BMC Medical Informatics and Decision Making

RESEARCH

Open Access

Data privacy-aware machine learning approach in pancreatic cancer diagnosis

Ömer Faruk AKMEŞE[1*]

Abstract

Problem Pancreatic ductal adenocarcinoma (PDAC) is considered a highly lethal cancer due to its advanced stage diagnosis. The five-year survival rate after diagnosis is less than 10%. However, if diagnosed early, the five-year survival rate can reach up to 70%. Early diagnosis of PDAC can aid treatment and improve survival rates by taking necessary precautions. The challenge is to develop a reliable, data privacy-aware machine learning approach that can accurately diagnose pancreatic cancer with biomarkers.

Aim The study aims to diagnose a patient's pancreatic cancer while ensuring the confidentiality of patient records. In addition, the study aims to guide researchers and clinicians in developing innovative methods for diagnosing pancreatic cancer.

Methods Machine learning, a branch of artificial intelligence, can identify patterns by analyzing large datasets. The study pre-processed a dataset containing urine biomarkers with operations such as filling in missing values, cleaning outliers, and feature selection. The data was encrypted using the Fernet encryption algorithm to ensure confidentiality. Ten separate machine learning models were applied to predict individuals with PDAC. Performance metrics such as F1 score, recall, precision, and accuracy were used in the modeling process.

Results Among the 590 clinical records analyzed, 199 (33.7%) belonged to patients with pancreatic cancer, 208 (35.3%) to patients with non-cancerous pancreatic disorders (such as benign hepatobiliary disease), and 183 (31%) to healthy individuals. The LGBM algorithm showed the highest efficiency by achieving an accuracy of 98.8%. The accuracy of the other algorithms ranged from 98 to 86%. In order to understand which features are more critical and which data the model is based on, the analysis found that the features "plasma_CA19_9", REG1A, TFF1, and LYVE1 have high importance levels. The LIME analysis also analyzed which features of the model are important in the decision-making process.

Conclusions This research outlines a data privacy-aware machine learning tool for predicting PDAC. The results show that a promising approach can be presented for clinical application. Future research should expand the dataset and focus on validation by applying it to various populations.

Keywords Pancreas cancer, Diagnosis, Machine learning, Classification, Data privacy-aware

Introduction

Pancreatic cancer has high mortality rates due to its often advanced stage at diagnosis. The incidence of pancreatic cancer ranks last among the top ten cancers. Nonetheless, the prognosis for survival remains significantly poor [1]. In 2022, cancer statistics indicate that pancreatic cancer is the third leading cause of cancer-related deaths

*Correspondence:
Ömer Faruk AKMEŞE
ofarukakmese@hitit.edu.tr
[1] Department of Computer Engineering, Hitit University Çorum, Çorum 19030, Türkiye

worldwide [2]. By 2030, it is projected that pancreatic cancer will be the second most prevalent cause of cancer-related mortality on a global scale [3]. Early detection of cancer significantly enhances patient survival rates. Data suggest that patients diagnosed with early-stage pancreatic cancer exhibit a 5-year survival rate of 73.3% and a median survival time of 9.8 years. Conversely, the median survival time for unscreened pancreatic cancer patients is only 1.5 years [4].

Pancreatic cancer treatment options encompass surgery, chemotherapy, radiation therapy, or a combination of these modalities [5]. Surgery is regarded as the most effective treatment option, providing the most significant potential for curing pancreatic cancer and markedly enhancing patient survival rates compared to other therapeutic modalities [6, 7]. However, the efficacy of surgical treatment is primarily determined by the stage at which the disease is diagnosed. Current techniques for the early diagnosis of pancreatic cancer are limited. Pancreatic cancer is usually already advanced by the time symptoms appear because there are no obvious symptoms in the early stages, and the tumor is located deep in the abdominal cavity. In this case, the chance of surgical intervention also decreases. Data indicate that in the majority of patients, tumors are significantly large, with the involvement of surrounding lymph nodes, blood vessels, and nerves. Therefore, only 10–15% of patients are suitable for radical surgery [1]. Targeted therapies are a modality of cancer treatment. However, effective and widely available targeted drugs for pancreatic cancer remain insufficient. Recent research indicates that, unlike lung and breast cancers, pancreatic cancer lacks a specific molecular variant. In addition, some researchers propose that pancreatic cancer may involve multiple molecular variants. The nonspecific symptoms of pancreatic cancer often lead to misdiagnosis as other abdominal diseases, resulting in ineffective treatment plans and delays in appropriate therapy [1, 8]. Therefore, developing more sensitive and early detection methods for pancreatic cancer is critical in the fight against pancreatic cancer [9].

Recent studies include studies on determining specific biomarkers for pancreatic cancer [10, 11]. When augmented by artificial intelligence, biomarkers can be pivotal in the early diagnosis of diseases. While blood has traditionally been the primary medium for detecting biomarkers, urine now stands out as a promising alternative. The kidneys' continuous ultrafiltration process can lead to accumulating and higher concentrations of specific biomarkers in urine. Additionally, urine sampling is non-invasive, enabling large-volume collection and repeatable measurements [12–14].

Additionally, urine has a less complex proteome and has less dynamic range, unlike blood [14–16]. Improving survival rates in pancreatic cancer patients may require in-depth research into early diagnosis and treatment [17]. Large amounts of biomedical data will emerge in such research processes. Effectively organizing, integrating, understanding, and analyzing big data is one of today's scientific problems. This challenge can be overcome with artificial intelligence methods [18].

The advancement of artificial intelligence techniques has led to promising developments in identifying, treating, and predicting outcomes for pancreatic cancer [19]. Between 1997 and 2021, 587 publications related to the application of artificial intelligence for pancreatic cancer were discovered in the WoS database. Following 2018, the number of publications increased significantly, culminating in 188 by 2021. The cumulative number of documents published between 2017 and 2021 constitutes 72.4% of all publications. It is predicted that the number of studies in the literature in this field will continue to grow, and artificial intelligence applications in pancreatic cancer will become one of the most popular methods [20].

Despite technological advances in medical research, there are still some problems in the early diagnosis of pancreatic cancer. One of these is that the diagnosis is at an advanced stage due to the location of the pancreatic tumor, which limits treatment options. In addition, the lack of symptoms and their comparison with other diseases may delay accurate diagnosis. Another is that although biomarkers are critical for early diagnosis, the specificity and sensitivity of existing biomarkers may be insufficient [1, 8, 9]. Finally, sharing medical data may bring about significant privacy issues. For this reason, patient privacy should be considered in medical data analysis. This study proposes a machine learning approach with data privacy sensitivity and high diagnostic accuracy to overcome the abovementioned problems. As evidenced by the high accuracy rates obtained in the study, diagnostic accuracy was significantly increased using the LGBM algorithm. An innovative and effective solution was presented, ensuring patient data privacy and security with the Fernet encryption algorithm. Thus, this study aims to improve patient outcomes and contribute to ongoing cancer research and treatment efforts.

This article proposes a new model for diagnosing pancreatic disease using urinary biomarkers, which offers significant potential for the early identification of PDAC. The paper was conducted with a publicly available dataset for pancreatic cancer diagnosis collected by Debernardi et al. [21]. This research examines the impact of technologies on clinical practice, explores the contribution of early diagnosis to disease management, and reviews the existing literature in this domain. The article also aims to guide researchers and clinicians in developing new approaches to diagnosing pancreatic cancer. It also

highlights the significance of these technologies, which have the potential to enhance patient well-being and prolong life expectancy. Few studies use machine learning methods to diagnose PDAC using urinary biomarkers data. However, this article's accuracy rate is higher than that of studies in the literature.

The main contributions of this study are as follows:

- Machine learning methods have been proposed to aid in accurately diagnosing PDAC based on urine biomarkers.
- During the pre-processing phase, data quality was evaluated and improved, which greatly impacted the accuracy of the diagnosis.
- Personal data confidentiality is ensured by encrypting the data set.
- The developed models achieved superior classification performance.
- It offers a privacy-preserving framework in the machine learning pipeline, allowing the developer to work without direct access to the data.
- This comprehensive study includes many algorithms for pancreatic cancer diagnosis in the data set used in the study, compares their performances, and aims to find the best among them.
- It is thought that PDAC, which is one of the most lethal cancers when diagnosed late, can increase the chances of survival of patients if diagnosed early with new methods.

The other sections of the study are organized as follows: Sect. " Materials and methods": Materials and Methods: This section includes data pre-processing, data encryption, data visualization, and measurement methods. Sect. " Result and discussion": Results and Discussion: This section presents the applied machine learning models' performance results and discusses the findings. Sect. " Conclusion": Conclusion: This section summarizes the research's general results, mentions its limitations, and suggests future studies.

