Predictive Modelling Using Urinary Biomarkers in Combination with the Serum Biomarker Carbohydrate Antigen (CA) 19-9 for Non-Invasive and Reliable Detection of Pancreatic Cancer

Rishab Perati

Received August 01, 2024
Accepted February 11, 2025
Electronic access February 28, 2025

Objective: Pancreatic ductal adenocarcinoma (PDAC), commonly known as pancreatic cancer is one of the rare cancers for which no significant improvements in diagnosis and therapy have been made in the last 30 years. Despite considerable progress in our understanding of the disease at the molecular level, novel findings have not yet translated into clinical benefit. PDAC has the highest mortality rate of all major cancers. Despite many years of experimental research and clinical trials, the 5-year survival rate for pancreatic cancer is still 13%. The major reason for the poor survival is due to late detection. By the time the cancer is detected, it is usually locally advanced or metastasized. Considering these dire statistics, reliable clinical markers to identify high-risk individuals is the key to improved PC patient survival. Identifying a panel of biomarkers will allow clinicians to further evaluate high-risk individuals with immediate and periodic surveillance with CT scans, resulting in early detection and timely therapeutic intervention, improving patient prognosis. This study aimed to determine if the use of a panel of urinary biomarkers, along with clinical markers already in use, can lead to reliable detection of PDAC.
Method: The study used data from Kaggle, comprising samples from 590 individuals. 183 of these samples were from healthy controls (group 1), 208 from patients with benign diseases (group 2), and 199 from PDAC patients (group 3). The machine learning models: Logistic regression, Decision trees, Random Forest (RF), and Support vector machine (SVM) models were first trained using patients with known labels (N=472). Following the training, all the models were applied to the test group (N=118), to determine the disease risk or the exact prognosis.
Results & Conclusion: The study used an improved panel of five urinary biomarkers REG1A, REG1B, LYVE1, TFF1, and creatinine together with plasma CA 19-9 showing a percent accuracy ranging from 53% for the SVM model to 75% for the RF model to discriminate PDAC patients from cancer-free controls.

Keywords: Machine learning, Clinical markers, Urinary biomarkers, Early Detection, Accuracy, Prognosis, Cancer Survival, Late Detection, Improved Patient Prognosis

Introduction

Pancreatic ductal adenocarcinoma (PDAC) is the fourth leading cause of cancer-related mortality in the United States[1]. PDAC is one of the most aggressive malignancies. It accounts for 55,550 deaths in the United States. It is expected to become the second-leading cause of cancerrelated deaths nationally by 2030[2]. The risk factors for PDAC include smoking, diabetes, chronic pancreatitis, obesity, inherited genetic mutations, pancreatic cysts, and race. African Americans have a higher incidence of pancreatic cancer compared to Caucasians, Hispanics, and Asian Americans. While only 30–40% of Patients with PDAC present with localized disease and undergo potentially curative surgical resection after diagnosis or following neoadjuvant chemotherapy, most develop recurrences and succumb to the disease[3–5]. Patients with PDAC is one of the most lethal cancers-related, with a 5-year survival rate of 15%. The poor outcomes of this disease are due to late diagnosis; however, if the disease is detected at an early stage when tumors are still small and resectable, 5-year survival can increase significantly.

Despite considerable progress in our understanding of the disease at the molecular level, novel findings have not yet translated into clinical benefit, and the 5-year survival rate for pancreatic cancer (PC) mortality rate of all major cancers, despite many years of experimental research and clinical trials. The main reason for the poor overall survival is late diagnosis, partially due to the lack of tools for early-stage detection. In addition, several challenges exist in evaluating response to treatment and predicting prognosis. While the five-year survival rate of patients with localized PC is 34.3%, unfortunately, only 10% of total PC patients are diagnosed early. Approximately 52% of cases are diagnosed at the late/metastasized stage, with a wors-

ened five-survival rate of only 2.7%[6]. There is a pressing need to discover biomarkers that will allow for noninvasive methods for diagnosis of PDAC, detect early recurrence with prognostic impact, and tailor therapy.

