VAWKUM Transactions on Computer Sciences
http://vfast.org/journals/index.php/VTCS@ 2025, ISSN(e): 2308-8168, ISSN(p): 2411-6335
Volume 13, Number 1, January-June 2025 *pp:161-177*

Development of a Diagnostic Model for Pancreatic Ductal Adenocarcinoma Using Nature-Inspired Optimization Algorithm and Machine Learning Techniques

Abbas Raza[1*], **Muhammad Jawwad**[2*], **Khadija Batool**[4], **Muhammad Sajdain**[1], **Ali Raza**[3]

[1]Department of Computer Information System Engineering, NED University of Engineering & Technology, Karachi City, Sindh, Pakistan; [2]Department of Computer Science & Information Technology, NED University of Engineering & Technology, Karachi City, Sindh, Pakistan; [3]Department of Computer Science, FAST National University Karachi Campus, St-4, Sector 17-D, NH 5, Karachi City, Sindh, Pakistan; [4]Bachelors of Medicine Bachelors of Surgery, Shaheed Mohtarma Benazir Bhutto Medical College Lyari, Lyari Hospital Rd, Rangiwara Lyari, Karachi, Karachi City, Sindh, Pakistan

Keywords: *Machine Learning, Urinary Biomarkers, Swarm Intelligence, Nature Inspired Algorithm, Random Forest, Support Vector Machine, Logistic Regression.*

Journal Info:
Submitted:
Feburary 26, 2025
Accepted:
April 29, 2025
Published:
May 09, 2025

Abstract

PDAC is one of the most harmful cancer causes due to late diagnosis, its rapid progression, and an 11% survival rate of 5 years. Current methods for diagnosis are very costly, uncomfortable, and unreliable, However, better and more accurate solutions are needed. This study proposes a diagnostic model using urinary biomarkers and machine learning techniques for early detection. Key urinary biomarkers, including LYVE-1, REG1B, TFF1, and plasma CA19-9 are used with patient data. Particle Swarm Optimization is used here for feature selection and hyperparameter tuning, optimizes the machine learning classifiers like Support Vector Machine, Logistic Regression, and Random Forest. Accuracy, precision, recall, and F1-score are used as evaluation metrics; however, random forest achieves the highest accuracy of 89.83%. This study shows how PDAC detection changes after combining molecular diagnostics with machine learning. Future research could explore the study of hybrid swarm intelligence algorithms and increase the data set to make further enhancements to diagnostic capabilities. This model shows a great step toward a quick and accurate diagnosis of PDAC and improves patient outcomes and survival rates.

***Correspondence author email address:** cis.abbasraza@gmail.com
DOI: 10.21015/vtcs.v13i1.2079

This work is licensed under a Creative Commons Attribution 3.0 License.

1 Introduction

One of the most lethal cancers is Pancreatic cancer, particularly pancreatic ductal adenocarcinoma (PDAC), as only 11% survive 5 years after its detection because of its late detection and rapid progression. PDAC accounts for most pancreatic cancer cases. PDAC accounts for most pancreatic cancer cases. Although it isn't as common as other cancer diseases, its harmful nature and silent symptoms highly outweigh its low incidence rate. The current diagnostic tools are severely limited and lack precision for detecting lethal and subtle diseases like PDAC. Traditional diagnostic tools were invasive, expensive, and inaccurate, delaying the detection of PDAC until it progressed to its lethal form when the resources to treat it became severely limited and ineffective, highlighting the urgent need for an innovative, advanced, and accurate diagnostic method.

The study of pancreatic disease has undergone innovative advancements through computational technology. The recent research signifies the efficiency of machine learning in analyzing biomarkers for early identification of PDAC. Research like the novel deep learning model introduced by Mohamed Esmail Karar et al. (2023) [1] using long short-term memory (LSTM) and one-dimensional convolutional neural networks (1D-CNNs) networks achieves 97% accuracy in the identification of urinary biomarkers. Similarly, the PancRISK score developed by Silvana Debernardi et al. (2020) [2] achieved promising results in early PDAC detection by combining urinary biomarkers, age, and creatinine levels. These algorithms are more efficient than traditional diagnosis tools, signifying the importance of machine learning in diagnostics.

Recent scientific research has identified urinary biomarkers as a reliable diagnostic sample. Urinary biomarkers like lymphatic vessel endothelial HA receptor (LYVE-1), regenerating islet-derived proteins (REG1B), trefoil factor family (TFF1), and plasma CA19-9 have been proven to have significant diagnostic potential by various researchers. These biomarkers have proved to distinguish early-stage PDAC patients from healthy people, with the sensitivity rate being reported up to 96% and the specificity rate 100%. The research paper uses unique and detailed techniques to diagnose PDAC that are:

1. Accurate and comprehensive dataset of PDAC patients and healthy people.
2. Particle Swarm Optimization (PSO) for feature selection and hyperparameter tuning.
3. Usage of machine learning algorithms like:
    - Logical Regression
    - Random Forest
    - Support Vector Machine

The unique methodology of this research lies in its diverse approach, using a machine-learning model with biomarkers. Our aim is to develop an advanced, non-invasive, and efficient diagnostic tool. The urinary sample is non-invasive, easily accessible, and collectible, making it the most relevant choice.

The effect of this study stretches far beyond just early diagnosis; a non-invasive, efficient, and early detection method can significantly improve pancreatic cancer by:

1. Allowing early treatment.
2. Providing personalized treatments.
3. Improving patient survival rate.
4. Increasing knowledge about PDAC.

By combining computational techniques and molecular diagnostics, this research helps not only in the early diagnosis of highly complex cancer disease but also in a better understanding of its structure and behavior, providing a path for future research and innovations. Present artificial intelligence advancements have revolutionized the

medical field. Based on these advancements, we're building a machine-learning model that will detect PDAC at its early stages through urinary samples. Urine samples are non-invasive and easily accessible, making them the perfect choice.

The unique integration of machine learning algorithms with molecular biomarkers overcomes the limitations in pancreatic diagnosis, allowing us to make an early diagnosis model to replace the current late and lethal stage diagnosis methodologies. The research introduces the integration of machine learning with medical research to defeat the highly lethal cancer disease, trying to help the global effort against the innovations and research against it.

2 Literature Review

Pancreatic cancer (PC) is the third most deadly cancer globally as per cancer statistics in 2022. [3] Among pancreatic cancers, PDAC is the most prevalent type affecting the exocrine pancreas. Although PDAC is only the 12th most common cancer, its combative nature and subtle symptoms contribute to its significant health challenge. Unluckily, PDAC has been owing the worst prognosis primarily due to late-stage diagnosis which is only 11%.

