LightGBM, AdaBoost, and GBC for healthy controls, pancreatic disorders, and PDAC patients.

(3) While the most successful result in the classification of the three groups is CV-5, it is seen that the GBC (72.91%) classifier is more successful than the other models according to the accuracy rate metric.

(4) While the most successful result is CV-10 in classifying healthy controls and patients with PDAC, the GBC (92.99%) classifier seems to be more successful than other models according to the accuracy rate metric.

(5) While the most successful result is CV-10 in classifying patients with pancreatic disorders and patients

with PDAC, it is seen that the LightGBM (86.37%) classifier is more successful than other models according to the accuracy rate metric.

In the second part of the study, information about the data set is given, and the methods mentioned above are explained. In the third chapter, the findings obtained from these methods are presented. In the last chapter, the results obtained in the study are discussed together with the literature.

2. Materials and methods

2.1. Data set

The Kaggle Urinary Biomarkers for Pancreatic Cancer (2020) open-access dataset was used in this study [18]. The data set used consists of 590 participants. The dataset includes three groups: healthy controls, patients with noncancerous pancreatic disorders such as chronic pancreatitis, and patients with pancreatic ductal adenocarcinoma. The distribution of the data used is shown in Table 1.

Table 1. Sample number distribution.

Label	*Classes*	*Number of data*	Sex	Mean age (±SD)
1	Healthy controls	183	115 Female (F)-68 Male (M)	56.33 (± SD 12. 20)
2	Pancreatic patients	208	101 F-107 M	54.70 (± SD 13.34)
3	Pancreatic ductal adenocarcinoma	199	83 F-116 M	66.18 (± SD 10.51)
				*SD: Standard deviation

Eight features were used in the dataset, including age, sex, five urinary biomarkers, and the FDA-approved CA19-9 plasma biomarker. The graph showing the correlation between the features is shown in Figure 1. Figure 1 expresses the degree of dependence between a variable used in the study and every other variable. To best generalize the dataset, it is desired that the dataset covers most of the feature space. This means the selection of dissimilar features. As seen in Figure 1, the features used in the study were chosen in a way to generalize the data set, that is, features that were not similar to each other were selected. According to the color scale, the correlation relationship between the columns close to blue is high, while the correlation relationship decreases gradually in the colors toward red.

The key features are four urinary biomarkers: LYVE1, REG1B, REG1A, TFF1 and urine creatinine.

- Creatinine is a protein often used as an indicator of kidney function.
- LYVE1 is lymphatic vessel endothelial hyaluronan receptor 1, a protein that may play a role in tumor metastasis.
- The REG1A (regenerating islet-derived 1 alpha) gene product, a member of the REG glycoprotein family, has been associated with pancreatitis [8, 19].
- REG1B (regenerating islet-derived 1 beta) is a protein that may be associated with pancreatic regeneration [20].
- TFF1 is trefoil factor 1, which plays a role in protecting epithelial cells and the regeneration and repair of the urinary tract [21].

2.2. Classifications

Classification is a supervised learning approach where the program learns from the given data entry and then uses this learning to classify new observations according to their characteristics. The increase in medical data

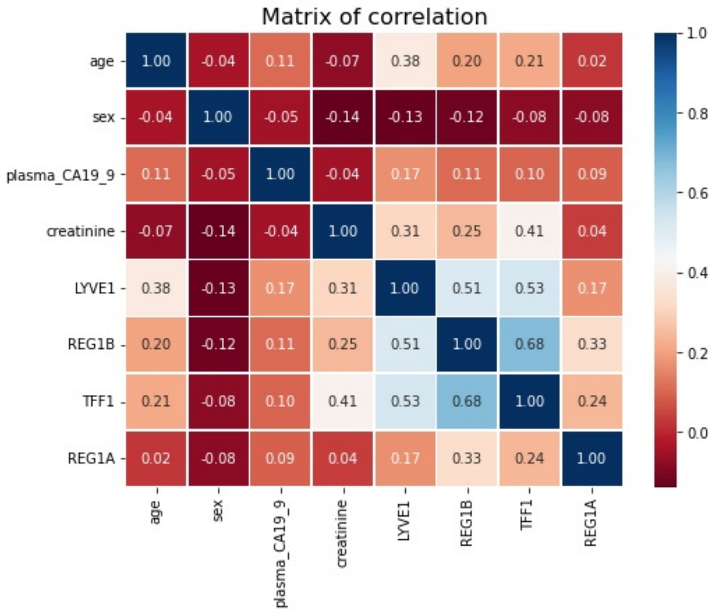


Figure 1. Feature correlation matrix.

also increases the need for effective classifiers in the field of machine learning. In this study, healthy controls-PDAC patients, patients with pancreatic disease-PDAC patients, and all three groups were classified using the most widely used machine learning algorithms (SVM, KNN, NB) and ensemble learning classifiers (RF, GBC, Light GBM, AdaBoost). Classification processes were carried out using the Python programming language. Traditional algorithms such as KNN, SVM, and NB have a lower performance, while ensemble algorithms have a relatively higher accuracy. As a preprocessing step, missing values in the data were filled with the mean. Data was validated in our study using 5–10-fold crossvalidation. The methodology applied in our study is explained in Figure 2. Information about the classifiers used in the study is as follows:

