HHS Public Access

Author manuscript

Abdom Radiol (NY). Author manuscript; available in PMC 2025 July 23.

Published in final edited form as:

Abdom Radiol (NY). 2024 March ; 49(3): 964–974. doi:10.1007/s00261-023-04127-1.

Assessing the robustness of a machine-learning model for early detection of pancreatic adenocarcinoma (PDA): evaluating resilience to variations in image acquisition and radiomics workflow using image perturbation methods

Sovanlal Mukherjee[1], **Panagiotis Korfiatis**[1], **Nandakumar G. Patnam**[1], **Kamaxi H. Trivedi**[1], **Aashna Karbhari**[1], **Garima Suman**[1], **Joel G. Fletcher**[1], **Ajit H. Goenka**[1]

[1]Divisions of Abdominal and Nuclear Imaging, Nuclear Radiology Fellowship, Nuclear Radiology Research Operations, Enterprise PET/MR Research and Development, Department of Radiology, Mayo Clinic, 200 First St SW, Charlton 1, Rochester, MN 55905, USA

Abstract

Purpose: To evaluate robustness of a radiomics-based support vector machine (SVM) model for detection of visually occult PDA on pre-diagnostic CTs by simulating common variations in image acquisition and radiomics workflow using image perturbation methods.

Methods: Eighteen algorithmically generated-perturbations, which simulated variations in image noise levels (σ, 2σ, 3σ, 5σ), image rotation [both CT image and the corresponding pancreas segmentation mask by 45° and 90° in axial plane], voxel resampling (isotropic and anisotropic), gray-level discretization [bin width (BW) 32 and 64)], and pancreas segmentation (sequential erosions by 3, 4, 6, and 8 pixels and dilations by 3, 4, and 6 pixels from the boundary), were introduced to the original (unperturbed) test subset (n = 128; 45 pre-diagnostic CTs, 83 control CTs with normal pancreas). Radiomic features were extracted from pancreas masks of these additional test subsets, and the model's performance was compared vis-a-vis the unperturbed test subset.

Results: The model correctly classified 43 out of 45 pre-diagnostic CTs and 75 out of 83 control CTs in the unperturbed test subset, achieving 92.2% accuracy and 0.98 AUC. Model's performance was unaffected by a three-fold increase in noise level except for sensitivity declining to 80% at 3σ (p = 0.02). Performance remained comparable vis-a-vis the unperturbed test subset despite variations in image rotation (p = 0.99), voxel resampling (p = 0.25–0.31), change in gray-level BW to 32 (p = 0.31–0.99), and erosions/dilations up to 4 pixels from the pancreas boundary (p = 0.12–0.34).

Conclusion: The model's high performance for detection of visually occult PDA was robust within a broad range of clinically relevant variations in image acquisition and radiomics workflow.

✉Ajit H. Goenka, goenka.ajit@mayo.edu.

Competing interest Authors have no conflict of interest.

Keywords

Pancreas; Machine-learning; Radiomics; Pancreatic ductal carcinoma; X-ray computed tomography

Introduction

Early detection of pancreatic ductal adenocarcinoma (PDA) is one of the interventions with the greatest potential to improve outcomes of patients with PDA [1]. Efforts to detect PDA earlier have been driven by the identification of high-risk cohorts, such as individuals with new-onset diabetes and weight loss, who may benefit from screening [2]. However, despite rigorous screening, an alarming majority of such high-risk individuals are being diagnosed with advanced stages of PDA [3–5]. Thus, the success of screening is dependent on the ability of imaging to detect early PDA [6]. However, early PDA is often visually occult on cross-sectional imaging and the pancreas demonstrates a visually normal appearance [1, 7–10]. Unfortunately, PDA can rapidly progress from such an undetectable state to widely metastatic disease in a matter of months [2]. Therefore, there is an urgent need to develop new strategies for the detection of early cancer in the seemingly normal pancreas at the asymptomatic stage.

Recently, we demonstrated that radiomics-based machine-learning (ML) models could detect and quantify the imaging signature of early pancreatic carcinogenesis from volumetrically segmented pancreas on standard-of-care CTs [11]. Specifically, the support vector machine (SVM) model could detect visually occult PDA on pre-diagnostic CTs, which are defined as incidental CT scans performed for unrelated indications between 3 and 36-months prior to its eventual clinical PDA diagnosis, at a median of 386 (97–1092) days prior to clinical diagnosis. Importantly, these CTs had been interpreted to be negative for PDA during routine clinical interpretation. The model's high specificity was generalizable to both an independent internal as well as the public National Institutes of Health-Pancreas CT (NIH-PCT) dataset. Finally, the model outperformed radiologists by a substantial margin in identifying PDA on these pre- diagnostic CTs.

Ensuring the robustness of radiomics-based machine-learning (ML) models such as SVM across the large potential range of variations that may affect identification of target features is critical prior to clinical translation. Variations in image acquisition (e.g., patient positioning and image noise) and radiomics workflow (e.g., image pre-processing and target organ segmentation) can substantially alter target images and image features resulting in degraded performance, sub-optimal reproducibility, and poor generalizability of radiomics-based models [12–15]. The ideal method to evaluate the robustness of radiomics-based models against such variations is through validation against test–retest data [16, 17], which involves obtaining multiple scans of the same patient within a short interval. However, this approach is resource-intensive, clinically impractical, and ethically unjustifiable given the need for ionizing radiation. To address this challenge, a novel method [12, 18] has recently emerged—algorithmically generated-perturbations of original images to simulate additional image sets. This method has been validated against test–retest datasets and

provides a promising alternative to traditional validation methods. In this study, our goal was to evaluate robustness of the radiomics-based SVM model for detection of early PDA on pre-diagnostic CTs by simulating common variations in image acquisition and radiomics workflow using image perturbation methods.

Materials and methods

Study participants and procedures

This Health Insurance Portability and Accountability Act (HIPAA)-compliant, case–control retrospective study was approved by our institutional review board, which had waived the requirement for written informed consent. Previously described [11] study participants and procedures are summarized below.

The study included 155 pre-diagnostic CTs of 155 patients (90 men, 65 women; median (range) age: 69 (34–88) years) who developed PDA. Of these, 49 CTs (31.6%) had been performed at our institution whereas 106 (68.4%) CTs were performed at other institutions. The median (range) time interval between pre-diagnostic CTs and the subsequent histopathological diagnosis of PDA was 398 (93–1092) days. The median (range) CT slice thickness was 3 (0.5–5) mm. CTs were acquired on CT systems from four different vendors [81 Siemens (52.3%), 41 GE (26.4%), 25 Toshiba (16.1%), and 8 Philips (5.2%)].

The controls included an age-matched set of patients (n = 265) [140 men, 125 women; median (range) age: 67 (30–89) years] who did not develop PDA after 3-years or more of clinical follow-up. Of this control cohort, 87 CTs (32.8%) were from our institution whereas 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 three different vendors [219 Siemens (82.6%), 19 GE (7.2%), 27 Toshiba (10.2%)].

Volumetric pancreas segmentations were performed by three fellowship-trained radiologists in 3D Slicer (Version 4.11.20210226) using the boundary-points-based segmentation mode of the organ segmentation module (NVIDIA). The dataset was randomly divided into training- validation (n = 292), and test subsets (n = 128; control: 83, pre-diagnostic: 45) using a 70%–30% split.

Radiomic feature extraction, selection and dimensionality reduction

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 images was rescaled to the range of 0 to 255. A total of 88 features were extracted from each segmentation mask using the PyRadiomics library (version 3.0.1) [19], 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) logistic

regression [20]. Two features, gray- level run length matrices run length nonuniformity and first-order energy, were dependent upon CT slice thickness and were, therefore, removed [11]. Thus, using the 32 selected radiomic features, the SVM model was developed on the training subset and its performance was evaluated on the test subset. The SVM model was generated only using the original (unperturbed) images of the training-validation subset.

