Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Toward Early Detection of Pancreatic Cancer: A Machine Learning Study on Urinary and Plasma Biomarker Data

Shivani Navgire

shivaninavgire2004@gmail.com

Sanjivani University https://orcid.org/0009-0000-5549-3242

Research Article

Keywords: Keywords—pancreatic cancer, early detection, machine learning, urinary biomarkers, CA19-9, random forest, class imbalance, ROCAUC

Posted Date: October 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11474245/v1

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

Additional Declarations: The authors declare no competing interests.

Toward Early Detection of Pancreatic Cancer: A Machine Learning Study on Urinary and Plasma Biomarker Data

Shivani Ravikant Navgire
Department of Artificial Intelligence and Data Science (AIDS), Sanjivani University, Kopargaon, Maharashtra, India
Email: shivaninavgire2004@gmail.com
ORCID: 0009-0000-5549-3242

***Abstract*—*Pancreatic ductal adenocarcinoma (PDAC) is usually diagnosed late, and the only widely used blood marker, CA19-9, is neither sensitive nor specific enough for screening. This paper asks whether a machine learning model that combines four urinary proteins (LYVE1, REG1A, REG1B, TFF1), urinary creatinine, plasma CA19-9, age and sex can separate PDAC from non-PDAC samples (healthy controls and benign hepatobiliary disease) in the public Debernardi et al. dataset of 590 participants (199 PDAC, 391 non-PDAC). The pipeline combines median imputation with missingness indicators, three engineered features, an 80/20 stratified split, SMOTE oversampling of the training data, and a comparison of six classifiers by five-fold cross-validated ROC-AUC. Random Forest performed best (CV ROC-AUC 0.967 ± 0.014) and was tuned by grid search. On a held-out test set of 118 participants the tuned model reached accuracy 0.924, precision 0.878, recall 0.900, F1-score 0.889, ROC-AUC 0.967 and average precision 0.948, detecting 36 of 40 PDAC cases with 5 false positives. Plasma CA19-9, urinary LYVE1 and a CA19-9/creatinine ratio had the highest feature importance. These results come from a single split of a single-source dataset, were not stratified by tumor stage, and were not compared with a CA19-9-only model; they are therefore a proof of concept, not evidence of early-stage screening performance. The limitations and the validation still required are discussed.*

Keywords*—*pancreatic cancer, early detection, machine learning, urinary biomarkers, CA19-9, random forest, class imbalance, ROC-AUC

I. INTRODUCTION

Pancreatic ductal adenocarcinoma (PDAC) is among the most lethal solid tumors. In the United States the five-year relative survival for pancreatic cancer is about 13% overall, but it is 44% when the disease is localized and only 3% once it has spread to distant organs, and only about 17% of patients are diagnosed at the localized stage [1].

The pancreas lies deep in the abdomen, early tumors cause few or nonspecific symptoms, and no screening test is accepted for the general population. The only blood biomarker in routine use, carbohydrate antigen 19-9 (CA19-9), is useful for monitoring but is a weak screening tool: it is elevated in benign conditions such as pancreatitis and biliary obstruction, and it is not produced by people who are Lewis-antigen negative [2]. Urinary proteins have therefore attracted interest as a noninvasive complement. A three-protein urinary panel (LYVE1, REG1A, TFF1) was proposed for early detection [3], and REG1B was later shown to improve on REG1A when the panel was combined with CA19-9 and age in the PancRISK score [4].

Because these markers are continuous, correlated and partly missing in practice, machine learning classifiers are a natural way to combine them. This paper implements and evaluates such a pipeline on the public urinary biomarker dataset of Debernardi *et al.* [4]. Its contributions are: (i) a documented preprocessing pipeline that preserves missingness information and adds three engineered features; (ii) a comparison of six classifiers by cross-validated ROC-AUC; (iii) an evaluation of the tuned model on a held-out test set, reported with exact confidence intervals; (iv) a feature-importance analysis; and (v) an explicit account of what the study does not show, including that no CA19-9-only baseline and no stage-stratified (early-stage) evaluation has yet been run.

Section II reviews related work, Section III describes the data and methods, Section IV reports the results, Section V discusses the findings and limitations, and Section VI concludes.

II. RELATED WORK

Five recent studies are most relevant here. They differ in biomarker type, modeling approach and validation strategy.