Blood biomarkers are the most accessible and well-characterized biomarkers used for pancreatic cancer. Peripheral blood analysis, however, is heavily dependent on the degree of tumor burden. The yield is prohibitively low until the disease is metastatic, which limits the use of blood biomarkers for reliable detection of early-stage disease diagnosis[7]. Carbohydrate antigen (CA) 19.9 is the most extensively validated PDAC biomarker in clinical practice. Serum CA19.9, the only PDAC biomarker in widespread clinical use, suffers from false negative results in patients with Lewis-negative genotype, low sensitivity (79%-81%) in symptomatic patients, and its levels may be elevated in various other benign and malignant pancreatic and hepatobiliary diseases, as well as in unrelated cystic and inflammatory diseases[8]. Serum CEA levels of CEA are high in 30%–60% of PC patients however, CEA having low sensitivity and specificity is not a good marker for diagnosis. CEA is often used as a prognostic tool, as increased levels can be associated with a higher tumor burden and worse prognosis[9,10]. Like blood, urine contains proteomic biomarkers and is a promising alternative body fluid for biomarker discovery. It is an ideal fluid for diagnostic screening tests because patients may easily provide a significant volume of it in an entirely non-invasive inexpensive way[11]. A prior study by Radon et. al measured urinary biomarkers like Regenerating Protein 1A (REG1A), Trefoil factor 1 (TFF1), and Lymphatic Vessel Endothelial Hyaluronan Receptor 1 (LYVE1) to distinguish patients with earlystage PDAC from healthy individuals (H)[12]. The diagnostic performance of the biomarker panel in Radon et al's study needs to be further validated: as the healthy controls in the study were younger on average than the cancer patients; and an older control group would thus be more relevant. In addition, further comparison of the performance of urine markers with CA19.9 was needed.

Machine learning (ML) and deep learning (DL) techniques have become central to computer-aided diagnosis (CAD), leveraging clinical data, medical images, genomics, and biomarkers. ML models can analyze patient data in supervised and unsupervised ways to predict pancreatic health. Advanced DL methods can extract complex, interrelated, and non-linear features from medical datasets to enhance diagnostic accuracy.

Although numerous studies have investigated the role of individual biomarkers in determining patient prognosis, apart from Radon et Al's study, which had limitations, no studies have combined multiple biomarkers along with serum CA19.9 to diagnose PC. This study chose to investigate five urinary biomarkers: creatinine, LYVE1, REG1A, Regenerating isletderived 1Beta (REG1B), and TFF1. These proteins were chosen as they all play a role in promoting tumor growth and metastasis. The Urinary biomarkers LYVE1, REG1B, and TFF1 are elevated in the urine of PDAC patients two years prior to diagnosis. The proteins creatine and CA19-9 were added to the panel of biomarkers to improve diagnostic accuracy. This study hypothesizes that integrating urinary biomarkers with CA 19-9 will improve the diagnostic accuracy of PDAC. The use of these diagnostic biomarkers will result in early diagnosis timely therapeutic intervention and improved patient prognosis.

The logistic regression model was chosen for the study as it is a powerful and versatile tool due to its simplicity and usability with binary outcomes, i.e., the presence and absence of pancreatic cancer. Logistic regression models while versatile have the limitation that they are sensitive to class imbalances and can result in poor performance accuracy with minority classes. The support vector machine model (SVM) was chosen for the study as it is a powerful tool that recognizes subtle patterns in complex datasets, which could make it a valuable tool as a cancer classifier. The decision tree and random forest models were chosen as they are known to better handle class imbalances compared to other models.

Results

Logistic Regression Analysis

Logistic regression is a statistical analysis method to predict binary outcomes. The prediction of the dependent data variable (PDAC, Control or Benign Tumor) is made by analyzing the relationship between one or more existing independent variables i.e., the urinary markers REG1A, REG1B, LYVE1, TFF1, and creatinine in combination with CA19-9. The data was visualized using a heatmap.

