scientific reports

OPEN

Pancreas segmentation using AI developed on the largest CT dataset with multi-institutional validation and implications for early cancer detection

Sovanlal Mukherjee[1], Ajith Antony[1], Nandakumar G. Patnam[1], Kamaxi H. Trivedi[1], Aashna Karbhari[1], Madhu Nagaraj[1], Murlidhar Murlidhar[1] & Ajit H. Goenka[2✉]

Accurate and fully automated pancreas segmentation is critical for advancing imaging biomarkers in early pancreatic cancer detection and for biomarker discovery in endocrine and exocrine pancreatic diseases. We developed and evaluated a deep learning (DL)-based convolutional neural network (CNN) for automated pancreas segmentation using the largest single-institution dataset to date (n = 3031 CTs). Ground truth segmentations were performed by radiologists, which were used to train a 3D nnU-Net model through five-fold cross-validation, generating an ensemble of top-performing models. To assess generalizability, the model was externally validated on the multi-institutional AbdomenCT-1K dataset (n = 585), for which volumetric segmentations were newly generated by expert radiologists and will be made publicly available. In the test subset (n = 452), the CNN achieved a mean Dice Similarity Coefficient (DSC) of 0.94 (SD 0.05), demonstrating high spatial concordance with radiologist-annotated volumes (Concordance Correlation Coefficient [CCC]: 0.95). On the AbdomenCT-1K dataset, the model achieved a DSC of 0.96 (SD 0.04) and a CCC of 0.98, confirming its robustness across diverse imaging conditions. The proposed DL model establishes new performance benchmarks for fully automated pancreas segmentation, offering a scalable and generalizable solution for large-scale imaging biomarker research and clinical translation.

Keywords Pancreas, Artificial intelligence, Volumetric segmentation, Computed tomography

Pancreatic ductal adenocarcinoma (PDA) remains one of the deadliest malignancies, with a five-year survival rate of approximately 10–12%[1]. In 2024 alone, an estimated 66,440 individuals were diagnosed with PDA, and over 51,750 patients succumbed to the disease[2]. This high incidence-to-mortality ratio is driven by the current paradigm of symptom-driven diagnosis, which results in over 85% of cases being detected at an unresectable stage, limiting therapeutic options to palliative care[3–5]. While hereditary PDA benefits from targeted surveillance, sporadic PDA (~ 85% of cases) lacks an equivalent early detection approach[6]. Recently, glycemically-defined new-onset diabetes (gNOD) with an Enriching NOD for Pancreatic Cancer (ENDPAC) score ≥ 3 has emerged as a validated high-risk cohort for sporadic PDA, with a three-year risk (~ 3.6%) nearly 20-fold higher than the general population[7,8]. This subgroup represents an actionable opportunity for targeted early detection, which aligns with emphasis of the United States Preventive Services Task Force (USPSTF) on risk-based screening strategies. The effectiveness of screening strategies, even in these high-risk, well-defined cohorts, hinges on the ability of imaging to detect PDA at its earliest, potentially curable stage.

Conventional imaging, however, is fundamentally limited in detection of PDA at its pre-diagnostic, asymptomatic stage (stage 0). The International Cancer of the Pancreas Screening (CAPS) Consortium, spanning 2552 high-risk individuals across 16 surveillance programs, found 75% of PDA cases diagnosed at the metastatic stage, with nearly 50% showing no abnormalities on MRI or endoscopic ultrasound (EUS) a year prior[9]. A

[1]Department of Radiology, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, USA. [2]Professor of Radiology, Consultant, Divisions of Abdominal and Nuclear Radiology, Co-Chair, Nuclear Radiology Research Operations, Chair, Enterprise PET/MR Research, Education and Executive Committee, Program Co-Leader, Risk Assessment, Early Detection and Interception (REDI), Mayo Clinic Comprehensive Cancer Center (MCCCC), 200 First St SW, Charlton 1, Rochester, MN 55905, USA. ✉email: goenka.ajit@mayo.edu

meta-analysis of 13 studies reported that over 58% of detected PDAs in high-risk individuals undergoing serial surveillance imaging were already advanced[10]. These findings underscore a significant barrier to early detection—the pancreas appears normal on standard imaging at the prediagnostic stage, which makes visual identification unreliable. Given the aggressive nature of PDA, which rapidly progresses from being subclinical to stage IV, this limitation significantly restricts the window for timely intervention[11].

Emerging evidence suggests that radiomic features extracted from normal-appearing pancreas can detect PDA at least a year before clinical diagnosis. Recent studies have demonstrated that machine learning-based analysis of pre-diagnostic CTs can identify subtle imaging biomarkers that are imperceptible to radiologists[12–16], with one study reporting an average lead time of 398 days before clinical detection[12,13]. Such advancements offer a paradigm shift in early PDA detection, enabling risk stratification and intervention before symptoms emerge. However, the clinical translation of these biomarkers is entirely contingent on precise, reproducible, and scalable volumetric pancreas segmentation.