A. Urinary Biomarker Panels with Classical Machine Learning

Acer *et al.* [5] compared seven classifiers (SVM, naive Bayes, k-NN, random forest, LightGBM, AdaBoost and gradient boosting) on the same public dataset used here, with five- and ten-fold cross-validation and mean imputation of missing values. Gradient boosting performed best for healthy controls versus PDAC (accuracy 92.99%, AUC 0.976), LightGBM for benign pancreatic disorders versus PDAC (accuracy 86.37%, AUC 0.935), and gradient boosting for the three-class problem (accuracy 72.91%). Ensemble methods outperformed the classical algorithms in all three settings. The evaluation was not restricted to early-stage cases.

B. Deep and Explainable Models

Karar *et al.* [6] combined a one-dimensional convolutional network with an LSTM on four urinary biomarkers (creatinine, LYVE1, REG1B and TFF1) and reported better performance than earlier classical classifiers on this data. Almisned *et al.* [7] trained six machine learning algorithms and an ensemble voting classifier on clinical characteristics and used SHAP [8] to show which features drove each prediction, arguing that interpretability is needed for clinical acceptance.

C. Blood-Based miRNA and Multimodal Panels

Kawai *et al.* [9] sequenced serum microRNA in 425 samples from patients with pancreatic cancer (14 hospitals) and healthy controls, and built automated machine-learning ensembles from

the 100 most highly expressed miRNAs, alone and combined with CA19-9. The miRNA + CA19-9 model reached an AUC of 0.99 (sensitivity 90%, specificity 98%) in the validation cohort; in an independent asymptomatic stage 0–I cohort the AUC was 0.97, but sensitivity fell to 67% at 98% specificity. Strong aggregate discrimination therefore need not translate into high sensitivity for early-stage disease. Yao *et al.* [10] combined 2,096 microRNAs, 125 metabolites and CA19-9 in an age-matched multicenter cohort of 203 participants (46 healthy controls, 36 pancreatitis controls, 121 cases), selected a compact marker subset without supervision, and reported that the panel generalized to validation samples and outperformed recent miRNA-based approaches.

D. Gap Addressed

Studies that explicitly address early-stage or asymptomatic cases [9], [10] rely on costly assays (miRNA sequencing, multi-omic profiling), whereas the studies built on inexpensive urinary panels [5], [6] report aggregate performance. The sensitivity drop reported in [9] shows why that distinction matters. The present work does not close this gap. It provides a transparent, reproducible baseline on the public urinary dataset, with uncertainty quantified and limitations stated, on which stage-stratified analyses can later be built.

III. MATERIALS AND METHODS

A. Dataset

The data come from Debernardi *et al.* [4] and are distributed publicly on Kaggle [11]. The file contains 590 participants and 14 columns: sample identifier, patient cohort (Cohort 1 or 2), sample origin, age, sex, diagnosis, tumor stage, benign diagnosis, plasma CA19-9, and urinary creatinine, LYVE1, REG1B, TFF1 and REG1A. Diagnosis is coded as healthy control (1), benign hepatobiliary disease (2) or PDAC (3). We defined the binary target as PDAC (diagnosis = 3) versus non-PDAC (diagnosis = 1 or 2), giving 199 PDAC (33.7%) and 391 non-PDAC samples. The non-PDAC class therefore includes patients with benign conditions such as chronic pancreatitis, which are harder to separate from cancer than healthy controls [5]. REG1A was measured only for Cohort 1, and about 41% of plasma CA19-9 values are missing. This is a secondary analysis of a public dataset; no new patient data were collected.

B. Preprocessing and Feature Engineering

Missing REG1A and plasma CA19-9 values were replaced by the median, and a binary "measured" indicator was kept for each so that the information that a test was not performed is not lost. Sex was label-encoded (F = 0, M = 1). Three features were engineered: the CA19-9/creatinine ratio, computed as CA19-9 divided by (creatinine + 0.001); the sum LYVE1 + REG1B + TFF1; and an age group coded 0–3 for the bins (0, 40], (40, 55], (55, 70] and (70, 100] years. The final feature set has 13 variables: eight base variables (age, sex, plasma CA19-9, creatinine, LYVE1, REG1B, TFF1, REG1A), two measured indicators and three engineered features. The medians were computed on the full dataset before the train/test split (see Section V-D).