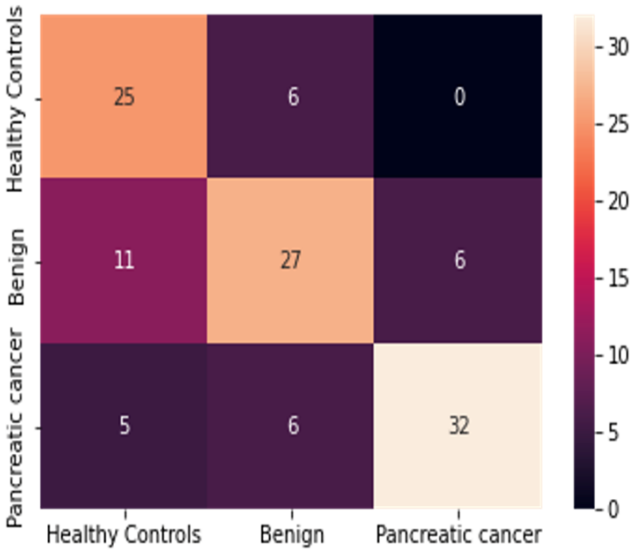

Fig. 1 Shows a Heatmap for Logistic Regression Analysis

In Figure 1 the columns correspond to the predicted diagnosis

and the rows correspond to the true diagnosis. Thirty-two cases were predicted to be pancreatic cancer and this diagnosis was correct. Six cases were incorrectly predicted to be pancreatic cancer, the true diagnosis for these individuals was a benign tumor. Five cases were predicted to be healthy individuals, but the true diagnosis was pancreatic cancer.

Support Vector Machine Analysis

SVM is an extremely popular machine learning algorithm based on the statistical learning theory concept of decision planes that define decision boundaries. This model works well when the data has clear margins of separation. SVM algorithms are not suitable for large data sets or datasets with overlapping values and take a long time to train. In Figure 2 as in Figure 1 the columns correspond to the predicted diagnosis and the rows correspond to the true diagnosis. Using the SVM analysis twenty-five cases were predicted to be pancreatic cancer and this diagnosis was correct. Ten cases were however incorrectly predicted to be pancreatic cancer when the actual diagnosis was benign tumors. Twelve of the cases were predicted to be tumor-free but the true diagnosis was pancreatic cancer.

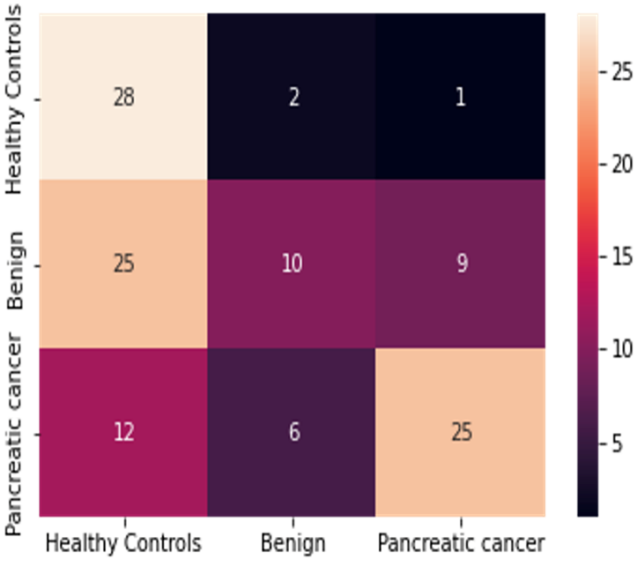

Fig. 2 Shows a Heatmap of the Support Vector Machine Analysis

Decision Tree Analysis

Decision tree models are widely used in cancer diagnosis due to their ability to classify patients based on multiple variables. The straightforward interpretation and visualization capabilities of decision tree models make them valuable tools for understanding the relationships between various risk factors and cancer outcomes.

The heatmap shown in Figure 3 shows twenty-six cases of pancreatic cancer predicted using the decision tree model. This classification was correct and matched the actual diagnosis. Ten cases were incorrectly predicted to be pancreatic cancer when these individuals had benign tumors. Seven cases were predicted to be tumor-free healthy individuals when they had pancreatic cancer.