For the diagnosis of pancreatic cancer, a previous study automated classification of urine biomarkers 1-D convolutional neural networks (Mohamed Esmail Karar et al, 2023) [1] proposed an innovative deep learning model integrating long short-term memory (LSTM) and one-dimensional convolutional neural networks (1DCNNs) networks to automate the classification of urinary biomarkers for PDAC diagnosis. In the aforementioned model, the patients were divided into three groups: benign hepatobiliary disease, healthy pancreas, and PDAC cases. A public dataset of 590 urine samples was selected for the conduction of evaluation, and displayed superior performance of the 1DCNN+LSTM model, achieving a precision of 97% and an area under the curve (AUC) of 98%, surpassing modern machine learning classifiers. Despite this another paper Non-invasive biomarkers for early diagnosis of pancreatic cancer risk: metabolite genome-wide association study based on the KCPS-II cohort (Youngmin Han et al,2023) [4] carried a comprehensive metabolite-genomewide association study (GWAS) to detect non-invasive biomarkers that could ease in early detection of pancreatic cancer, a disease with dreadful prognosis. Their model employed serum samples from the Korean Cancer Prevention Study-II (KCPS-II) cohort, consisting of 128 pancreatic cancer cases and 256 matched controls. By utilizing non-targeted metabolomics, they found out that there are 11 significant metabolites associated with pancreatic cancer risk, using a stringent threshold of 4.0 using the XGBoost algorithm for feature selection (Brezočnik, I. Fister Jr., and V. Podgorelec). [5] On further studying they find out genetic correlations by carrying out a genome-wide association study (GWAS) on these metabolites, pointing out five single nucleotide polymorphisms (SNPs) majorly chained to pancreatic cancer risk. Comparatively, SNP-metabolite pairs such as rs2370981, rs55870181, and rs72805402 show distinct network patterns when compared between allelic carriers and non-carriers, proposing the important role they are carrying in the manifestation of disease. According to Mediation analysis, there is also a secondary effect of rs59519100 on pancreatic cancer risk, which is facilitated by γ-glutamyl tyrosine levels, influenced by smoking status. After the integration of these identifies SNPs and metabolites along with conventional risk factors into a predictive model achieving an area under the curve (AUC) of 0.738, depicting its potential for clinical application. The above findings suggest the utility of combining metabolomics with genetic analysis to unveil biomarkers that can help in the early diagnosis and risk assessment of pancreatic cancer. On the other hand, no non-invasive biomarkers are available as revealed by (Youngmin Han et al, 2023) [4] for the early detection of pancreatic cancer. However, continuous efforts are being made in biomarker research to identify non-invasive biomarkers. A recent Urinary biomarkers analysis as a diagnostic tool for early detection of pancreatic adenocarcinoma: Molecular quantification approach study by (Safia Samir, 2024) [6] is thought to detect urinary biomarkers to distinguish early-stage PDAC patients from healthy individuals. The expression levels of three biomarkers: lymphatic vessel endothelial HA receptor (LYVE-1),

regenerating islet-derived 1 alpha (REG1A), and trefoil factor family (TFF1) were measured by using quantitative real-time PCR (qPCR) for the analysis of urine samples of 50 healthy controls and 75 PDAC patients. The resultant experiment elucidated significantly raised expression levels of LYVE-1, REG1A, and TFF1 in PDAC patients compared to healthy controls ($p < 0.05$). The sensitivity of these markers for detecting early-stage PDAC was 96%, 100%, and 73.33%, respectively, while their specificity was 100%, 82%, and 100%, respectively. These findings highlight the potential of these biomarkers for early tumor diagnosis using non-invasive urine samples. Notably, this study is among the first to investigate the mRNA expression levels of LYVE-1, REG1A, and TFF1 in urine specimens within the Egyptian population. These results underscore the need for further validation and development of molecular quantification approaches to improve the early detection and management of PDAC.

Prominent study Urine biomarkers enable pancreatic cancer detection up to 2 years before diagnosis (Silvana Debernardi et al,2023) [7] revealed the predicting capacity of the urine biomarkers LYVE1, REG1B, and TFF1, previously identified as promising markers for PDAC detection. This study utilized a nested case-control design, consisting of 99 PDAC cases and 198 matched controls, along with five years before diagnosis urine samples collection. The samples were sourced from multiple cohorts, including the Shanghai Women's Health Study (SWHS), Shanghai Men's Health Studies (SMHS), and the Southern Community Cohort Study (SCCS). The quantitative analysis of biomarkers was done by using ELISA, while plasma CA19-9 levels were measured with Luminex technology. This study introduced the PancRISK score, an algorithm integrating urinary biomarkers, urine creatinine levels, and age. This resulted in an achievement of a score of an area under the curve (AUC) of 0.79 for samples collected up to one year before diagnosis. After integration with plasma CA19-9, the predictive performance improved significantly, achieving an AUC of 0.89. The study affirms good differential ability up to two years before diagnosis, with AUC values of 0.77. At one year before diagnosis, the combined model had a sensitivity of 72% at 90% specificity, while at two years, it showed a sensitivity of 60% at 80% specificity.

Body mass index (BMI) inclusion as clinical information, has an outstanding result in enhancement of the PancRISK model's performance. These outcomes emphasized the potential of combining urine biomarkers with traditional markers like CA19-9 and clinical data to develop a reliable and non-invasive tool for early PDAC detection, that will further make us capable of identifying cases up to two years before clinical diagnosis. In the study, A Combination of Urinary Biomarker Panel and PancRISK Score for Earlier Detection of Pancreatic Cancer (Silvana Debernardi et al, 2020) [2] authenticate a refined urinary biomarker panel and a corresponding risk stratification algorithm, PancRISK, for earlier PDAC detection. The inclusion of the study is 590 urine specimens, samples collected from PDAC patients at various stages, benign hepatobiliary disease patients, and healthy controls. By replacing REG1A with REG1B, the study is further accelerated, hypothesizing enhanced diagnostic performance, particularly for respectable stage I-II PDAC cases. As per the results, the panel's diagnostic accuracy is improved by substituting REG1A with REG1B. Despite this, the new panel, comprising LYVE1, REG1B, and TFF1, achieved areas under the curve (AUC) of 0.936 in both training and validation datasets, with sensitivity and specificity exceeding 85% for detecting respectable PDAC. When plasma CA19-9 is combined with the results, the panel's performance is further refined, achieving an AUC of 0.992, sensitivity of 96.3%, and specificity of 96.7%. These findings illuminate the complementary nature of urinary biomarkers and CA19- 9 in early-stage PDAC diagnosis.