C. Data Splitting and Class Balancing

The data were split 80/20 with stratification on the target (random seed 42), giving 472 training samples (313 non-PDAC, 159 PDAC) and 118 test samples (78 non-PDAC, 40 PDAC). Features were standardized to zero mean and unit variance using statistics from the training set only. The synthetic minority over-sampling technique (SMOTE) [12] was applied to the standardized training set only, balancing it to 313 samples per class (626 in total). The test set was never resampled.

D. Classifiers and Model Selection

Six classifiers were compared: logistic regression, Random Forest [13] (300 trees), gradient boosting, XGBoost [14], a support vector machine with RBF kernel, and k-nearest neighbors (k = 7), all with scikit-learn [15] defaults except where stated. Each was scored by five-fold stratified cross-validation (shuffled, seed 42) on the resampled training set, using ROC-AUC. The best model, Random Forest, was tuned by exhaustive grid search over the number of trees {200, 300, 500}, maximum depth {4, 6, 8, unlimited} and minimum split size {2, 5, 10} (36 configurations), again by five-fold ROC-AUC.

E. Evaluation

The tuned model was evaluated once on the held-out test set at the default probability threshold of 0.5. We report accuracy, precision, recall (sensitivity), F1-score, ROC-AUC and average precision (area under the precision–recall curve). Recall is emphasized because a false negative is a missed cancer, which is a worse error than a false positive that triggers a follow-up test. Exact Clopper–Pearson 95% confidence intervals [16] were computed for proportions from the confusion matrix, and an analytic Hanley–McNeil interval [17] for the AUC. These intervals were computed after the experiment from the reported counts and AUC; the test-set predictions were not bootstrapped.

F. Implementation and Reproducibility

All code was run in Python in Google Colab using pandas, NumPy, scikit-learn [15], imbalanced-learn [18], XGBoost, matplotlib and seaborn. A fixed seed (42) was used for the split, SMOTE, cross-validation and models. The trained model, scaler, imputer, sex encoder and feature list were serialized with joblib. The notebook and saved artifacts are available at [repository URL to be added before submission].

G. Prototype Interface

A Gradio web form loads the saved artifacts, reproduces the same preprocessing and feature engineering for one patient entry, and returns the estimated PDAC probability, labeling it higher risk at or above 0.5. The interface is labeled a research demonstration, not a diagnostic tool, and has not been clinically validated. Fig. 5 and Table IV (Section IV-D) show the interface with two hypothetical inputs.

IV. RESULTS

A. Model Comparison and Tuning

Table I gives the five-fold cross-validated ROC-AUC of the six classifiers on the resampled training set. Random Forest ranked first (0.9666 ± 0.0135), followed by XGBoost (0.9618), gradient boosting (0.9600), k-NN (0.9316), the RBF SVM (0.9286) and logistic regression (0.9103). The three tree ensembles are separated by less than 0.007, well inside the fold-to-fold standard deviation of about 0.014, so their ranking should not be over-interpreted; the clear separation is between the tree ensembles (0.960 or higher) and the linear, kernel and distance-based models (0.932 or lower).

TABLE I. FIVE-FOLD CROSS-VALIDATED ROC-AUC BY CLASSIFIER (TRAINING DATA, AFTER SMOTE)

Classifier	CV ROC-AUC (mean ± SD)
Logistic regression	0.9103 ± 0.0166
Random forest	**0.9666 ± 0.0135**
Gradient boosting	0.9600 ± 0.0149
XGBoost	0.9618 ± 0.0140
SVM (RBF kernel)	0.9286 ± 0.0199
k-NN (k = 7)	0.9316 ± 0.0097

Grid search selected 500 trees, unlimited depth and a minimum split size of 5, with a cross-validated ROC-AUC of 0.9668 against 0.9666 for the untuned 300-tree forest. A gain of 0.0002 is far below the fold-to-fold variability, so tuning changed little: the default forest was already close to what this feature set supports.

B. Held-Out Test Performance

On the 118 held-out participants the tuned Random Forest detected 36 of 40 PDAC cases and missed 4; of 78 non-PDAC cases, 73 were correctly classified and 5 were falsely flagged (Fig. 1). Table II gives the metrics with 95% confidence intervals, and Table III the per-class report. Because the test set contains only 40 PDAC cases, the estimates are imprecise: the exact interval for recall is 76.3%–97.2%, so the data are compatible with a true sensitivity anywhere from the mid-seventies to the high nineties.