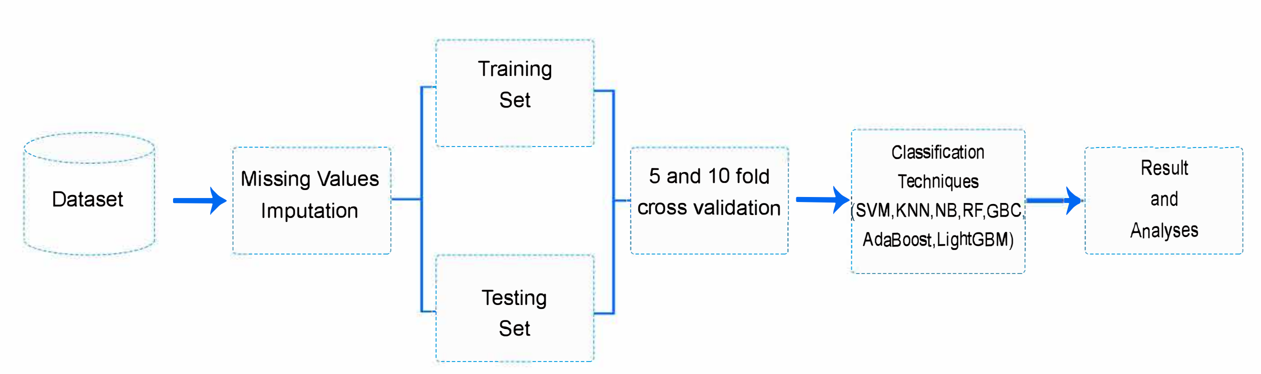

Dataset
Missing Values Imputation
Training Set
Testing Set
5 and 10 fold cross validation
Classification Techniques (SVM,KNN,NB,RF,GBC, AdaBoost,LightGBM)
Result and Analyses

Figure 2. The methodology of this research.

SVM: It was first proposed by Cortes and Vapnik in 1995 [22]. It can be used in classification and regression problems. Its development in both practical and theoretical fields is related to its solid mathematical background, widely accessible software applications, and ease of use [23]. Its purpose is to find a hyperplane that can separate the class of given features with a maximal margin between the categories. The SVM algorithm steps in on how to draw this boundary [24]. The line's location is intended to be at the furthest point between the two groups, allowing the two classes to be separated as widely as possible.

KNN: It was first proposed by Fix and Hodges in 1951. It is a simple and effective method for classification [25]. It finds the nearest neighbors among the variables for classification. The most important point is the distance between the data points. k is a parameter used to determine distances. To classify the data set, there are nearest neighbors between the variables [26].

NB: It was introduced by Thomas Bayes in 1812 [27]. The algorithm calculates the probability of each situation for an element and classifies it according to the highest probability value. Very successful classifications can be made with a small number of training data.

Ensemble learning can provide more consistent and accurate predictions than a single prediction model. It is a current field of machine learning research that first trains a set of individual learners and then combines them with some strategies to improve overall performance. The goal is to improve accuracy and significantly reduce classification errors by combining estimates [28, 29]. In this study, RF, Light GBM, AdaBoost and GBC were used as ensemble learning classifiers.

RF: RF is one of the best known, widely used techniques for classification in a variety of applications. It is an ensemble learning method consisting of several decision tree classifiers to obtain the optimal solution and increase the classification value [30].

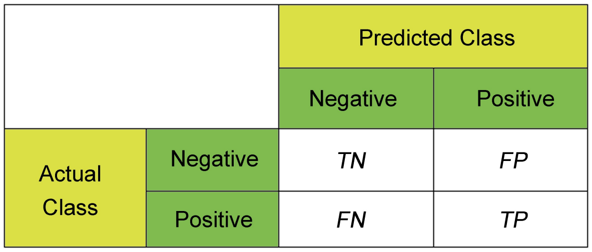
Light GBM: LightGBM is a decision tree-based ensemble learning method which is a vertical tree growth procedure. It generalizes well and combines the predictions of multiple decision trees to make the final prediction [31].

AdaBoost: In the AdaBoost algorithm, the training set is first trained with a weak learner. At each posttraining stage, it is retrained by rerunning the classifier by increasing the weight of the wrong predictions made as a result of the previous stage. With these processes, it is aimed to focus on the wrong predictions and to increase the accuracy rate of the created model in classification [32].

GBC: GBC is a decision tree-based estimation algorithm used for classification problems. It uses an amplification technique that combines a series of weak tutorials to build a strong model. It is specified with a loss function and a weak base classifier. The goal here is to estimate an aggregate model that minimizes the loss function [33].

2.3. Performance measures

The performances of the classification algorithms used in the study were calculated according to the confusion matrix given below in Figure 3 for standard performance criteria such as accuracy (Acc.), recall, precision (Prec.), and area under the curve (AUC) values.



		Predicted Class	
		Negative	Positive
Actual Class	Negative	*TN*	*FP*
	Positive	*FN*	*TP*

Figure 3. Confusion matrix.

