Published in final edited form as:

Gastroenterology. 2022 November ; 163(5): 1435–1446.e3. doi:10.1053/j.gastro.2022.06.066.

Radiomics-based Machine-learning Models Can Detect Pancreatic Cancer on Prediagnostic Computed Tomography Scans at a Substantial Lead Time Before Clinical Diagnosis

Sovanlal Mukherjee[1], **Anurima Patra**[2], **Hala Khasawneh**[1], **Panagiotis Korfiatis**[1], **Naveen Rajamohan**[1], **Garima Suman**[1], **Shounak Majumder**[3], **Ananya Panda**[1], **Matthew P. Johnson**[4], **Nicholas B. Larson**[1,4], **Darryl E. Wright**[1], **Timothy L. Kline**[1], **Joel G. Fletcher**[1], **Suresh T. Chari**[3,5], **Ajit H. Goenka**[1]

[1]Department of Radiology, Mayo Clinic, Rochester, Minnesota

[2]Department of Radiology, Tata Medical Centre, Kolkata, India

[3]Department of Gastroenterology, Mayo Clinic, Rochester, Minnesota

[4]Department of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, Minnesota

Correspondence Address correspondence to Ajit H. Goenka, MD, Department of Radiology, Mayo Clinic, 200 First Street SW, Charlton 1, Rochester, Minnesota 55905. goenka.ajit@mayo.edu.

Data Transparency

All the authors had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analyses. All authors were responsible for the critical revision of the manuscript for important intellectual content.

CRediT Authorship Contributions

Order of Authors (with Contributor Roles):

Sovanlal Mukherjee, PhD (Data curation: Equal; Formal analysis: Lead; Investigation: Equal; Methodology: Equal; Software: Lead; Validation: Lead; Writing – original draft: Equal; Writing – review & editing: Equal).

Anurima Patra, MD (Conceptualization: Supporting; Data curation: Lead; Investigation: Supporting; Methodology: Supporting; Writing – original draft: Supporting; Writing – review & editing: Supporting).

Hala Khasawneh, MD (Conceptualization: Supporting; Data curation: Supporting; Investigation: Supporting; Methodology: Supporting; Writing – original draft: Supporting; Writing – review & editing: Supporting).

Panagiotis Korfiatis, PhD (Conceptualization: Supporting; Data curation: Supporting; Formal analysis: Supporting; Investigation: Supporting; Methodology: Supporting; Supervision: Supporting; Writing – review & editing: Supporting).

Naveen Rajamohan, MD (Data curation: Supporting; Investigation: Supporting; Methodology: Supporting; Writing – review & editing: Supporting).

Garima Suman, MD (Conceptualization: Supporting; Data curation: Supporting; Investigation: Supporting; Methodology: Supporting; Writing – review & editing: Supporting).

Shounak Majumder, MD (Investigation: Supporting; Supervision: Supporting; Writing – review & editing: Supporting).

Ananya Panda, MD (Investigation: Supporting; Methodology: Supporting; Writing – review & editing: Supporting).

Matthew P. Johnson, MS (Formal analysis: Supporting; Investigation: Supporting; Writing – review & editing: Supporting).

Nicholas B. Larson, PhD (Formal analysis: Supporting; Investigation: Supporting; Writing – review & editing: Supporting).

Darryl E. Wright, PhD (Formal analysis: Supporting; Investigation: Supporting; Writing – review & editing: Supporting).

Timothy L. Kline, PhD (Formal analysis: Supporting; Investigation: Supporting; Supervision: Supporting; Writing – review & editing: Supporting).

Joel G. Fletcher, MD (Investigation: Supporting; Methodology: Supporting; Supervision: Supporting; Writing – review & editing: Supporting).

Suresh T. Chari, MD (Conceptualization: Supporting; Investigation: Supporting; Supervision: Supporting; Writing – review & editing: Supporting).

Ajit Harishkumar Goenka, MD (Conceptualization: Lead; Data curation: Supporting; Formal analysis: Supporting; Funding acquisition: Lead; Investigation: Equal; Methodology: Equal; Project administration: Lead; Resources: Lead; Supervision: Lead; Writing – original draft: Lead; Writing – review & editing: Lead).

Conflicts of interest

The authors disclose no conflicts.

[5]Department of Gastroenterology, Hepatology, and Nutrition, The University of Texas MD Anderson Cancer Center, Houston, Texas

Abstract

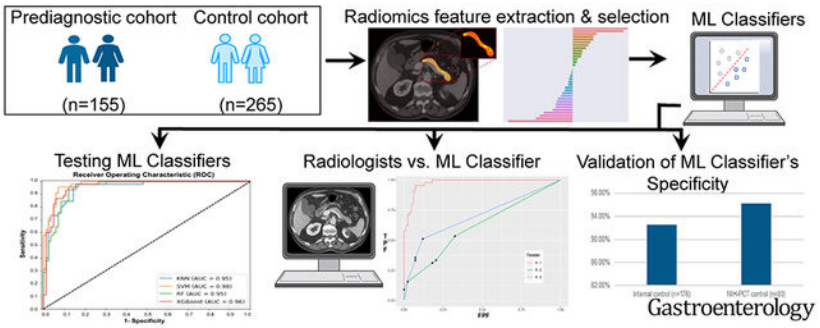
BACKGROUND & AIMS: Our purpose was to detect pancreatic ductal adenocarcinoma (PDAC) at the prediagnostic stage (3–36 months before clinical diagnosis) using radiomics-based machine-learning (ML) models, and to compare performance against radiologists in a case-control study.

METHODS: Volumetric pancreas segmentation was performed on prediagnostic computed tomography scans (CTs) (median interval between CT and PDAC diagnosis: 398 days) of 155 patients and an age-matched cohort of 265 subjects with normal pancreas. A total of 88 first-order and gray-level radiomic features were extracted and 34 features were selected through the least absolute shrinkage and selection operator–based feature selection method. The dataset was randomly divided into training (292 CTs: 110 prediagnostic and 182 controls) and test subsets (128 CTs: 45 prediagnostic and 83 controls). Four ML classifiers, k-nearest neighbor (KNN), support vector machine (SVM), random forest (RM), and extreme gradient boosting (XGBoost), were evaluated. Specificity of model with highest accuracy was further validated on an independent internal dataset (n = 176) and the public National Institutes of Health dataset (n = 80). Two radiologists (R4 and R5) independently evaluated the pancreas on a 5-point diagnostic scale.

RESULTS: Median (range) time between prediagnostic CTs of the test subset and PDAC diagnosis was 386 (97–1092) days. SVM had the highest sensitivity (mean; 95% confidence interval) (95.5; 85.5–100.0), specificity (90.3; 84.3–91.5), F1-score (89.5; 82.3–91.7), area under the curve (AUC) (0.98; 0.94–0.98), and accuracy (92.2%; 86.7–93.7) for classification of CTs into prediagnostic versus normal. All 3 other ML models, KNN, RF, and XGBoost, had comparable AUCs (0.95, 0.95, and 0.96, respectively). The high specificity of SVM was generalizable to both the independent internal (92.6%) and the National Institutes of Health dataset (96.2%). In contrast, interreader radiologist agreement was only fair (Cohen's kappa 0.3) and their mean AUC (0.66; 0.46–0.86) was lower than each of the 4 ML models (AUCs: 0.95–0.98) ($P < .001$). Radiologists also recorded false positive indirect findings of PDAC in control subjects (n = 83) (7% R4, 18% R5).

CONCLUSIONS: Radiomics-based ML models can detect PDAC from normal pancreas when it is beyond human interrogation capability at a substantial lead time before clinical diagnosis. Prospective validation and integration of such models with complementary fluid-based biomarkers has the potential for PDAC detection at a stage when surgical cure is a possibility.