TABLE II. TUNED RANDOM FOREST ON THE HELD-OUT TEST SET (N = 118)

Metric	Value	95% CI
Accuracy	0.9237	0.860–0.965
Precision (PPV)	0.8780	0.738–0.959
Recall (sensitivity)	0.9000	0.763–0.972
Specificity*	0.9359	0.857–0.979
Negative predictive value*	0.9481	0.872–0.986
F1-score	0.8889	—
ROC-AUC	0.9667	0.927–1.000†
Average precision	0.948	—

Proportions: exact Clopper–Pearson intervals. †Analytic Hanley–McNeil interval, truncated at 1.0. *Derived from the confusion matrix; not printed by the original notebook.

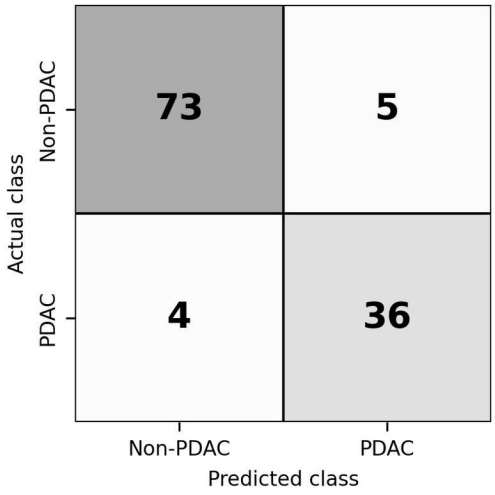


Fig. 1. Confusion matrix of the tuned Random Forest on the held-out test set (N = 118).

TABLE III. PER-CLASS CLASSIFICATION REPORT (TEST SET)

Class	Precision	Recall	F1	Support
Non-PDAC	0.95	0.94	0.94	78
PDAC	0.88	0.90	0.89	40
Macro average	0.91	0.92	0.92	118

The ROC curve (Fig. 2) gives an AUC of 0.967. The precision–recall curve (Fig. 3) gives an average precision of 0.948, against a no-skill baseline equal to the PDAC prevalence in the test set (40/118 = 0.339); precision stayed high over most of the recall range and fell to about 0.6 only as recall approached 1.0.

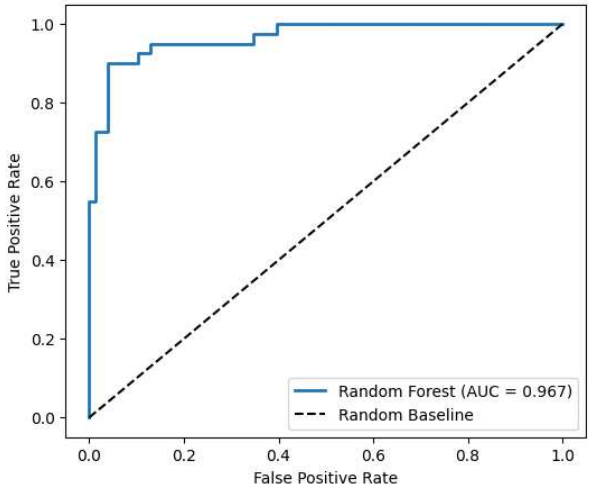


Fig. 2. ROC curve of the tuned Random Forest on the held-out test set (AUC = 0.967).

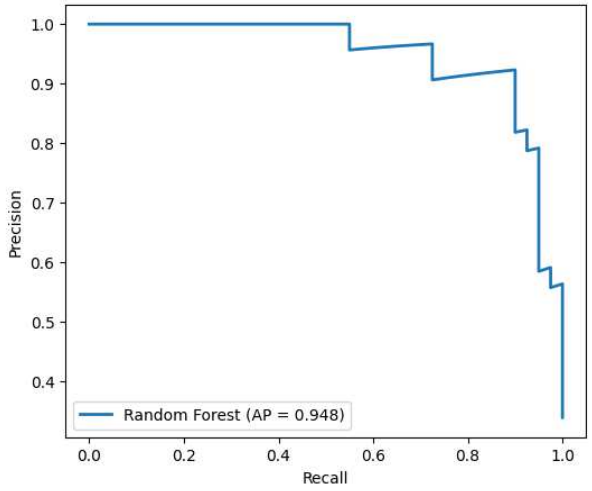


Fig. 3. Precision–recall curve of the tuned Random Forest on the held-out test set (AP = 0.948).