Accuracy means the percentage of samples that make the correct guess among all available samples. The recall is indicated how successfully positive states were predicted. Precision is defined as a corrective prediction in the positive class of samples. In Equations 1 − 3, true positive (TP) means that the true patient class and the class found by the algorithm match correctly, while false positive (FP) means that the nonpatient class is predicted as the patient. Likewise, true negative (TN) means that the class that is not sick is predicted correctly by the algorithm, while a false negative (FN) means that the class that is sick is incorrectly predicted as the class that is not sick. The receiver operating characteristic (ROC) curve is used to describe the accuracy of the diagnostic test itself and to allow a reliable comparison between tests. AUC is an indicator of how well machine learning models can distinguish classes. As the area under the curve increases, the discrimination performance between classes increases. Kappa statistical value is a statistical criterion used to test the reliability of algorithms.

Crossvalidation was set as 5 and 10 in the classification processes and reported according to the highest value. Cohen's kappa coefficient is a statistical measure used to test the reliability of two raters. It was stated that the closer the kappa statistical value was to 0, the worse the agreement, and the closer to 1, the better the agreement [34].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \times 100 \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \times 100 \tag{3}$$

3. Results

In this study, pancreatic cancer detection was performed with SVM, KNN, NB, and ensemble learning (RF, Light GBM, AdaBoost ve GBC) machine learning algorithms on the Kaggle Urinary Biomarkers for Pancreatic Cancer (2020) open-access dataset. The results of the machine learning techniques used are given comparatively. The classification results of the study as healthy controls and patients with pancreatic ductal adenocarcinoma, patients with noncancerous pancreatic disorders such as chronic pancreatitis, and patients with pancreatic ductal adenocarcinoma and 3 groups together are given below. All models were trained with 5- and 10-fold crossvalidation techniques, and results of accuracy, recall, precision, and AUC scores were obtained.

3.1. Classification of healthy controls and patients with PDAC

The evaluation of the results of the classification algorithms used is given in Table 2. It has been seen that ensemble learning algorithms are more successful in classification than classical classification algorithms. In the results, it is seen that the GBC classifier is more successful than other models according to the accuracy rate metric from seven different machine learning models trained with 5- and 10-fold crossvalidation techniques in classifying healthy controls and patients with PDAC. The most successful result was given by the GBC (92.99%) model when CV-10.

3.2. Classification of patients with pancreatic disorders and PDAC

The evaluation of the results of the classification algorithms used is given in Table 3. As in the other group, ensemble learning algorithms were more successful than classical classification algorithms. The results show that the LightGBM classifier is more successful than the other models according to the accuracy rate metric from 7 different machine learning models trained with the 5- and 10-fold crossvalidation technique in the classification of patients with pancreatic cancer disorders and PDAC. The LightGBM (86.37%) model gave the most successful result when CV-10.

Table 2. Classification results of healthy controls and patients with PDAC.

CV			*CV-5*					*CV-10*		
Model	Acc. (%)	AUC	Recall	Prec.	Kappa	Acc. (%)	AUC	Recall	Prec.	Kappa
GBC	92.10	0.9734	0.9335	0.9242	0.8408	**92.99**	**0.9761**	**0.9245**	**0.9368**	**0.8598**
LightGBM	92.10	0.9742	0.9213	0.9347	0.8412	92.13	0.9731	0.9134	0.9315	0.8425
AdaBoost	89.45	0.9620	0.8963	0.9112	0.7879	91.24	0.9615	0.9013	0.9233	0.8248
RF	89.46	0.9578	0.9029	0.9041	0.7882	90.38	0.9726	0.8961	0.9142	0.8076
KNN	89.14	0.9292	0.8904	0.9107	0.7816	86.57	0.9252	0.85	0.8812	0.7317
NB	77.34	0.9218	0.6298	0.9341	0.5569	79.88	0.9274	0.6474	0.9309	0.5987
SVM	71.12	0.00	0.7092	0.8068	0.4145	68.20	0.00	0.7340	0.7147	0.3626

Table 3. Classification results of patients with pancreatic disorders and PDAC.

CV			*CV-5*					*CV-10*		
Model	Acc. (%)	AUC	Recall	Prec.	Kappa	Acc. (%)	AUC	Recall	Prec.	Kappa
GBC	81.25	0.9034	0.7942	0.8215	0.6232	84.97	0.9205	0.8392	0.8486	0.6983
LightGBM	83.72	0.9144	0.8329	0.8381	0.6733	**86.37**	**0.9348**	**0.8451**	**0.8759**	**0.7261**
AdaBoost	78.47	0.8650	0.7562	0.7952	0.5668	78.70	0.8671	0.7768	0.7807	0.5725
RF	80.00	0.8886	0.8071	0.7916	0.5988	81.41	0.90	0.7748	0.8377	0.6253
KNN	76.63	0.8186	0.7246	0.7724	0.5296	78.20	0.8206	0.7771	0.7725	0.5625
NB	71.06	0.85	0.4946	0.8197	0.4098	69.12	0.8278	0.4448	0.8667	0.3677
SVM	62.71	0.00	0.7325	0.6055	0.2594	64.21	0.00	0.5725	0.7131	0.2807