Image perturbations to create multiple additional test subsets

A total of 18 additional test subsets were created by introducing algorithmically generated-perturbations to the images and/or segmentation masks on the original (unperturbed) test subset ($n = 128$) to simulate a large range of variations in the image acquisition and radiomics workflow in a univariate paradigm. The univariate strategy was chosen because exploring every possible permutation would have been impractical due to the overwhelming number of possible configurations. Image acquisition categories included image noise level and image rotation, as described below. Radiomics workflow categories included voxel resampling, gray-level discretization, and pancreas segmentation (Fig. 1). All individual 18 perturbations are listed in Tables 1–3.

Image noise level

To assess the impact of variability in image noise during CT acquisition, random Gaussian noise was added to generate additional test subsets with noisier images [12]. The noise was drawn from a normal distribution with a mean of 0 and a standard deviation (σ) equal to the estimated standard deviation of the noise present in the original (unperturbed) test subset, as recently described [12]. The noise level (σ) was then gradually increased by a factor of 2, 3 and 5 to evaluate the limits of the model's performance vis-à-vis increasing image noise (Fig. 2).

Image rotation

To evaluate the impact of different patient positions during image acquisition, both the CT image and the corresponding pancreas segmentation mask were rotated in the axial (x, y) plane, i.e., around the z-axis, at two different rotational angles (45° and 90°) (Fig. 3).

Both these image-based perturbations, i.e., noise and rotations, were introduced prior to windowing and rescaling of the CT images.

Voxel resampling

Voxel resampling tends to standardize the voxel size (in x, y, and z directions) across the entire imaging cohort. Voxel resampling had not been deployed on the original (unperturbed) test subset. To assess robustness of the SVM model across variations in voxel resampling parameters, three different resampling configurations were evaluated: (1) isotropic resampling ($1 \times 1 \times 1$ mm^3), (2) anisotropic resampling-1 ($0.75 \times 0.75 \times 2$-mm^3), and (3) anisotropic resampling-2 ($0.75 \times 0.75 \times 3$-mm^3). With isotropic resampling, the voxels have the same dimension (1-mm) along the x-, y-, and z-axes. With anisotropic resampling, the voxels have the same dimension (0.75-mm) along x- and y- axes but different dimensions (2-mm or 3-mm) along the z-axes. B-spline interpolation was used for both isotropic and anisotropic resampling.

Gray-level discretization

Gray-level discretization entails clustering of the image pixels according to their gray-level intensity values for extraction of higher-order textural features. The latter provide a measurement of intrinsic tissue heterogeneity by quantifying the spatial variations in gray levels within an image. In the discretization process, the number of clusters or gray levels is determined by a fixed parameter called bin width (BW). An increasing BW results in decreasing gray levels (Fig. 4). The default BW is 25, which had been used to process the original (unperturbed) test subset. To assess robustness of the SVM model against variations in gray-level discretization, two different BWs of 32 and 64 were applied to the test subset CTs in the current study. Radiomics texture maps derived from GLCM-difference entropy were generated at a BW of 64 (or gray-level of 4) on both the control and pre-diagnostic CTs for a visual evaluation of the impact on model's performance (Fig. 5).

Segmentation variations

To evaluate robustness of SVM model against segmentation variability, pancreas segmentation masks were sequentially eroded by 3, 4, 6, and 8 pixels from the segmentation boundary. The erosion was done slice-by-slice along the *z*-axes (Fig. 6). Likewise, to simulate over- segmentation, the pancreas masks were sequentially dilated by 3, 4 and 6 pixels from the segmentation boundary (Fig. 7).

Following the generation of new test subsets encompassing each category of the aforementioned perturbations, radiomic features were extracted from the volumetric pancreas segmentation masks within these additional test subsets. The performance of the SVM model, utilizing the selected features, was then individually assessed, and compared to its performance when employing features from the original (unperturbed) test subset.

Statistical analyses

Statistical analyses were conducted using R (version 4.0.4) and Python software, utilizing scipy (version 1.8.0) and statsmodels (version 0.13.2). Categorical variables were compared using Fisher's exact test, while continuous variables were compared using the Mann–Whitney *U*-test.

The performance of the SVM model on the test subset was evaluated based on accuracy, sensitivity/recall, specificity, and the area under the curve (AUC). The accuracy assessed the model's ability to classify CT scans as either pre-diagnostic for pancreatic ductal adenocarcinoma (PDA) or as normal control pancreas. Sensitivity and specificity measured the proportion of correctly classified pre-diagnostic and control CT scans, respectively, using a case probability cut-off value of 0.5. To compare the AUCs of the SVM model between the original (unperturbed) test subset and the test subsets generated with algorithmically generated-perturbations, DeLong's test [21] was employed. Additionally, McNemar's test was used to compare sensitivity, specificity, and accuracy. Statistical significance was defined as p-values less than 0.05.

Results

Dataset characteristics

The demographic characteristics [age, gender, body mass index (BMI) and race] of patients with pre-diagnostic CTs and control subjects with normal pancreas were comparable ($p > 0.05$). Most of the CT scans in the test subset (83 out of 128; 64.8%) were from external institutions. The median time interval (range) between pre-diagnostic CTs and histopathological diagnosis of PDA was 386 (97–1092) days. The mean (SD) volume of the pancreas in the pre-diagnostic CT cohort was 89.4 (40.9) cc, which was higher compared to the pancreas volume [75.7 (26.1) cc] in control CT scans, although this difference was not statistically significant ($p = 0.10$).

Model performance on the original (unperturbed) test subset

The SVM model accurately classified 43 out of 45 (95%) pre-diagnostic CTs and 75 out of 83 (90.3%) control CTs, resulting in a sensitivity of 95.5% [95% confidence interval (CI) 85.5–100.0], specificity of 90.3% (95% CI 84.3–91.5), accuracy of 92.2% (95% CI 86.7–93.7) and an AUC of 0.98 (95% CI 0.94–0.98).

Robustness against common variations in image acquisition

The overall standard deviation (σ) of the noise of the entire original (unperturbed) test subset was 5.4 HU. The model's specificity, accuracy, and AUC remained robust against noise levels of σ, 2σ and 3σ ($p > 0.05$) when compared to its performance on the original (unperturbed) test subset. Sensitivity was retained at 2σ [89.0% (95% CI 74.4–96.7), $p = 0.25$], but decreased to 80% (95% CI 69.9–91.1) ($p = 0.02$) at a noise level of 3σ. At the higher noise level of 5σ all performance metrics declined [specificity (95% CI) 85.4%, (75.1–89.0), sensitivity (95% CI) 60% (42.1–83.3), accuracy (95% CI) 76.5% (70.0–89.1), and AUC (95% CI) 0.80 (0.75–0.96); $p < 0.05$] (Table 1).

The model's performance metrics were not impacted by the rotations of the CT image and the corresponding pancreas segmentation mask by 45° and 90° in the axial (x–y) plane (Table 1).

Robustness against common variations in radiomics workflow

For both isotropic and anisotropic voxel resampling, the model's performance was comparable to its performance on the original (unperturbed) test subset ($p = 0.25$–0.31). The model's performance was also robust against change in the gray-level BW from the default value of 25 to 32 ($p = 0.31$–0.99). However, the BW of 64 resulted in extreme decrease in the gray levels and decline in model's performance with AUC, accuracy, and sensitivity decreasing to 0.80, 74.2%, and 51.1%, respectively ($p < 0.001$–0.03), and specificity tending to decline ($p = 0.06$) (Table 2). With a BW of 64 or a gray-level setting of 4, we observed a notable decrease in GLCM- difference entropy—a measure of textural heterogeneity—on pre-diagnostic CT (Fig. 5). This reduction in textural heterogeneity corresponded with a marked drop in the model's sensitivity, down to 51.1%. Conversely, control CT maintained more textural complexity, contributing to a higher specificity of 86.4% (Table 2).