Graphical Abstract



Keywords

Pancreas; Artificial Intelligence; Biomarkers; Pancreatic Ductal Carcinoma; X-Ray Computed Tomography

Pancreatic ductal adenocarcinoma (PDAC) is the third leading cause of cancer-related deaths in the United States. In fact, PDAC is an almost uniformly fatal disease, with the number of deaths almost equal to the incidence. Despite its overall grim prognosis, there are substantial differences in survival of patients with stage I vs stage IV disease (median disease-specific survival: 26 vs 4.8-months, respectively).[1] These differences highlight the urgent unmet need for early detection of PDAC to improve outcomes. Early detection offers survival benefit beyond lead time, as tumors are smaller in volume and more likely to be resectable when detected earlier.[2-4] Second, early detection before the decline in performance status from cancer-induced cachexia increases the prospect of surgical resection in eligible patients.[2,5]

Recently, cohorts at sufficiently high risk of PDAC to experience a potential net benefit from screening have been identified. One such cohort is subjects with new-onset diabetes and a high score (≥3) on the Enriching New-Onset Diabetes for Pancreatic Cancer (END-PAC) model.[4] A computed tomography (CT) scan performed at the onset of new-onset diabetes in such a high-risk group has the potential to increase the proportion of resectable PDACs to 3 times higher than the current resectability rate.[2] A risk-based screening strategy using CT in such cohorts has societal economic value across a range of analyses.[1] A large prospective trial, the Early Detection Initiative (EDI) (NCT04662879), sponsored by the Pancreatic Cancer Action Network, is under way to evaluate outcomes of a screening strategy in 12,500 participants by using the END-PAC model and CT.[6] These exciting developments underscore the necessity for simultaneous investigations into novel imaging approaches for detection of PDAC at a time when the disease is potentially curable.

In patients with PDAC, subtle indirect findings suggestive of cancer (eg, pancreatic duct cutoff or dilatation, focal pancreatic atrophy) can be present on CTs at the prediagnostic stage (ie, 3–36 months before clinical diagnosis).[2,7,8] However, these findings are often overlooked in clinical practice, leading to delayed diagnosis.[9,10] Second, these findings are not specific for early PDAC because they are also seen in control subjects.[2,11] Importantly, the pancreas tends to be morphologically normal on CT at prediagnostic stage in many patients. This is a major challenge because the cancer can rapidly advance from being

subclinical on imaging to stage IV.[2,11] Thus, there is a critical need for novel methods such as artificial intelligence (AI) for detection of PDAC at a stage when the cancer is subclinical. We hypothesized that the subclinical pancreatic changes at the prediagnostic stage of PDAC can be detected through advanced computational techniques such as radiomics. Radiomics entails extraction and quantification of high-throughput imaging biomarkers that are beyond human perceptible range. These biomarkers are combined with various machine-learning (ML) techniques to identify imaging signatures of subtle yet complex tissue changes for a range of clinical applications.[12-15] Our purpose was to detect PDAC at the prediagnostic stage using radiomics-based ML models, and to compare the performance of such ML models against radiologists in a case-control study.

Materials and Methods

Study Participants

This Health Insurance Portability and Accountability Act–compliant, case-control retrospective study was approved by our institutional review board, which waived the requirement for written informed consent.

Patients with prediagnostic CTs. Using our electronic medical record, patients diagnosed with biopsy-proven PDAC between January 2006 and December 2020 (n = 3000) were identified. The radiology records of these patients were searched to select prediagnostic contrast-enhanced CT abdomen studies, which were defined as incidental CT scans performed for unrelated indications between 3 and 36 months before PDAC diagnosis (n = 581). In patients who had more than 1 prediagnostic CT, the CT temporally closest to the date of PDAC diagnosis was selected. The curation process led to a dataset of 155 prediagnostic CTs (90 men, 65 women; median [range] age: 69 [34–88] years) (Figure 1). Of these, 49 CTs (31.6%) had been performed at our institution, whereas the other 106 (68.4%) CTs were performed at other institutions. The median (range) time interval between prediagnostic CTs and histopathological diagnosis of PDAC was 398 (93–1092) days.

The median (range) CT slice thickness was 3 (0.5–5) mm. CTs were acquired on CT systems from 4 different vendors (81 Siemens, Munich, Germany [52.3%]; 41 GE, Boston, MA [26.4%]; 25 Toshiba, Tokyo, Japan [16.1%]; and 8 Philips, Amsterdam, The Netherlands [5.2%]). All these CTs had been previously interpreted to be negative for PDAC during routine clinical interpretation at our institution. Each CT was rereviewed by 1 of 3 radiologists (R1, R2, R3 with 2–4 years of post-radiology residency experience) who confirmed optimal image quality (eg, lack of motion artifacts compromising visual interpretation), portal venous phase of enhancement, absence of pancreatitis, focal solid or cystic lesion, and biliary or pancreatic duct stent (Figure 2). In addition, radiologists in consensus evaluated each prediagnostic CT for any indirect or secondary imaging signs of early PDAC, such as focal contour abnormality or attenuation difference, biliary or pancreatic duct dilatation or cutoff, and focal parenchymal atrophy.

Control subjects with normal pancreas. We created an age-matched control cohort of CTs with normal pancreas, which was randomly drawn using our Radiology Information System. From this cohort, we selected an age-matched set of control patients (n = 265)

(140 men, 125 women; median [range] age: 67 [30–89] years) who were not diagnosed subsequently with PDAC during 3 years or more of follow-up (Figure 1). Of these, 87 CTs (32.8%) were from our institution, whereas the other 178 (67.2%) CTs were from other institutions. Median (range) CT slice thickness was 2 (0.5–3) mm. CTs had been performed on CT systems from 3 different vendors (219 Siemens [82.6%], 19 GE [7.2%], 27 Toshiba [10.2%]). These control CTs were also curated by radiologists to confirm optimal image quality, portal venous phase, and normal pancreas morphology. Of the 265 control subjects, 178 (67.2%) were outpatients, 81 (30.6%) were from the emergency department, and 6 (2.2%) were in-patients at the time that their CTs had been acquired.

The preceding dataset of prediagnostic CTs (n = 155) and age-matched control CTs with normal pancreas (n = 265) was randomly divided into training-validation (n = 292) and test (n = 128) subsets using a 70% / 30% split for ML model development and testing. For external validation of the model with highest accuracy, we stratified our test subset according to the institution where CTs had been done: 45 (35.2%) were from our institution (17 [37.8%] prediagnostic CTs and 28 [62.2%] control CTs), whereas 83 (64.8%) were from external institutions (28 [33.7%] prediagnostic CTs and 55 [66.3%] control CTs).

Independent internal validation of specificity. For additional independent validation of the specificity, we created another internal hold-out set of 176 CTs (68 men, 108 women; median [range] age: 45 [19–94] years) who had normal pancreas and did not develop PDAC during at least 3 years or more of follow-up. Of these, 68 CTs (38.6%) were performed our institution, whereas the other 108 (61.4%) CTs had been performed at other institutions. Median (range) CT slice thickness was 3 (0.6–5) mm and the CTs had been acquired on CT systems from 3 different vendors (167 Siemens [94.9%], 8 GE [4.5%], and 1 Toshiba [0.6%]). These CTs were used to validate the specificity of the ML model that had highest area under the curve (AUC) on the test subset.

External validation of specificity. There is no public dataset of prediagnostic CTs. Therefore, to externally validate its specificity, the classifier with the highest AUC on the internal test subset was tested on the National Institutes of Health-Pancreas CT (NIH-PCT) dataset. This public dataset consists of 80 abdomen CTs acquired in portal venous (53 men, 27 women; mean [standard deviation (SD)] age: 46.8 [16.7] years) from healthy subjects without pancreas pathology.[16] All CTs have morphologically normal pancreas. The CTs had been acquired from 2 different vendors (Siemens and Philips) with slice thickness range of 1.5 to 2.5 mm. The public dataset includes pancreas segmentation masks, which we reviewed to ensure the accuracy of segmentation boundaries. The ML classifier was tested on this dataset by measuring the proportion of CTs that it correctly classified as normal (ie, specificity).