3.3. Classification of healthy controls, pancreatic disorders, and patients with PDAC

The evaluation of the results of the classification algorithms used is given in Table 4. It was observed that ensemble learning algorithms were more successful in classification than classical classification algorithms in all groups. In the results, it is seen that the GBC classifier is more successful than other models according to the accuracy rate metric from 7 different machine learning models trained with 5 crossvalidation techniques in classifying healthy controls, pancreatic disorders, and patients with PDAC. The GBC model (72.91%) gave the most successful result when CV-5. Figure 4 shows the GBC ROC curve when CV-5, which is the most successful result. In Figure 5, the feature importance graph in this classification is given. According to the accuracy rate metric from 7 different machine learning models trained with 10 crossvalidation techniques, the LightGBC classifier was found to be more successful than other models in the classification of healthy controls, pancreatic disorders, and patients with PDAC.

Table 4. Classification results of healthy controls, pancreatic disorders, and patients with PDAC.

CV			*CV-5*					*CV-10*		
Model	Acc. (%)	AUC	Recall	Prec.	Kappa	Acc. (%)	AUC	Recall	Prec.	Kappa
GBC	**72.91**	**0.8733**	**0.7788**	**0.7310**	**0.5954**	70.83	0.8698	0.7108	0.7125	0.5614
LightGBM	69.31	0.8793	0.6902	0.6983	0.5377	72.33	0.8717	0.7251	0.7302	0.5840
AdaBoost	61.69	0.7353	0.6170	0.6259	0.4233	55.95	0.7210	0.5594	0.5772	0.3353
RF	68.68	0.8584	0.6857	0.3890	0.5282	67.82	0.8582	0.6794	0.6831	0.5156
KNN	64.45	0.7955	0.6471	0.6544	0.4664	62.94	0.7805	0.6330	0.6402	0.4444
NB	52.35	0.7459	0.5424	0.5752	0.2979	51.79	0.7305	0.5348	0.5733	0.2872
SVM	43.87	0.00	0.44	0.4233	0.1559	45.79	0.00	0.4677	0.3895	0.1933

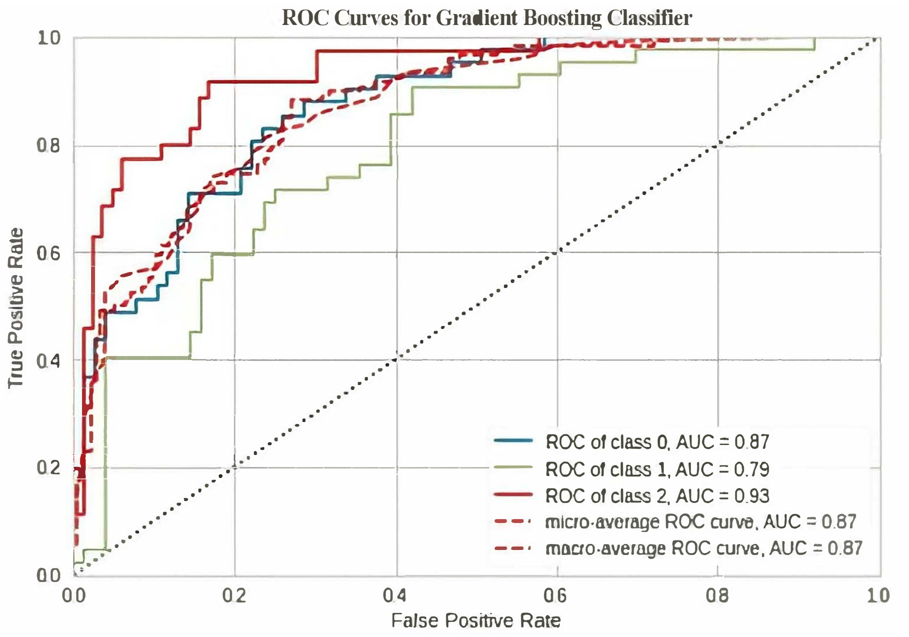


Figure 4. Three-class gradient boosting classifier ROC curve (CV-5).

4. Conclusion and discussion

PDAC is the most common type of pancreatic cancer. Diagnosing PDAC at an early stage is very important for increasing the survival rate. It is important to distinguish PDAC from other nonmalignant benign gastrointestinal diseases. Early diagnosed PDAC patients survive longer than patients diagnosed at a more advanced stage. Therefore, it is essential to diagnose the disease at an early stage. The lack of clinically approved biomarkers with the required performance criteria is associated with the lack of early detection of PDAC [35]. Evaluation of biomarkers in body fluids such as blood, tissue, urine, saliva, or pancreatic juice is partially invasive for diagnosing PDAC. However, it can enable early diagnosis and intervention of PDAC and is somewhat less expensive compared to existing methods. Another challenge with early diagnosis is differentiating PDAC from other nonmalignant benign gastrointestinal diseases such as chronic pancreatitis. It is difficult to make a differential diagnosis because the imaging and clinical patterns are generally similar [36, 37]. This study aimed to determine the best machine learning classifier by analyzing the performance of machine learning models in determining the classes of healthy controls, pancreatic disorders, and patients with PDAC and to contribute to early diagnosis.