C. Feature Importance

Gini importance from the tuned forest is shown in Fig. 4. Plasma CA19-9 ranked first (0.203), followed by LYVE1 (0.185) and the CA19-9/creatinine ratio (0.157); the LYVE1 + REG1B + TFF1 sum (0.102) and TFF1 (0.088) completed the top five. Grouped by origin, the two CA19-9-based features (raw value and ratio) account for 0.361 of total importance, and the four features built from LYVE1, REG1B and TFF1 (individually and as the sum) account for 0.443. The two measured indicators together account for 0.031. A SHAP analysis [8] was also attempted, but its output was not informative and is not reported.

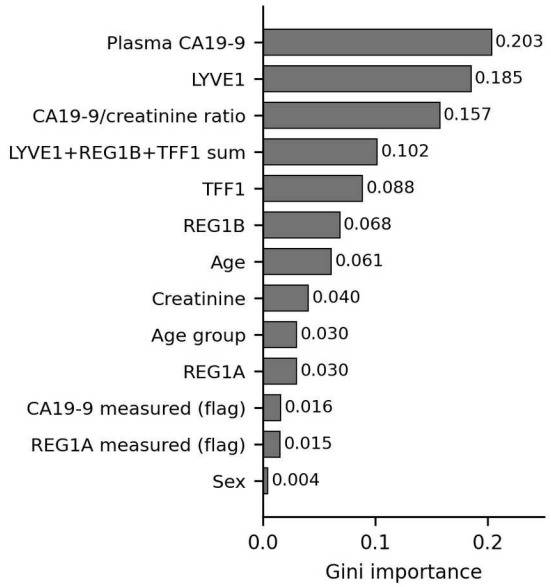


Fig. 4. Gini feature importance of the tuned Random Forest (values sum to 1).