Beyond oncology, quantitative morphometric and radiomic analysis of the pancreas has significant implications for other endocrine and exocrine disorders. Pancreas volume and radiomic signatures have been linked with diabetes mellitus, metabolic syndrome, and early chronic pancreatitis[17–20]. Longitudinal studies have demonstrated that pancreatic volumetric changes serve as indicators of treatment response in bariatric surgery, anti-diabetic therapies, and immunotherapy[21–23]. Therefore, the development of an accurate, automated pancreas segmentation tool would not only accelerate early PDA detection but also facilitate broader clinical and research applications in pancreatic diseases.

Manual pancreas segmentation is inherently challenging due to the organ's irregular shape, variable anatomic positioning, and complex interface with adjacent structures. Even among experienced radiologists, manual segmentation is time-consuming and subject to high inter-reader and intra-reader variability, which leads to inconsistencies that hinder biomarker reproducibility[24]. While previous deep learning (DL)-based pancreas segmentation studies have made notable progress, they have largely been constrained by small, single-institution datasets, widely variable performance, limited external validation, and segmentation pipelines that require substantial manual intervention[25,26]. Given these limitations, there is a critical need for a large-scale, high-performance, and generalizable AI-based pancreas segmentation model capable of fully automated volumetric analysis in heterogeneous real-world datasets. This study aimed to address this unmet need by developing and validating a fully automated DL-based model for volumetric pancreas segmentation using the largest reported dataset of CT scans. Furthermore, we evaluate its generalizability across an independent international multi-institutional public dataset to ensure its robustness for potential clinical implementation.

Methods
Study participants
This Health Insurance Portability and Accountability Act (HIPAA)-compliant retrospective study was approved by the Institutional Review Board (IRB) under Protocol Number 19-003018. The study was classified as minimal risk and was, therefore, granted a waiver by the IRB for written informed consent. All methods were conducted in accordance with relevant guidelines and regulations, including the principles outlined in the Declaration of Helsinki.

Portal venous phase CTs were identified from our Radiology Information System by selecting those with radiology reports indicating a normal pancreas. We deliberately chose not to impose additional exclusion criteria to enhance real-world applicability. Our approach was intended to ensure that the dataset represented the full spectrum of anatomical and physiological variations encountered in routine clinical practice. Given that minor incidental pancreatic findings—such as tiny cystic lesions, lipomatosis, or calcifications—are not consistently documented by radiologists in their reports, their presence within the dataset mirrors the inherent variability and diversity seen in standard clinical imaging. Consequently, this allowed the model to be trained on data that more accurately reflected the heterogeneity of pancreas morphology encountered in real-world patient populations. Furthermore, only one CT scan was selected per individual. All CT scans were acquired and reconstructed using standardized protocols for portal venous phase CTs, as previously outlined[27]. Figure 1 shows the study design and structure of the datasets.

Each CT study was downloaded and de-identified by anonymizing the Digital Imaging and Communication in Medicine (DICOM) tags using Clinical Trial Processor[28]. Metadata elements, such as scanner vendor and CT slice thickness, were extracted from the DICOM headers. These anonymized CT datasets were then converted into Neuroimaging Informatics Technology Initiative (NIfTI) format. Fellowship-trained radiologists confirmed portal venous acquisition phase and verified optimal image quality. CTs were excluded if they exhibited significant image quality issues, including motion artifacts (e.g., breathing-related blurring), streak artifacts from metallic implants or spinal hardware, or inadequate contrast enhancement impairing delineation of the pancreatic parenchyma. This curation process resulted in a dataset of 3031 CTs from 3031 patients (Fig. 1).