Volumetric Pancreas Segmentation

Standard acquisition and reconstruction parameters as per our previously published protocols[17] had been followed for all CTs performed at our institution. All CT studies were downloaded and de-identified by anonymization of Digital Imaging and Communication in Medicine (DICOM) tags using Clinical Trial Processor.[18] Metadata elements such

as scanner vendor and CT slice thickness were extracted from DICOM headers. These anonymized CTs were then converted into the Neuroimaging Informatics Technology Initiative (NIfTI) format. The radiologists performed volumetric pancreas segmentations using the boundary-points based segmentation mode of the organ segmentation module (NVIDIA) in 3D Slicer (Version 4.11.20210226) as previously described.[17] Any tissue with Hounsfield Unit (HU) less than 0 was eliminated by using the "logical operators" tool in 3D Slicer to eliminate the peripancreatic fat from the segmentation mask. Pancreas volumes were extracted from the volumetric segmentations and were compared between the prediagnostic and the control CTs.

Feature Extraction and Reduction

Radiomic analyses were performed with software written in Python (version 3.8.5; Python Software Foundation, Wilmington, DE) and leveraging the PyRadiomics library (version 3.0.1).[19] Images were pre-processed with a modified soft tissue CT window (level 50 HU, width 500 HU) and a bin width of 25. Intensity of the images was rescaled to the range of 0 to 255. A total of 88 radiomics features were extracted from each segmentation mask, which included 18 first-order and 70 texture features. The latter included 24 gray-level cooccurrence matrices (GLCMs), 16 gray-level run length matrices (GLRLMs), 16 gray-level size zone matrices (GLSZMs), and 14 gray-level dependence matrices (GLDMs). All features were normalized using a z-score normalization.

Feature selection was done with least absolute shrinkage and selection operator (LASSO)[20] logistic regression, which is an L1-regularization–based method that penalizes the L1-norm of the feature weight coefficients. Therefore, it reduces the model's complexity by eliminating some of the coefficients (ie, coefficients become zero) and the corresponding features to minimize overfitting and improve generalization. Hyperparameter for LASSO was evaluated using stratified 5-fold cross-validation–based grid search method on the training set (n = 292). The parameter that provided the highest cross-validation AUC was selected.

Exclusion of CT Slice Thickness Dependent Radiomics Features

To reduce the potentially confounding effect of slice thickness on the ML models, we used a recently described process[21] to identify and then account for potential dependence of the extracted features on slice thickness: First, all CTs were divided into 2 groups: slice thickness ≥3 mm and slice thickness <3 mm. The selection of 3 mm as a cutoff was based on the current clinical practice at our institution. Second, AUC was calculated for each feature identified through the LASSO selection process based on the slice thickness group labels. The AUC provided a goodness-of-fit measure of a given feature with the binary outcome (groups of slice thickness ≥3 mm and <3 mm). Features with a high AUC (> 0.8) were deemed to be biased by CT slice thickness and were, therefore, removed.

Model Development and Testing

After feature selection through LASSO and exclusion of CT slice thickness dependent features, 4 independent ML classifiers based on k-nearest neighbor (KNN), support vector machine (SVM), random forest (RF), and extreme gradient boosting (XGBoost) were

trained. A 5-fold stratified cross-validation–based randomized search was applied to evaluate each model's hyperparameters. The parameters that yielded the highest cross-validation AUC were selected for each classifier model. Each classifier model was evaluated on the CTs in the test subset (n = 128). As mentioned previously, external validation of the model with highest accuracy was performed by stratifying the test subset according to the institution (35.2% CTs from our institution and 64.8% CTs from external institutions). Its specificity was also evaluated on an independent internal validation set (n = 176) and on the public NIH-PCT dataset (n = 80) with normal pancreas. All the incorrectly classified CTs were reviewed in consensus by R1, R2, and R3 to identify any potential causes for misclassification. Feature selection process and the 3 ML models (KNN, SVM, and RF) were implemented using scikit-learn library (version 0.24.2). A scikit-learn–compatible Python-based package was used for the XGBoost model.[22]

Finally, to identify features that had the highest influence on the sensitivity of the best performing model, we used an ablation-type methodology. In this approach, 1 feature was removed from the selected set of features and sensitivity of the model without this feature was compared against the model's sensitivity with all the features. This procedure was repeated for each feature to identify those features whose removal resulted in the highest drop in the model's sensitivity.

Multireader Evaluation of the Test Subset

Two radiologist readers (R4 with 9 years and R5 with 3 years of post-residency experience) who were blinded to the outcomes and the distribution of prediagnostic vs control CTs independently evaluated the pancreas in the test subset CTs. Interpretation was performed under identical ambient conditions with calibrated DICOM–compliant monitors at a PACS workstation (Visage Imaging, San Diego, CA). Each reader independently rated their assessment of the pancreas on a 5-point scale: 1, definitely normal; 2, probably normal; 3, equivocal; 4, probably abnormal; and 5, definitely abnormal. For CTs with score 4 or 5, readers recorded the imaging findings that supported the score. The classification performance of the 2 readers (R4 and R5) was compared against the radiomics-based ML models. In addition, the imaging findings noted by the 2 readers (R4 and R5) to support assignment of score 4 or 5 to a prediagnostic CT were compared one-on-one with the imaging findings documented by the 3 radiologists (R1, R2, R3). Any findings missed by the 2 readers (R4 and R5) on the prediagnostic CTs were also noted.

Statistical Analyses

Statistical analyses were performed with R (version 4.0.4) and with Python software using the scikit-learn library. Fisher exact test and Mann-Whitney *U* test were used to compare categorical and continuous variables, respectively. The performance of classifier models on the test subset was evaluated by the mean and 95% confidence intervals (CIs) of the accuracy, sensitivity/recall, specificity, and precision based on a case probability cutoff value of 0.5, as well as the F-score metric and AUC. Accuracy measured the model's performance in classifying the CTs as prediagnostic for PDAC or normal control pancreas. Sensitivity and specificity measured the proportion of prediagnostic and control CTs that were correctly classified, respectively. Precision was the fraction of true positive instances among the total

positive instances. The F-score measured the model's accuracy and was the harmonic mean of precision and recall. A DeLong's test[23] was performed to compare AUCs of different ML models. Performance of the ML model with the highest AUC was subanalyzed by stratifying the test subset according to different CT vendors and the 2 groups of slice thickness (≥3 mm and <3 mm). Interreader agreement of the 2 radiologist readers (R4 and R5) was evaluated with weighted Cohen's kappa value. To account for the interobserver variability, we used statistical method for multireader multicase (MRMC) receiver operating characteristics studies[24] to compare the readers' AUCs against the models' AUCs with adjustments for clustered data. MRMC analysis was completed using the RJafroc package in R. The analysis used an adaptation of the single-treatment multiple-reader Obuchowski Rockette analysis of variance model described by Hillis.[25] Both readers and cases were treated as random effects. For the MRMC analysis, the ML model was considered as reader 1, and R4 and R5 were considered as reader 2 and 3, respectively. The test subset CTs were considered as cases. We constructed 95% CIs for the difference in AUCs for each comparison. *P* values smaller than .05 indicated statistical significance.

Results

Patient Characteristics

The demographic characteristics were comparable between patients with prediagnostic CTs and the control subjects with normal pancreas (Table 1). The median (range) time interval between prediagnostic CTs and histopathological diagnosis of PDAC was 398 (93–1092) days in the entire dataset (n = 155) and 386 (97–1092) days in the test subset (n = 45). The location of PDAC on subsequent diagnostic CTs was in the head (n = 107; 69%), body (n = 22; 14.2%), and tail (n = 26; 16.8%) of the pancreas.

Based on the consensus review of 3 radiologists (R1, R2, R3), most CTs in the prediagnostic cohort had normal pancreas (n = 89, 57.4%). The remaining CTs had the following indirect imaging findings: subtle biliary or pancreatic duct dilatation (n = 24, 15.5%), focal attenuation differences (n = 24, 15.5%), focal contour abnormality (n = 10, 6.4%), regional pancreatic parenchymal atrophy (n = 2, 1.3 %), or 2 or more of these findings (n = 6, 3.9%). In the test subset of prediagnostic CTs, the majority of the pancreas were also normal (n = 25, 55.6%). Other CTs had indirect imaging findings such as subtle biliary or pancreatic duct dilatation (n = 7, 15.6 %), focal attenuation differences (n = 4, 8.9 %), focal contour abnormality (n = 4, 8.9 %), or 2 or more of these findings (n = 5, 11%).