Furthermore, the biomarkers are significantly stable under various conditions, demonstrating no daily alteration, and can remain stable for up to 5 days at room temperature. The PancRISK score, integrating biomarker data with clinical variables, provided a binary risk output ("elevated" or "normal") for stratifying individuals at risk of developing PDAC. Regardless of the limitations that the study has, including deficiency stage I-IIA PDAC samples and the absence of specimens from people with genetic tendencies, it authenticates a clinically applicable and robust diagnostic tool. In a nutshell, the biomarkers panel, combined with the PancRISK score, represents a significant step toward precision surveillance for PDAC patients. The researchers propose further testing in

prospective clinical studies, such as the UroPanc study, to evaluate its efficacy in broader clinical settings.

Lacking accurate and accessible risk prediction models, in the assessment of PDAC, hinders early detection of pancreatic ductal adenocarcinoma (PDAC), a disease with a dismal prognosis. To address this gap, development of PancRISK: A Urine Biomarker-Based Risk Score for Stratified Screening of Pancreatic Cancer Patients (Oleg Blyuss et al, 2020) [8] developed the PancRISK score, a urine biomarker-based risk stratification tool. This study comprises a dataset of 379 samples from PDAC patients and healthy controls, measuring three urinary biomarkers—LYVE1, REG1B, and TFF1—along with creatinine levels and age. Samples were divided into training and validation sets, with machine learning algorithms applied to develop and assess risk prediction models.

Multiple machine learning approaches are compared in the study, including logistic regression M. P. LaValley [9], emphasizing their predictive performance using the area under the receiver operating characteristic (ROC) curve (AUC) and sensitivity at clinically relevant specificity. Although no algorithm supervenes another, logistic regression was selected for its simplicity and ease of interpretation. The resulting PancRISK score effectively stratified patients based on their risk of developing PDAC. According to the study, the incorporation of CA199 in the algorithm can further enhance the PancRISK model's predictive performance. It improved the model's accuracy, making it a promising tool for non-invasive, urine-based patient stratification.

Currently, to establish its clinical utility the PancRISK score is undergoing further validation. If successful, this tool could be very helpful in the early detection and stratified screening of individuals at risk of PDAC, potentially improving patient outcomes through earlier intervention. In their study, Early Diagnosis of Pancreatic Cancer by Machine Learning Methods Using Urine Biomarker Combinations (İrem ACER et al, 2023) [10] proposed the use of urine biomarkers combined with carbohydrate antigen 19-9 (CA19-9) and advanced machine learning techniques to develop a reliable diagnostic framework for PDAC.

Using the Kaggle Urinary Biomarkers for Pancreatic Cancer (2020) dataset [11] comprising 590 participants, utilizing 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). These classifiers were evaluated using binary and multi-class classification to distinguish healthy controls, individuals with pancreatic disorders, and PDAC patients. The study utilized 5- and 10-fold cross-validation methods to ensure robust performance evaluation. Results revealed that ensemble learning models consistently outperformed other approaches across all classification tasks. To be more specific, the highest precision was achieved by the gradient boosting classifier (GBC) that is (92.99%) and an area under the curve (AUC) of 0.9761 in distinguishing between healthy controls and PDAC patients using 10-fold cross-validation. LightGBM model is superior in performance with an accuracy of 86.3% and an AUC of 0.9348 for the binary classification of pancreatic disorder and PDAC patients. When classifying all three categories—healthy controls, pancreatic disorders, and PDAC patients—the best performance with an accuracy of 72.91% and an AUC of 0.8733 was achieved by the GBC model. In this study, the potential combination of non-invasive urinary biomarkers with advanced machine learning is utilized for the early detection of PDAC. The integration of ensemble learning models, particularly GBC and LightGBM, demonstrates significant promise in improving diagnostic accuracy. For future research for the development of reliable non-invasive diagnostic tool for PDAC, these findings provide a cornerstone.
The key studies summarizing machine learning approaches for PDAC diagnosis are listed in Table 1.

3 Methodology

3.1 Planning and Data Collection

The study is designed to make a model for Pancreatic Ductal Adenocarcinoma PDAC by making full use of urinary biomarkers and machine learning models. The data set is taken from Debernardi et al. (2020) [2] and includes 8 features: age, diagnosis, plasma_CA19_9, creatinine, LYVE1, REG1B, TFF1, and REG1A. The data is divided

Table 1. Summary of Key Studies on PDAC Diagnosis Using Biomarkers and Machine Learning

Study (Year)	Focus	Method/Approach	Key Findings
Karar et al. (2023) [1]	Automated urine biomarker classification	1DCNN + LSTM (590 urine samples)	97% precision, 98% AUC; outperformed ML models
Han et al. (2023) [4]	Non-invasive biomarkers (metabolites/SNPs)	Metabolite-GWAS + XGBoost (128 PC cases)	11 metabolites, 5 SNPs; combined AUC = 0.738
Samir (2024) [6]	Urinary biomarkers (LYVE-1, REG1A, TFF1)	qPCR (75 PDAC vs. 50 controls)	Sensitivity: 96-100%, Specificity: 82-100%
Debernardi et al. (2023) [7]	Urine biomarkers (LYVE1, REG1B, TFF1)	PancRISK + CA19-9 (99 PDAC vs. 198 controls)	AUC = 0.89 (1 yr prior), 0.77 (2 yrs prior)
Debernardi et al. (2020) [2]	Refined urine panel (LYVE1, REG1B, TFF1)	590 samples + CA19-9 integration	AUC = 0.992 (with CA19-9), sensitivity > 96%
Blyuss et al. (2020) [8]	PancRISK risk stratification	Logistic regression (379 samples)	Enhanced by CA19-9 integration
Acer et al. (2023) [10]	ML for PDAC diagnosis	GBC/LightGBM (Kaggle: 590 samples)	GBC: 92.99% precision, AUC = 0.9761

into three groups, first, those who are healthy control, second who have non-cancerous pancreatic conditions, and third who are diagnosed with PDAC. Age and sex were matched where possible to control for demographic changes.

3.2 Data Preprocessing

For data preprocessing, missing values in the dataset [11] are (in plasma_CA19_9 and REG1A) computed through median values to maintain the quality of the dataset, the diagnosis variable is converted to binary language that if it is diagnosed PDCA then 1 otherwise 0, and continuous features were adjusted to an average of zero to help algorithm work better.

3.3 Feature Selection and Hyperparameter Tuning

Particle Swarm Optimization (PSO) was working to identify feature selection and hyperparameter tuning. PSO algorithm is invented due to the social behavior of warms. PSO adjusted particle position repeatedly within the search space to identify two things, first most predictive subset of biomarkers for model training, and the second favorable hyperparameter for machine learning algorithm, including the number of trees in Random Forest, kernel parameters in Support Vector Machine, and order strength in Logistic Regression.