Materials and methods

The implementation of the proposed method involves a series of steps. All of these steps are shown in Fig. 1. This study used a publicly available Debernardi et al. [21] dataset collected from multiple centers. This dataset includes urinary biomarkers. The dataset was pre-processed to estimate the classifications, provide reliable and acceptable results, and obtain better results in terms of model performance. Processes such as filling in missing values, cleaning outlier data, and feature selection were performed in pre-processing. Thus, the data was ready for analysis. Data quality greatly affects the estimation result, and pre-processing is an important step in modeling [22]. After pre-processing, data visualization was performed to explain the variable interactions and transform complex data into clear, understandable visuals. After the data set was divided into an 80% training set and a 20% test set, it was encrypted with the Fernet encryption algorithm to protect data privacy. The proposed method is based on the principle that encrypted data is sent to the data analyst who knows the key, and after decryption, analysis is performed using cross-validation. A total of 26 machine learning algorithms were used to perform the analysis, and the results of the ten most commonly used algorithms that gave the best results were included in the study. The modeling process was evaluated using various classifier models with performance measures such as F1 score, recall, precision, and accuracy. Finally, the results obtained were compared with those of other studies in the current literature to demonstrate the superiority of the proposed method. This process provides a safe, effective, and highly accurate method for early diagnosis of pancreatic cancer. Data analysis for this study was performed using the Python programming language (version 3.10).

Pre-processing

Raw data often possesses various imperfections, such as inconsistencies, missing values, noise, and redundancies. Therefore, the performance of subsequent algorithms may be adversely impacted when dealing with low-quality data. In this case, appropriate pre-processing steps need to be applied. Implementing these pre-processing steps can substantially enhance the quality and reliability of subsequent automated analyses and decision-making processes [23]. These methods offer several benefits, including a more rapid and precise learning process and a better-organized raw data structure. It is important to recognize that data pre-processing often constitutes the most time-consuming and labor-intensive phase of the data analysis workflow [24]. Data visualization was also performed in the research to understand the data better. Figure 1 shows the proposed model.

During the pre-processing phase, it was noted that certain attribute records in the dataset contained outliers while others exhibited missing values. Figure 2 shows a graph that captures structural defects related to the randomness of missing values. A total of 3 variables (stage, sample_id, benign_sample_diagnosis) that had too many missing values and did not contribute anything were removed from the data set. Gaps in variables with acceptable missing values were filled with the mean. The average values within each class were considered while filling the gaps in the data set with the average method. Outliers were detected using the IQR (Interquartile Range)

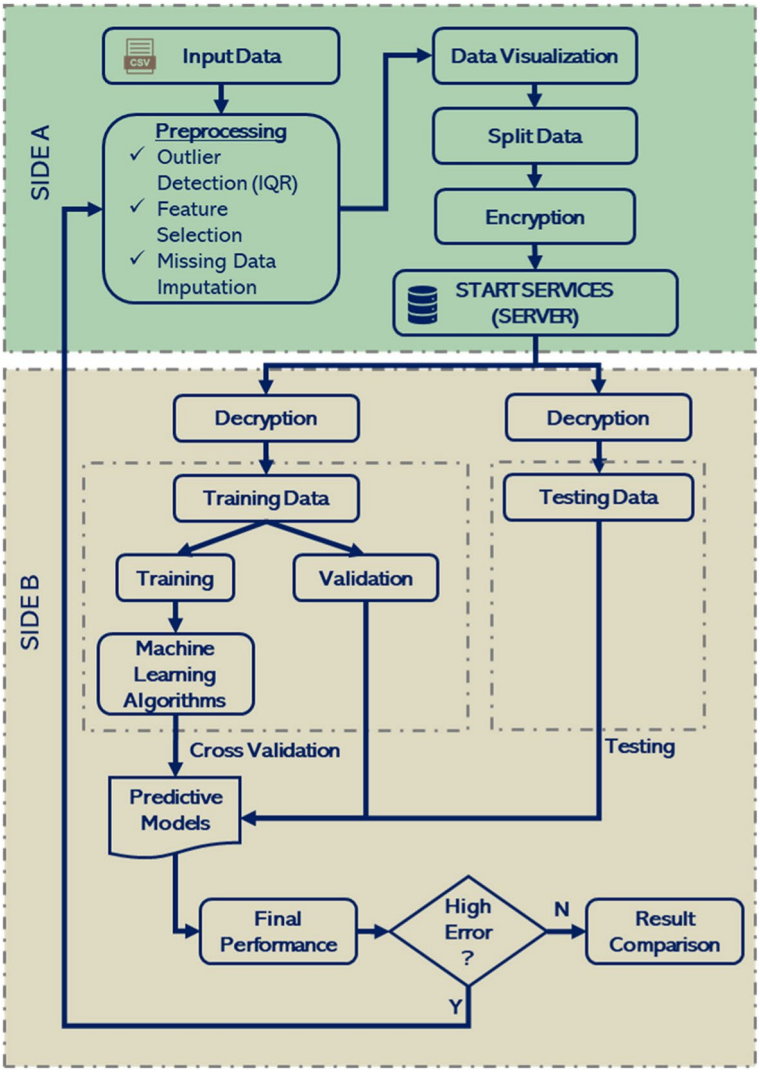

Fig. 1 The proposed model

method. Only approximately 19% of the plasma CA19_9 variable was changed from the mean.

Data privacy

Machine learning offers a data-driven approach. This approach aims to develop the ability to make decisions in unknown test scenarios by using the problems of a specific task and data sets obtained from previous experiences. This process aims to discover existing patterns to increase the ability to make correct decisions in new situations. This method is effective in theoretical and practical applications and strengthens the capacity to apply learned knowledge to new and unpredictable situations. However, this traditional problem-solving process requires collaboration between machine learning practitioners and data providers, which requires technical expertise. This collaboration often requires sharing significant amounts of data, which can raise privacy concerns because there is a risk of disclosure of shared sensitive data [25].

Figure 1 is based on two physically separated areas. The first of these, "Side A," takes place after the steps of data storage, data pre-processing, data visualization, data division, encryption, and storage on the server. In the "Side B" development area, the implementer receives the data, and the transmitted decoding code evaluates the model, developing the machine learning application. In the study, the dataset was encrypted using the Fernet encryption algorithm. According to this algorithm, data is decoded and converted into a data frame object. Fernet, a symmetric encryption algorithm, is utilized to securely and efficiently encrypt and decrypt data. This algorithm is a variant of the AES (Advanced Encryption Standard) algorithm and provides key-based solid encryption.

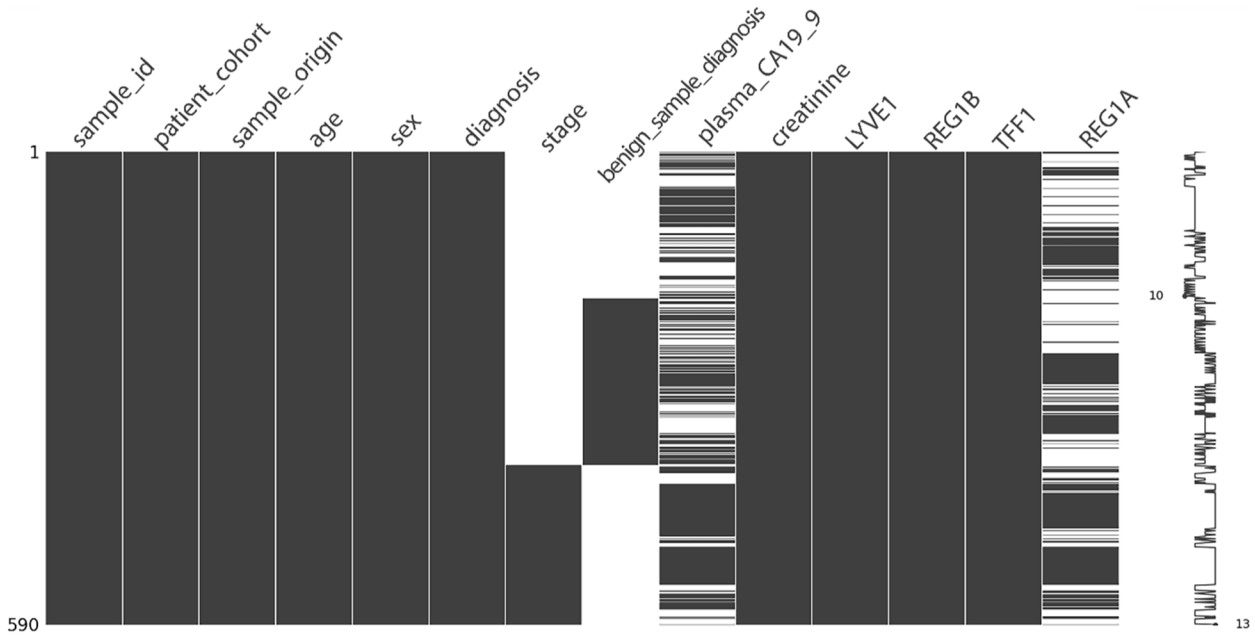

Fig. 2 Structural defect plot of missing values

Fernet encrypts data using the specified key and decrypts data using the same key. This way, while the data is transmitted securely, it can only be read if the receiving party knows the key. The Fernet algorithm is widely used to meet industry-standard secure data encryption requirements [26].