The mean (SD) volume of pancreas in the prediagnostic CTs was 89.4 (40.9) mL, which was higher compared with the pancreas volume (75.7 [26.1] mL) (*P* = .10) in control CTs. For the test subset CTs from our institution, the mean (SD) pancreas volume for control and prediagnostic cohorts was 73.4 (21.2) mL and 79.7 (32.8) mL (*P* = .8), respectively. For the test subset CTs from external institutions, the mean (SD) pancreas volume for control and prediagnostic cohorts was 76.9 (28.2) mL and 95.2 (44.1) mL (*P* = .06), respectively. The mean (SD) pancreas volume was 81.0 (23.4) mL and 68.7 (17.5) mL for the independent 176 controls and NIH-PCT datasets respectively.

Radiomics-based ML Models

A total of 88 features were extracted from each of the pancreas segmentation masks from the prediagnostic and control CTs (Supplementary Table 1). Application of LASSO-based feature selection method resulted in 34 features (7 first-order and 27 gray-level features) (Supplementary Figure 1). Of these, gray-level run length matrices run length nonuniformity (AUC = 0.87) and first-order energy (AUC = 0.86) were biased by CT slice thickness based on their AUCs >0.8 and were, therefore, removed. Subsequently, 32 features were used to construct the 4 optimized ML classifiers, which were evaluated on the test subset.

The SVM model had the highest AUC (0.98; 95% CI, 0.94–0.98) on the test subset (Table 2, Figure 3, Supplementary Figure 2). It correctly classified 43 of 45 (95%) prediagnostic CTs and 75 of 83 (90.3%) control CTs, yielding an overall accuracy of 92.2%. A review of the 10 CTs misclassified by the SVM model by the 3 nonreader radiologists (R1, R2, and R3) did not identify a structural aberration in the pancreas to explain the cause of misclassification. Further, there was no statistically significant difference in accuracy of SVM model when the cases in the test subset were stratified by different vendors (Siemens, Toshiba, GE, and Philips were 92.7%, 92.1%, 88.2%, and 100%, respectively; $P = .30$). The performance of the SVM model for CTs with slice thicknesses ≥3 mm was higher compared with CTs with <3-mm slice thickness (98.4% and 85.7%, respectively; $P = .02$) (Supplementary Table 2). Because the SVM model had the highest AUC, it was the reference standard for comparison of AUC of other models. Other 3 models (KNN, RF, and XGB) had comparable AUCs (0.95, 0.95, and 0.96, respectively) to SVM (0.98) (Table 2). The selected hyperparameters for each individual model are listed in Supplementary Table 3.

Because the SVM model had the highest AUC on the test subset, it was externally validated by stratifying the test subset (n = 128) according to institution (35.2% CTs from our institution vs 64.8% from external institutions). Of the 45 CTs from our institution, the model correctly classified 16 of 17 (94%) prediagnostic CTs and 27 of 28 (96%) control CTs, yielding an overall accuracy of 95%. Of the 83 CTs from external institutions, the model correctly classified 27 of 28 (96%) prediagnostic CTs and 48 of 55 (87%) control CTs, yielding an overall accuracy of 90.4% (Table 3). Further, of the 176 CTs of healthy subjects from the independent internal set, the SVM model correctly classified the pancreas as normal in 163 CTs, yielding a specificity of 93%. The specificity was also generalizable on the public NIH-PCT dataset (n = 80) where the SVM model correctly classified the pancreas as normal in 77 CTs, yielding a specificity of 96%.

Finally, the ablation-based methodology demonstrated that the top 3 features with the highest influence on the SVM model's sensitivity were GLSZM high gray-level zone emphasis (HGLZE), GLDM dependence entropy, and GLCM difference variance. The sensitivity of the SVM model without these features was 92%, 92%, and 93%, respectively, vs 95% with all the 32 features.

Multireader Evaluation of the Test Subset

Evaluation by R4. On the prediagnostic CTs (n = 45), R4 rated 30 pancreas as normal (grades 1, 2, 3) and 15 pancreas as abnormal (grades 4, 5). Subtle biliary or pancreatic duct

dilatation was the finding commonly detected (n = 12 of 15, 80%). In contrast, the most missed finding was focal contour abnormality (n = 4 of 6; 66.7%). False positive biliary or pancreatic duct dilatation was noted in 1 patient. Of the control CTs with normal pancreas (n = 83), 77 were graded as normal and 6 as abnormal. Focal attenuation difference was the most common reported false positive finding (n = 4 of 6, 66.7%). Other findings included duct dilatation (n = 1) and duct dilation with hypodensity at the ampulla (n = 1).

Evaluation by R5. On the prediagnostic CTs (n = 45), R5 rated 31 pancreas as normal and 14 pancreas as abnormal. Subtle biliary or pancreatic duct dilatation was the finding commonly detected (n = 9 of 14, 64.3%). In contrast, the most missed finding was focal attenuation difference (n = 5 of 9; 55.5%). False positive focal attenuation difference was noted in 2 patients. In the control test subset (n = 83), 68 were graded as normal and 15 as abnormal. Subtle biliary or pancreatic duct dilatation was the most common false positive finding (n = 5 of 15, 33.3%). Other false positive findings included focal attenuation difference (n = 3 of 15, 20%), focal hypodensity (n = 2 of 15, 13.3%), peripancreatic fat stranding (n = 1 of 15, 6.7%) or a combination of the preceding findings (n = 4 of 15, 26.7%).

The strength of agreement between the 2 readers for pancreas assessment on the 5-point ordinal scale was only fair (weighted Cohen's kappa = 0.3). Readers had modest sensitivity (R4: 33.3%, R5: 31.1%) and predictive values (PPV R4: 71.4%, R5: 48.3%; NPV R4: 72.0%, R5: 68.7%) (Table 4). The mean classification AUC of the 2 readers was 0.66 (CI, 0.46–0.86), which was significantly lower than each of the 4 ML classifiers (AUCs: 0.95–0.98) ($P < .001$) (Figure 4).

Discussion

In our study, ML models based on first-order intensity and second-order texture features from volumetrically segmented normal pancreas detected the imaging signature of PDAC at a substantial lead time (median [range] 398 [93–1092] days) before its clinical diagnosis.

These models had high discrimination performance (AUC 0.98; CI, 0.94–0.98) for classification of pancreas into normal vs prediagnostic for PDAC. The performance was robust across different CT systems and slice thicknesses. The high AUC (0.97) of the SVM model on test subset CTs from our institution was identical to its AUC on the CTs from external institutions. Its high specificity was generalizable to an independent internal set as well to the external NIH-PCT dataset. In contrast, 2 radiologist readers demonstrated a significantly lower discrimination performance (AUC 0.66; CI, 0.46–0.86) and only fair interreader agreement (Cohen's kappa = 0.3). Besides, the readers recorded false positive indirect findings of PDAC in control subjects with normal pancreas (n = 83) (7% R4, 18% R5). Thus, radiomics-based ML classifiers can detect the imaging signatures of early PDAC from morphologically normal pancreas. Integration of such ML approaches with blood or other fluid-based biomarker strategies of early PDAC detection[26] could diagnose subclinical disease when surgical cure is possible. Finally, such models can be deployed to detect early cancer in ongoing clinical trials such as the Early Detection Initiative (NCT04662879).