3.4 Machine Learning Models

Each machine learning model was trained on (80-20) split of the data set and advanced by PSO to enhance performance three machine learning models were used in the study:

1. Logistic regression is used for binary classification as well as to explain the contribution of each respective biomarker of PDAC diagnosis.
2. Random forest is used to upgrade predictive accuracy while combining different multiple decision trees.
3. Support vector machine separated the data into categories by finding the best boundary line.

3.5 Evaluation Metrics

The model evaluation done by some metrics that are Accuracy means the proportion of correct prediction among all predictions, precision means the proportion of true positives among that are predicted positives and it also

concludes the ability of a model to decrease false prediction, recall means the proportion of actual positives correctly identified by the model, and the final one is F1 score that is the harmonic mean of precision and recall. These are used to evaluate how the diagnosis is good.

3.6 Feature Importance Analysis

The evaluation of feature importance is done by random forest to find how much each biomarker contributes to PDAC diagnosis. In the study, plasma_CA19_9 and REG1A are biomarkers that contribute the most to the PDAC pathogenesis.

The overall workflow of the methodology is illustrated in Figure 1.

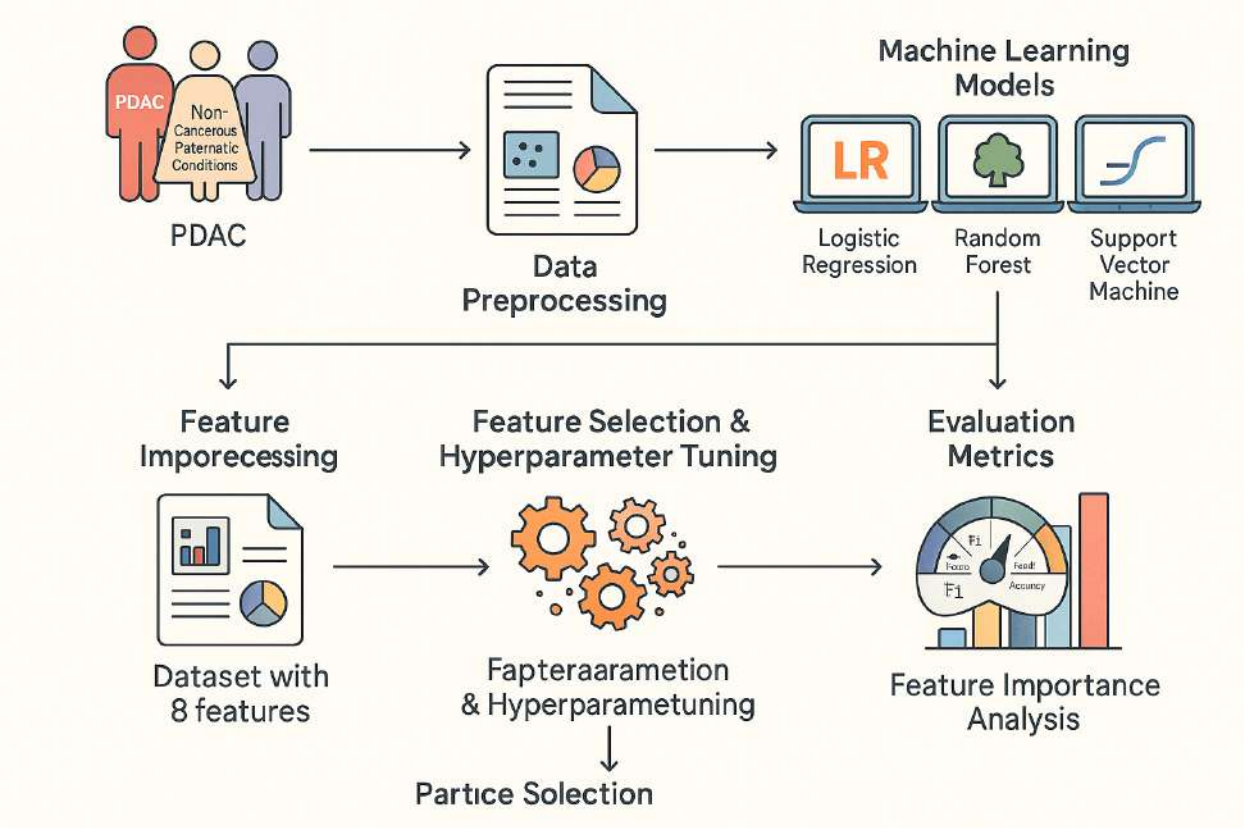

Figure 1. Central diagram illustrating the overall workflow.

4 Simulation and Results

This section explains the outcome and performance evaluation of machine learning classifiers such as Logistic Regression (LR), Support Vector Machine (SVM), and Random Forest (RF), which were optimized using Particle Swarm Optimization (PSO) for the diagnosis of pancreatic ductal adenocarcinoma (PDAC).

4.1 Experimental Setup

This work aims to develop an efficient diagnostic model for Pancreatic Ductal Adenocarcinoma (PDAC), using machine learning and urinary biomarkers. We implemented three classifiers Logistic Regression (LR), Support Vector Machine (SVM), and Random Forest (RF) with enhanced feature selection using Particle Swarm Optimization (PSO) and hyperparameter tuning. The key urinary biomarkers included in the used dataset [11] are (plasma_CA19_9,

```
C:\Users\HP\anaconda3\lib\site-packages\pandas\core\series.py:4463: SettingWithCopyWarni
ng:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_
guide/indexing.html#returning-a-view-versus-a-copy
  return super().fillna(
Starting Particle Swarm Optimization (PSO) for feature selection...
Stopping search: maximum iterations reached --> 10

Best Individual (Selected Features): ['plasma_CA19_9', 'creatinine', 'LYVE1', 'REG1B']
```

Figure 2. Output from PSO results for Logistic Regression.

```
C:\Users\HP\anaconda3\lib\site-packages\pandas\core\series.py:4463: SettingWithCopyWarni
ng:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_
guide/indexing.html#returning-a-view-versus-a-copy
  return super().fillna(
Starting Particle Swarm Optimization (PSO) for feature selection...
Stopping search: maximum iterations reached --> 10

Best Individual (Selected Features): ['plasma_CA19_9', 'LYVE1', 'REG1B']
```

Figure 3. Output from PSO results for SVM.

creatinine, LYVE1, REG1B, and TFF1) and patient age, categorized into three diagnostic classes: healthy, non-cancerous pancreatic conditions, and PDAC.