Ensemble learning algorithms showed higher success in all classifications than classical classification algorithms. This result parallels with the research result of Gupta et al. [38] on cancer diagnosis. Urine biomarkers used in combination with CA19-9 are a promising approach for earlier detection of PDAC. In one study, it was stated that a combination of urinary biomarkers (LYVE-1, REG1A, and TFF1) could be used in the diagnosis of pancreatic cancer [8]. In a later study, it was stated that using REG1B instead of REG1A increased the detection performance of PDAC [18]. When we look at the feature importance graph in Figure 5, it is seen that REG1B is a more important feature than REG1A for classification.

The most successful classification method in classifying healthy controls and patients with PDAC was CV-10, while the GBC (Acc. = 92.99%) model was (AUC = 0.976). In Figure 6, the results of all classification algorithms according to the accuracy metric are given in the graph.

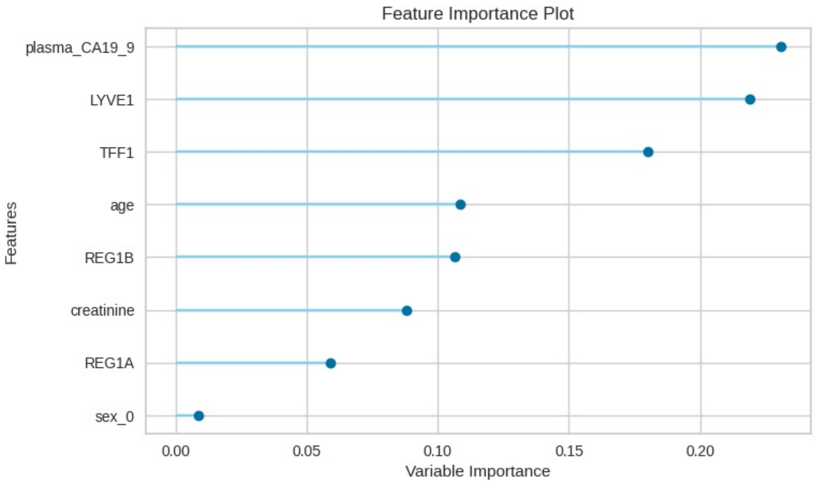


Figure 5. Feature importance graph.

The most successful classification method in classifying patients with pancreatic disorders and PDAC was CV-10, while the LightGBM (Acc. = 86.37%) model was (AUC = 0.935). In Figure 7, the results of all classification algorithms according to the accuracy metric are given in the graph.

In the classification of healthy controls, pancreatic disorders, and patients with PDAC, the most successful classification method was CV-5, while the GBC (Acc. = 72.91%) model was (AUC = 0.873). In Figure 8, the results of all classification algorithms according to the accuracy metric are given in the graph.

CA19-9 has been studied as a biomarker in many cancers since its discovery, but the highest sensitivity and specificity were obtained in pancreatic cancer patients [39, 40]. Recent studies show that the combination of C19-9 and markers can improve diagnostic accuracy [18, 41]. As seen in Figure 8 as a result of classification, the most important feature is CA 19-9, which has been studied extensively and is the gold standard [10]. The next important features are LYVE-1, TFF1, age, REG1B, creatinine, REG1A, and sex. Less than 10% of total pancreatic cancer cases occur under the age of 55. The average age of onset of pancreatic cancer is 71 [42]. Therefore, age is an important feature in classification. Studies in the literature using biomarkers obtained from various clinical data and different classification algorithms to detect pancreatic cancer were reviewed. In Table

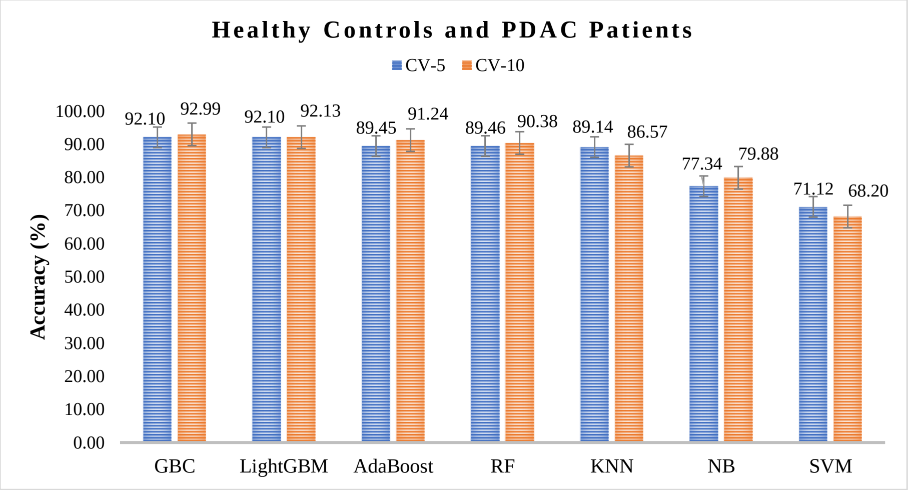


Figure 6. Accuracy of each classification algorithm (healthy controls and patients).