To maintain consistency and minimize annotation variability, each CT scan was randomly assigned to one out of five fellowship-trained radiologists, each with 3–5 years of post-residency experience in abdominal imaging and specific expertise in pancreatic CT interpretation. These radiologists followed a standardized pancreas segmentation protocol using the boundary points-based segmentation mode of the organ segmentation module from NVIDIA in an established segmentation platform (3D Slicer, version 4.11.0). Rather than implementing independent multi-reader segmentation, which can introduce considerable variability in the absence of harmonization protocols, we adopted a structured single-reader workflow. This approach is particularly relevant in pancreas segmentation, where the organ's variable morphology and poorly defined boundaries already pose significant challenges. Introducing multiple independent annotations without reconciliation would likely have amplified this ambiguity and increased noise in the training data.
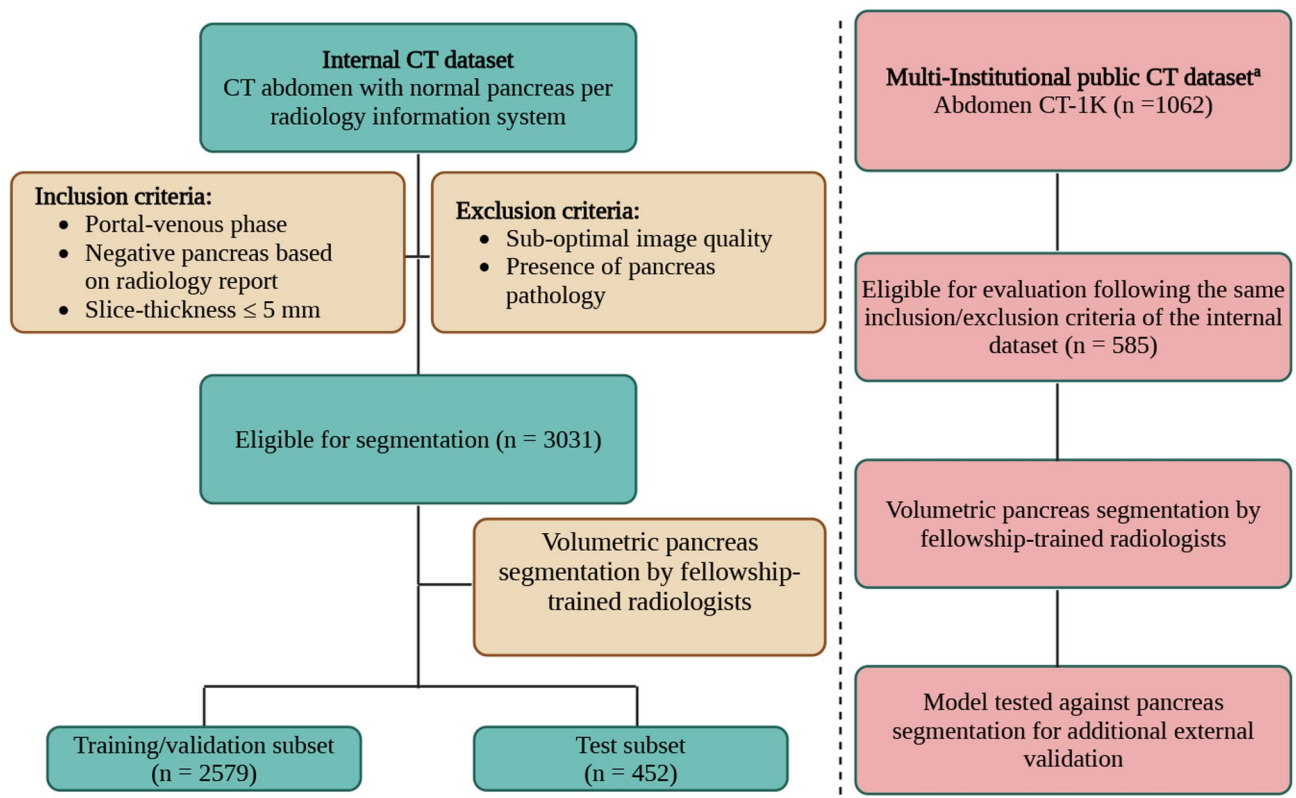
Fig. 1. Study design and structure of datasets.

Inter-reader variability has been well-documented in prior pancreas segmentation studies[29], and our strategy aligns with the prevailing methodological standards in the field, where single-reader annotations are the norm to ensure consistency and scalability. While consensus-based techniques such as STAPLE can be effective in synthesizing multiple segmentations, they are primarily suited to smaller datasets or studies explicitly designed around multi-reader inputs. In the context of our large-scale dataset, implementing such methods would have been logistically impractical and computationally intensive. Instead, we prioritized expert-driven, protocol-adherent segmentation, performed by five fellowship-trained expert radiologists. The validity of these segmentations is further supported by the model's robust performance on large-scale internal and external datasets, underscoring the reliability of this single-annotator framework.

The resulting segmentations served as ground truth for training and evaluating the 3D CNN model. The dataset was randomly divided into training/validation (n = 2579) and test (n = 452) subsets.

CNN architecture

The CNN was developed using a 3D nnU-Net model[30], which is a widely used CNN model that demonstrated state-of-the-art performance for biomedical image segmentation. The basic configuration of 3D nnU-Net resembles the original 3D U-Net[31], which consists of contracting and expanding stages, where feature maps from the contracting stage are combined with feature maps from the expanding stage via long skip connections. However, individual U-Net models are task-specific and require a tedious manual selection of model design elements for optimized performance. The nnU-Net tries to improve the generalizability of the U-Net model by automatically configuring the preprocessing, network architecture, training, and postprocessing for any new task of biomedical image segmentation using a dynamic network adaptation strategy. Due to this, nnU-Net has seen many applications in image segmentations from multiple imaging modalities, including CTs[32]. In this study, the model was trained with the Python package nnunet-V2, using fivefold cross-validation, where each fold ran for 1000 epochs. The final model was an ensemble of the five models generated from this fivefold training process. The ensembling process employed a "soft" majority voting strategy, where the predicted probabilities from each fold were averaged voxel-wise across the five models. The final prediction for each voxel was determined based on the average probability, with the class corresponding to the highest average probability selected as the predicted label.