Erosion of the pancreas segmentation mask by 3 and 4 pixels did not impact the model's performance vis-à-vis its performance on the original (unperturbed) test subset [AUC (95% CI) 0.96 (0.91–0.97) and 0.94 (0.89–0.95); accuracy (95% CI) 87.5% (83.9–91.4) and 88.3% (81.2–89.8); sensitivity: 93.3% (78.8–96.8) and 93.3% (73.3–95.5); and specificity: 84.3% (80.7–92.8), 85.5% (78.3–91.5), respectively; $p = 0.12–0.34$]. However, erosions of 6 and 8 pixels resulted in decline in some performance metrics (Table 3).

Likewise, dilation of the pancreas segmentation mask by 3 and 4 pixels did not impact the model's performance vis-à-vis its performance on the original (unperturbed) test subset [AUC (95% CI) 0.95(0.90–0.96) and 0.90 (0.85–0.91); accuracy (95% CI) 87.5% (83.2–89.6) and 85.1% (78.2–87.8); sensitivity: 93.1% (78.2–95.8) and 90.1% (72.2–94.1); and specificity: 84.1% (80.2–92.4), 83.1% (75.3–85.5), respectively; $p = 0.10–0.38$]. However, dilation of 6 pixels resulted in a significant decline in the performance metrics ($p < 0.05$) (Table 3).

Discussion

Recently, radiomics-based ML models have shown promise in identifying visually occult PDA at the pre-diagnostic stage, enabling the potential for earlier detection through screening asymptomatic high-risk individuals such as those with new-onset diabetes and weight loss [11, 22]. By leveraging quantitative analyses of radiomic features extracted from volumetrically segmented pancreas, these models offer the prospect of identifying subtle patterns that may signify the presence of early PDA, even when it is visually imperceptible. To optimize generalizability, evaluating the robustness of these models is crucial to understand the limits of variations in image acquisition and radiomics workflow that preserve model performance.

Our study demonstrated that the SVM model for early PDA detection on pre-diagnostic CTs maintained comparable performance up to a three-fold increase in original noise level (σ), except for sensitivity declining to 80% at 3σ. The model's performance was unaffected by image or segmentation mask rotation, voxel size resampling, and a bin-width (BW) of up to 32. However, performance declined with a BW of 64 and erosion/dilation beyond 4 pixels from the pancreas boundary. We found that image perturbation is a useful technique to investigate the limits of variations in image acquisition and radiomics workflow that degrade the model's performance.

Previously, we had employed an ablation approach to identify the radiomic features that had the highest influence on the model's performance [11]. The identified features—High Gray-Level Zone Emphasis (HGLZE) from the GLSZM, Dependence Entropy from the GLDM, and Difference Variance from the GLCM—reflect spatial heterogeneity and randomness on images due to heterogeneity at the tissue level. The mechanism of the observed impact of image perturbations on the model in the current study is likely alteration of these essential radiomic features. This underscores the necessity for standardized image acquisition and radiomics workflow to ensure the reliability of radiomics-based AI models.

Image noise has significant implications for radiomic features, including decreased signal-to- noise ratio (SNR), obscuring subtle details, and compromising accuracy. Image noise can arise from decreased radiation output from the CT system (e.g., tube current or potential) or from different kernels or denoising methods in image reconstruction. Texture features are particularly vulnerable to noise-induced variations, impacting their interpretability [23]. However, our experiments demonstrate the robustness of the SVM model despite a doubling in image noise. We attribute the model's resilience to the careful choice of feature extraction and selection methods, which influence the sensitivity of radiomic analysis to image noise, as well as the nature of the analyzed volume-of-interest (VOI). Specifically, our analysis focused on the pancreas gland, which exhibits relatively homogeneous characteristics compared to a tumor- derived VOI. Additionally, the model's robustness against rotations in CT images and corresponding segmentations is because these rotations did not alter the VOI or the segmentation mask, ensuring consistency in the analysis.

Image preprocessing in the radiomics workflow aims to enhance the quality and standardization of the images before feature extraction and analysis. Preprocessing allows for the harmonization of images, ensuring that the radiomic features are extracted from a consistent and comparable scale. By minimizing confounding factors, preprocessing ensures that the extracted features are more indicative of the underlying biological information and less influenced by technical or image-related biases [24]. However, it is equally important to evaluate robustness of radiomics models when confronted with variations in the image analyses pipeline. To that end, we introduced a wide range of variations in voxel resampling and gray-level discretization.

Voxel resampling is used to standardize voxel resolution in the radiomics workflow across imaging cohorts. Anisotropic voxel resampling, which considers the combination of pixel size and slice thickness commonly found in reconstructed CT images, is clinically more relevant [25].

We tested two anisotropic configurations, $0.75 \times 0.75 \times 2\ mm^3$ and $0.75 \times 0.75 \times 3\ mm^3$, reflecting the median slice thicknesses of the control and pre-diagnostic CT cohorts in our dataset. Preserved SVM performance across voxel sampling schemes may be due to feature extraction methods that are less sensitive to preprocessing variations, prioritized stable features, appropriate feature normalization techniques, and the incorporation of diverse feature sets.

We investigated the impact of varying bin width (BW) on the model's performance in gray-level discretization. BW group image pixels into distinct gray levels, influencing texture and heterogeneity representation. Starting from a default BW of 25, we increased it gradually up to 64. The model's performance remained comparable ($p > 0.05$) up to a BW of 32. However, a significant drop in performance ($p < 0.05$) was observed with a BW of 64. With only 4 gray levels, images became more homogeneous, losing discriminative features essential for accurate classification. Optimal performance requires selecting an appropriate BW that preserves textural heterogeneity while avoiding excessive reduction of gray levels, maintaining the model's discriminative power.

Organ segmentation variability is a limitation in radiomics studies, especially for small organs like the pancreas [26–28]. We performed algorithmic erosion on the pancreatic segmentation mask, eroding it by 3, 4, 6, and 8 pixels and dilation by 3,4 and 6 pixels from the pancreas boundary. The model consistently performed well with up to 4 pixels of erosion/dilation but showed a decline in certain metrics with erosions/dilations beyond 6 pixels. This decline reflects the challenge of capturing complete tissue heterogeneity with a higher level of erosion and the contamination of radiomics features with spurious extra-pancreatic features with a higher level of dilation. Our study simulates inter/intra-reader variability, highlighting the challenges in manual segmentation workflows. The development of AI models for automated high-fidelity pancreas segmentation [29–31] holds promise for enhancing the scalability of radiomics-based ML approach.

Our study has some limitations. Firstly, we were unable to examine all possible combinations and perturbations of configurations in image acquisition and radiomics workflow. Instead, we focused on evaluating the 18 configurations individually. Exploring every possible permutation would have led to an overwhelming number of configurations, rendering the study impractical and hindering our ability to draw meaningful conclusions. Secondly, we were unable to evaluate the impact of different reconstruction kernels on the robustness of the model's performance, as we did not have access to the raw CT data in view of the retrospective nature of this study. Thirdly, we did not investigate the stability of individual radiomic features, although robustness of the model is a surrogate indicator of the robustness of the underlying features [32]. Besides, it important to note that even if a feature is reproducible, it may not necessarily be clinically informative [13], which underscores the importance of evaluating the robustness of the model.

Finally, although algorithmically generated-perturbations offer a promising alternative to traditional validation methods, they may not capture the entire complexity of clinical imaging workflow. Therefore, the next phase investigation necessitates prospective clinical trials involving multiple institutions with diverse patient populations and varying imaging protocols.