The 3 features that had highest influence on the SVM model's sensitivity were GLSZM HGLZE, GLDM dependence entropy, and GLCM difference variance. Of these, GLSZM-HGLZE and GLCM difference variance were also among the highest absolute weight-coefficient features selected by LASSO. GLSZM-HGLZE is a regional texture feature that provides information about the proportion of high gray-level zones in a voxel. GLCM difference variance is a measure of heterogeneity between gray-level intensity pairs. These 2 features capture the intrinsic spatial heterogeneity due to biological variations in tissues.[27,28] The other feature, GLDM dependence entropy, measures the degree of randomness or nonuniformity in an image. It had higher influence on the model's sensitivity than the first-order interquartile range, which was the third-highest feature as per LASSO. First-order features such as interquartile range do not take the spatial relationship between voxels into account and, therefore, unlike the second-order features, fail to capture the intrinsic tissue heterogeneity.[29,30] These observations support the biologic insights from prior studies that the prediagnostic stage of PDAC is marked by substantial cellular activity and infiltration, which results in marked tissue heterogeneity.[31] Our study suggests that this tissue heterogeneity is beyond the human perceptive ability but can be captured and leveraged for actionable insights through computational post-processing techniques such as radiomics.

The prediagnostic CT dataset used in this study is the largest reported in the literature. We combined radiomics feature selection and 4 well-established supervised ML techniques (KNN, SVM, RF, XGBoost) to identify the image-based biomarkers of the subclinical changes that preceded clinical diagnosis of PDAC. Further, we followed a recently described ablation study paradigm,[21] that is, investigating factors sequentially, from feature redundancy to imaging parameters (ie, CT slice thickness), to exclude their effect on the final radiomics signature. For instance, CT slice thickness can variably influence radiomics signatures.[32] Therefore, we tested and accounted for the impact of the slice thickness both at the stage of feature selection and subsequently through stratifying model performance as per slice thicknesses. Because our models have been developed on standard-of-care portal venous phase CTs, their performance is not contingent on any custom imaging protocol, and they can also process previously acquired CTs. Once validated, such radiomics-based ML models also non-invasively evaluate the longitudinal changes of pancreatic carcinogenesis that precede the clinical diagnosis of PDAC.

In one prior study,[12] radiomics-based classifiers had high discrimination accuracy between CTs with PDAC at the diagnostic stage and healthy controls. The mean (SD) tumor size in that study was 4.1 (1.7) cm. In contrast, our cohort consists of CTs at the prediagnostic stage (ie, 3–36 months before tumor development). Importantly, none of these CTs had a focal mass and most of the CTs in the test subset had normal pancreas on visual inspection. Other prediagnostic CTs (n = 66, 41.8%) had indirect findings and the frequency of these findings is concordant with a recent study.[11] Moreover, 20.4% of healthy subjects in that study also had indirect findings, which shows that indirect findings per se are not specific for early PDAC detection. Besides, evaluation of such indirect findings can be quite subjective, as evident from the low interreader agreement between the radiologist readers in our study. In fact, the observed reader performance could be an overestimate because the readers were exclusively focused on the pancreas. In clinical practice, assessment of the pancreas often

does not occur with the same degree of thoroughness, which is evident because inattentional blindness is one of the main factors responsible for missed PDAC on CTs.[10]

Further, we validated the high specificity of the SVM classifier on an independent set of CTs with normal pancreas as well as on the public NIH-CT dataset. One of the next steps is validation of these classifiers on external datasets that include both cases and controls. However, there are no public datasets of prediagnostic CTs, which is one major barrier for investigation of CT for early detection of PDAC. We are in the process of designing external validation studies through prediagnostic datasets from other institutions. In addition, the Imaging Working Group of the Alliance of Pancreatic Cancer Consortia (APaCC), a virtual network of researchers from multiple consortia, is addressing these challenges through the collection of relevant imaging datasets acquired during routine clinical practice.[33] Finally, in our study, the volumetric pancreas segmentations were done by radiologists. Such manual segmentation is a time-consuming and cumbersome process.[34] Recently, AI models for fully automated high-fidelity pancreas segmentation have been developed.[35] Such models can automate and further extend the scalability of our radiomics-based ML approach.

Our study has limitations. The retrospective nature of the study is generally prone to selection bias. As with other radiomics studies, the precise pathologic correlates of the radiomic features that constitute the ML classifiers are not entirely known. We did not investigate the impact of differences in all the acquisition or post-processing parameters (eg, voxel width, bin width) on the classifiers, which will be the subject of the next phase of our ongoing investigation. Although we validated the high specificity of the SVM classifier on an independent internal cohort of control CTs as well as on the public NIH-PCT dataset, the sample size of these cohorts was small and the subjects in these cohorts were relatively younger. Thus, prospective larger cohorts with both cases and controls are warranted for further validation. Such prospective studies would also help determine the optimal operating point for the models to avoid a high false positive rate in the context of a screening paradigm.

In conclusion, we detected and quantified the imaging signature of early pancreatic carcinogenesis from volumetrically segmented normal pancreas on standard-of-care CTs. The radiomics-based ML classifiers had high discrimination accuracy for classification of pancreas into prediagnostic for PDAC vs normal. The high accuracy of the SVM model was validated on CTs from external institutions. Its high specificity was generalizable on an independent internal cohort and on an external public dataset. In contrast, radiologist readers had low interreader agreement, sensitivity, and discrimination accuracy, which shows that novel AI-based approaches can detect PDAC at a subclinical stage when it is beyond the scope of the human interrogation. Prospective validation of these ML models and their integration with complementary blood and other fluid-based biomarkers has the potential to further improve cancer prediction capabilities at the prediagnostic or symptom-free stage. Such models also have the potential to elucidate the longitudinal changes of carcinogenesis that precede the clinical diagnosis of PDAC. Finally, such models can be deployed to detect early cancer in ongoing clinical trials such as the Early Detection Initiative that seeks to evaluate outcomes of a screening strategy by using clinical risk-prediction models and CT in cohorts at high risk for PDAC.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Funding

Ajit H. Goenka gratefully acknowledges a research grant from the Champions for Hope Pancreatic Cancer Research Program of the Funk Zitiello Foundation, Advance the Practice Award from the Department of Radiology, Mayo Clinic, Rochester, Minnesota, and the Centene Charitable Foundation. Unrelated to this work: Ajit H. Goenka is the principal investigator (PI) and supported by CA190188, Department of Defense, Office of the Congressionally Directed Medical Research Programs. Ajit H. Goenka is also the co-PI and supported by R01CA256969, National Cancer Institute of the National Institutes of Health. Ajit H. Goenka is also on the Advisory Board (ad hoc), BlueStar Genomics.

Data Availability

Access to datasets from the Mayo Clinic Foundation should be requested directly via their data access request forms. Subject to the institutional review boards' ethical approval, deidentified data would be made available. All experiments and implementation details are described thoroughly in the Materials and Methods section so they can be independently replicated with nonproprietary libraries.

Abbreviations used in this paper:

AI	artificial intelligence
AUC	area under the curve
CI	confidence interval
CT	computed tomography
GLCM	gray-level cooccurrence matrices
GLDM	gray-level dependence matrices
GLSZMs	gray-level size zone matrices
HGLZE	high gray-level zone emphasis
KNN	k-nearest neighbor
LASSO	least absolute shrinkage and selection operator
ML	machine learning
MRMC	multireader multicase
NIH-PCT	National Institutes of Health-Pancreas CT
PDAC	pancreatic ductal adenocarcinoma
RF	random forest
SD	standard deviation