The original nnUNet model has some built-in fixed design parameters, including the combination of cross-entropy and dice loss functions, the stochastic gradient descent (SGD) optimizer with Nesterov momentum, polynomial decay learning rate scheduler, and a comprehensive set of data augmentation techniques, which we utilized throughout the training process. The integration of both dice and cross-entropy loss functions ensures an optimal trade-off between improving volumetric overlap and maintaining voxel-level classification accuracy. The use of SGD as the optimizer facilitates efficient and scalable training, particularly for DL models applied

to large-scale medical imaging datasets, while Nesterov momentum accelerates convergence by enhancing the efficiency of gradient updates. The learning rate is progressively reduced during training via polynomial decay to minimize the risk of overfitting and ensure greater model generalizability.

Furthermore, the model utilized an extensive set of data augmentation strategies to improve its robustness and capacity to generalize across diverse datasets. These techniques include spatial transformations such as rotation, scaling, and flipping, which ensure invariance to changes in image orientation and variations in object size; addition of Gaussian noise that enhances the model's ability to generalize in the presence of noisy data; Gaussian blurring to improve its robustness to slight variations in image resolution; random adjustments to brightness and contrast to simulate the effect of different acquisition settings, and gamma correction to account for variations in scanning protocols and imaging artifacts. These augmentation methods collectively enhance the model's resilience and performance across a wide range of real-world imaging scenarios. The entire training of the model was conducted on a Nvidia GV100 graphical processing unit (GPU) with 32 GB of memory, operating in Linux (Ubuntu version 22.04.3) environment.

Multi-institutional public datasets

The model's generalizability was further evaluated using CTs from the international multi-institutional public dataset AbdomenCT-1K[33], which consists of 1062 CT scans from 12 medical centers. This dataset includes abdomen CT scans from five distinct benchmarking subsets: Liver Tumor Segmentation (LiTS)[34], Kidney Tumor Segmentation (KiTS)[35], Medical Segmentation Decathlon (MSD)[36], National Institutes of Health-Pancreas CT (NIH-PCT)[37], and a dataset from Nanjing University in China, highlighting its broad diversity and international composition. Given that this public dataset encompasses multi-phase, multi-vendor, and multi-disease cases, it offers a comprehensive representation of real-world clinical scenarios. Utilizing such a heterogeneous dataset allows for a rigorous evaluation of the model's robustness and generalizability under highly variable imaging conditions and patient demographics to ascertain its applicability across diverse clinical settings. After study radiologists curated the dataset according to the inclusion and exclusion criteria of the intramural dataset, excluding 477 cases, volumetric pancreas segmentations were independently performed by them on the remaining 585 cases to establish the external ground truth. To encourage further research and contribute to the community efforts, all these ground truth segmentations will be made publicly available.

NIH-PCT dataset

To further evaluate the generalizability of the model on an independently annotated public dataset, we tested its performance on the NIH-Pancreas CT (NIH-PCT) dataset. This dataset comprises 80 abdominal CT scans acquired in the portal venous phase from individuals with morphologically normal pancreas. Pancreas segmentations are available for all subjects. As per the documentation provided by the NIH, these segmentations were initially generated by a medical student using manual slice-by-slice tracing and subsequently reviewed and finalized by an experienced radiologist.

Ablation study on individual model performance

Finally, because the final prediction from our nnU-Net model was derived from an ensemble of five models generated through fivefold cross-validation, we conducted an ablation study to assess the individual contribution of each model. Specifically, we evaluated the segmentation performance of each of the five independently trained models on the test dataset. The results from the best-performing individual model were then compared against the performance of the ensemble model to quantify the added benefit of ensembling.

Analyses of cases in the lower range of model performances

Cases in the lower range of the model's performance, defined as those with a Dice Similarity Coefficient (DSC) below the 5th percentile compared to the ground truth segmentation, were analyzed. The CT scans were reviewed, and the model's predicted segmentations were compared with the ground truth to identify potential reasons for the lower performance. The model's predicted segmentations were then categorized as either under-segmentation, over-segmentation, or both. For under-segmentations, the portions of the pancreas not included in the model-predicted segmentations were noted. For over-segmentations, the structures incorrectly included in the segmentations were identified.

Comparison with a pre-trained public DL model

To benchmark the performance of our model against an established public solution, we conducted a head-to-head comparison with TotalSegmentator[32], a widely used, open-source DL model for anatomical segmentation. TotalSegmentator was trained on a diverse set of CT images and provides automated volumetric segmentations for 104 anatomical structures, including the pancreas. We evaluated its pancreas segmentation performance on both our intramural dataset and the multi-institutional AbdomenCT-1K public dataset and compared the results directly with our ensemble model.