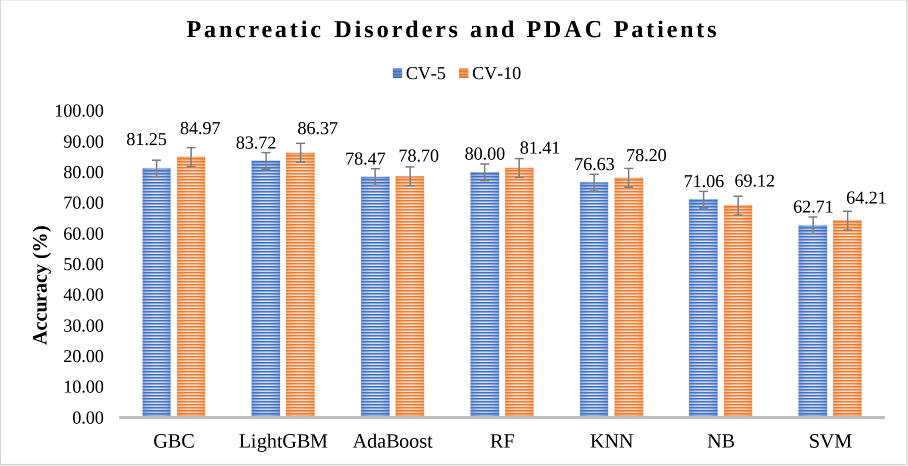


Figure 7. Accuracy of each classification algorithm (pancreatic disorders and PDAC patients).

5, the data set in this study [18, 43] and the results obtained in studies using biomarkers with different data sets are given.

Detecting PDAC at an early stage for effective treatment is currently the most important challenge. In this study, the classification success of machine learning models for early-stage PDAC diagnosis with the combination of invasive urine biomarkers and CA19-9 was analyzed. It is aimed to determine the best classifying machine learning classifiers among the models used. It is a remarkable finding that ensemble learning classifiers were more successful in all our groups. The successful results of ensemble learning classifiers support the idea stated in [30] that ensemble methods such as GBC, LightGBM give superior results compared to other widely used machine learning algorithms. As a result, healthy controls, pancreatic disorders, and individuals with PDAC disease have been successfully classified. Further prospective validation studies are needed before machine learning can be applied to clinical use for PDAC diagnosis with biomarkers.

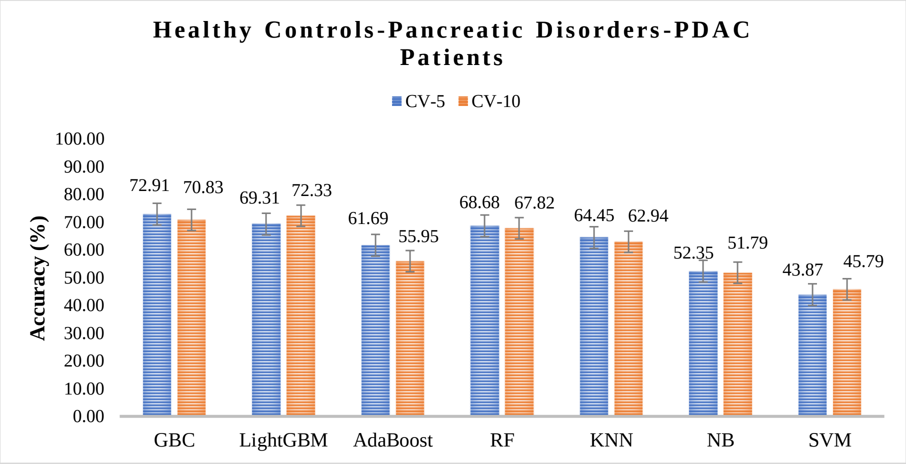


Figure 8. Accuracy of each classification algorithm (healthy controls-pancreatic disorders-PDAC patients).

Table 5. Various biomarkers, classifier models and evaluation metrics of different classification algorithms are given in the literature to detect pancreatic cancer.

Study	*Biomarkers*	*Best classifier models*	*Evaluation results%*

Almeida et al. [11]	Genetic biomarkers	Artificial neural networks (ANN)	Acc. = 85.71, Sens. = 87.6
			Spec. = 83.1

Honda et al. [15]	Plasma biomarkers	SVM	Acc. = 91

Khatri et al. [17]	Tissue and blood biomarkers	SVM	**Tissue**: Sens. = 88
			Spec. = 80
			Blood: Sens. = 94
			Spec. = 97

Debernardi et al. [18]	Urine biomarkers	LR	Control-PDAC: AUC = 99.2 Benign-PDAC: AUC = 91.9
Narayanan et al. [43]	Urine biomarkers	RF and LR	Prec. = 91, F1-score = 91
This study	Urine biomarkers	GBC-Light GBM	**Control-PDAC** (GBC, Acc. = 92.99 AUC = 97.61, Prec. = 93.68) **PD-PDAC** (Light GBM, Acc. = 86.37 AUC = 93.48, Prec. = 87.59) **Control-PD-PDAC** (GBC, Acc. = 72.91 AUC = 87.33, Prec. = 73.10)

References

[1] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A cancer journal for clinicians 2018; 68 (6): 394-424. https://doi.org/10.3322/caac.21492

[2] Franck C, Müller C, Rosania R, Croner RS, Pech M et al. Advanced pancreatic ductal adenocarcinoma: moving forward. Cancers 2020; 12 (7): 1955. https://doi.org/10.3390/cancers12071955

[3] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. CA: A cancer journal for clinicians 2019 Jan; 69 (1): 7-34. https://doi.org/10.3322/caac.21551

[4] Korc M. Pathogenesis of pancreatic cancer-related diabetes mellitus: Quo Vadis? Pancreas 2019; 48 (5): 594.