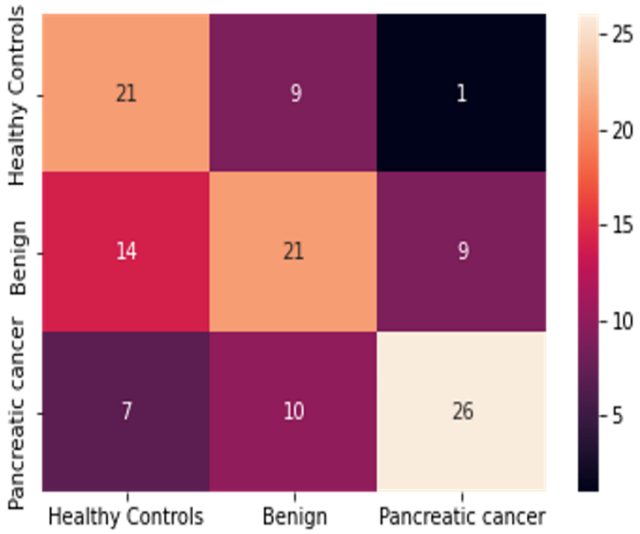

Fig. 3 Shows a Heatmap of the Decision Tree Analysis

Random Forest Analysis

Random forest models are highly effective for cancer classification. They consist of multiple decision trees, which work together to improve predictive accuracy and reduce overfitting. By aggregating the predictions from each tree, random forests provide robust classifications, making them particularly useful for handling complex datasets with numerous variables. Their ability to maintain accuracy even with noisy inputs makes random forest models a powerful tool for identifying cancerous patterns and classifying different types of cancer. This study explored the use of a random forest model for diagnosis of pancreatic cancer.

The heatmap in Figure 4 shows thirty-four correctly predicted cases of pancreatic cancer using the random forest model. The model however identified four cases of pancreatic cancer that were individuals with benign tumors. The model also identified three individuals as healthy when the true diagnosis was that of pancreatic cancer.

The data was visualized using a heatmap to obtain the importance of individual features; the five urinary markers REG1A, REG1B, LYVE1, TFF1, and creatinine and serum CA19-9. As can be seen from the heatmap the combination of features is strongly correlated to the dependent variable.

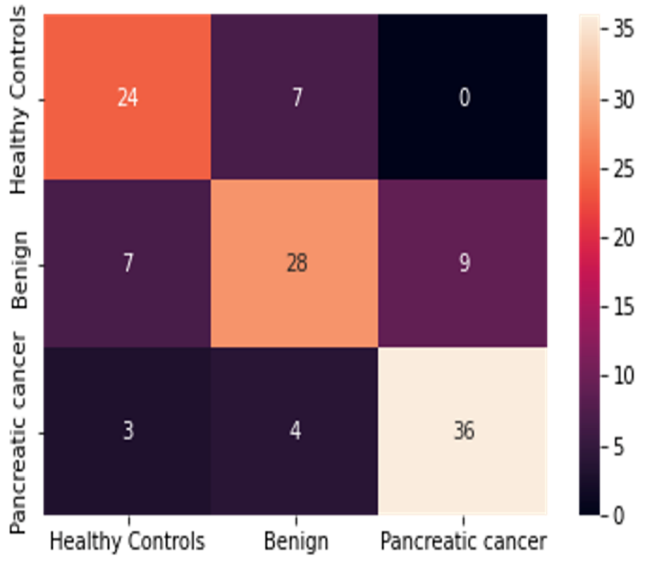

Fig. 4 Shows a Heatmap of the Random Forest Analysis

As seen in Figure 5 the biomarker LYVE1 has the highest correlation of 0.54 for the diagnosis of pancreatic cancer. The biomarker creatinine has a very low correlation of 0.075 for the diagnosis of pancreatic cancer. The correlation ranges from -1.0 to 1.0, the closer the correlation is to 1.0 the higher the correlation. The biomarkers REG1B and TFF1 have a high correlation of 0.69, and since they are highly correlated only one of them can be used for further feature analysis for future analysis.