Statistical analyses

Statistical analyses were conducted using R (version 4.0.4) and Python software with scipy (version 1.8.0). Categorical variables were compared using the Fisher exact test, while continuous variables were compared using the Mann–Whitney U-test. The Kruskal–Wallis test was used for comparing continuous variables across multiple groups. The model's performance on the test subset was evaluated using the Dice Similarity Coefficient (DSC) and Hausdorff Distance (HD). The DSC quantifies the similarity between the ground truth and AI-predicted segmentations. It ranges from 0 (no overlap) to 1 (perfect overlap). A higher DSC indicates greater agreement between the segmentations, while a lower DSC reflects less overlap and, consequently, poorer

segmentation performance. Conversely, HD, measured in millimeters (mm), represents the maximum distance between the closest boundary voxels of the ground truth and AI-predicted segmentation, and serves as an indicator of segmentation error. A higher HD value corresponds to a greater segmentation discrepancy, whereas a lower HD (approaching 0 mm) signifies a more accurate segmentation. Performance was further analyzed based on different CT vendors, slice thicknesses, and patient gender. The correlation between model-predicted and reference pancreas volumes was assessed using the Concordance Correlation Coefficient (CCC). Absolute pancreas volumes were compared by evaluating the mean volume error between model-predicted and reference segmentations using Bland–Altman analysis. Statistical significance was indicated by p-values less than 0.05.

Results

Study participants

The entire cohort (n = 3031) included 1388 males (46%) and 1643 females (54%) with a mean age of 56.8 ± 16.8 years. The majority of CTs were acquired using Siemens scanners (n = 2266, 75%) with slice thicknesses between 3.0 and 5.0 mm (n = 1077, 36%) (Table 1). There was no significant difference in patient demographics or CT acquisition parameters between the training-validation and test subsets ($p > 0.05$) (Table 1).

CNN performance

In the test subset (n = 452), the model achieved a DSC of 0.94 ± 0.05 and a HD of 1.70 ± 3.04 mm. Figure 2 shows the comparison of ground truth and the AI model-predicted pancreas segmentation, depicting a high level of agreement.

There was no significant difference between the model-predicted pancreas volumes (75.8 ± 23.6 cc) and the reference ground truth volumes (78.2 ± 23.7 cc) ($p = 0.15$). The mean volume difference between the reference and model-predicted segmentations was 2.41 cc, with limits of agreement ranging from 16.3 to −11.5 cc, which aligns with the known inter-individual pancreas volume variability documented in large-scale studies[38]. This range is also well below clinically significant thresholds, as pancreas volume reductions exceeding 15–20 cc have been correlated with metabolic dysfunction, exocrine insufficiency, and increased risk for diabetes progression[39]. Additionally, pancreas volume alterations associated with bariatric surgery-induced diabetes remission typically range from 6 to 12 cc, which reinforces that segmentation errors within our reported limits are unlikely to have a clinically meaningful impact. Moreover, a strong correlation was observed between the reference and model-predicted pancreas volumes (CCC = 0.95). Figure 3 presents a Bland–Altman analysis alongside a correlation plot, demonstrating strong volume agreement between the ground truth and AI-predicted segmentations.

Finally, the model's performance was consistent across different CT acquisition characteristics and patient gender ($p = 0.06$–0.72) (Table 2).

Multi-institutional public dataset

On the international multi-institutional AbdomenCT-1K public dataset (n = 585), the model achieved a DSC of 0.96 ± 0.04 and a HD of 2.2 ± 8.8 mm. There was no significant difference between the model-predicted pancreas volumes (77.7 ± 24.1 cc) and the reference ground truth volumes (79.6 ± 24.4 cc) ($p = 0.22$). The mean volume difference was 1.85 cc, with limits of agreement ranging from 9.8 to −6.2 cc. A strong correlation was observed between the pancreas volumes (CCC = 0.98). Figure 4 presents a Bland–Altman analysis alongside a correlation plot, illustrating strong volume agreement between the ground truth and AI-predicted segmentations.

NIH-PCT dataset

On the NIH-PCT dataset, the model achieved a DSC of 0.87 ± 0.04 and an HD of 4.3 ± 2.9 mm when compared against the original segmentations ($p < 0.05$). Notably, the original pancreas segmentations were initially generated by a medical student and later reviewed by a radiologist. As previously reported in other studies[40], this segmentation workflow may introduce delineation inaccuracies, particularly along subtle anatomical boundaries,

	Training/validation subset (n = 2579)	Test subset (n = 452)
Age (mean ± SD**) (years)	56.7 ± 16.9	57.4 ± 16.1
Sex (male:females)	0.84:1	0.87:1
Slice thickness (mm)*		
0.5–1.25	759	130
1.25–3.0	895	170
3.0–5.0	925	152
CT vendor		
Siemens	1943	323
GE	155	36
Toshiba	469	92
Philips	11	1
Canon	1	0