Understanding and visualizing data

Transforming abstract data into physical images, utilizing attributes such as length, location, shape, and color, provides compelling data narratives for individuals who prefer visual representation. Currently, organizations possess unprecedented amounts of data. Consequently, many organizations utilize data and sophisticated analytics to inform their strategic and tactical decision-making processes. In addition to providing a comprehensive overview of big data, data visualization is a natural fit to facilitate the interpretation of data analytics results by data scientists [27]. Data visualizations include illustrative and comparative graphs and tables to effectively convey concrete and abstract concepts. Various visualizations have been made to simplify complex information in the data set, visually reveal relationships, and highlight patterns and trends. The name, data type, definition, and role of the variables are presented in Table 1.

Figure 3a shows the graph of the clinical records for the target variable. Of the 590 clinical records examined in the data set, 199 (33.7%) were with pancreatic cancer (PDAC samples) (3), 208 (35.3%) were with non-cancerous pancreatic condition (benign hepatobiliary disease samples) (2), and 183 (31%) consists of healthy (control samples) (1) individuals. Figure 3b shows the gender distribution of clinical records according to the target variable. Accordingly, although the control group had a higher number of women, the incidence of PDAC diagnoses was observed to be higher in men. The number of patients with benign hepatobiliary disease samples is close to each other in both groups. Additionally, the % of female individuals in the data set is 50.7%.

Figure 4a shows the graph of the average age according to the target variable, depending on gender. While the average age of women is higher in the control and PDAC sample groups, the average age of men is higher in patients with benign hepatobiliary disease samples. Figure 4b shows a box age plot for gender and target variable. Accordingly, although there are some differences for gender in the age distribution box plots of each diagnostic group, a similar situation is observed.

sA correlation matrix is a matrix that measures relationships between variables in a data set, shows the importance of variables, and helps discover relationships and patterns. Figure 5 shows the correlation matrix of attributes. While some are weak, nearly all attributes exhibit linear correlations. Therefore, the correlation matrix is important for data analysis and model-building processes. The correlation matrix examines the relationships between the target feature (diagnosis) and other features. The order of correlation with the dataset target variable 'diagnosis' in descending order is 'plasma_CA19_9', 'LYVE1', 'TFF1', 'REG1B', sample_origin,

Table 1 Dataset description

Name	Type	Description	Role
Patient Cohort	Categorical	Cohort1 / Cohort2	Input
Sample Origin	Categorical	BPTB / LIV / ESP / UCL	Input
Age	Numerical	Age (years)	Input
Sex	Categorical	Female / Male	Input
Plasma CA19_9	Numerical	A tumor marker used in the detection of gastrointestinal system cancers	Input
Creatinine	Numerical	A protein used to evaluate kidney function	Input
LYVE1	Numerical	A gene representing a protein in the lymphatic system	Input
REG1B	Numerical	The gene name for the regenerating islet-derived 1 beta protein	Input
TFF1	Numerical	Trefoil factor 1 plays a role in the regenerative and reparative processes of the urinary tract	Input
REG1A	Numerical	The gene name for the regenerating islet-derived protein 1 alpha	Input
Diagnosis	Categorical	1(Healthy controls) / 2 (Pancreatic patients) / 3(Pancreatic ductal adenocarcinoma)	Target

Categorical data pertains to information that cannot be categorized or represented numerically. In contrast, numerical data is expressed in numeric form, not as letters or words, and isn't amenable to grouping. The target is the estimated output variable; input refers to attributes or features. *BPTB* Barts Pancreas Tissue Bank, *LIV* University of Liverpool, *ESP* Spanish National Cancer Research Centre, Madrid, Spain, *UCL* University College London

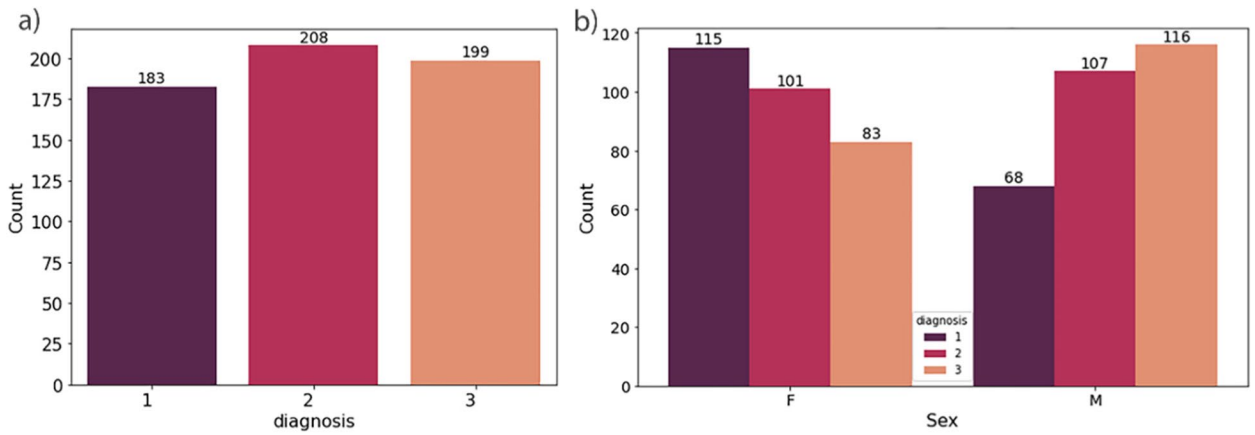

Fig. 3 **a** Count of the target variable (diagnosis) **b** Count of Diagnosis by Sex

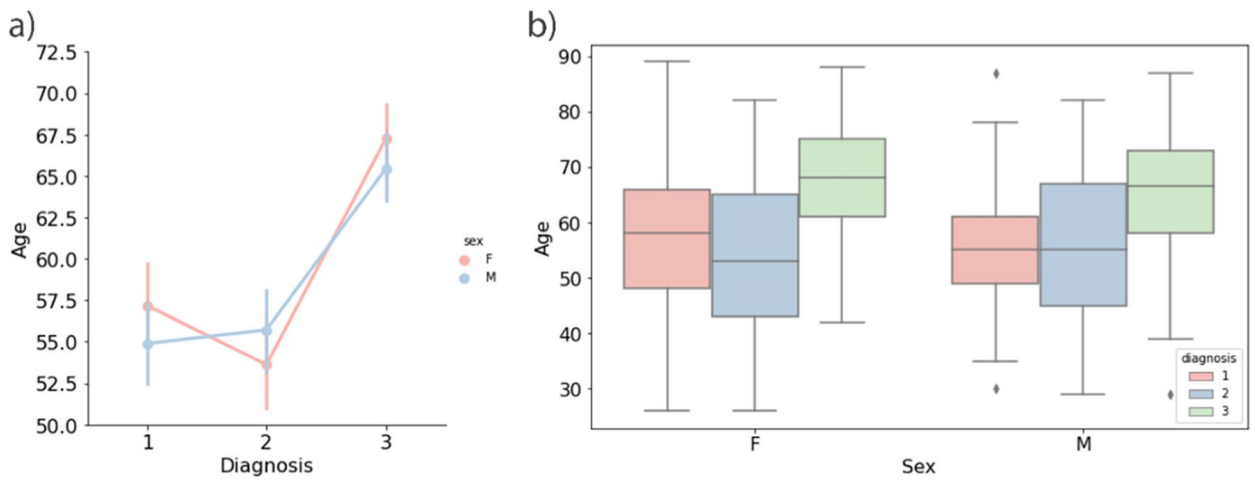

Fig. 4 **a** Pointplot of Age by Sex and Diagnosis **b** Boxplot of Age by Sex and Diagnosis