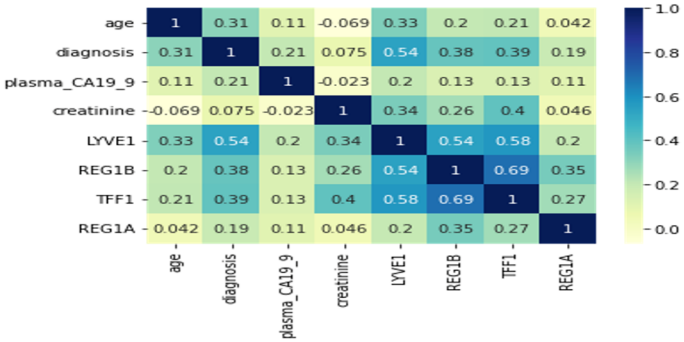

Fig. 5 A Heatmap of each of the variables REG1A, REG1B, LYVE1, TFF1, creatinine and CA19-9

Table 1 shows the accuracy scores for each of the models. The Percent accuracy ranges from 53% for the SVM model to 78% for the RF model.

Type of Model used	Accuracy Score
Logistic Regression	72%
Support Vector Machine	53%
Decision Tree	57%
Random Forest	75%

Table 1. Shows the Accuracy Score of each of the Models.

Methods

A publicly available dataset from Kaggle consisting of samples from 590 individuals was used. This prelabeled dataset contains 590 urine samples and is divided into three patient groups: healthy patients, benign and PDAC cases of 183, 208 and 199 samples, respectively, as illustrated in Table 1. The column with the different stages of pancreatic cancer was dropped and not considered in this study, as it contained over 50% null values. Logistic regression, decision trees, RF, and SVM models were first trained using patients with known labels (N=472). Following the training, all the models were applied to new patients (N=118), to determine the risk of the disease or the exact prognosis.

All the models were fitted for the training set using the six predictors – the five urinary biomarkers REG1A, REG1B, LYVE1, TFF1, and creatinine together with plasma CA 19-9 values and followed the training process shown in Figure 6.

Logistic regression Model

The first model used in the study was a logistic regression model. Logistic regression models are beneficial because they predict the probability of binary outcomes, such as the presence or absence of disease. They help understand the impact of various risk factors by providing clear coefficients and odds ratios. Additionally, performance metrics like the receiver operating characteristic (ROC) curve and area under the curve (AUC) assess the accuracy of these models, making them a reliable tool in predictive modeling. In recent years, logistic regression has become a key tool in cancer research. It helps model the probability of cancer and understand the relationships between risk factors and cancer occurrence. As cancer research evolves, reviewing the basics, methods, and interpretations of logistic regression is crucial[13].

Support Vector Machine Analysis

SVMs are powerful algorithms for data classification and regression. They use a subset of the training data, called support vectors, to create a hypersurface that separates input data effectively. SVMs work through training, testing, and performance evaluation. During training, the algorithm optimizes a cost

Sample Type	Control Group			Benign Group			PDAC Group		
	Sample (n)	Gender	Age Range (Median)	Sample (n)	Gender	Age Range (Median)	Sample (n)	Gender	Age Range (Median)
Urine (N=590)	183	F = 115	26-89 (58)	208	F = 101	26-82 (53)	199	F = 83	42-88 (68)
		M = 68	30-87 (55)		M = 107	29-82 (55)		M = 116	29-87 (67)
Plasma (N= 350)	92	F = 58	26-84 (60)	108	F = 57	26-77 (52)	150	F = 66	42-82 (68)
		M = 34	30-87 (53)		M = 51	29-73 (54)		M = 84	29-83 (67)

Table 2: Breakdown of the Samples Collected with Gender, Diagnosis, and Age Details

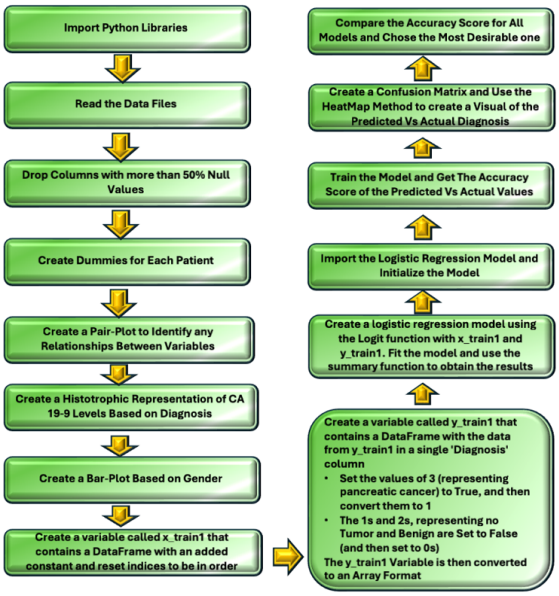

Fig. 6 The Model Training Process

function without local minima, making learning straightforward. Testing involves using the support vectors to classify new data [14].