Table 1. comparison of training/validation and test subsets* *There was no significant difference between training-validation and test subsets for any of the categories ($p > 0.05$). **SD: Standard deviation.
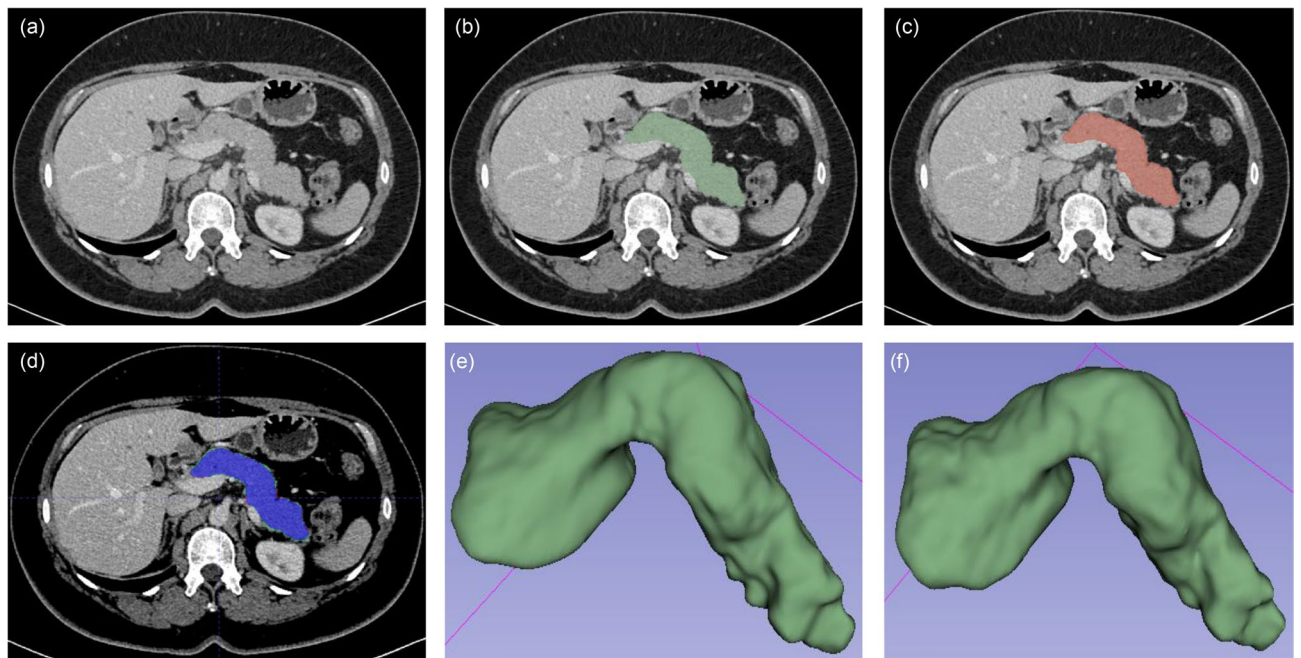
Fig. 2. Comparison of pancreas segmentations by radiologists and AI model: (**a**) Portal venous phase axial slice of a patient depicting the pancreas, (**b**, **c**) Radiologists' and model-predicted segmentation respectively, (**d**) Overlaid segmentation where the overlap, over- and under-segmentation are represented by blue, red and green, respectively. The Dice similarity coefficient (DSC) for this case was 0.97, indicating a high level of agreement. The model's segmentation demonstrates excellent fidelity despite the pancreas's lobulated contours and curvature. Three-dimensional views of the radiologists' and model-predicted segmentations are presented in panels (**e**) and (**f**), respectively.
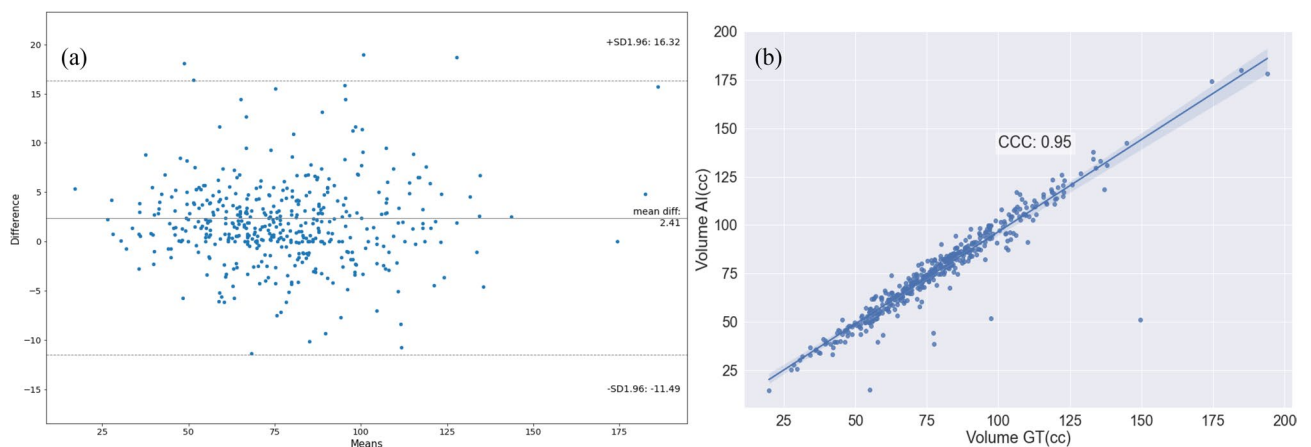
Fig. 3. Comparison of radiologists and AI model-predicted pancreas volumes for internal test subset: (**a**) Bland–Altman analyses show that the mean pancreas volume difference between reference and model-predicted segmentation 2.41 cc (limits of agreement 1.96 SD of 16.3 cc to − 11.5 cc). (**b**) The plot shows high correlation between reference and model-predicted volumes with the concordance correlation coefficient (CCC) of 0.95.

when compared to expert-driven, protocol-adherent segmentation, directly performed by fellowship-trained expert radiologists, as was done in our study. In fact, in several cases, our model produced segmentations that were qualitatively more accurate than the provided ground truth, as illustrated in Fig. 5.