Key experimental configurations included:

1. Data preprocessing, which includes handling missing values and normalization using StandardScaler.
2. A train-test split of 70%-30%, ensuring sufficient data for validation.
3. PSO is used to select optimal feature subsets and optimize hyperparameters, e.g., the number of trees for RF or kernel parameters for SVM.
4. Accuracy, precision, recall, F1-score, and area under the curve (AUC) are the metrics used for model evaluation.

4.2 Feature Selection by PSO

For each model, PSO identified the most predictive biomarkers, which helped to improve the model's efficiency (see Table 2).

Table 2. The selected features by PSO for each model.

Models	Selected Features by PSO
Logistic Regression	plasma_CA19_9, creatinine, LYVE1, REG1B
SVM	plasma_CA19_9, LYVE1, REG1B
Random Forest	plasma_CA19_9, creatinine, LYVE1, REG1B

The consistency was observed in the selection biomarker including plasma_CA19_9, LYVE1, and REG1B across each model, showing the relevance in the diagnosis of PDAC.

4.3 Model Performance

The performance matrices of all three classifiers are summarized in below tables:

```
C:\Users\HP\anaconda3\lib\site-packages\pandas\core\series.py:4463: SettingWithCopyWarni
ng:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_
guide/indexing.html#returning-a-view-versus-a-copy
  return super().fillna(
Starting Particle Swarm Optimization (PSO) for feature selection...
Stopping search: maximum iterations reached --> 10

Best Individual (Selected Features): ['plasma_CA19_9', 'creatinine', 'LYVE1', 'REG1B']
```

Figure 4. Output from PSO results for Random Forest.

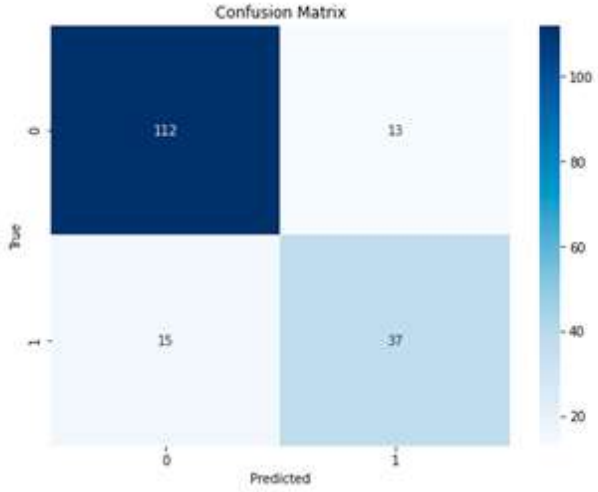

Figure 5. Confusion matrix for logistic regression.

4.3.1 Logistic Regression

Logistic regression is a traditional, analytical grounded method for binary classification problems, where the result is modelled as probability using a logistic function. It predicts the types by calculating the chances that something resembles a certain type. It assumes a straight line between the input feature and the chance of a result [12]. In the study, LR was applied to detect PDAC diagnosis based on selected feature features however its accuracy is 84.18% lower than other models but it is known for simplicity.

Table 3. Results of Logistic Regression

Accuracy	Precision	Recall	F1-Score	AUC
84.18%	84.02%	84.18%	84.09%	0.841

The key strength of this model is that it is a straightforward model with balanced performance across metrics. Ideal for interpretability and identifying contributions of individual biomarkers (see Table 3).

Confusion Matrix

Indicates strong performance in identifying both PDAC cases and controls (see Figure 5).

Precision-Recall Curve

Reflects the model's ability to minimize false positives while maintaining high sensitivity (see Figure 6).

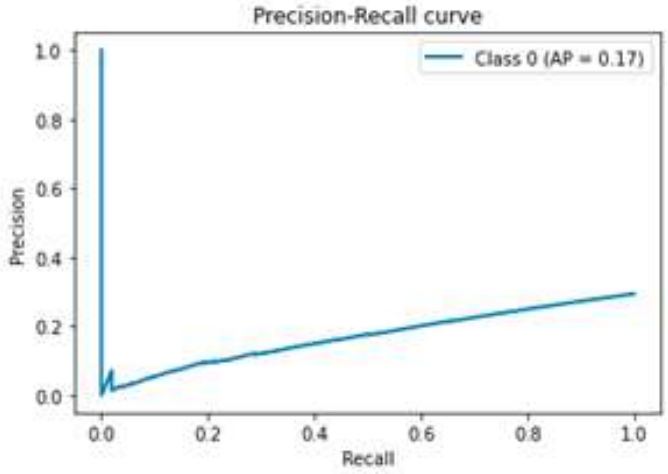

Figure 6. Precision recall curve for logistic regression.

ROC Curve

Demonstrates effective class separability, achieving an AUC of 0.841 (see Figure 7).

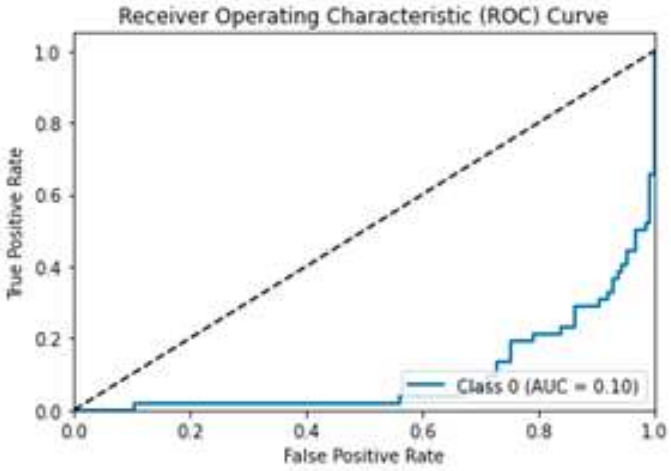

Figure 7. ROC curve for logistic regression.

Feature Importance

The kernel coefficients further highlighted the predictive power of 'LYVE1' and 'plasma_CA19_9' (see Figure 8).

4.3.2 Support Vector Machine

Support vector machine is an efficient supervised learning model that aims to find the ideal hyperplane that divides the data points of different classes with maximum border. SVM utilizes the kernel function to predict the data into a higher dimensional space where linear dividers are not found in the classification of nonlinear data [13]. It performs slightly better accuracy 84.75% than logistic regression and shows a high capacity to classify difficult biomarker data.