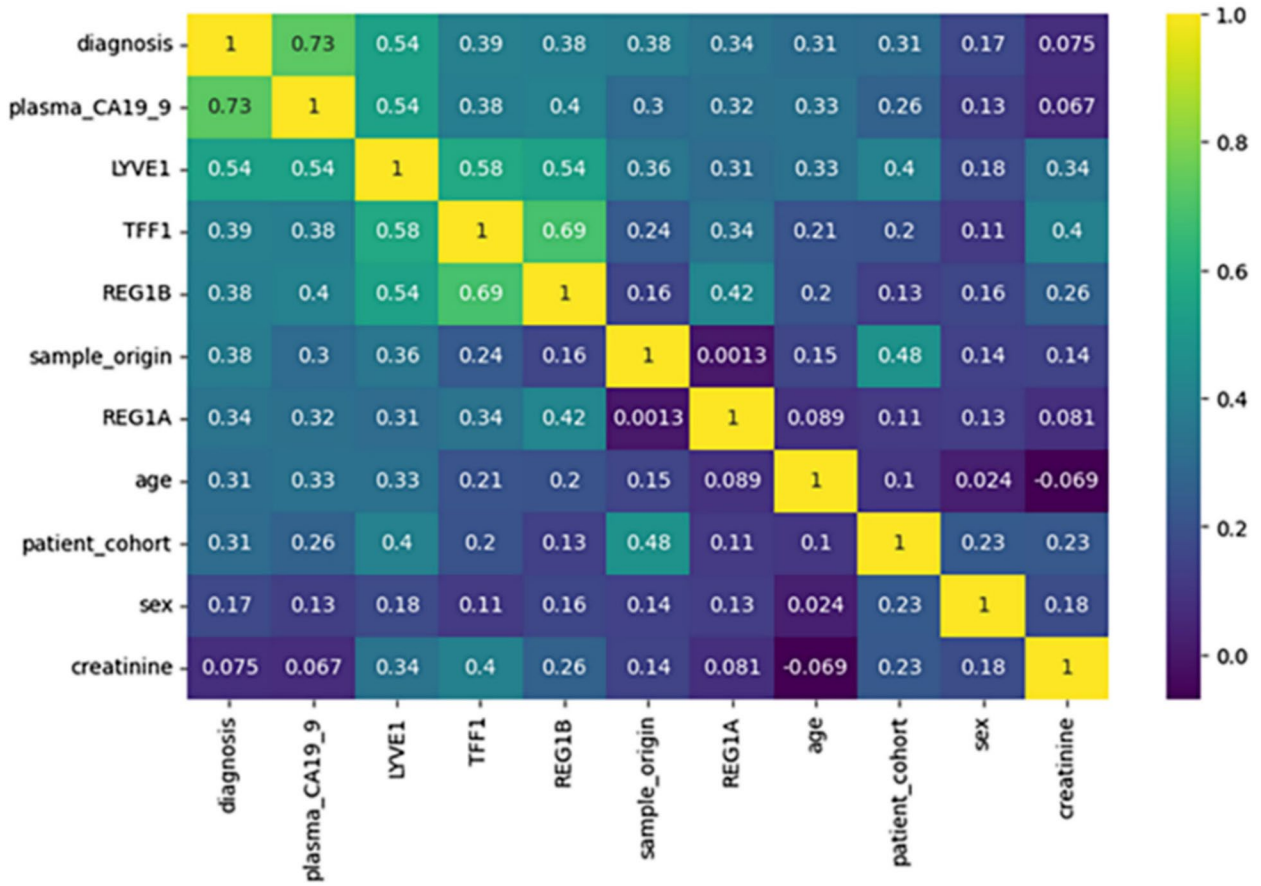

Fig. 5 Correlation matrix

'REG1A', 'age', 'patient_cohort', 'sex' and creatinine' can be seen. Features with strong correlations are expected to facilitate the prediction of the target class and yield more meaningful results. The correlation matrix is an important tool for understanding the relationships between variables and using them in modeling processes. However, correlation coefficients only measure linear relationships; Therefore, it is necessary to use other analytical methods to evaluate non-linear relationships or causal relationships.

Measurement

The following metrics were utilized to gauge the classification performance [28]. In classification metrics, FP denotes false positives, TP represents true positives, TN stands for true negatives, and FN indicates false negatives. P (Positive): It is called the positive class. N (Negative): It is called the negative class. P′: Total number of samples the model predicts as positive (TP+FP). N′: Total number of samples the model predicts as negative (TN+FN).

Accuracy provides an overall measure of the model's ability to make accurate predictions over the entire data set.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} = \frac{TP + TN}{P + N} \quad (1)$$

Precision refers to the ratio of the units the model defines as positive to those actually positive.

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{TP}{P'} \quad (2)$$

Recall measures the model's prediction accuracy for the positive class and evaluates how well it intuitively measures its ability to find all positive units in the dataset.

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{TP}{P} \quad (3)$$

F1 Score combines Precision and Recall measurements under the concept of harmonic mean.

$$\text{F1} - \text{score} = \frac{2.Precision.Recall}{Precision + Recall} \quad (4)$$

Machine learning models include the parameters by which they are trained to perform a specific task. These parameters affect the model's performance. Hyperparameters are the parameters that control the training process of machine learning algorithms and need to be adjusted. Hyperparameters are crucial values that influence the model's complexity, the speed of the training process, and its overall performance. Different hyperparameter values can lead to different performance results of the same model. Therefore, finding the best hyperparameter values through hyperparameter tuning is important. In this study, hyperparameter optimization was made, and the values that gave the best results were investigated. The Grid Search method is used for hyperparameter optimization. Grid Search evaluates each parameter combination's effect on the model performance and selects the combination with the highest accuracy. This process aims to increase the model's accuracy, ensure its generalizability, and optimize the training time. Table 2 lists the parameters used in the ten machine-learning techniques.

Result and discussion

Table 3 presents the outcomes of the ten machine-learning algorithms utilized in the study. The LGBM algorithm exhibited the highest accuracy, achieving a rate of 98.8%. The table provides each model's Accuracy percentage, F1 score, Precision, and Recall values. The LGBM algorithm proved the most successful method according to the accuracy percentage.

The main reasons behind the superior performance of the LGBM algorithm compared to other tested algorithms may be as follows. With the tree-based structure and boosting technique, each new tree is optimized to correct the errors of previous trees. With feature selection and automatic feature engineering, LGBM can automatically ignore unimportant features, preventing the model from focusing on less important information and improving overall performance. Its good performance and fast operation on imbalanced data sets are also important advantages. Providing wide control over hyperparameters helps the model best fit the data set to optimize its performance. In addition, the LGBM algorithm provides high accuracy while keeping the model's overfitting tendency under control. The combination of these factors has contributed to LGBM being an effective tool in the diagnosis of critical and complex conditions such as pancreatic cancer.

According to Table 4, the model correctly predicted that those with diabetes were 98.8%.

The importance of the features was determined using the LGBM algorithm on the dataset. Figure 6 can be used to understand which features are more critical for early

Table 3 Accuracy, F1 score, precision, and recall values of models using the dataset

No	Method	Accuracy (%)	F1	Precision	Recall
1	LGBM	98.8	0.99	0.99	0.99
2	Bagging Classifier	98.6	0.99	0.99	0.99
3	CatBoost	95.9	0.96	0.96	0.96
4	Gradient Boosting Machines	95.5	0.96	0.96	0.96
5	Random Forest	93	0.93	0.93	0.93
6	CART	92.7	0.93	0.93	0.93
7	AdaBoost	90	0.91	0.91	0.91
8	Logistic Regression	89.4	0.89	0.90	0.90
9	SVC	87.4	0.88	0.88	0.88
10	kNN	86.7	0.87	0.87	0.87

Accuracy reflects the overall correctness of the model, while the F1-score serves as the harmonic mean of precision and recall. Precision measures correctly identified positive instances among all predicted positives, whereas recall denotes the proportion of actual positive instances the model correctly identifies