In conclusion, we assessed the SVM model's robustness against common variations in image acquisition and radiomics workflow using image perturbations prior to prospective evaluation.

The model demonstrated comparable performance for detection of visually occult PDA at the pre-diagnostic stage within a broad range of clinically relevant variations in image acquisition and radiomics workflow. Prospective validation and integration of the model with complementary biomarkers could further enhance cancer prediction at the pre-diagnostic stage in high-risk cohorts. The model is being considered for deployment in ongoing clinical trials, such as the Early Detection Initiative [33], which aims to evaluate the outcomes of a screening strategy for sporadic PDA utilizing risk-prediction models and CT in 12,500 high-risk subjects.

Funding

Dr. Goenka acknowledges grants from non-profit entities such as the Champions for Hope Pancreatic Cancer Research Program of the Funk Zitiello Foundation, the Centene Charitable Foundation, and the Advance the Practice Award from the Department of Radiology, Mayo Clinic, Rochester, Minnesota. Unrelated to this work (Dr. Goenka): CA190188, Department of Defense (DoD), Office of the Congressionally Directed Medical Research Programs (CDMRP); R01CA256969, National Cancer Institute (NCI) of the National Institutes of Health (NIH); R01CA272628- 01, National Cancer Institute (NCI) of the National Institutes of Health (NIH); Institutional research grant from Sofie Biosciences and Clovis Oncology; Advisory Board (ad hoc), BlueStar Genomics; Consultant, Bayer Healthcare, LLC; Consultant, Candel Therapeutics; Consultant, UWorld.

References

1. Singh DP, Sheedy S, Goenka AH, Wells M, Lee NJ, Barlow J, et al. (2020) Computerized tomography scan in pre-diagnostic pancreatic ductal adenocarcinoma: Stages of progression and potential benefits of early intervention: A retrospective study. Pancreatology 20(7):1495–501 10.1016/j.pan.2020.07.410. [PubMed: 32950386]
2. Chari ST, Kelly K, Hollingsworth MA, Thayer SP, Ahlquist DA, Andersen DK, et al. (2015) Early detection of sporadic pancreatic cancer: summative review. Pancreas 44(5):693–712 10.1097/MPA.0000000000000368. [PubMed: 25931254]
3. Chhoda A, Vodusek Z, Wattamwar K, Mukherjee E, Gunderson C, Grimshaw A, et al. (2022) Late-Stage Pancreatic Cancer Detected During High-Risk Individual Surveillance: A Systematic Review and Meta-Analysis. Gastroenterology 162(3):786–98 10.1053/j.gastro.2021.11.021. [PubMed: 34813861]
4. Klatte DCF, Boekestijn B, Onnekink AM, Dekker FW, van der Geest LG, Wasser M, et al. (2023) Surveillance for Pancreatic Cancer in High-Risk Individuals Leads to Improved Outcomes: A Propensity Score-Matched Analysis. Gastroenterology 164(7):1223–31 e4 10.1053/j.gastro.2023.02.032. [PubMed: 36889551]
5. Overbeek KA, Goggins MG, Dbouk M, Levink IJM, Koopmann BDM, Chuidian M, et al. (2022) Timeline of Development of Pancreatic Cancer and Implications for Successful Early Detection in High-Risk Individuals. Gastroenterology 162(3):772–85 e4 10.1053/j.gastro.2021.10.014. [PubMed: 34678218]
6. Kurita Y, Kuwahara T, Hara K, Mizuno N, Okuno N, Matsumoto S, et al. (2019) Diagnostic ability of artificial intelligence using deep learning analysis of cyst fluid in differentiating malignant from benign pancreatic cystic lesions. Sci Rep 9(1):6893 10.1038/s41598-019-43314-3. [PubMed: 31053726]
7. Kang J, Clarke SE, Abdolell M, Ramjeesingh R, Payne J, Costa AF (2021) The implications of missed or misinterpreted cases of pancreatic ductal adenocarcinoma on imaging: a multi-centered population-based study. Eur Radiol 31(1):212–21 10.1007/s00330-020-07120-0. [PubMed: 32785768]
8. Kang JD, Clarke SE, Costa AF (2021) Factors associated with missed and misinterpreted cases of pancreatic ductal adenocarcinoma. Eur Radiol 31(4):2422–32 10.1007/s00330-020-07307-5. [PubMed: 32997176]
9. Dewitt J, Devereaux BM, Lehman GA, Sherman S, Imperiale TF (2006) Comparison of endoscopic ultrasound and computed tomography for the preoperative evaluation of pancreatic cancer: a systematic review. Clin Gastroenterol Hepatol 4(6):717–25; quiz 664 10.1016/j.cgh.2006.02.020. [PubMed: 16675307]
10. Toshima F, Watanabe R, Inoue D, Yoneda N, Yamamoto T, Sasahira N, et al. (2021) 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. AJR Am J Roentgenol:1–12 10.2214/AJR.21.26014.
11. Mukherjee S, Patra A, Khasawneh H, Korfiatis P, Rajamohan N, Suman G, et al. (2022) Radiomics-based Machine-learning models can detect pancreatic cancer on prediagnostic computed tomography scans at a substantial lead time before clinical diagnosis. Gastroenterology 163(5):1435–46 e3 10.1053/j.gastro.2022.06.066. [PubMed: 35788343]

12. Zwanenburg A, Leger S, Agolli L, Pilz K, Troost EGC, Richter C, et al. (2019) Assessing robustness of radiomic features by image perturbation. Sci Rep 9(1):614 10.1038/s41598-018-36938-4. [PubMed: 30679599]

13. Zhao B (2021) Understanding Sources of Variation to Improve the Reproducibility of Radiomics. Front Oncol 11:633176 10.3389/fonc.2021.633176. [PubMed: 33854969]

14. Mackin D, Fave X, Zhang L, Fried D, Yang J, Taylor B, et al. (2015) Measuring Computed Tomography Scanner Variability of Radiomics Features. Invest Radiol 50(11):757–65 10.1097/RLI.0000000000000180. [PubMed: 26115366]

15. Yip SS, Aerts HJ (2016) Applications and limitations of radiomics. Phys Med Biol 61(13):R150–66 10.1088/0031-9155/61/13/R150. [PubMed: 27269645]

16. Tixier F, Hatt M, Le Rest CC, Le Pogam A, Corcos L, Visvikis D (2012) Reproducibility of tumor uptake heterogeneity characterization through textural feature analysis in 18F-FDG PET. J Nucl Med 53(5):693–700 10.2967/jnumed.111.099127. [PubMed: 22454484]

17. Leijenaar RT, Carvalho S, Velazquez ER, van Elmpt WJ, Parmar C, Hoekstra OS, et al. (2013) Stability of FDG-PET Radiomics features: an integrated analysis of test-retest and inter- observer variability. Acta Oncol 52(7):1391–7 10.3109/0284186X.2013.812798. [PubMed: 24047337]

18. Teng X, Zhang J, Zwanenburg A, Sun J, Huang Y, Lam S, et al. (2022) Building reliable radiomic models using image perturbation. Sci Rep 12(1):10035 10.1038/s41598-022-14178-x. [PubMed: 35710850]

19. van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. (2017) Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Res 77(21):e104–e7 10.1158/0008-5472.CAN-17-0339. [PubMed: 29092951]

20. Tibshirani R (1997) The lasso method for variable selection in the Cox model. Stat Med 16(4):385–95 10.1002/(sici)1097-0258(19970228)16:4<385::aid-sim380>3.0.co;2-3. [PubMed: 9044528]

21. DeLong ER, DeLong DM, Clarke-Pearson DL (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44(3):837–45 [PubMed: 3203132]