[5] Rahib L, Smith BD, Aizenberg R, Rosenzweig AB, Fleshman JM et al. Projecting cancer incidence and deaths to 2030: the unexpected burden of thyroid, liver, and pancreas cancers in the United States. Cancer research 2014; 74 (11): 2913-2921. https://doi.org/10.1158/0008-5472.CAN-14-0155

[6] Brezgyte G, Shah V, Jach D, Crnogorac-Jurcevic T. Non-Invasive Biomarkers for Earlier Detection of Pancreatic Cancer—A Comprehensive Review. Cancers 2021 ;13 (11): 2722. https://doi.org/10.3390/cancers13112722

[7] Young MR, Wagner PD, Ghosh S, Rinaudo JA, Baker SG et al. Validation of biomarkers for early detection of pancreatic cancer: summary of the alliance of pancreatic cancer consortia for biomarkers for early detection workshop. Pancreas 2018; 47 (2): 135-141.

[8] Radon TP, Massat NJ, Jones R, Alrawashdeh W, Dumartin L et al. Identification of a three-biomarker panel in urine for early detection of pancreatic adenocarcinoma. Clinical Cancer Research 2015 ; 21 (15): 3512-3521.

[9] Pereira SP, Oldfield L, Ney A, Hart PA, Keane MG et al. Early detection of pancreatic cancer. The lancet Gastroenterology & hepatology 2020;5 (7): 698-710.

[10] Kriz D, Ansari D, Andersson R. Potential biomarkers for early detection of pancreatic ductal adenocarcinoma. Clinical and Translational Oncology 2020;(12): 2170-2174. https://doi.org/10.1007/s12094-020-02372-0

[11] Almeida PP, Cardoso CP, de Freitas LM. PDAC-ANN: an artificial neural network to predict pancreatic ductal adenocarcinoma based on gene expression. BMC cancer 2020; 20 (1): 1-11.

[12] Muhammad W, Hart GR, Nartowt B, Farrell JJ, Johung K et al. Pancreatic cancer prediction through an artificial neural network. Frontiers in Artificial Intelligence 2019; 2 (2): 1-10. https://doi.org/10.3389/frai.2019.00002

[13] Barat M, Chassagnon G, Dohan A, Gaujoux S, Coriat R et al. Artificial intelligence: a critical review of current applications in pancreatic imaging. Japanese Journal of Radiology 2021; 39 (6) : 514-523. https://doi.org/10.1007/s11604-021-01102-y

[14] Kaissis G, Ziegelmayer S, Lohöfer F, Algül H, Eiber M et al. A machine learning model for the prediction of survival and tumor subtype in pancreatic ductal adenocarcinoma from preoperative diffusion-weighted imaging. European radiology experimental 2019; 3 (1): 1-9. https://doi.org/10.1186/s41747-019-0119-0

[15] Honda K, Hayashida Y, Umaki T, Okusaka T, Kosuge T et al. Possible detection of pancreatic cancer by plasma protein profiling. Cancer research 2005 ; 65 (22): 10613-10622.

[16] Hsieh MH, Sun LM, Lin CL, Hsieh MJ, Hsu CY et al. Development of a prediction model for pancreatic cancer in patients with type 2 diabetes using logistic regression and artificial neural network models. Cancer management and research 2018; 10: 6317-6324. https://doi.org/10.2147/CMAR.S180791

[17] Khatri I, Bhasin MK. A Transcriptomics-Based Meta-Analysis Combined With Machine Learning Identifies a Secretory Biomarker Panel for Diagnosis of Pancreatic Adenocarcinoma. Frontiers in genetics 2020; 11: 572284. https://doi.org/10.3389/fgene.2020.572284

[18] Debernardi S, O'Brien H, Algahmdi AS, Malats N, Stewart GD et al. A combination of urinary biomarker panel and PancRISK score for earlier detection of pancreatic cancer: A case–control study. PLoS Medicine 2020 ; 17 (12): e1003489. https://doi.org/10.1371/journal.pmed.1003489

[19] Zhang YW, Ding LS, Lai MD. Reg gene family and human diseases. World journal of gastroenterology: WJG 2003 Dec 15; 9 (12): 2635-2641. https://doi.org/10.3748/wjg.v9.i12.2635

[20] Makawita S, Dimitromanolakis A, Soosaipillai A, Soleas I, Chan A et al. Validation of four candidate pancreatic cancer serological biomarkers that improve the performance of CA19. 9. BMC cancer 2013; 13 (1): 1-11. https://doi.org/10.1186/1471-2407-13-404

[21] O'Neill RS, Stoita A. Biomarkers in the diagnosis of pancreatic cancer: Are we closer to finding the golden ticket?. World Journal of Gastroenterology 2021 ; 27 (26): 4045-4087.