SVM	support vector machine
XGBoost	extreme gradient boosting

References

1. Schwartz NRM, Matrisian LM, Shrader EE, et al. Potential cost-effectiveness of risk-based pancreatic cancer screening in patients with new-onset diabetes. J Natl Compr Canc Netw 2021;20:451–459. [PubMed: 34153945]
2. Singh DP, Sheedy S, Goenka AH, et al. Computerized tomography scan in pre-diagnostic pancreatic ductal adenocarcinoma: stages of progression and potential benefits of early intervention: a retrospective study. Pancreatology 2020;20:1495–1501. [PubMed: 32950386]
3. Vasen H, Ibrahim I, Ponce CG, et al. Benefit of surveillance for pancreatic cancer in high-risk individuals: outcome of long-term prospective follow-up studies from three european expert centers. J Clin Oncol 2016;34:2010–2019. [PubMed: 27114589]
4. Yuan C, Babic A, Khalaf N, et al. Diabetes, weight change, and pancreatic cancer risk. JAMA Oncol 2020;6:e202948. [PubMed: 32789511]
5. Hart PA, Chari ST. Is screening for pancreatic cancer in high-risk individuals one step closer or a fool's errand? Clin Gastroenterol Hepatol 2019;17:36–38. [PubMed: 30268560]
6. Chari ST, Maitra A, Matrisian LM, et al. Early detection initiative: a randomized controlled trial of algorithm-based screening in patients with new onset hyperglycemia and diabetes for early detection of pancreatic ductal adenocarcinoma. Contemp Clin Trials 2021;113:106659. [PubMed: 34954100]
7. Gangi S, Fletcher JG, Nathan MA, et al. Time interval between abnormalities seen on CT and the clinical diagnosis of pancreatic cancer: retrospective review of CT scans obtained before diagnosis. Am J Roentgenol 2004;182:897–903. [PubMed: 15039161]
8. Pelaez-Luna M, Takahashi N, Fletcher JG, et al. Resectability of presymptomatic pancreatic cancer and its relationship to onset of diabetes: a retrospective review of CT scans and fasting glucose values prior to diagnosis. Am J Gastroenterol 2007;102:2157–2163. [PubMed: 17897335]
9. Kang J, Clarke SE, Abdolell M, et al. The implications of missed or misinterpreted cases of pancreatic ductal adenocarcinoma on imaging: a multi-centered population-based study. Eur Radiol 2021;31:212–221. [PubMed: 32785768]
10. Kang JD, Clarke SE, Costa AF. Factors associated with missed and misinterpreted cases of pancreatic ductal adenocarcinoma. Eur Radiol 2021;31:2422–2432. [PubMed: 32997176]
11. Toshima F, Watanabe R, Inoue D, et al. CT abnormalities of the pancreas associated with the subsequent diagnosis of clinical stage I pancreatic ductal adenocarcinoma more than 1 year later: a case-control study. Am J Roentgenol 2021;217:1353–1364. [PubMed: 34161128]
12. Chu LC, Park S, Kawamoto S, et al. Utility of CT radiomics features in differentiation of pancreatic ductal adenocarcinoma from normal pancreatic tissue. Am J Roentgenol 2019;213:349–357. [PubMed: 31012758]
13. Kim BR, Kim JH, Ahn SJ, et al. CT prediction of resectability and prognosis in patients with pancreatic ductal adenocarcinoma after neoadjuvant treatment using image findings and texture analysis. Eur Radiol 2019;29:362–372. [PubMed: 29931561]
14. Li J, Lu J, Liang P, et al. Differentiation of atypical pancreatic neuroendocrine tumors from pancreatic ductal adenocarcinomas: using whole-tumor CT texture analysis as quantitative biomarkers. Cancer Med 2018;7:4924–4931. [PubMed: 30151864]
15. Rigiroli F, Hoye J, Lerebours R, et al. CT radiomic features of superior mesenteric artery involvement in pancreatic ductal adenocarcinoma: a pilot study. Radiology 2021;301:610–622. [PubMed: 34491129]
16. Suman G, Patra A, Korfiatis P, et al. Quality gaps in public pancreas imaging datasets: Implications & challenges for AI applications. Pancreatology 2021;21:1001–1008. [PubMed: 33840636]
17. Panda A, Garg I, Truty MJ, et al. Borderline resectable and locally advanced pancreatic cancer: FDG PET/MRI and CT tumor metrics for assessment of pathologic response to neoadjuvant therapy and prediction of survival. Am J Roentgenol 2021;217:730–740. [PubMed: 33084382]

18. Medical Imaging Resource Center-Clinical Trial Processor. Available at: https://mircwiki.rsna.org/index.php?title=MIRC_CTP. Accessed September 20, 2022.
19. van Griethuysen JJM, Fedorov A, Parmar C, et al. computational radiomics system to decode the radiographic phenotype. Cancer Res 2017;77:e104–e107. [PubMed: 29092951]
20. Tibshirani R. The lasso method for variable selection in the Cox model. Stat Med 1997;16:385–395. [PubMed: 9044528]
21. Lu L, Ahmed FS, Akin O, et al. Uncontrolled confounders may lead to false or overvalued radiomics signature: a proof of concept using survival analysis in a multicenter cohort of kidney cancer. Front Oncol 2021;11:638185. [PubMed: 34123789]
22. XGBoost. Available at: https://xgboost.readthedocs.io/en/stable/. Accessed September 20, 2022.
23. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44:837–845. [PubMed: 3203132]
24. Obuchowski NA. New methodological tools for multiple-reader ROC studies. Radiology 2007;243:10–12. [PubMed: 17392244]
25. Hillis SL. A comparison of denominator degrees of freedom methods for multiple observer ROC analysis. Stat Med 2007;26:596–619. [PubMed: 16538699]
26. Brezgyte G, Shah V, Jach D, et al. Non-invasive biomarkers for earlier detection of pancreatic cancer-a comprehensive review. Cancers (Basel) 2021;13:2722. [PubMed: 34072842]
27. Chaddad A, Kucharczyk MJ, Niazi T. Multimodal radiomic features for the predicting Gleason score of prostate cancer. Cancers (Basel) 2018;10:249. [PubMed: 30060575]
28. Narang S, Kim D, Aithala S, et al. Tumor image-derived texture features are associated with CD3 T-cell infiltration status in glioblastoma. Oncotarget 2017;8:101244–101254. [PubMed: 29254160]
29. Sandrasegaran K, Lin Y, Asare-Sawiri M, et al. CT texture analysis of pancreatic cancer. Eur Radiol 2019;29:1067–1073. [PubMed: 30116961]
30. Zhang Y, Lobo-Mueller EM, Karanicolas P, et al. Prognostic value of transfer learning based features in resectable pancreatic ductal adenocarcinoma. Front Artif Intell 2020;3:550890. [PubMed: 33733206]
31. Zheng L, Xue J, Jaffee EM, et al. Role of immune cells and immune-based therapies in pancreatitis and pancreatic ductal adenocarcinoma. Gastroenterology 2013;144:1230–1240. [PubMed: 23622132]
32. Zhao B. Understanding sources of variation to improve the reproducibility of radiomics. Front Oncol 2021;11:633176. [PubMed: 33854969]
33. Kenner B, Chari ST, Kelsen D, et al. artificial intelligence and early detection of pancreatic cancer: 2020 summative review. Pancreas 2021;50:251–279. [PubMed: 33835956]
34. Suman G, Panda A, Korfiatis P, et al. Development of a volumetric pancreas segmentation CT dataset for AI applications through trained technologists: a study during the COVID 19 containment phase. Abdom Radiol (NY) 2020;45:4302–4310. [PubMed: 32939632]
35. Panda A, Korfiatis P, Suman G, et al. Two-stage deep learning model for fully automated pancreas segmentation on computed tomography: comparison with intra-reader and inter-reader reliability at full and reduced radiation dose on an external dataset. Med Phys 2021;48:2468–2481. [PubMed: 33595105]

WHAT YOU NEED TO KNOW

BACKGROUND AND CONTEXT

Inability of imaging to detect early pancreatic ductal adenocarcinoma is a major barrier to improving outcomes, which necessitates novel methods for early pancreatic ductal adenocarcinoma detection at a stage when surgical cure is possible.

NEW FINDINGS

Artificial intelligence can detect subclinical cancer in normal pancreas at a substantial lead time (median 386 days) before clinical diagnosis even when it is beyond the scope of human perception.

LIMITATIONS

Although we validated the artificial intelligence approach on computed tomography scans from external institutions, prospective larger multicenter studies are warranted before potential clinical translation.

IMPACT

Artificial intelligence approaches can diagnose subclinical cancer at a stage when surgical cure is possible, as well elucidate the longitudinal changes of carcinogenesis that precede the clinical diagnosis of pancreatic ductal adenocarcinoma.