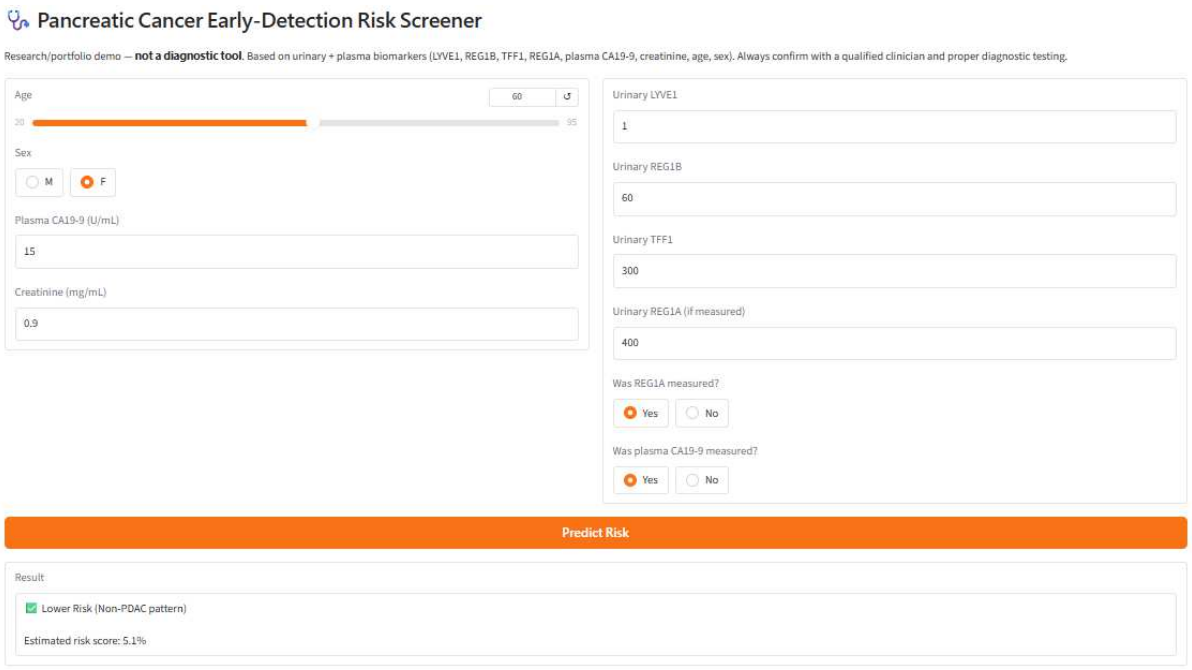
D. Prototype Demonstration

To confirm that the saved artifacts run end to end, the Gradio prototype was exercised with two hypothetical, author-chosen entries (Fig. 5, Table IV). These are not patient records, and each is a single prediction, so they illustrate the interface behavior and do not measure model accuracy. Example (a), a 60-year-old female entry with plasma CA19-9 of 15 U/mL, returned a lower-risk (non-PDAC) pattern with an estimated risk score of 5.1%. Example (b), a 60-year-old male entry with higher urinary TFF1 (600) and the plasma CA19-9 indicator set to “No”, returned a higher-risk (PDAC) pattern with a score of 82.1%. Both outputs are consistent with the 0.5 decision threshold described in Section III-G, and example (b) shows that the interface still returns a score when CA19-9 is flagged as not measured.

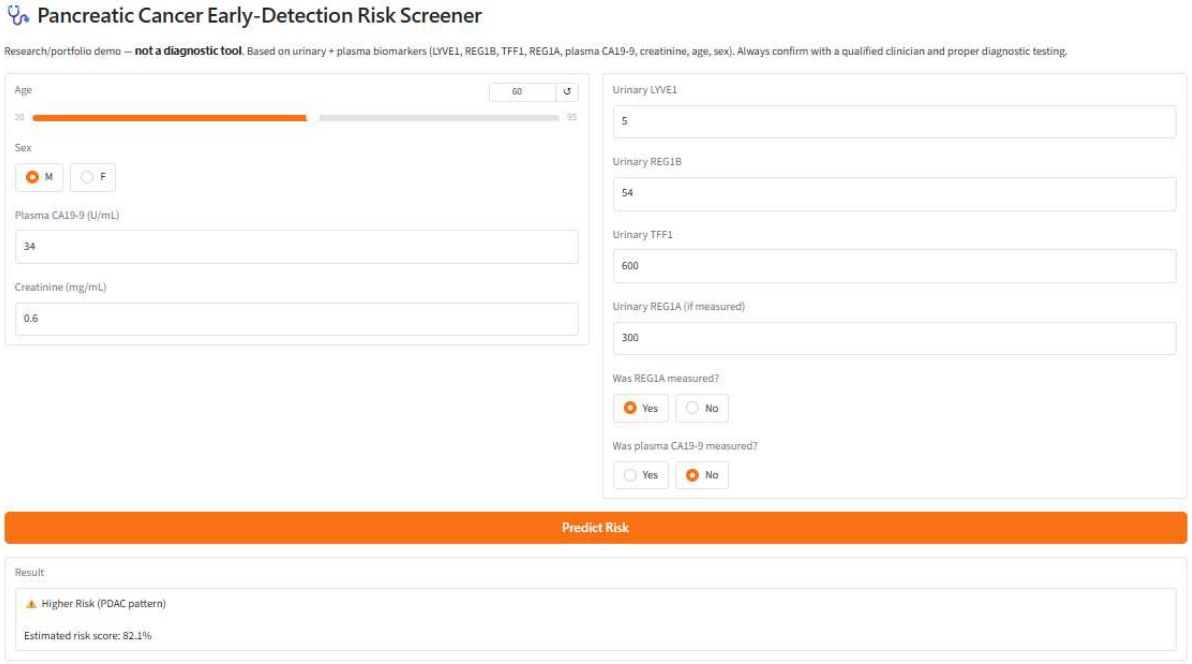
TABLE IV. HYPOTHETICAL INPUTS AND PROTOTYPE OUTPUTS SHOWN IN FIG. 5

Field	**Example (a)**	**Example (b)**
Age (years)	60	60
Sex	F	M
Plasma CA19-9 (U/mL)	15	34
CA19-9 measured	Yes	No
Creatinine (mg/mL)	0.9	0.6
Urinary LYVE1	1	5
Urinary REG1B	60	54
Urinary TFF1	300	600
Urinary REG1A	400	300
REG1A measured	Yes	Yes
Output label	Lower risk (non-PDAC pattern)	Higher risk (PDAC pattern)
Estimated risk score	5.1%	82.1%

Author-chosen illustrative values, not patient data. Output labels and scores are as displayed by the prototype.



(a)



(b)

Fig. 5. Screenshots of the Gradio prototype with hypothetical, author-chosen inputs (not patient data; values in Table IV): (a) a lower-risk output (estimated risk score 5.1%); (b) a higher-risk output (82.1%). The interface is a research demonstration and not a diagnostic tool.