Decision Tree Analysis

Decision tree analysis is a commonly used data mining method for establishing classification systems based on multiple covariates to develop prediction algorithms. This method classifies a population into branch-like segments that construct an inverted tree with a root node, internal nodes, and leaf nodes. When the sample size is large, the study data can be divided into training and validation datasets. The training dataset was used to build a decision tree model and a validation dataset to decide on the appropriate tree size needed to achieve the optimal final model. The decision tree technique can detect similarities and differences that a human analyst may not notice and therefore create and introduce more accurate/useful categories [15].

Random Forest Analysis

Random Forest is based on the bagging algorithm and uses a collection (ensemble) of decision trees. It is a popular ensemble technique in pattern recognition that creates as many trees as possible on the subset of the data and combines the output of all the trees. As a result, this method reduces overfitting and variance thereby improving the model accuracy. In theory among all the available classification methods, random forest provides the highest accuracy. The random forest technique can also handle big data with numerous variables [16].

A Grid Search was used to choose the best model out of parameters for the model ranging from 50-700 estimators (the number of trees), 2-20 max_depth (the maximum depth of each tree), and 2-10 min_samples_split (minimum number of samples to split a decision node).

Discussion & Conclusion

In this study, we successfully developed and validated four different models to classify patients with pancreatic cancer from those with benign disease as well as healthy controls. The random forest model had the highest accuracy score of 78%, closely followed by the logistic regression model with an accuracy score of 72%. The support vector machine and decision tree model had a less-than-optimal accuracy score of 53% and 57% respectively. The Biomarker CA19-9 is a non-specific inflammatory marker elevated in the benign and PDAC groups. The lack of clear margins separating the classes and the slight class imbalance could have been the cause of the low accuracy of the SVM and decision tree models. SVM models take a long time to train, especially when the features are not well-defined. Clinicians can use the panel successfully validated in this study to non-invasively identify individuals at increased risk of PDAC and monitor these individuals further with immediate and periodic surveillance CT scans. This method would ensure early detection of PDAC in individuals at increased risk of developing the disease.

This improved panel using the five urinary biomarkers REG1A, REG1B, LYVE1, TFF1, and creatinine together with

plasma CA 19-9 showed a % accuracy of 75% to discriminate PDAC patients from controls.

Study	Technique Used	Tool Used	Accuracy %
Chen et al.	Convolution Neural Network Model	Retrospectively collected contras-enhanced CT	98.5%
Radon et al.	Logistic Regression Model	Urine Proteomic biomarkers LYVE-1, REG1A, and TFF1	75.7%

Table 3: Shows the Accuracy of the Models Used in Other Studies

Table 3 shows the comparative accuracy of different methods used to classify pancreatic cancer using urinary biomarkers. Chen et. al's study used retrospectively collected contrast-enhanced CT scan images from a total of 546 patients with pancreatic cancer (mean age, 65 years ± 12 [SD], 297 men) and 733 control subjects were randomly divided into training, validation, and test sets[17]. This study developed an end-to-end deep learning–based computer-aided detection (CAD) tool to accurately and robustly detect PCs on contrast-enhanced CT scans. The CAD tool may be a useful supplement for radiologists to enhance the detection of already diagnosed pancreatic cancer patients. Though Chen et al.'s study sheds light on the potential use of deep learning models on CT scans to detect pancreatic cancer, the reliance of this study on already diagnosed cases and the lack of pre-diagnostic samples limit its applicability in a diverse clinical setting. Future prospective studies, including high-risk and asymptomatic populations, will be necessary in establishing this method's clinical utility.

Radon et. al study measured the urinary biomarkers REG1A, TFF1, and LYVE1 using a random forest model to distinguish patients with early-stage PDAC from healthy patients. The model had an accuracy percentage close to the random forest model used in our study. Early detection is the most important strategy to reduce mortality rates in pancreatic cancer. The lack of biomarkers other than CA 19-9 with clinical utility is a major problem. The panel of biomarkers in this study can be used to detect early-stage pancreatic cancer.