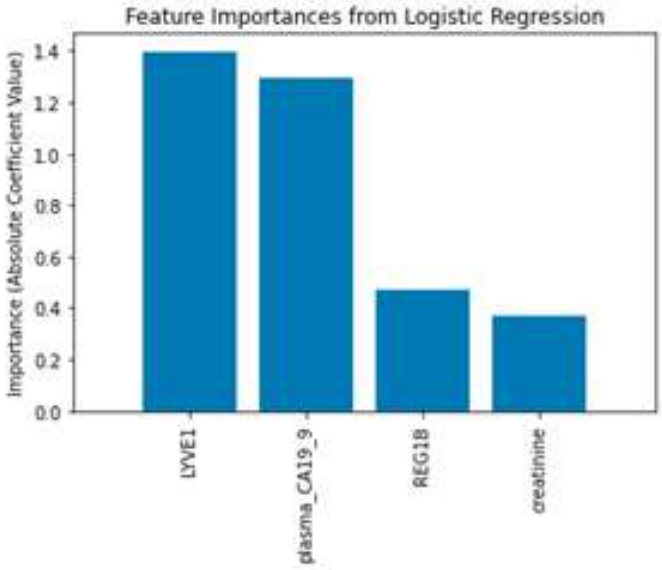

Figure 8. Feature Importance for logistic regression.

Table 4. Results of Support Vector Machine

Accuracy	Precision	Recall	F1-Score	AUC
84.75%	84.66%	84.75%	84.70%	0.847

The key strength of this model is that it outperformed LR in both precision and recall, indicating better robustness for the diagnosis of PDAC (see Table 4).

Confusion Matrix

High classification accuracy with slight misclassification of healthy and non-cancerous pancreatic cases (see Figure 9).

Precision-Recall Curve

Reflects the model's ability to minimize false positives while maintaining high sensitivity (see Figure 10).

ROC Curve

Slight improvement in AUC compared to LR (see Figure 11).

Feature Importance

'plasma_CA19_9' and 'LYVE1' contributed the most in the classification as you can see in the below figure (see Figure 12).

4.3.3 Random Forest

Random forest is a collection of learning methods that is it makes multiple decision trees during training and merges the results to improve the accuracy of the model. In working, it selects random subset data and features to make each tree that makes the model more strong and less likely to overfit [14]. In this study, the random

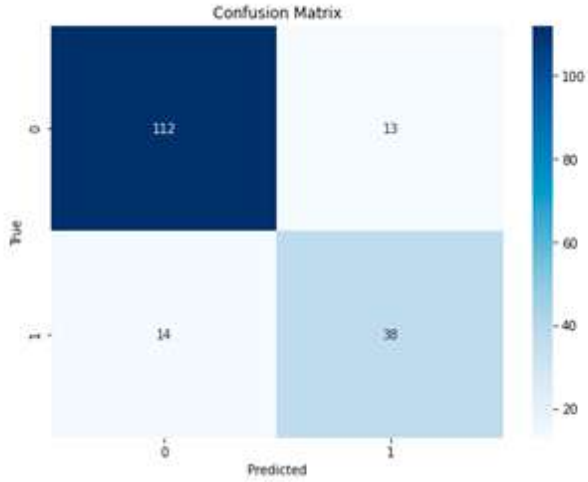

Figure 9. Confusion matrix for SVM.

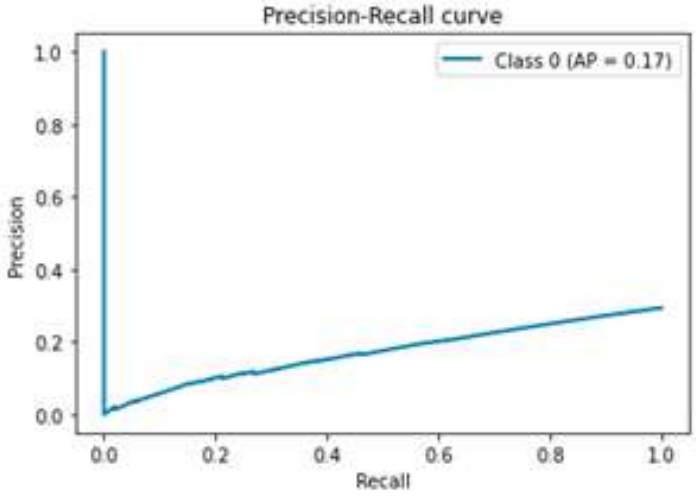

Figure 10. Precision recall curve for SVM.

forest was enhanced using PSO however it achieved the highest accuracy of 89.83% in all the models. It shows a strong ability to manage high-dimensional biomarker data and show important features.

Table 5. Results of Random Forrest

Accuracy	Precision	Recall	F1-Score	AUC
89.83%	89.97%	89.83%	89.89%	0.898

The key strength of this model is that it showed superior overall performance across all metrics, making it the best classifier for this study. Feature importance analysis confirmed the critical role of plasma_CA19_9 and LYVE1 (see Table 5).

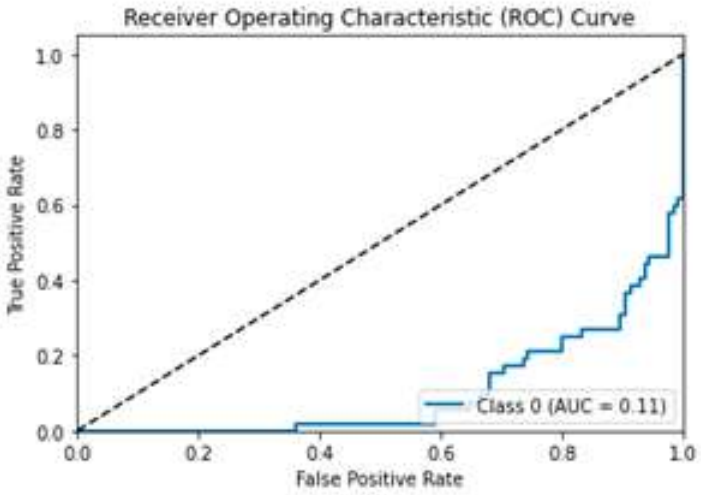

Figure 11. ROC curve for SVM.