V. DISCUSSION

A. Main Finding

On a held-out set of 118 participants, a tuned Random Forest separated PDAC from non-PDAC with ROC-AUC 0.967, recall 0.900 and precision 0.878, missing 4 of 40 cancers and raising 5 false alarms among 78 non-PDAC cases. Given the small test set these are point estimates with wide intervals (for example recall 0.76–0.97), so the evidence supports promising discrimination on this dataset, not a specific operating performance.

B. Context in the Literature

Acer *et al.* [5] report AUCs of 0.976 for healthy controls versus PDAC and 0.935 for benign disorders versus PDAC on the same data. Our task pools both kinds of non-PDAC samples and uses a different feature set, split and validation scheme, so the numbers cannot be ranked directly; our test AUC of 0.967 lies between those two results. The original study derived its PancRISK score from logistic regression [4], and [5] tabulate AUCs of 0.992 (control versus PDAC) and 0.919 (benign versus PDAC) for it. These are not directly comparable to our

pooled task, but they show that a simple linear model can perform well on these markers, so the benefit of non-linear ensembles has to be tested directly under one protocol rather than assumed.

C. Contribution of CA19-9 and Urinary Markers

CA19-9-based features carry about 36% of the forest's importance and the LYVE1/REG1B/TFF1-based features about 44%, which is consistent with the premise that urinary proteins add information beyond CA19-9. Importance is not incremental value, however. Demonstrating it requires a CA19-9-only model trained under the same protocol and compared with the full model using paired tests such as DeLong's test for AUC [19] and McNemar's test [20]. Gini importance is also biased toward continuous features and splits credit among correlated or derived features [21], as the ratio and sum features here are.

D. Limitations

The study has the following limitations, which bear directly on how far the results can be generalized.

1) **Single split, small test set.** All test results come from one 80/20 split with 40 PDAC cases; repeated cross-validation or bootstrapping was not run.
2) **Resampling before cross-validation.** SMOTE was applied before the cross-validation used for model comparison and tuning, so synthetic samples derived from validation-fold neighbors can appear in training folds and inflate the cross-validated scores [22]. The test-set results are not affected, because SMOTE was applied to the training data only, and the cross-validated and test AUCs agree closely (0.967 and 0.967), which is reassuring but does not rule out bias in model selection. Resampling should be placed inside each cross-validation fold.
3) **Imputation.** The medians used to fill REG1A and CA19-9 were computed on the full dataset before splitting, which lets a small amount of test-set information into training. Median filling of a large share of values is also crude, and the measured indicators may encode the cohort or protocol rather than biology (they carry 3.1% of importance).
4) **No baseline or significance testing.** No CA19-9-only, logistic-regression or PancRISK-style model was trained, and no paired test between models was run.
5) **No stage-stratified evaluation.** The tumor-stage field is populated for only a minority of records and was not used. Whether the model detects stage I–II or asymptomatic cancers, where [9] observed a sensitivity drop, is unknown. For this reason the results should not be read as evidence of early-detection performance.
6) **Case–control data and prevalence.** PDAC makes up 34% of this dataset, far above any screening population. Predictive values do not transfer: with the point estimates for sensitivity (0.90) and specificity (0.936), the positive predictive value would be only about 12% at an illustrative prevalence of 1%.
7) **Single source, no external validation.** All samples come from one study of 590 participants, and no independent cohort was used.
8) **Interpretability.** Gini importance is a limited explanation, and the attempted SHAP analysis did not yield usable output; a corrected SHAP analysis is still needed.
9) **Fixed threshold and prototype.** The 0.5 threshold was not tuned for a sensitivity target, and the Gradio prototype is a demonstration that has not been clinically validated.

VI. CONCLUSION AND FUTURE WORK

A Random Forest trained on urinary proteins, plasma CA19-9, creatinine, age and sex separated PDAC from non-PDAC samples in the Debernardi dataset with a held-out ROC-AUC of 0.967, recall of 0.900 and precision of 0.878. Because of the single split, the absence of a CA19-9-only baseline, and the lack of stage-specific evaluation, this is a proof of concept and not evidence that the approach detects early-stage disease.