The metric that was used in this study was accuracy, as the primary goal was identifying the presence or absence of pancreatic cancer, other metrics such as recall, and f1-score will be used, analyzed, and compared in future work. Future work will also remove biomarkers with low correlation (e.g. creatinine) to improve the accuracy.

Additionally, as the urinary biomarkers REG1B and TFF1 show a high correlation of 0.69, future models will be developed using fewer biomarkers to see if the accuracy scores can be further increased by feature analysis in the future.

Acknowledgments

I would like to thank Mr. Scott DeRuiter & Mr. Diego Iriarte Sainz for their guidance and support during this project.

Abbreviation

PDAC: Pancreatic Ductal Adenocarcinoma
CA 19-9: Carbohydrate Antigen 19-9
REG1A: Regenerating Protein 1A
REG1B: Regenerating Islet-Derived 1 Beta
LYVE1: Lymphatic Vessel Endothelial Hyaluronan Receptor 1
TFF1: Trefoil Factor 1
ML: Machine Learning
DL: Deep Learning
SVM: Support Vector Machine
RF: Random Forest
CAD: Computer-Aided Diagnosis
ROC: Receiver Operating Characteristic
AUC: Area Under the Curve

Author Information

Corresponding Author : *Rishab Perati, MONTA VISTA HIGH SCHOOL, 21840 McClellan Rd, Cupertino, CA, 95014

References

1 L. Rahib, *Projecting Cancer Incidence and Deaths to 2030: The Unexpected Bur-den of Thyroid, Liver, and Pancreas Cancers in the United States*.

2 R. Siegel, K. Miller, N. Wagle and A. Jemal, *Cancer statistics, 2023*.

3 M. Gostimir, S. Bennett, T. Moyana, H. Sekhon and G. Martel, *Complete pathological re-sponse following neoadjuvant FOLFIRINOX in borderline resectable pancreatic cancer - a case report and review*.

4 D. Pietrasz, *Pathologic Major Response After FOLFIRINOX is Prognostic for Pa-tients Secondary Resected for Borderline or Locally Advanced Pancreatic Adenocarcino-ma: An AGEO-FRENCH, Prospective, Multicentric Cohort*.

5 W. Park, A. Chawla and E. O'Reilly, *Pancreatic Cancer*.

6 M. Arnold, *Progress in cancer survival, mortality, and incidence in seven high-income countries 1995–2014 (ICBP SURVMARK-2): a population-based study*.

7 M. Capello, *Sequential Validation of Blood-Based Protein Biomarker Candidates for Early-Stage Pancreatic Cancer*.

8 U. Ballehaninna and R. Chamberlain, *The clinical utility of serum CA 19-9 in the di-agnosis, prognosis, and management of pancreatic adenocarcinoma: An evidence-based appraisal*.

9 E. Rofi, *The Emerging Role of Liquid Biopsy in Diagnosis, Prognosis and Treatment Monitoring of Pancreatic Cancer*.

10 Q. Meng, *Diagnostic and prognostic value of carcinoembryonic antigen in pancreatic cancer: a systematic review and meta-analysis*.

11 E. Lepowsky, F. Ghaderinezhad, S. Knowlton and S. Tasoglu, *Paper-based assays for urine analysis*.

12 T. Radon, *Identification of a Three-Biomarker Panel in Urine for Early Detection of Pancreatic Adenocarcinoma*.

13 S. Kumar and V. Gota, *Logistic regression in cancer research: A narrative review of the concept, analysis, and interpretation*.

14 N. Sweilam, A. Tharwat and N. Abdel Moniem, *Support vector machine for diag-nosis cancer disease: A comparative study*.

15 N. Al-Salihy and T. Ibrikci, *Classifying breast cancer by using decision tree algorithms*.

16 D. T. Mathew, *An Improvised Random Forest Model for Breast Cancer Classification*.

17 P.-T. Chen, *Pancreatic Cancer Detection on CT scans with Deep Learning: A Nation-wide Population-based Study*.