22. Qureshi TA, Gaddam S, Wachsman AM, Wang L, Azab L, Asadpour V, et al. (2022) Predicting pancreatic ductal adenocarcinoma using artificial intelligence analysis of pre- diagnostic computed tomography images. Cancer Biomark 33(2):211–7 10.3233/CBM-210273. [PubMed: 35213359]

23. Oliver JA, Budzevich M, Hunt D, Moros EG, Latifi K, Dilling TJ, et al. (2017) Sensitivity of Image Features to Noise in Conventional and Respiratory-Gated PET/CT Images of Lung Cancer: Uncorrelated Noise Effects. Technol Cancer Res Treat 16(5):595–608 10.1177/1533034616661852. [PubMed: 27502957]

24. Suman G, Patra A, Mukherjee S, Korffiatis P, Goenka AH (2022) Radiomics for Detection of Pancreas Adenocarcinoma on CT Scans: Impact of Biliary Stents. Radiol Imaging Cancer 4(1):e210081 10.1148/rycan.210081. [PubMed: 34939849]

25. Huang K, Rhee DJ, Ger R, Layman R, Yang J, Cardenas CE, et al. (2021) Impact of slice thickness, pixel size, and CT dose on the performance of automatic contouring algorithms. J Appl Clin Med Phys 22(5):168–74 10.1002/acm2.13207.

26. van Timmeren JE, Cester D, Tanadini-Lang S, Alkadhi H, Baessler B (2020) Radiomics in medical imaging-“how-to” guide and critical reflection. Insights Imaging 11(1):91 10.1186/s13244-020-00887-2. [PubMed: 32785796]

27. Haarburger C, Muller-Franzes G, Weninger L, Kuhl C, Truhn D, Merhof D (2020) Radiomics feature reproducibility under inter-rater variability in segmentations of CT images. Sci Rep 10(1):12688 10.1038/s41598-020-69534-6. [PubMed: 32728098]

28. Wright DE, Mukherjee S, Patra A, Khasawneh H, Korfiatis P, Suman G, et al. (2022) Radiomics-based machine learning (ML) classifier for detection of type 2 diabetes on standard-of-care abdomen CTs: a proof-of-concept study. Abdom Radiol (NY) 47(11):3806–16 10.1007/s00261-022-03668-1. [PubMed: 36085379]

29. Panda A, Korfiatis P, Suman G, Garg SK, Polley EC, Singh DP, et al. (2021) 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 48(5):2468–81 10.1002/mp.14782. [PubMed: 33595105]

30. Khasawneh H, Patra A, Rajamohan N, Suman G, Klug J, Majumder S, et al. (2022) Volumetric Pancreas Segmentation on Computed Tomography: Accuracy and Efficiency of a Convolutional Neural Network Versus Manual Segmentation in 3D Slicer in the Context of Interreader Variability of Expert Radiologists. J Comput Assist Tomogr 46(6):841–7 10.1097/RCT.0000000000001374. [PubMed: 36055122]
31. Korfiatis P, Suman G, Patnam NG, Trivedi KH, Karbhari A, Mukherjee S, et al. (2023) Automated Artificial Intelligence Model Trained on a Large Data Set Can Detect Pancreas Cancer on Diagnostic Computed Tomography Scans As Well As Visually Occult Preinvasive Cancer on Prediagnostic Computed Tomography Scans. Gastroenterology 10.1053/j.gastro.2023.08.034.
32. Teng X, Zhang J, Ma Z, Zhang Y, Lam S, Li W, et al. (2022) Improving radiomic model reliability using robust features from perturbations for head-and-neck carcinoma. Front Oncol 12:974467 10.3389/fonc.2022.974467. [PubMed: 36313629]
33. Chari ST, Maitra A, Matrisian LM, Shrader EE, Wu BU, Kambadakone A, et al. (2022) 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 113:106659 10.1016/j.cct.2021.106659. [PubMed: 34954100]