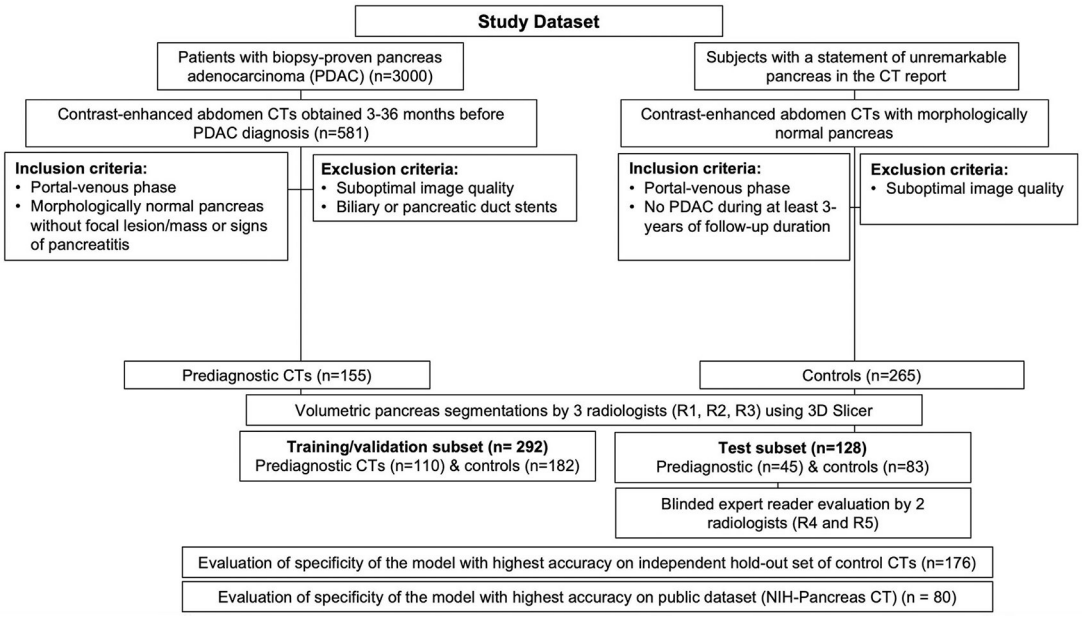


Figure 1.
Study design and structure of datasets.

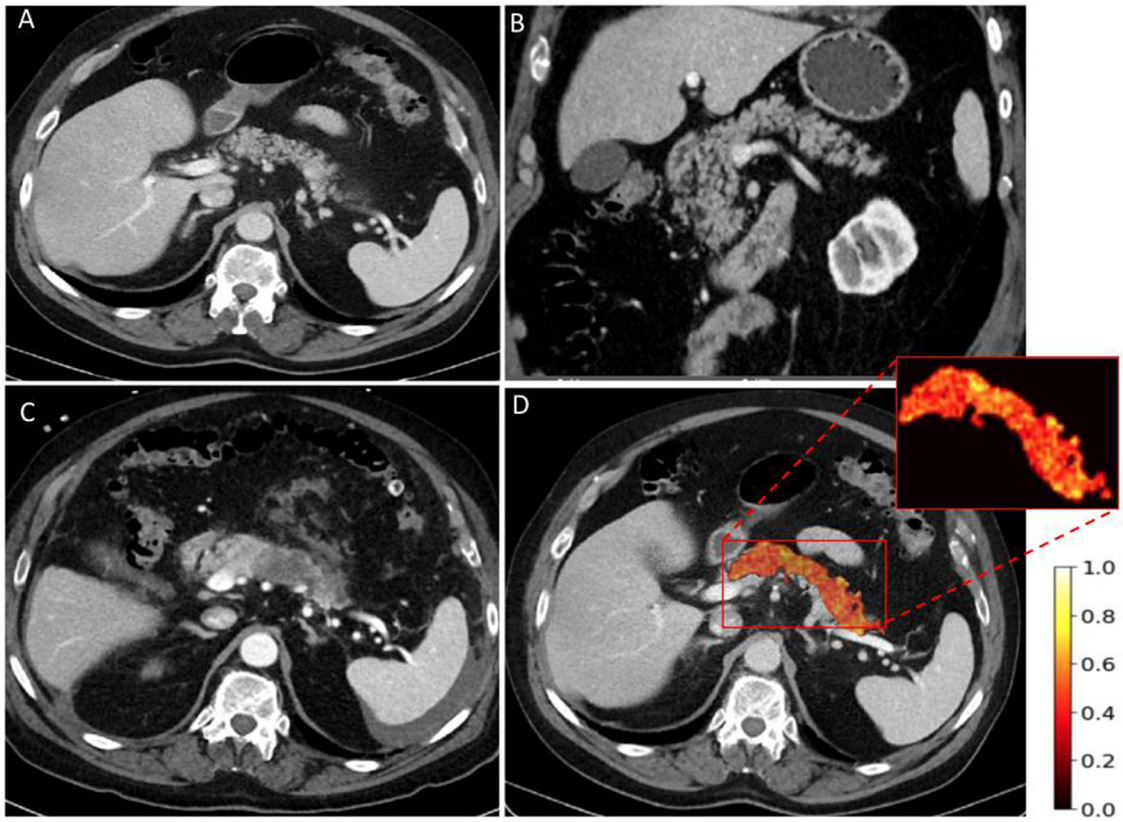


Figure 2.

Prediagnostic and diagnostic CT: 75-year-old man with PDAC. Pancreas was normal on the prediagnostic CT (*A* and *B*). Approximately 2.5 years later, the patient presented with PDAC in the body with peritoneal carcinomatosis (*C*). Color-coded radiomics texture map overlaid on the prediagnostic CT (*D*) shows the distribution of gray-level size zone matrix-small area emphasis (GLSZM-SAE) (measurement of fine gray-level texture) over a single slice of pancreas on the prediagnostic CT.

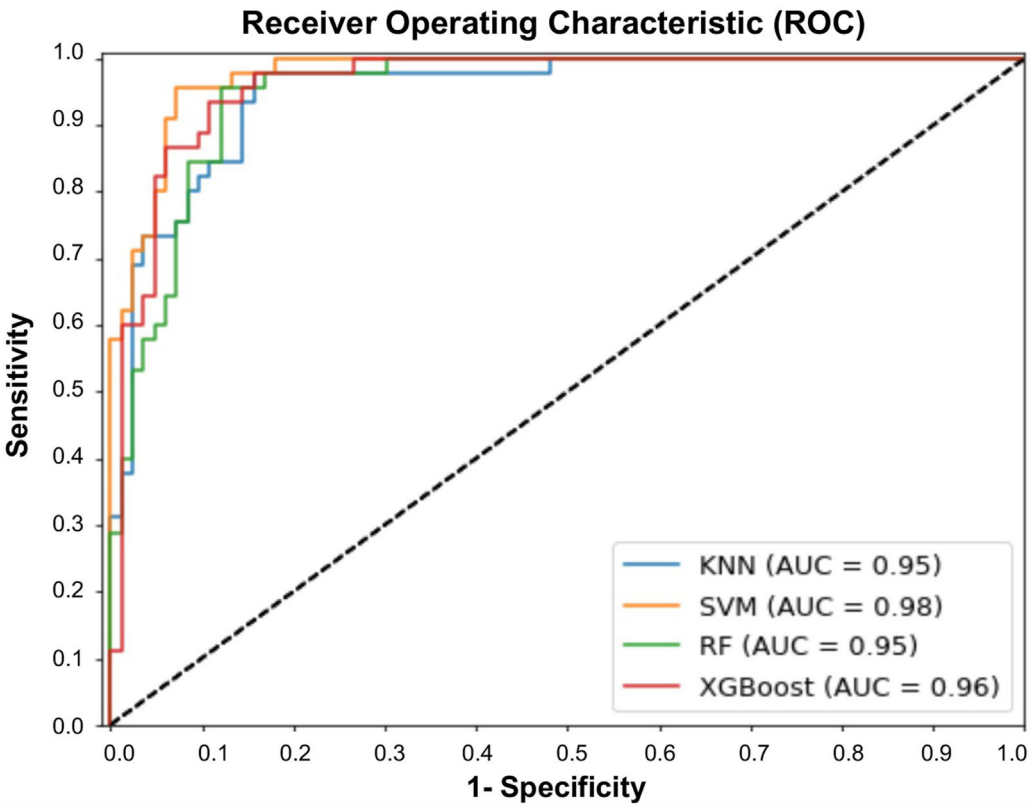


Figure 3.