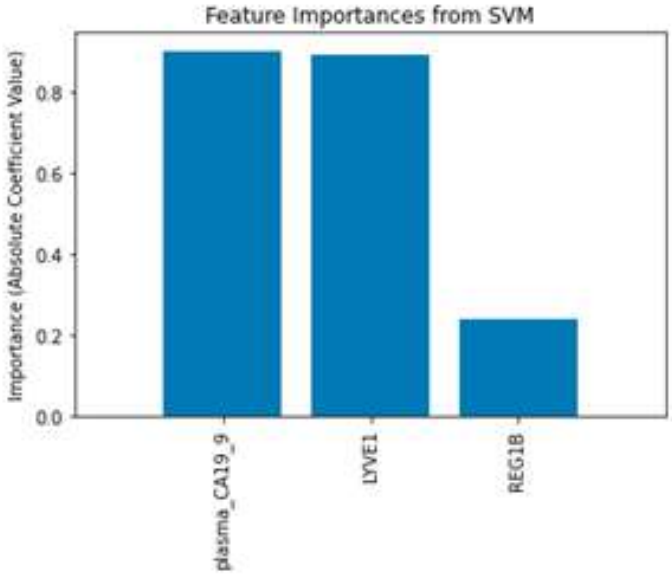

Figure 12. Feature Importance for SVM.

Confusion Matrix

The confusion matrix demonstrates high classification accuracy, with slight misclassifications between the two classes (see Figure 13).

Precision-Recall Curve

The highest precision across all classifiers, indicating its strength in minimizing false positives (see Figure 14).

ROC Curve

Achieved the highest AUC, reinforcing its robustness in distinguishing PDAC from other conditions (see Figure 15).

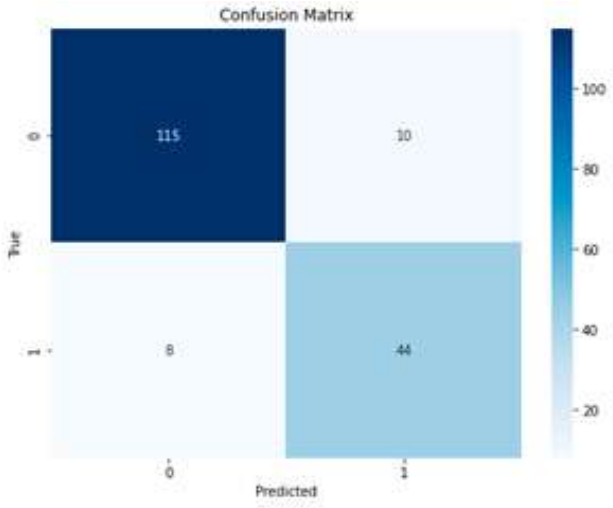

Figure 13. Confusion Matrix for Random Forest.

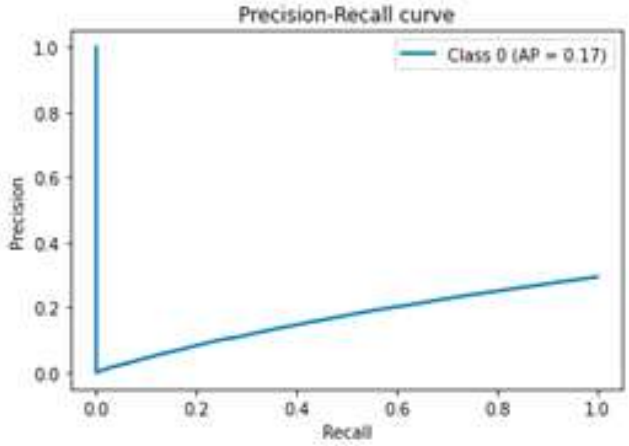

Figure 14. Precision-Recall curve for Random Forest.

Feature Importance

'plasma_CA19_9' contributed the most to classification accuracy, followed by 'LYVE1' (see Figure 16).

Summary of Findings

According to the above results, you can see a successful combination of Machine Learning and urinary biomarkers to design an efficient model for the diagnostic of early PDAC detection. Out of the three evaluated models, Random Forest outperformed in terms of accuracy, recall, and AUC, however, LR balanced simplicity with performance and SVM interpretability better. A consistent choice of plasma_CA19_9 and LYVE1 biomarkers is shown by the PSO. These findings showed the potential of biomarkers integration with ML algorithms, with the help of swarm intelligence (J. Tang, G. Liu, and Q. Pan) [15] the improved efficiency opens a new direction for future research.

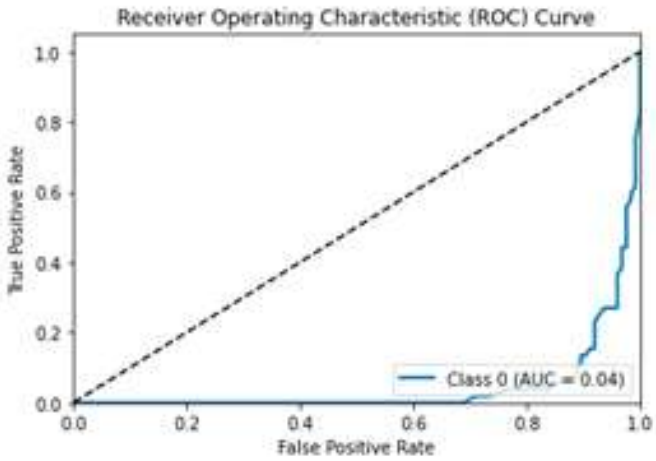

Figure 15. ROC curve for Random Forest.

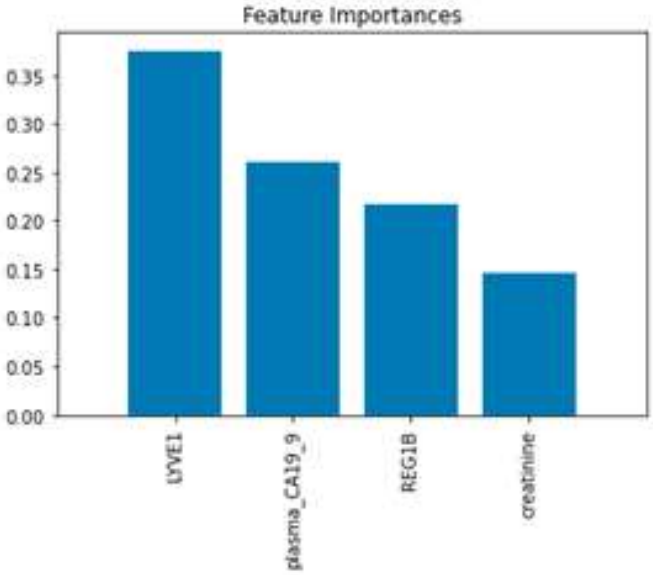

Figure 16. Feature Importance Plot for Random Forest.

5 Comparison

To evaluate the proposed method, the comparative analysis with previous studies is analyzed that (irem Acer et al.2023) developed a machine learning framework including a support vector machine, random forest, light GBM, and gradient boosting with urinary biomarkers with CA19-9 [10]. The accuracy of the gradient boosting classifier is 92.99% in binary classification however its accuracy decreased to 72.91% in a multi-class setting. (Ruben Berreby 2023) developed a logistic regression and random forest model for the detection of urinary biomarker-based pancreatic cancer. The accuracy of logistic regression and random forest is 91% and 89%, respectively, but both models face challenges in detecting minor cancer stages [16].