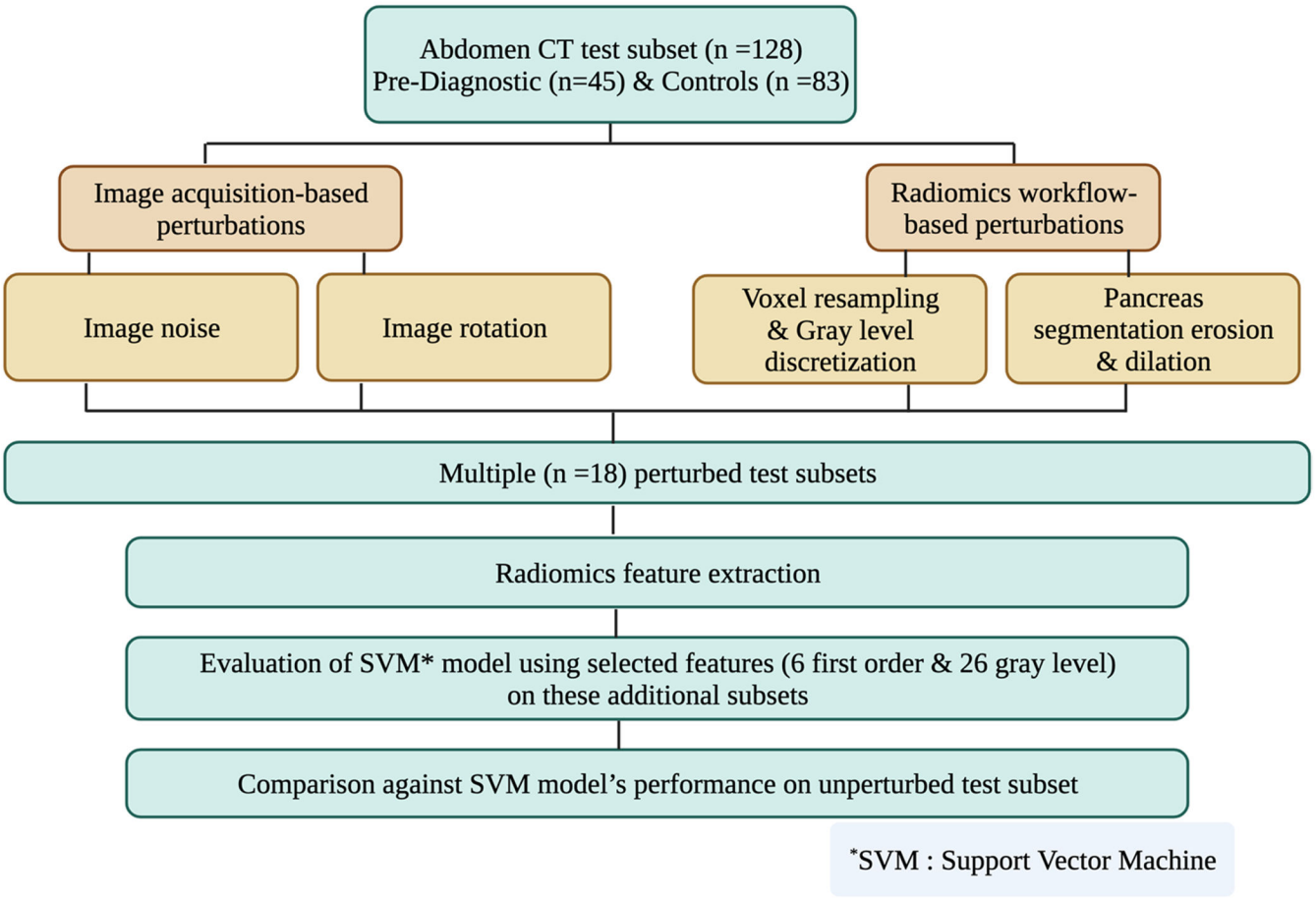

Fig. 1.
Study design

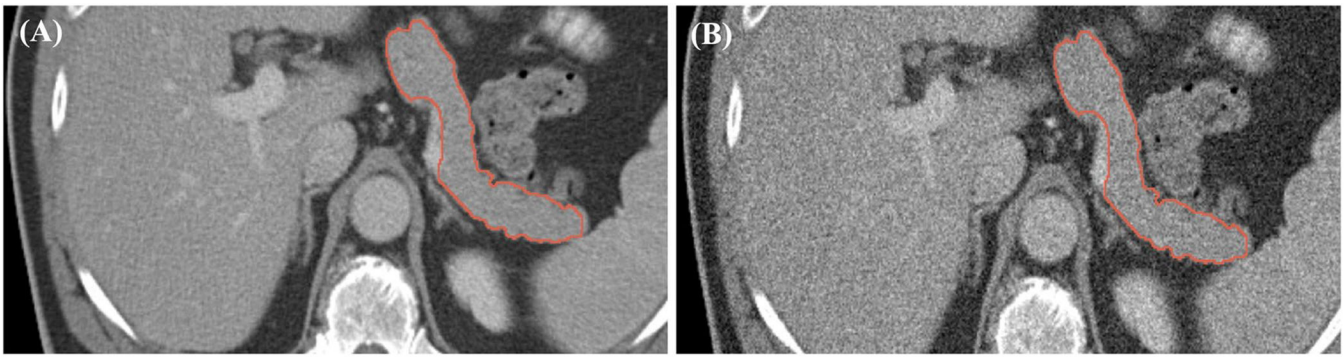

Fig. 2.

Addition of image noise: **A** axial portal venous phase image with the pancreas segmentation, **B** algorithmically generated random Gaussian noise (level of 5σ) was added to the image

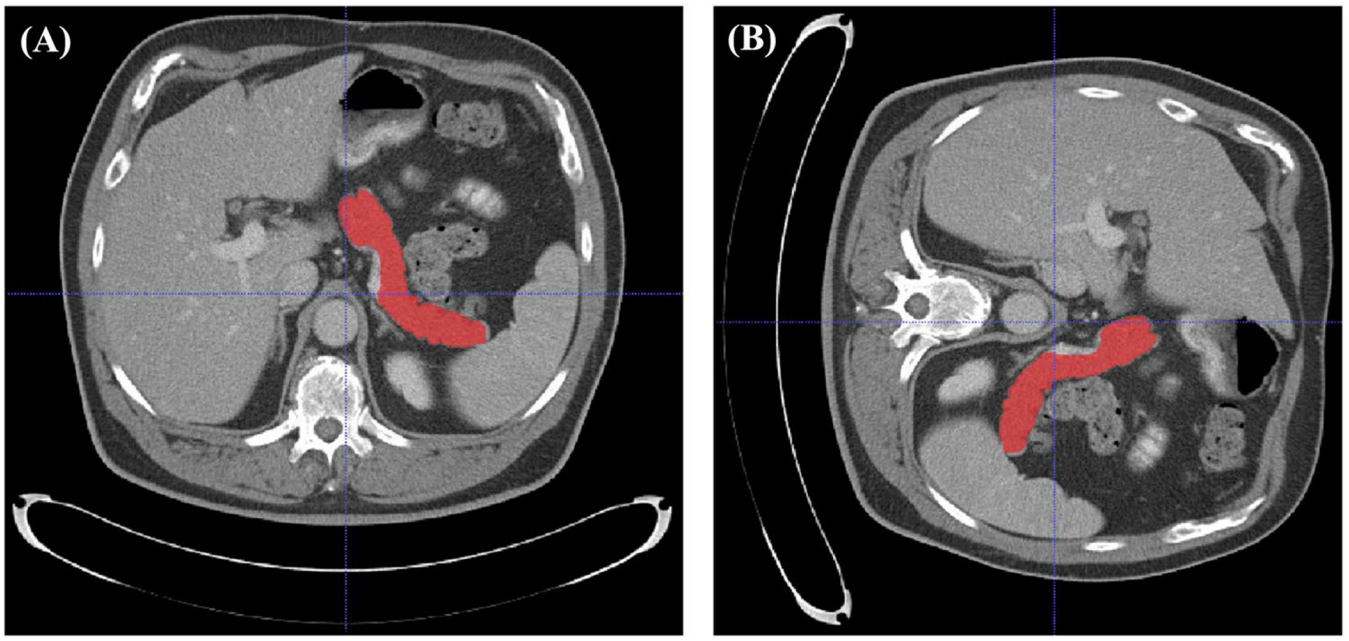

Fig. 3.

Image rotation: **A** axial portal venous phase image with the pancreas segmentation, **B** the same image and pancreas segmentation rotated by 90° in the axial (*x*–*y*) plane

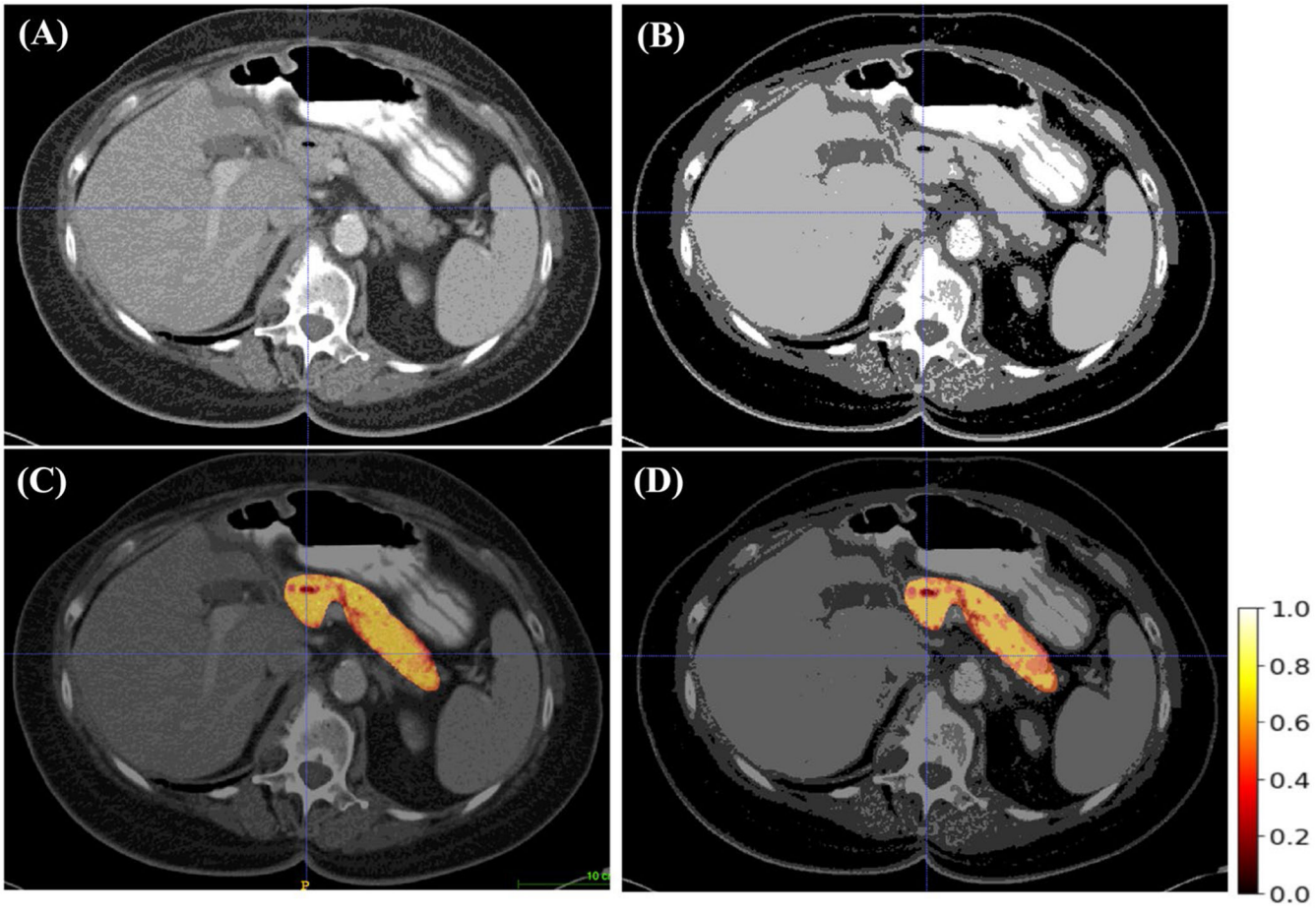

Fig. 4.