Table 2 Parameters in algorithms

No	Algorithm	Parameters
1	LGBM	{'colsample_bytree': 0.8, 'learning_rate': 0.1, 'max_depth': 15, 'n_estimators': 50, 'subsample': 0.8}
2	Bagging Classifier	{'bootstrap': True, 'bootstrap_features': False, 'max_features': 0.5, 'max_samples': 0.7, 'n_estimators': 100}
3	CatBoost	{'depth': 8, 'iterations': 40, 'learning_rate': 0.1}
4	Gradient Boosting Machines	{'learning_rate': 0.1, 'max_depth': 3, 'min_samples_split': 2, 'n_estimators': 30}
5	Random Forest	{'max_depth': 9, 'max_features': 'sqrt', 'max_leaf_nodes': 9, 'n_estimators': 150}
6	CART	{'criterion': 'gini', 'max_depth': None, 'min_samples_leaf': 4, 'min_samples_split': 2, 'splitter': 'best'}
7	AdaBoost	{'algorithm': 'SAMME', 'learning_rate': 0.5, 'n_estimators': 100}
8	Logistic Regression	{'C': 100, 'penalty': 'l2', 'solver': 'newton-cg'}
9	SVC	{'C': 100, 'degree': 2, 'gamma': 'scale', 'kernel': 'rbf'}
10	kNN	{'metric': 'manhattan', 'n_neighbors': 5, 'p': 1, 'weights': 'uniform'}

Table 4 Results of LGBM

Accuracy: 98.8%	True 1	True 2	True 3	Total	Class Precision	
Pred. 1	183	0	0	183	1	0.99 (weighted avg)
Pred. 2	0	203	2	205	0.99	
Pred. 3	0	5	197	202	0.98	
Total	183	208	199	590		
Class Recall	1	0.98	0.99			
	0.99 (weighted avg)					

Precision denotes the proportion of correctly identified positive instances among all predicted positives, whereas recall represents the ratio of correctly identified positive instances to all actual positives

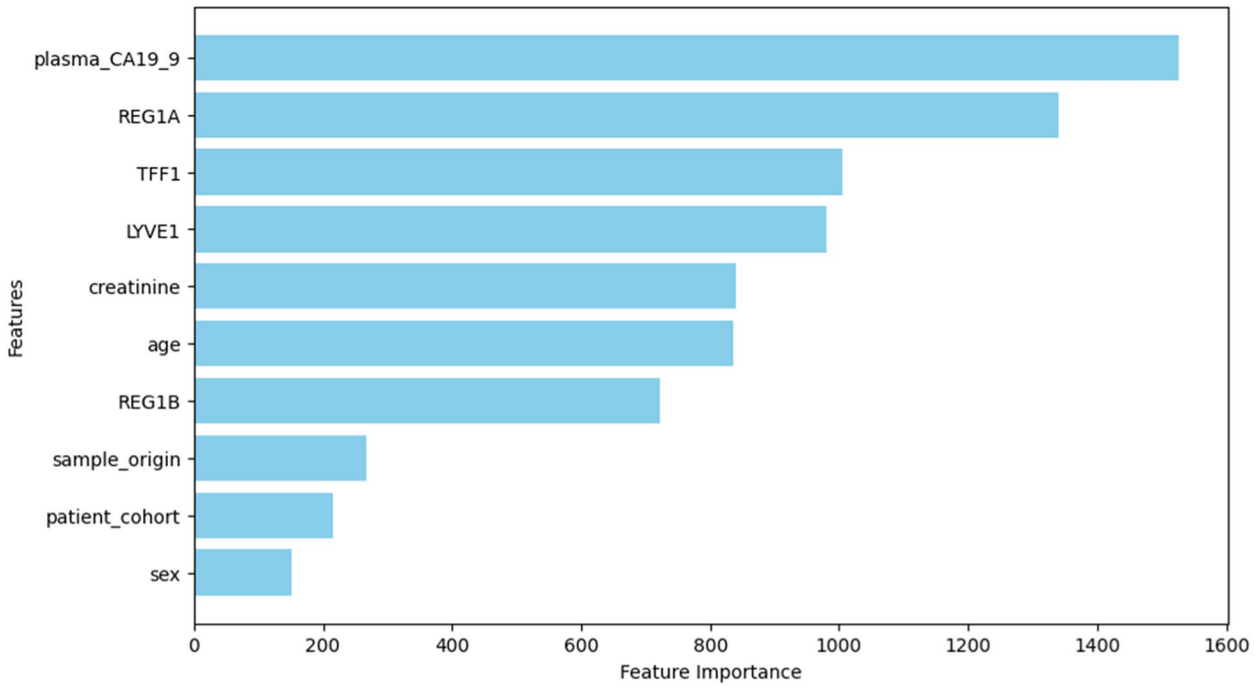

Fig. 6 Feature Importance Distribution with LGBM

diagnosis of pancreatic cancer and which data the model is based on. In particular, while "plasma_CA19_9" is the variable with the highest feature importance, REG1A, TFF1, and LYVE1 features also have high importance levels and play a critical role in the model making accurate predictions. This analysis can be useful both to increase the explainability of the model and to determine which biomarkers should be focused more for clinical applications.

Figure 7 describes the decision-making processes of the LGBMClassifier model using the LIME (Local Interpretable Model-agnostic Explanations) method. Figure 7 shows the model's prediction probabilities in the upper left corner; here, the model predicts "Class 2" with 99% probability. The upper right corner shows the feature values the model considers in the decision-making process. The LIME plot in the middle section shows the features contributing to the model's "Class 2" prediction. Each feature and its contribution to the classification decision is shown. The contributions favoring Class 2 and other classes are shown in different colors. The features contributing to the "Class 2" prediction are: "plasma_CA19_9" value is 0.31, "LYVE1" value is 0.25, "age" value is 0.20, "REG1B" value is 0.04 and "TFF1" value is 0.04. It appears that biomarkers such as "plasma_CA19_9", "LYVE1", and "age" play an important role in classifying the sample as "Class 2". Figure 7 provides a detailed breakdown of which features of the model are important in the decision-making process and how the values of these features contribute to the model's prediction.