In contrast, our model study uses particle swarm optimization for selecting a feature and hyper-tuning, which has not been done in any previous studies. Moreover, our model uses a support vector machine with 84.75% accuracy, logistic regression with 84.18% accuracy, and random forest with 89.83% accuracy. In conclusion, the

random forest model showed good performance aligned with the previous study while using PSO for feature selection.

5.1 Difference

1. The utilization of PSO for both selecting feature and hyperparameter tuning forwarded to enhance the model generalization without overfitting.
2. The proposed techniques escape the difficulty of deep learning models like 1D-CNN + LSTM while managing competitive performance.
3. In all the metrices, random forest model manifest consistency, unlike different methods where performance decreased in multiclass or minor stage detection prediction challenges.
4. Random forest easily shows which biomarkers were most important for diagnosis while other uses complex models.

6 Discussion and conclusions

PDAC is one of the most lethal cancers worldwide with a dismal prognosis. Early-stage diagnosis is very challenging due to subtle symptoms and the clinical pattern that correlates with benign gastrointestinal disease. This study determines the performance of machine-learning models to analyze the best machine-learning classifier in the early detection of pancreatic cancer using urine biomarkers. Among different classifiers, RF supervenes with the highest accuracy of 89.83%. After that PSO and SVM with an accuracy of 84.75% and at last LR with 84.18%.

7 Future Work

In future, the further exploration of other swarm intelligence algorithms such as ant colony optimization, artificial bee colony and grey wolf optimizer for feature selection could heighten the effectiveness of the features and lead to the identification of the model with the best accuracy. Despite of this hybridization of multiple swarm algorithms with traditional machine learning can be studied. By expanding the dataset, the model may further be refined.

Author Contributions

Abbas Raza: Conceptualization, Methodology **Muhammad Jawwad**: Data curation, Writing- Original draft preparation. **Khadija Batool**: Visualization, Field knowledge. **Muhammad Sajdain**: Implementation **Ali Raza**: Writing, Reviewing and Editing

Compliance with Ethical Standards

It is declare that all authors don't have any conflict of interest. It is also declare that this article does not contain any studies with human participants or animals performed by any of the authors. Furthermore, informed consent was obtained from all individual participants included in the study.

References

[1] M. E. Karar, N. El-Fishawy, and M. Radad, "Automated classification of urine biomarkers to diagnose pancreatic cancer using 1-d convolutional neural networks," *Journal of Biological Engineering*, vol. 17, no. 1, p. 28, 2023.

[2] S. Debernardi, H. O'Brien, A. S. Algahmdi, N. Malats, G. D. Stewart, M. Plješa-Ercegovac, E. Costello, W. Greenhalf, A. Saad, R. Roberts *et al.*, "A combination of urinary biomarker panel and pancrisk score for earlier detection of pancreatic cancer: A case–control study," *PLoS medicine*, vol. 17, no. 12, p. e1003489, 2020.

[3] T. Kamisawa, L. D. Wood, T. Itoi, and K. Takaori, "Pancreatic cancer," *The Lancet*, vol. 388, no. 10039, pp. 73–85, 2016.

[4] Y. Han, K. J. Jung, U. Kim, C. I. Jeon, K. Lee, and S. H. Jee, "Non-invasive biomarkers for early diagnosis of pancreatic cancer risk: metabolite genomewide association study based on the kcps-ii cohort," *Journal of Translational Medicine*, vol. 21, no. 1, p. 878, 2023.

[5] L. Brezočnik, I. Fister Jr, and V. Podgorelec, "Swarm intelligence algorithms for feature selection: a review," *Applied Sciences*, vol. 8, no. 9, p. 1521, 2018.

[6] S. Samir, M. El-Ashry, W. Soliman, and M. Hassan, "Urinary biomarkers analysis as a diagnostic tool for early detection of pancreatic adenocarcinoma: molecular quantification approach," *Computational Biology and Chemistry*, vol. 112, p. 108171, 2024.

[7] S. Debernardi, O. Blyuss, D. Rycyk, K. Srivastava, C. Y. Jeon, H. Cai, Q. Cai, X.-O. Shu, and T. Crnogorac-Jurcevic, "Urine biomarkers enable pancreatic cancer detection up to 2 years before diagnosis," *International journal of cancer*, vol. 152, no. 4, pp. 769–780, 2023.

[8] O. Blyuss, A. Zaikin, V. Cherepanova, D. Munblit, E. M. Kiseleva, O. M. Prytomanova, S. W. Duffy, and T. Crnogorac-Jurcevic, "Development of pancrisk, a urine biomarker-based risk score for stratified screening of pancreatic cancer patients," *British journal of cancer*, vol. 122, no. 5, pp. 692–696, 2020.

[9] M. P. LaValley, "Logistic regression," *Circulation*, vol. 117, no. 18, pp. 2395–2399, 2008.

[10] I. Acer, F. O. Bulucu, S. Içer, and F. Latifoğlu, "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, pp. 112–125, 2023.

[11] J. J. D. IV, "Urinary biomarkers for pancreatic cancer," https://www.kaggle.com/datasets/johnjdavisiv/urinary-biomarkers-for-pancreatic-cancer, 2024, accessed: 2025-04-28.

[12] P. A. Sunarya, U. Rahardja, S. C. Chen, Y.-M. Lic, and M. Hardini, "Deciphering digital social dynamics: A comparative study of logistic regression and random forest in predicting e-commerce customer behavior," *Journal of Applied Data Sciences*, vol. 5, no. 1, pp. 100–113, 2024.

[13] R. Guido, S. Ferrisi, D. Lofaro, and D. Conforti, "An overview on the advancements of support vector machine models in healthcare applications: a review," *Information*, vol. 15, no. 4, p. 235, 2024.

[14] H. A. Salman, A. Kalakech, and A. Steiti, "Random forest algorithm overview," *Babylonian Journal of Machine Learning*, vol. 2024, pp. 69–79, 2024.

[15] J. Tang, G. Liu, and Q. Pan, "A review on representative swarm intelligence algorithms for solving optimization problems: Applications and trends," *IEEE/CAA Journal of Automatica Sinica*, vol. 8, no. 10, pp. 1627–1643, 2021.

[16] R. Berreby, "Combining urinary biomarker panels and machine learning for earlier detection of pancreatic cancer," *Available at SSRN 4636409*, 2023.