Effect of bin width on gray levels and textural pattern: **A** gray-level discretization of an axial portal venous phase image of a pre-diagnostic CT with a BW of 25 shows 11 distinct gray levels, **B** the same image discretized with a BW of 64 shows only 4 Gy levels, **C**, **D** color-coded GLSZM-high gray-level zone emphasis (GLSZM-HGLZE) feature extracted from the volumetric pancreas segmentation and overlaid on the corresponding CTs. Reduction in fineness of the textural heterogeneity in **D** is apparent. Such extreme reduction in gray levels degraded model's performance. Images were normalized to (0, 255) prior to gray-level discretization

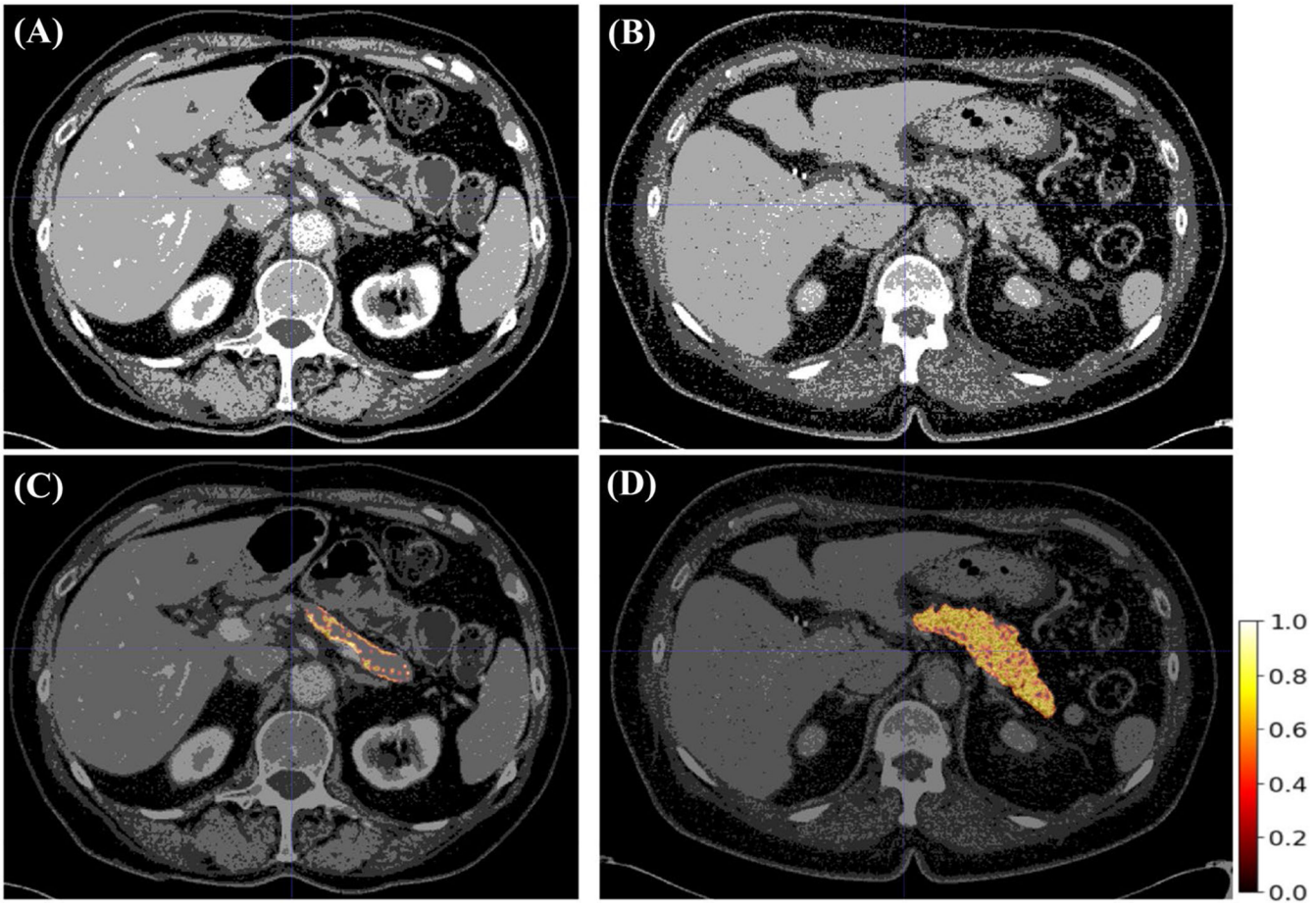

Fig. 5.

Texture maps at low gray level: **A**, **C** axial portal venous phase pre-diagnostic CT and corresponding color-coded GLCM-difference entropy textural map at a BW of 64 overlaid on the pancreas, **B**, **D** axial portal venous phase control CT and corresponding color-coded GLCM- difference entropy textural map at a BW of 64 overlaid on the pancreas. Images were normalized to (0, 255) prior to gray-level discretization. There is pronounced decrease in GLCM-difference entropy—a measure of textural heterogeneity—on pre-diagnostic CT, which corresponds with a marked drop in the model's sensitivity to 51.1%. Conversely, textural complexity was better preserved on the control CT, which corresponds with higher model's specificity (86.4%)

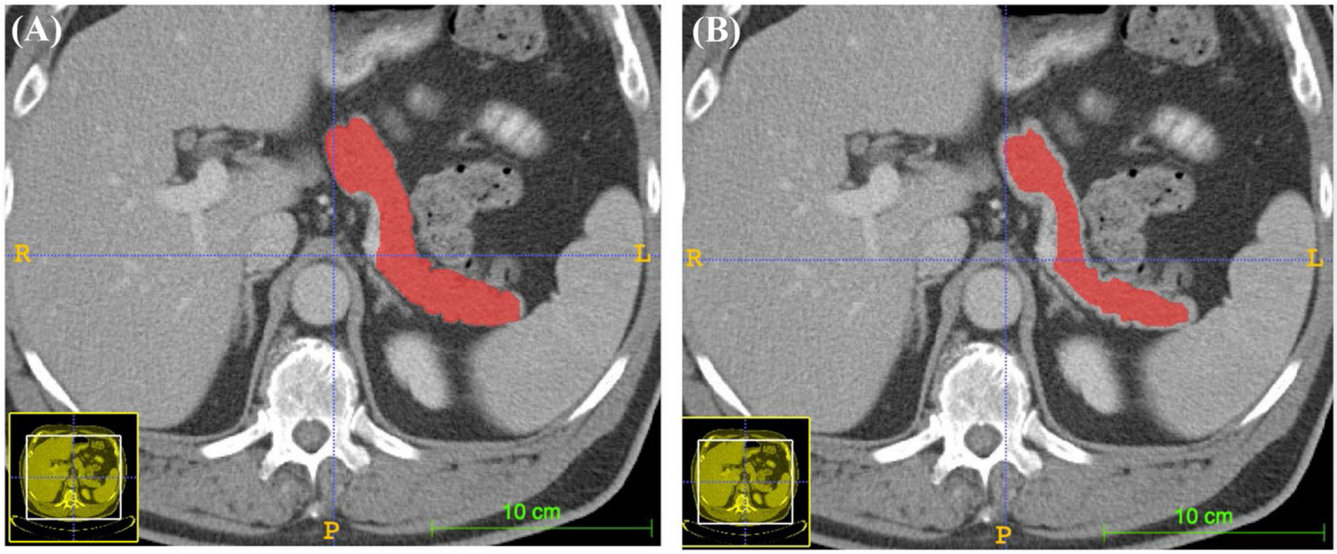

Fig. 6.