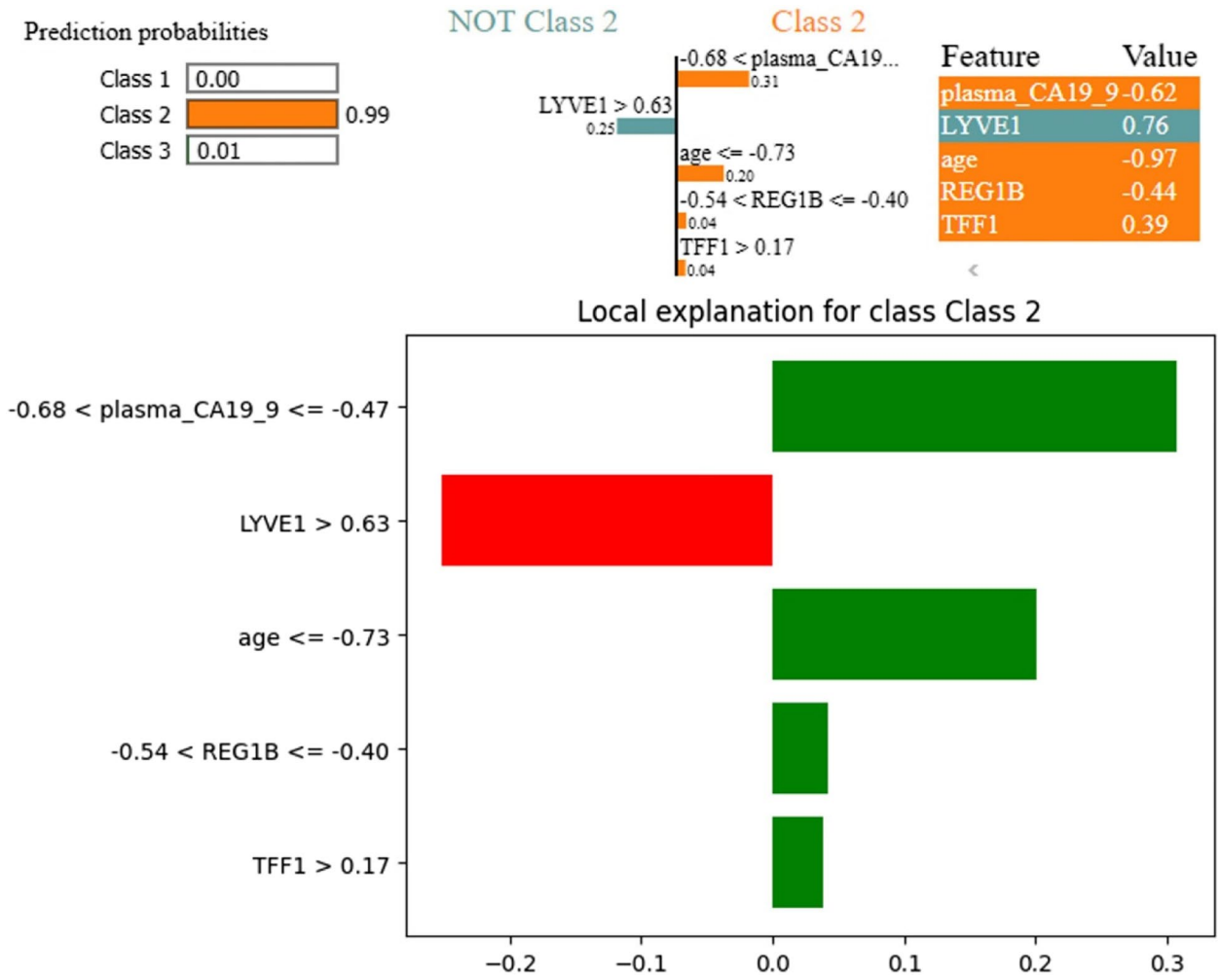

Fig. 7 Explaining model decision processes using LIME with LGBM

Machine learning and deep learning algorithms can increase cancer diagnosis accuracy, cost-effectiveness, and speed. As these algorithms develop, they are expected to play an increasing role in cancer treatment. Therefore, further research is needed to broaden the perspective on this research area and provide a comprehensive view by reviewing the literature [29]. There are many studies on artificial intelligence in the medical field. For example, in the Seyala and Abdullah studies, the effectiveness of penalty methods was emphasized by analyzing longitudinal data of kidney failure patients with nonparametric clustering methods [30]. Muhammed et al. designed the Multi-Cancer Multi-Omics Clinical Dataset Laboratories (MCMOCL) schemes that include federated auto-encoder and XGBoost methods for prediction. The aim is to collect different trained and tested cancer features while loading patient data and to determine the correct cancer types in the system. Also, a 256-bit advanced standard encryption (AES) based encryption and decryption process was performed [31].

A few researchers have used the data set used in the study in predictive analyses. However, there are many studies in the literature on PDAC prediction. Table 5 shows some studies conducted with different and identical data sets and various methods for predicting pancreas cancer diagnosis. Variations in prediction percentages arise from methodological differences in datasets, algorithms, and studies. The dataset size and the number of features can substantially impact the algorithm's performance. The reviewed studies show that the prediction accuracy is between 72.91% and 97%. This study achieved a very high success with 98.8% accuracy.

There are a limited number of studies using the dataset in this study. In the study by Acer et al. [32], seven machine learning classifiers (support vector machine (SVM), naive Bayes (NB), k-nearest neighbors (kNN), random forest (RF), light gradient boosting machine

Table 5 Reported classification accuracies of different classifiers in the literature

	Authors	Data	Year	Method	Best Accuracy
1	Acer et al. [32]	Urine biomarkers	2023	GBC	72.91%
2	Karar et al. [33]	Urine biomarkers	2023	1D CNN-LSTM	97%
3	Mallipudi et al. [34]	Urine biomarkers	2024	Random Forest	86.34%
4	Laxminarayanamma et al. [35]	Urine biomarkers	2022	CNN	95%
5	Baig et al. [36]	Urine biomarkers	2021	SVM	75%
6	Almeida et al. [37]	Genetic biomarkers	2020	ANN	85.71%
7	Lee et al. [38]	Urine Proteomic Biomarkers	2023	Logistic regression	77%
8	Lee et al. [39]	Identified (miRNA) biomarkers	2021	SVM	93%
9	Si et al. [40]	CT images	2021	FEE-DL	87.6%
10	Naito et al. [41]	Endoscopic ultrasound	2021	Deep Learning	94%
11	Wei et al. [42]	F-FDG PET/CT	2023	Deep Learning	90.1%
12	Akmeşe (This study)	Urine biomarkers	2024	LGBM	98.8%

(LightGBM), AdaBoost, and gradient boosting classifier (GBC)) were used to detect PDAC disease on the same dataset in Table 5. A result of 72.91% was achieved with GBC in triple classification. Karar et al. [33] used the same dataset to develop a new and efficient 1D CNN-LSTM model for early PDAC diagnosis, using four proteomic urinary biomarkers: creatinine, LYVE1, REG1B, and TFF1. In another study using the same data, Mallipudi et al. [34] classified pancreatic cancer using SVM, Extra Trees, Decision Trees, and Random Forest methods. When the values obtained from these methods were compared, Random Forest achieved the highest success with 86.34%. Finally, in the study conducted by Laxminarayanamma et al. [35] using urine biomarkers with a CNN model, 95% success was achieved.

In this study, machine learning algorithms were used for PDAC diagnosis. A total of ten different machine learning algorithms were applied for PDAC diagnosis. The LGBM algorithm achieved the highest success rate with an accuracy rate of 98.8%.

Table 5 also includes studies using different data sets on pancreatic cancer. In the study by Baig et al. [36], the SVM method was used to predict whether survival after surgery was less than two years, and results of 75% accuracy, 41.9% sensitivity, and 97.5% specificity were obtained. Almeida et al. [37] achieved an accuracy of 85.71% with the ANN method in their study using genetic biomarker data to diagnose PDAC. Lee et al. [38] conducted a study using the Taiwan Health Insurance Database (NHIRD) between 2000 and 2009 to predict pancreatic cancer, achieving an accuracy of 77% with the Logistic regression model. In their study, which aimed to identify non-invasive miRNA biomarkers and establish a model for PC diagnosis, Lee et al. [39] achieved an accuracy of 93% with SVM. Si et al. [40] achieved 87.6% accuracy in diagnosing PDAC using deep learning methods using a dataset obtained from CT abdominal images from 319 patients. Naito et al. [41] developed a model using endoscopic ultrasonography-guided fine-needle biopsy (EUS-FNB) that accurately detected difficult cases of isolated and low-volume cancer cells, achieving an accuracy of 94%. Wei et al. [42] constructed a novel multi-domain fusion model of radiomics and deep learning features based on F-fluorodeoxyglucose positron emission tomography/computed tomography (F-FDG PET/CT) images. This model demonstrated a diagnostic performance of 90.1% accuracy in noninvasively distinguishing PDAC and AIP using multi-domain features.

Pancreatic cancer can be challenging to diagnose because the pancreas is a complex and deeply located organ. The extensive vascularization around the pancreas facilitates the swift dissemination of cancer, enhancing its aggressiveness. Common symptoms of pancreatic cancer encompass abdominal pain, alterations in stool consistency, nausea, bloating, concurrent conditions like diabetes and jaundice, abnormal liver function test results, and weight loss [7]. While radiographic imaging-based studies are regarded as the primary method for pancreatic cancer screening, such screening is not recommended for asymptomatic individuals owing to the high costs and the relatively infrequent occurrence of pancreatic cancer [43].

PDAC has high mortality rates, primarily because it is often diagnosed at advanced stages of progression. Therefore, studies on early diagnosis of this disease with artificial intelligence can help treatment and increase survival rates. Although AI has unique advantages, many people are still concerned about its use in clinical trials. For instance, no single model can address every problem, as each model has its specific range of applicability [44].

However, the transparency and interpretability of artificial intelligence may be affected by patient privacy, interpretability of the algorithm, publication bias, etc., which makes it difficult due to factors. It is necessary to solve these problems for the future of artificial intelligence in clinical applications [45–47].

Machine learning techniques can predict whether patients have cancer using biomarkers as attributes. Biomarkers are molecules that indicate the presence or absence of disease. Large amounts of patient data have significant potential for early diagnosis of diseases. Utilizing biomarkers for diagnosing pancreatic cancer may be crucial in identifying the disease early, thereby improving patients' quality of life and increasing survival rates.

The medical diagnosis process is intricate and vital, necessitating real patient data, comprehensive knowledge of medical literature, and clinical expertise, as it encompasses numerous unpredictable scenarios. Clinical decisions are primarily guided by the perceptions and experiences of physicians [48]. Nevertheless, patients might not consistently articulate their symptoms accurately. Moreover, the exponential increase in data volume presents further challenges in the decision-making process. Urinary biomarkers (LYVE1, REG1B, TFF1, and Creatinine) provide a promising, non-invasive, and cost-effective approach for PDAC diagnosis [33]. This predictive model has the potential to raise awareness of pancreatic cancer risk and offer patients a straightforward tool for early screening during the critical period when the disease can still be effectively treated.

The main innovation of this study is that it presents a machine-learning approach that prioritizes data privacy for the diagnosis of pancreatic cancer. The use of urine biomarkers and the application of advanced machine learning algorithms on these biomarkers create a difference compared to the general studies in the literature. This study achieved high accuracy rates and considered the need to protect patient privacy.