Future work will (i) rebuild the pipeline so that imputation, scaling and resampling are fitted inside each cross-validation fold, with repeated stratified cross-validation and confidence intervals; (ii) train CA19-9-only and logistic-regression baselines under the same protocol and compare them using DeLong's and McNemar's tests; (iii) evaluate performance by tumor stage using the stage field and, where possible, asymptomatic cohorts; (iv) validate on an independent external cohort; (v) calibrate predicted probabilities and choose an operating threshold for a target sensitivity; (vi) produce a corrected SHAP analysis; and (vii) test whether panels built only from routine hospital blood tests (for example CEA, bilirubin, glucose, amylase and lipase) can approach the performance of urinary panels.

ACKNOWLEDGMENT

The author thanks the faculty of the Department of Artificial Intelligence and Data Science for guidance on the research methodology, the authors of the studies cited in Section II for making their methods and findings available, and Debernardi *et al.* for releasing their dataset publicly.

REFERENCES

[1] American Cancer Society, *Cancer Facts & Figures 2026*. Atlanta, GA, USA: American Cancer Society, 2026.

[2] K. S. Goonetilleke and A. K. Siriwardena, "Systematic review of carbohydrate antigen (CA 19-9) as a biochemical marker in the diagnosis of pancreatic cancer," *Eur. J. Surg. Oncol.*, vol. 33, no. 3, pp. 266–270, 2007, doi: 10.1016/j.ejso.2006.10.004.

[3] T. P. Radon *et al.*, "Identification of a three-biomarker panel in urine for early detection of pancreatic adenocarcinoma," *Clin. Cancer Res.*, vol. 21, no. 15, pp. 3512–3521, 2015.

[4] S. Debernardi *et al.*, "A combination of urinary biomarker panel and PancRISK score for earlier detection of pancreatic cancer: A case–control study," *PLoS Med.*, vol. 17, no. 12, Art. no. e1003489, 2020, doi: 10.1371/journal.pmed.1003489.

[5] İ. Acer, F. Orhanbulucu, S. İçer, and F. Latifoğlu, "Early diagnosis of pancreatic cancer by machine learning methods using urine biomarker combinations," *Turk. J. Electr. Eng. Comput. Sci.*, vol. 31, no. 1, pp. 112–125, 2023, doi: 10.55730/1300-0632.3973.

[6] M. E. Karar, N. El-Fishawy, and M. Radad, "Automated classification of urine biomarkers to diagnose pancreatic cancer using 1-D convolutional

neural networks," *J. Biol. Eng.*, vol. 17, Art. no. 28, 2023, doi: 10.1186/s13036-023-00340-0.

[7] F. A. Almisned, N. Usanase, D. U. Ozsahin, and I. Ozsahin, "Incorporation of explainable artificial intelligence in ensemble machine learning-driven pancreatic cancer diagnosis," *Sci. Rep.*, 2025.

[8] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, 2017, pp. 4765–4774.

[9] M. Kawai *et al.*, "Early detection of pancreatic cancer by comprehensive serum miRNA sequencing with automated machine learning," *Br. J. Cancer*, 2024, doi: 10.1038/s41416-024-02794-5.

[10] T.-H. Yao *et al.*, "Machine learning-based multimodal biomarkers enable accurate diagnosis and early detection of pancreatic ductal adenocarcinoma," *Sci. Rep.*, vol. 16, 2026, doi: 10.1038/s41598-025-29548-4.

[11] Kaggle, "Urinary Biomarkers for Pancreatic Cancer," dataset, 2020. [Online]. Available: https://www.kaggle.com/datasets/johnjdavisiv/urinary-biomarkers-for-pancreatic-cancer

[12] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[13] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[14] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 785–794.

[15] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[16] C. J. Clopper and E. S. Pearson, "The use of confidence or fiducial limits illustrated in the case of the binomial," *Biometrika*, vol. 26, no. 4, pp. 404–413, 1934.

[17] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

[18] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning," *J. Mach. Learn. Res.*, vol. 18, no. 17, pp. 1–5, 2017.

[19] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach," *Biometrics*, vol. 44, no. 3, pp. 837–845, 1988.

[20] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947.

[21] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution," *BMC Bioinformatics*, vol. 8, Art. no. 25, 2007.

[22] M. S. Santos, J. P. Soares, P. H. Abreu, H. Araujo, and J. Santos, "Cross-validation for imbalanced datasets: Avoiding overoptimistic and overfitting approaches," *IEEE Comput. Intell. Mag.*, vol. 13, no. 4, pp. 59–76, Nov. 2018.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- confusionmatrix.png
- gini.png
- precisionrecall.png
- ROC.png