Erosion of segmentation mask: **A** axial portal venous phase CT image with overlaid pancreas segmentation, **B** the segmentation mask with erosion of 6 pixels from the boundary

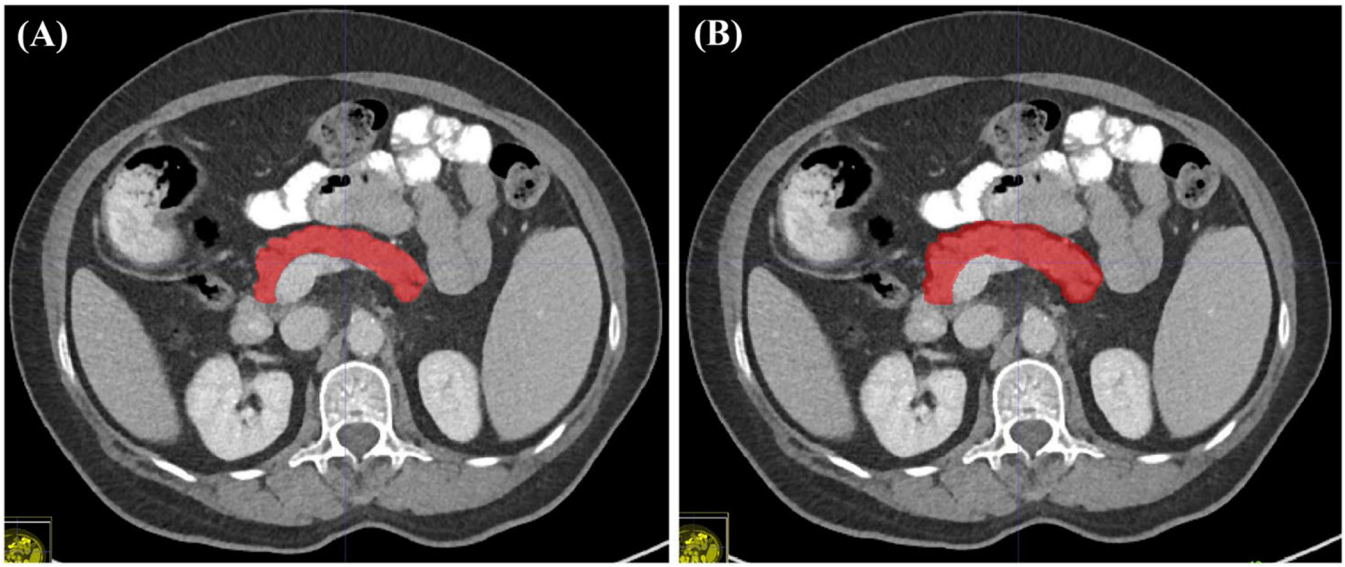

Fig. 7.

Dilation of segmentation mask: **A** axial portal venous phase CT image with overlaid pancreas segmentation, **B** the segmentation mask with dilation of 6 pixels from the boundary

Table 1

Model's performance at different variations in image acquisition workflow

	AUC [95% CI]	Accuracy[a] [95% CI]	Sensitivity[a] [95% CI]	Specificity[a] [95% CI]
Original	0.98 [0.94–0.98]	92.2 [86.7–93.7]	95.5 [85.5–100.0]	90.3 [84.3–91.5]
Noise-1σ	0.98 [0.94–0.98]	92.2 [87.0–93.7]	95.5 [83.2–100.0]	90.3 [85.5–92.7]
Noise-2σ	0.97 [0.93–0.98]	89.8 [85.1–92.6]	89.0 [74.4–96.7]	90.2 [85.5–92.7]
Noise-3σ	0.94 [0.91–0.98]	85.9 [84.7–91.4]	80.0* [69.9–91.1]	89.2 [87.9–96.4]
Noise-5σ	0.80* [0.75–0.96]	76.5* [70.0–89.1]	60.0* [42.1–83.3]	85.4* [75.1–89.0]
Rotation 45°	0.98 [0.94–0.98]	92.0 [86.7–93.7]	95.5 [85.5–100.0]	90.3 [84.3–91.5]
Rotation 90°	0.98 [0.94–0.98]	92.0 [87.0–93.7]	95.5 [85.5–100.0]	90.3 [84.3–91.5]

[a]Numbers shown in %

*$p < 0.05$

Table 2

Model's performance at variations in voxel resampling and gray-level discretization

	AUC [95% CI]	Accuracy[a] [95% CI]	Sensitivity[a] [95% CI]	Specificity[a] [95% CI]
Original	0.98 [0.94–0.98]	92.2 [86.7–93.7]	95.5 [85.5–100.0]	90.3 [84.3–91.5]
Isotropic resampling (1 × 1 × 1 mm^3)	0.96 [0.94–0.97]	89.0 [85.5–92.2]	89.0 [87.7–100.0]	89.2 [80.7–90.4]
Anisotropic resampling (0.75 × 0.75 × 2)	0.96 [0.94–0.97]	86.7 [84.4–91.4]	89.0 [86.6–97.7]	86.0 [78.8–89.1]
Anisotropic resampling (0.75 × 0.75 × 3)	0.91 [0.89–0.96]	86.7 [82.1–90.2]	89.0 [84.2–97.0]	85.8 [78.4–89.1]
Bin-width 32	0.98 [0.94–0.98]	92.2 [85.1–92.9]	95.5 [80.0–100.0]	90.2 [78.3–93.9]
Bin-width 64	0.80* [0.67–0.93]	74.2* [55.8–84.8]	51.1* [22.1–100.0]	86.4 [35.3–100.0]

[a]Numbers shown in %

*$p < 0.05$

Table 3

Model's performance at different erosions of pancreas segmentation

	AUC [95% CI]	Accuracy[a] [95% CI]	Sensitivity[a] [95% CI]	Specificity[a] [95% CI]
Original	0.98 [0.94–0.98]	92.2 [86.7–93.7]	95.5 [85.5–100.0]	90.3 [84.3–91.5]
Erosion-3	0.96 [0.91–0.97]	87.5 [84.0–91.4]	93.3 [78.8–96.8]	84.3 [80.7–92.8]
Erosion-4	0.94 [0.89–0.95]	88.3 [81.2–89.8]	93.3 [73.3–95.5]	85.5 [78.3–91.5]
Erosion-6	0.90 [0.85–0.92]	83.6* [76.0–85.0]	91.1 [64.4–95.5]	79.5* [69.8–88.0]
Erosion-8	0.87* [0.82–0.90]	81.2* [74.2–83.6]	89.0 [64.4–93.3]	77.1* [65.1–86.2]
Dilation-3	0.95 [0.90–0.96]	87.5 [83.2–89.6]	93.1 [78.2–95.8]	84.1 [80.2–92.4]
Dilation-4	0.90 [0.85–0.91]	85.1 [78.2–87.8]	90.1 [72.2–94.1]	83.1 [75.3–85.5]
Dilation-6	0.80* [0.66–0.87]	79.0* [57.8–84.8]	79.4* [52.1–83.2]	78.2* [53.2–83.5]

[a] Numbers shown in %

* $p < 0.05$