The dataset used in this study consists of data collected from specific centers. Although the analyses were performed with data obtained from different regions, collecting data from a wider geography and different populations may increase the generalizability of the findings. Future studies should validate the results using larger and more diverse datasets. While machine learning algorithms offer high accuracy rates, decision-making processes are often described as a "black box". This can make it difficult to understand how the model makes decisions in clinical applications. Although steps were taken towards the explainability of the model in this study, this issue needs to be addressed in more depth. Encryption techniques were used to protect the confidentiality of the data in the study. However, since the data is processed before encryption in the pre-processing phase, data confidentiality and ethical concerns cannot always be completely eliminated. This situation can become a bigger problem, especially when working with larger and more sensitive datasets. The biomarkers used in the study provide important information for the early diagnosis of pancreatic cancer. However, the lack of different biomarkers may limit the study's results. In addition, studies can be designed by combining different types of datasets. For example, including radiological images in the analysis may increase the accuracy of the model and its usefulness in clinical applications. In conclusion, these limitations indicate that the findings of the study should be interpreted with caution and that these limitations should be addressed in future research. Future studies are expected to overcome these limitations with larger data sets and improved explainability methods.

Conclusion

This study makes a significant contribution to using machine learning algorithms to diagnose pancreatic PDAC. The theoretical implications of this study, which focuses on urine biomarkers, reveal that advanced machine learning algorithms such as LGBM have great potential to increase the accuracy and early diagnosis of PDAC. A total of ten different machine learning algorithms were applied to diagnose PDAC. The LGBM algorithm achieved the highest success rate with an accuracy rate of 98.8%. The accuracy rates of other algorithms ranged from 86 to 98%. The findings show that ensemble learning models generally outperform traditional classifiers in the study dataset. This study is important because it is low-cost, provides the advantage of rapid diagnosis, can increase the recognition of pancreatic cancer risk, takes into account the need to protect patient confidentiality, and can offer early diagnosis advantages to patients with pancreatic cancer.

However, there are some limitations to this study. The fact that the dataset was collected from specific centers and covered narrow geography, the difficulty of understanding how machine learning models make decisions in clinical applications, the fact that ethical concerns about data privacy cannot always be completely eliminated, and the lack of different biomarkers and different datasets can be given as examples of these limitations.

In future research, several suggestions can be built on the findings of this study. First, expanding the dataset to include a wider geography, a larger population, and a wider range of biomarkers can increase the model's generalizability. Second, integrating these machine learning models with clinical decision support systems can enable their adoption in medical settings. Third, additions can be made to address concerns about data privacy. Finally,

developing interpretable machine learning models that make decision-making processes transparent can ensure that healthcare professionals use these tools reliably and effectively.

Acknowledgements
The data set used in this study was provided from the article titled "A combination of urinary biomarker panel and PancRISK score for earlier detection of pancreatic cancer: A case-control study" written by Debernardi et al. I would like to thank the authors of the article for permission to access the dataset and their scientific contributions.

Authors' contributions
OFA, the sole author of this manuscript, acknowledges responsibility for all aspects of the research presented.

Funding
The author received no specific funding for this study.

Availability of data and materials
Publicly available datasets were analyzed in this study. This data can be found here: https://doi.org/10.1371/journal.pmed.1003489.s009.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 5 July 2024 Accepted: 29 August 2024
Published online: 05 September 2024

References

1. Lee HA, Chen KW, Hsu CY. Prediction Model for Pancreatic Cancer—A Population-Based Study from NHIRD. Cancers 2022 14(4):882. Available from: https://www.mdpi.com/2072-6694/14/4/882/htm. Cited 2023 Sep 27.
2. Siegel R, Miller K, Fuchs H, Clin AJCCJ, 2021 undefined. Cancer statistics, 2021. medicine-opera.com. 2022. Available from: https://medicine-opera.com/wp-content/uploads/2022/01/CA-A-Cancer-J-Clinicians-2022-Siegel-Cancer-statistics-2022.pdf. Cited 2023 Sep 27
3. Rahib L, Smith BD, Aizenberg R, Rosenzweig AB, Fleshman JM, Matrisian LM. Projecting cancer incidence and deaths to 2030: The unexpected burden of thyroid, liver, and pancreas cancers in the united states. Cancer Res. 2014Jun 1;74(11):2913–21.
4. Dbouk M, Katona BW, Brand RE, Chak A, Syngal S, Farrell JJ, et al. The Multicenter Cancer of Pancreas Screening Study: Impact on Stage and Survival. J Clin Oncol. 2022 40(28). Available from: https://pubmed.ncbi.nlm.nih.gov/35704792/. Cited 2023 Sep 27.
5. Rustam Z, Zhafarina F, Saragih GS, Hartini S. Pancreatic cancer classification using logistic regression and random forest. IAES Int J Artif Intell IJ-AI. 2021;10(2):476–81.
6. McGuigan A, Kelly P, Turkington RC, Jones C, Coleman HG, McCain RS. Pancreatic cancer: A review of clinical diagnosis, epidemiology, treatment and outcomes. World J Gastroenterol. 2018Nov 21;24(43):4846–61.
7. Kamisawa T, Wood LD, Itoi T, Takaori K. Pancreatic cancer. Lancet. 2016Jul 2;388(10039):73–85.
8. Chang CL, Hsu MY. The study that applies artificial intelligence and logistic regression for assistance in differential diagnostic of pancreatic cancer. Expert Syst Appl. 2009Sep;36(7):10663–72.
9. Klein AP. Pancreatic cancer epidemiology: understanding the role of lifestyle and inherited risk factors. Nat Rev Gastroenterol Hepatol. 2021;18(7):493–502. Available from: https://doi.org/10.1038/s41575-021-00457-x.
10. Seyhan AA. Circulating microRNAs as Potential Biomarkers in Pancreatic Cancer—Advances and Challenges. Int J Mol Sci. 2023;24(17):13340.
11. Brezgyte G, Shah V, Jach D, Crnogorac-jurcevic T. Non-invasive biomarkers for earlier detection of pancreatic cancer—a comprehensive review. Cancers (Basel). 2021;13(11):1–25.
12. Thongboonkerd V. Recent progress in urinary proteomics. Proteomics - Clin Appl. 2007Aug;1(8):780–91.
13. Dinges SS, Hohm A, Vandergrift LA, Nowak J, Habbel P, Kaltashov IA, et al. Cancer metabolomic markers in urine: evidence, techniques and recommendations. Vol. 16, Nature Reviews Urology. 2019. p. 339–62. Available from: https://idp.nature.com/authorize/casa?redirect_uri=https://www.nature.com/articles/s41585-019-0185-3&casa_token=1XVMlLmt0ZkAAAAA:ntBGnTtUs4_LKxakPRKZ0_anhld-DLWLnJafewo2w1EfpNKR3Xa5IhxAxF4aHPqFJEcBQavQCtEDH95MCA. Cited 2024 Feb 11.
14. Debernardi S, O'Brien H, Algahmdi AS, Malats N, Stewart GD, Pljesa-Ercegovac M, et al. A combination of urinary biomarker panel and PancRISK score for earlier detection of pancreatic cancer: A case–control study. PLOS Med. 2020 Dec 10 ;17(12):e1003489. Available from: https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1003489. Cited 2022 Nov 11.
15. Biology JAG, 2006 undefined. The human urinary proteome contains more than 1500 proteins, including a large proportion of membranes proteins. cir.nii.ac.jp. ; Available from: https://cir.nii.ac.jp/crid/1573668925816041984. Cited 2024 Feb 11.
16. Good DM, Thongboonkerd V, Novak J, Bascands JL, Schanstra JP, Coon JJ, et al. Body fluid proteomics for biomarker discovery: lessons from the past hold the key to success in the future. ACS Publ. 2007 Dec; 6(12):4549–55. Available from: https://pubs.acs.org/doi/abs/10.1021/pr070529w. Cited 2024 Feb 11.
17. Pereira SP, Oldfield L, Ney A, Hart PA, Keane MG, Pandol SJ, et al. Early detection of pancreatic cancer. Lancet Gastroenterol Hepatol. 2020 5(7):698–710. Available from: http://www.thelancet.com/article/S2468125319304169/fulltext. Cited 2024 Feb 11.
18. Yin H, Zhang F, Yang X, Meng X, Miao Y, Noor Hussain MS, et al. Research trends of artificial intelligence in pancreatic cancer: a bibliometric analysis. Front Oncol. 2022Aug;2:12.
19. Hayashi H, Uemura N, … KMWJ of, 2021 undefined. Recent advances in artificial intelligence for pancreatic ductal adenocarcinoma. ncbi.nlm.nih.govH Hayashi, N Uemura, K Matsumura, L Zhao, H Sato, Y Shiraishi, Y Yamashita, H BabaWorld J Gastroenterol 2021•ncbi.nlm.nih.gov. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8613738/. Cited 2024 Jan 30.
20. Yin H, Zhang F, Yang X, Meng X, Miao Y, Noor Hussain MS, et al. Research trends of artificial intelligence in pancreatic cancer: a bibliometric analysis. Front Oncol. 2022Aug;2(12): 973999.
21. Debernardi S, O'Brien H, Algahmdi AS, Malats N, Stewart GD, Pljesa-Ercegovac M, et al. A combination of urinary biomarker panel and PancRISK score for earlier detection of pancreatic cancer: A case–control study. PLOS Med. 2020;17(12):e1003489. Available from: https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1003489. Cited 2023 May 22.
22. Kuhn M, Johnson K. Data Pre-processing BT - Applied Predictive Modeling. In: Kuhn M, Johnson K, editors. New York, NY: Springer New York; 2013. p. 27–59. Available from: https://doi.org/10.1007/978-1-4614-6849-3_3.
23. Ramírez-Gallego S, Krawczyk B, García S, Woźniak M, Herrera F. A survey on data pre-processing for data stream mining: Current status and future directions. Neurocomputing. 2017;239:39–57. Available from: https://www.sciencedirect.com/science/article/pii/S0925231217302631.
24. Pyle D, Cerra DD, Kaufmann M. Data preparation for data mining. 1999. Available from: https://books.google.com/books?hl=tr&lr=&id=hhdVr9F-JfAC&oi=fnd&pg=PR17&ots=6iaS7OLz9w&sig=_TtpKedEngDqKp-1Gognpxy_PF8. Cited 2024 Apr 25.
25. Pastorino J, Biswas AK. Data-Blind ML: Building privacy-aware machine learning models without direct data access. In: 2021 IEEE Fourth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE). 2021. p. 95–8.