[22] Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995 ; 20 (3): 273-297. https://doi.org/10.1007/BF00994018

[23] Hearst MA, Dumais ST, Osuna E, Platt J, Scholkopf B. Support vector machines. IEEE Intelligent Systems and their applications 1998; 13 (4): 18-28. https://doi.org/10.1109/5254.708428

[24] Chang CC, Lin CJ. LIBSVM: A library for support vector machines. ACM transactions on intelligent systems and technology (TIST) 2011; 2 (3): 1-27. https://doi.org/10.1145/1961189.1961199

[25] Jiang S, Pang G, Wu M, Kuang L. An improved K-nearest-neighbor algorithm for text categorization. Expert Systems with Applications 2012 ; 39 (1): 1503-1509. https://doi.org/10.1016/j.eswa.2011.08.040

[26] Xing W, Bei Y. Medical health big data classification based on KNN classification algorithm. IEEE Access 2019 ; 8: 28808-28819.

[27] Langley P, Iba W, Thompson K. An analysis of Bayesian classifiers. InAaai 1992 ; 90: 223-228.

[28] Zhou ZH. Machine learning: Ensemble learning. 1st ed. China: Springer Singapore Press 2021: pp. 181-210. https://doi.org/10.1007/978-981-15-1967-3_8

[29] Verma A, Mehta S. A comparative study of ensemble learning methods for classification in bioinformatics. In 2017 7th International Conference on Cloud Computing, Data Science & Engineering-Confluence 2017: 155-158. https://doi.org/10.1109/CONFLUENCE.2017.7943141

[30] Jafarzadeh H, Mahdianpari M, Gill E, Mohammadimanesh F, Homayouni S. Bagging and boosting ensemble classifiers for classification of multispectral, hyperspectral and PolSAR data: a comparative evaluation. Remote Sensing 2021; 13 (21): 4405. https://doi.org/10.3390/rs13214405

[31] Alzamzami F, Hoda M, El Saddik A. Light gradient boosting machine for general sentiment classification on short texts: a comparative evaluation. IEEE access 2020; 8: 101840-101858.

[32] Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences 1997; 55 (1): 119-139. https://doi.org/10.1006/jcss.1997.1504

[33] Bahad P, Saxena P. Study of adaboost and gradient boosting algorithms for predictive analytics. In International Conference on Intelligent Computing and Smart Communication 2019; Singapore 2020: 235-244.

[34] Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Physical therapy 2005 ; 85 (3): 257-268. https://doi.org/10.1093/ptj/85.3.257

[35] Singhi AD, Koay EJ, Chari ST, Maitra A. Early detection of pancreatic cancer: opportunities and challenges. Gastroenterology 2019 ; 156 (7): 2024-2040. https://doi.org/10.1053/j.gastro.2019.01.259

[36] Brand RE, Matamoros A. Imaging techniques in the evaluation of adenocarcinoma of the pancreas. Digestive Diseases 1998; 16 (4): 242-252. https://doi.org/10.1159/000016872

[37] De La Cruz MS, Young AP, Ruffin MT. Diagnosis and management of pancreatic cancer. American family physician 2014 ; 89 (8): 626-632.

[38] Gupta S, Gupta MK. A comprehensive data-level investigation of cancer diagnosis on imbalanced data. Computational Intelligence 2022 ; 38 (1): 156-186. https://doi.org/10.1111/coin.12452

[39] Goonetilleke KS, Siriwardena AK. Systematic review of carbohydrate antigen (CA 19-9) as a biochemical marker in the diagnosis of pancreatic cancer. European Journal of Surgical Oncology (EJSO) 2007; 33 (3): 266-270. https://doi.org/10.1016/j.ejso.2006.10.004

[40] Azizian A, Rühlmann F, Krause T, Bernhardt M, Jo P et al. CA19-9 for detecting recurrence of pancreatic cancer. Scientific reports 2020 ; 10 (1): 1-10. https://doi.org/10.1038/s41598-020-57930-x

[41] Majumder S, Taylor WR, Foote PH, Berger CK, Wu CW et al. High detection rates of pancreatic cancer across stages by plasma assay of novel methylated DNA markers and CA19-9. Clinical Cancer Research 2021; 27 (9): 2523-2532. https://doi.org/10.1158/1078-0432.CCR-20-0235

[42] Yadav D, Lowenfels AB. The epidemiology of pancreatitis and pancreatic cancer. Gastroenterology 2013 ; 144 (6): 1252-1261. https://doi.org/10.1053/j.gastro.2013.01.068

[43] Narayanan S, Balamurugan NM, Maithili K, Palas PB. Leveraging Machine Learning Methods for Multiple Disease Prediction using Python ML Libraries and Flask API. In IEEE 2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC) 2022: 694-701.