Ablation study on individual model performance

To evaluate the contribution of individual models to the ensemble framework, we assessed the performance of the best-performing model from the fivefold training process. This model achieved DSCs of 0.94 ± 0.06 and 0.96 ± 0.05, and HDs of 1.76 ± 3.00 mm and 2.22 ± 8.7 mm, on the intramural and AbdomenCT-1K datasets,

	Mean ± SD* Dice similarity coefficient (DSC)	Mean ± SD* Hausdorff distance [HD (mm)])
Slice thickness (mm)		
0.5–1.25	0.94 ± 0.05	2.1 ± 1.5
1.25–3.0	0.94 ± 0.06	1.8 ± 4.7
3.0–5.0	0.95 ± 0.04	1.2 ± 0.9
CT vendor		
Siemens	0.94 ± 0.05	1.7 ± 3.5
GE	0.95 ± 0.04	1.6 ± 1.4
Toshiba	0.94 ± 0.06	1.7 ± 1.5
Philips	0.93 ± 0.00	3.4 ± 0.0
Gender		
Male	0.94 ± 0.05	1.7 ± 1.3
Female	0.95 ± 0.05	1.7 ± 3.9

Table 2. Model's stratified performance on test subset (n = 452)[†] [†]Model's performance was comparable for CT slice thicknesses, vendors and patients' gender ($p > 0.05$). *SD: Standard deviation.
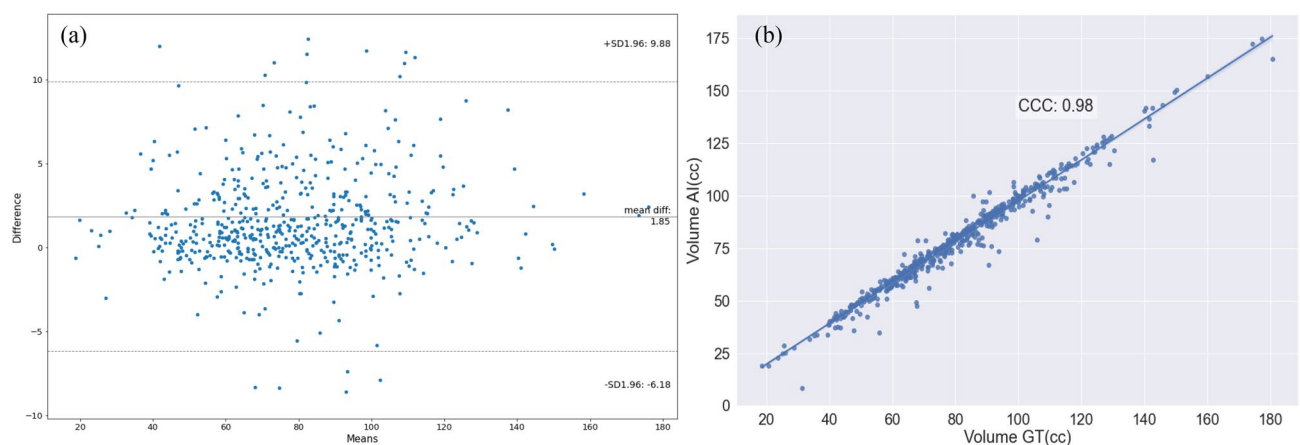
Fig. 4. Comparison of radiologists and AI model-predicted pancreas volumes for abdomen CT-1 K multi-institutional public dataset: (**a**) Bland–Altman analyses show that the mean pancreas volume difference between reference and model-predicted segmentation 1.85 cc (limits of agreement 1.96 SD of 9.8 cc to − 6.2 cc). (**b**) The plot shows high correlation between reference and model-predicted volumes with the concordance correlation coefficient (CCC) of 0.98.
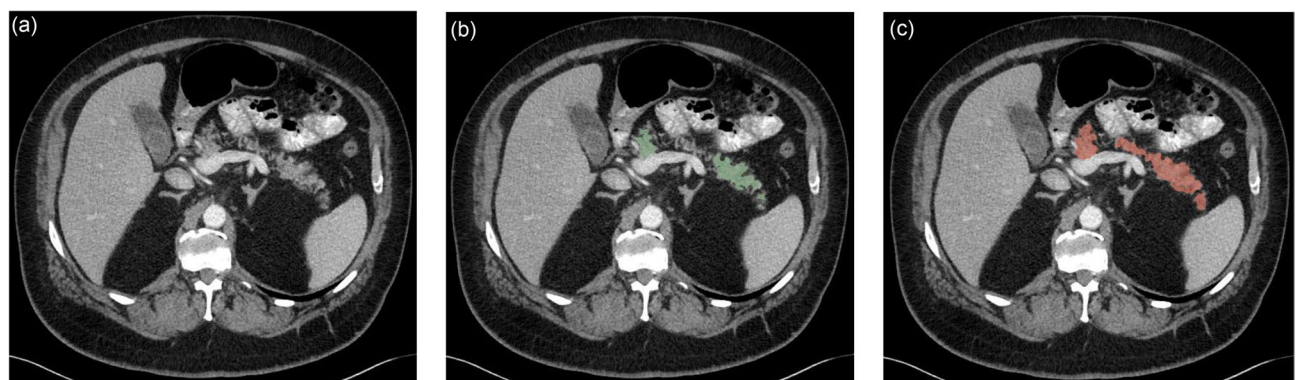