Receiver operating characteristics of the 4 ML models on the test subset. KNN, k-nearest neighbor; RF, random forest; SVM, support vector machine; XGBoost, extreme gradient boosting.

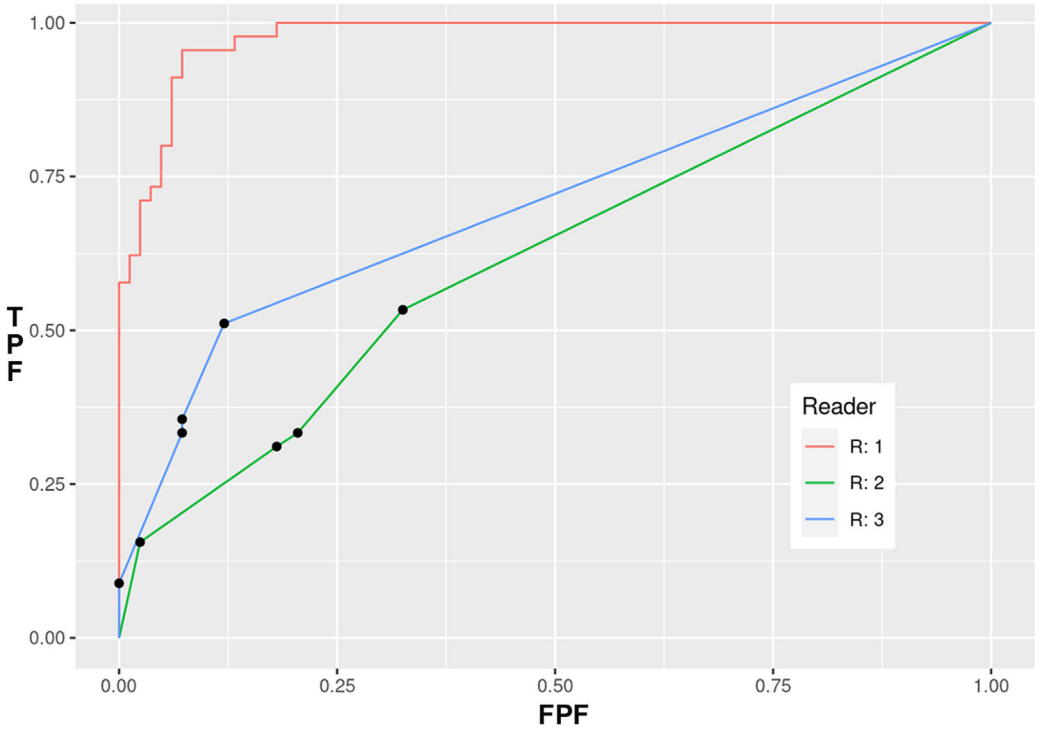


Figure 4.

Receiver operating characteristics of the SVM model and the 2 radiologist readers on the test subset. R1: SVM model; R2, R3: two radiologist readers. FPF, false positive fraction or 1 – Specificity; TPF, true positive fraction or sensitivity.

Table 1.

Demographic Characteristics of Patients With Prediagnostic CTs and Control Subjects

	Prediagnostic CTs			Controls CTs		
	Entire cohort (n = 155)	**Training subset (n = 110)**	**Test subset (n = 45)**	**Entire cohort (n = 265)**	**Training subset (n = 182)**	**Test subset (n = 83)**
Male-to-female ratio	1.4	1.4	1.2	1.1	1.2	0.93
Age,[a] *y*, median (range)	69 (34–88)	68 (34–88)	72 (45–84)	67 (30–89)	69 (30–89)	66 (38–89)
BMI[b]	29.3 (19.3–57.9)	28.8 (19.3–47.4)	30.4 (20.7- 57.9)	28.7 (15.2–49.6)	28.5 (15.9–45.7)	29.1 (15.2–49.6)
Race						
White	141	98	43	240	162	78
Asian	1	1	0	4	4	0
Black	2	2	0	2	2	0
Other	1	0	1	7	5	2
Not available	10	9	1	12	9	3

[a]Only age distribution between the control and prediagnostic test subset was significant ($P = .01$).

[b]BMI (body mass index): BMI was available for 117 (75.5%) patients with prediagnostic CTs and for 245 (92.5%) of the control subjects

Table 2.

Performance of the Radiomics-based ML Classifiers

	Sensitivity[a] (95% CI)	Specificity[a] (95% CI)	Precision[a] (95% CI)	F1-score[a] (95% CI)	Accuracy[a] (95% CI)	AUC (95% CI)	*P* [b]
KNN	80.0 (68.8–91.1)	91.5 (83.7–94.6)	83.7 (74.3–88.3)	81.8 (74.8–85.5)	87.5 (82.7–89.8)	0.95 (0.91–0.96)	.27
SVM	95.5 (85.5–100)	90.3 (84.3–91.5)	84.3 (76.3–86.2)	89.5 (82.3–91.7)	92.2 (86.7–93.7)	0.98[c] (0.94–0.98)	ref
RF	91.1 (78.8–93.3)	87.9 (84.3–91.5)	80.3 (75.2–84.8)	85.4 (79.3–87.2)	89.0 (84.7–90.6)	0.95 (0.92–0.96)	.24
XGB	91.1 (82.2–95.5)	89.1 (83.1–92.2)	82.0 (74.3–85.5)	86.3 (79.5–88.3)	89.8 (84.7–91.4)	0.96 (0.93–0.97)	0.60

NOTE. Precision is the fraction of true positive instances among total positive instances. F1-score is the harmonic mean of precision and recall/sensitivity.

Ref, reference; RF, random forest; XGB, extreme gradient boosting.

[a]The values are represented in %.

[b]*P* values are derived from the DeLong's test of AUCs where AUC of SVM is the reference standard for comparison.

[c]The highest AUC.

Table 3.

Performance of SVM Classifier on Test Subset Stratified by Institution

	Sensitivity[a] (95% CI)	Specificity[a] (95% CI)	Precision[a] (95% CI)	F1-score[a] (95% CI)	Accuracy[a] (95% CI)	AUC (95% CI)
Our institution (n = 45; 35.2%)	94.1 (82.3–100.0)	96.4 (87.4–97.0)	94.1 (82.0–94.4)	94.1 (84.8–97.1)	95.5 (88.8–97.7)	0.97 (0.96–0.99)
Outside institutions (n = 83; 64.8%)	96.4 (84.0–100.0)	87.2 (82.0–90.0)	79.4 (71.6–82.8)	87.1 (78.0–88.8)	90.4 (84.3–91.5)	0.97 (0.92–0.98)
Total (n = 128)	95.5 (85.5–100)	90.3 (84.3–91.5)	84.3 (76.3–86.2)	89.5 (82.3–91.7)	92.2 (86.7–93.7)	0.98 (0.94–0.98)

[a]The values are represented in %.

Table 4.

Performance of the Radiologist Readers on the Test Subset

	Reader 4	Reader 5
Metric	**Value[a] (95% CI)**	**Value[a] (95% CI)**
Sensitivity	33.3 (20.4–49.1)	31.1 (18.6–46.8)
Specificity	92.8 (84.4–97.0)	81.9 (71.6–89.2)
Positive predictive value	71.4 (47.7–87.8)	48.3 (29.9–67.1)
Negative predictive value	72.0 (62.3–80.1)	68.7 (58.5–77.4)
Accuracy	71.9 (63.1–79.3)	64.0 (55.1–72.2)
AUC	0.70 (0.62–0.78)	0.62 (0.52–0.71)

[a]The values are represented in %.