26. Pronika, Tyagi SS. Enhancing security of cloud data through encryption with AES and fernet algorithm through Convolutional-Neural-Networks (CNN). Int J Comput Networks Appl. 2021;8(4):288–99.
27. Qin X, Luo Y, Tang N, Li G. Making data visualization more efficient and effective: a survey. VLDB J. 2020;29(1):93–117. Available from: https://doi.org/10.1007/s00778-019-00588-3.
28. Grandini M, Bagli E, Visani G. Metrics for Multi-Class Classification: an Overview. 2020 Aug 13; Available from: https://arxiv.org/abs/2008.05756v1. Cited 2024 Apr 29.
29. Ali AM, Mohammed MA. A Comprehensive Review of Artificial Intelligence Approaches in Omics Data Processing: Evaluating Progress and Challenges. Int J Math Stat Comput Sci. 2023;2:114–67.
30. Seyala N, Abdullah SN. Cluster Analysis on Longitudinal Data of Patients with Kidney Dialysis using a Smoothing Cubic B-Spline Model. Int J Math Stat Comput Sci. 2023;2:85–95.
31. Mohammed MA, Lakhan A, Abdulkareem KH, Garcia-Zapirain B. Federated auto-encoder and XGBoost schemes for multi-omics cancer detection in distributed fog computing paradigm. Chemom Intell Lab Syst. 2023;241(August).
32. Acer İ, Orhanbulucu F, Içer S, Latifoglu F. Early diagnosis of pancreatic cancer by machine learning methods using urine biomarker combinations. Turkish J Electr Eng Comput Sci. 2023;31(1):112–25.
33. Karar ME, El-Fishawy N, Radad M. Automated classification of urine biomarkers to diagnose pancreatic cancer using 1-D convolutional neural networks. J Biol Eng. 2023 17(1):1–12. Available from: https://link.springer.com/articles/10.1186/s13036-023-00340-0. Cited 2023 Sep 27.
34. Devi M, Sai S, Prudhvi P, Ganeswara I, Sai N, Ram M, et al. Early Detection : Machine Learning Techniques in Pancreatic Cancer Diagnosis. 2024;26(5):175–82.
35. Laxminarayanamma K, Krishnaiah RV, Sammulal P. Enhanced CNN Model for Pancreatic Ductal Adenocarcinoma Classification Based on Proteomic Data. Available from: https://doi.org/10.18280/isi.270115. Cited 2024 May 8.
36. Baig Z, Abu-Omar N, Khan R, Verdiales C, Frehlick R, Shaw J, et al. Prognosticating Outcome in Pancreatic Head Cancer With the use of a Machine Learning Algorithm. Technol Cancer Res Treat. 2021 Nov 5;20. Available from: https://journals.sagepub.com/doi/full/10.1177/15330338211050767. Cited 2024 Apr 29.
37. 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. Available from: https://link.springer.com/articles/10.1186/s12885-020-6533-0. Cited 2024 Apr 29.
38. Lee HA, Chen KW, Hsu CY. Prediction Model for Pancreatic Cancer—A Population-Based Study from NHIRD. Cancers 2022 14(4):882. Available from: https://www.mdpi.com/2072-6694/14/4/882/htm. Cited 2023 Oct 9.
39. Lee J, Lee HS, Park SB, Kim C, Kim K, Jung DE, et al. Identification of Circulating Serum miRNAs as Novel Biomarkers in Pancreatic Cancer Using a Penalized Algorithm. Int J Mol Sci. 2021 22(3):1007. Available from: https://www.mdpi.com/1422-0067/22/3/1007/htm. Cited 2024 May 8.
40. Si K, Xue Y, Yu X, Zhu X, Li Q, Gong W, et al. Fully end-to-end deep-learning-based diagnosis of pancreatic tumors. Theranostics. 2021 11(4):1982. Available from: https://pmc/articles/PMC7778580/. Cited 2024 May 8.
41. Naito Y, Tsuneki M, Fukushima N, Koga Y, Higashi M, Notohara K, et al. A deep learning model to detect pancreatic ductal adenocarcinoma on endoscopic ultrasound-guided fine-needle biopsy. Sci Reports. 2021 11(1):1–8. Available from: https://www.nature.com/articles/s41598-021-87748-0. Cited 2024 May 8.
42. Wei W, Jia G, Wu Z, Wang T, Wang H, Wei K, et al. A multi-domain fusion model of radiomics and deep learning to discriminate between PDAC and AIP based on 18F-FDG PET/CT images. Jpn J Radiol. 2023 41(4):417–27. Available from: https://link.springer.com/article/10.1007/s11604-022-01363-1. Cited 2024 May 8.
43. Raman SP, Horton KM, Fishman EK. Multimodality imaging of pancreatic cancer-computed tomography, magnetic resonance imaging, and positron emission tomography. Cancer J (United States). 2012 18(6):511–22. Available from: https://journals.lww.com/journalppo/fulltext/2012/12000/multimodality_imaging_of_pancreatic.6.aspx. Cited 2024 May 8.
44. Gruson D, Helleputte T, Rousseau P, Gruson D. Data science, artificial intelligence, and machine learning: Opportunities for laboratory medicine and the value of positive regulation. Clin Biochem. 2019Jul;1(69):1–7.
45. Huang B, Huang H, Zhang S, Zhang D, Shi Q, Liu J, et al. Artificial intelligence in pancreatic cancer. Theranostics. 2022 12(16):6931. Available from: https://pmc/articles/PMC9576619/. Cited 2024 May 8.
46. Haibe-Kains B, Adam GA, Hosny A, Khodakarami F, Shraddha T, Kusko R, et al. Transparency and reproducibility in artificial intelligence. Nature. 2020 586(7829):E14–6. Available from: https://pubmed.ncbi.nlm.nih.gov/33057217/. Cited 2024 May 8.
47. Stupple A, Singerman D, Celi LA. The reproducibility crisis in the age of digital medicine. NPJ Digit Med. 2019 2(1). Available from: https://pubmed.ncbi.nlm.nih.gov/31304352/. Cited 2024 May 8.
48. Parthiban L, Biological RSIJ of, 2008 undefined. Intelligent heart disease prediction system using CANFIS and genetic algorithm. CiteseerL Parthiban, R SubramanianInternational J Biol Biomed Med Sci 2008•Citeseer. Available from: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=587432f342adab0b3cd00b392d474e8100d8df45. Cited 2024 Jan 30.

Publisher's Note