Fig. 5. Representative case highlighting segmentation discrepancies on a public dataset: (**a**) Axial portal venous phase CT demonstrating a diffusely fatty pancreas. (**b**) Ground truth segmentation from the NIH-PCT dataset, showing marked under-segmentation of the pancreatic parenchyma. (**c**) Segmentation predicted by our AI model, demonstrating more accurate and anatomically complete delineation of the pancreas.
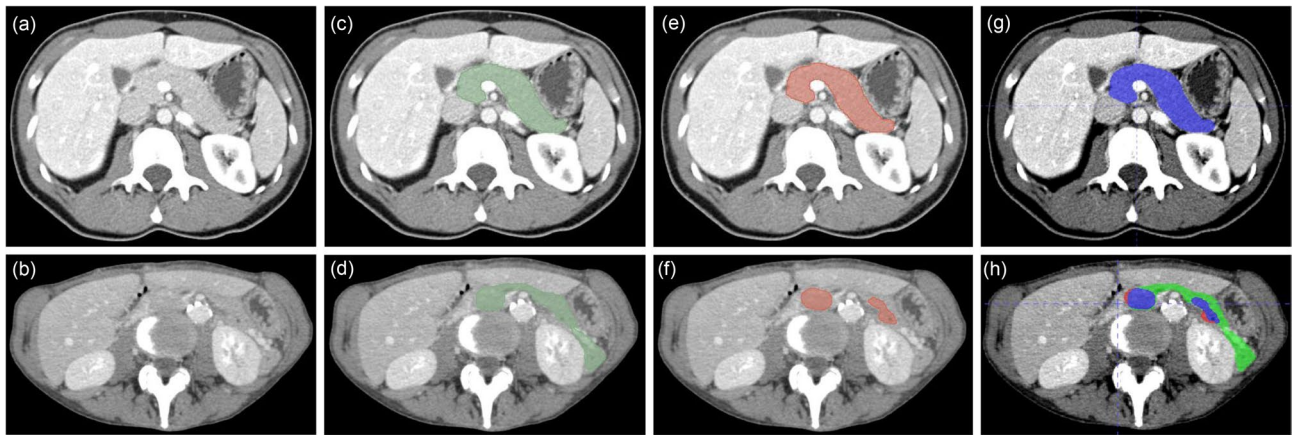
Fig. 6. AI-predicted pancreas segmentations in representative high- and low-performing cases: (**a**, **b**) Axial portal venous phase CT slices from two patients, corresponding to a high-performing case (**a**) and a low-performing case (**b**). (**c**, **d**) Ground truth segmentations provided by expert radiologists. (**e**, **f**) AI-generated segmentations. (**g**, **h**) Overlay maps highlighting agreement (blue), over-segmentation (red), and under-segmentation (green) between ground truth and AI predictions. The first case demonstrated excellent concordance, with a Dice Similarity Coefficient (DSC) of 0.99. The second case exhibited pronounced segmentation error (DSC = 0.36), associated with substantial paucity of peripancreatic fat and a diffusely atrophic pancreas.

respectively. These results were statistically comparable to the ensemble model ($p = 0.58$–0.67), supporting the robustness of individual fold models and confirming that the performance gain from ensembling was modest but consistent.

Analyses of cases in the lower range of model performance

In the test subset (n = 452), 23 model-predicted segmentations had a DSC below the 5th percentile (DSC < 0.89) (Fig. 6). Most of these cases (n = 14; 61%) exhibited a combination of minor over- and under-segmentation. In the AbdomenCT-1K dataset (n = 585), 30 model-predicted segmentations had a DSC below the 5th percentile (DSC < 0.91). Under-segmentation was more common (n = 18; 60%), followed by cases with both over- and under-segmentation (n = 11; 37%). Overall, common areas of over-segmentation included the duodenum at the pancreas head–duodenum interface and portions of the splenic vessel, jejunum, or stomach near the pancreatic tail. Common areas of under-segmentation were the portion of the pancreas head at its interface with the duodenum, or the terminal-most portion of the tail. While precise categorization was limited by the inherent subjectivity of interpretation, certain recurring patterns emerged as the potential sources of segmentation inaccuracies. These included diffuse fatty atrophy or focal glandular thinning (particularly of the pancreatic head or uncinate process), overlapping anatomy (e.g., bowel, duodenum, or vasculature), and poorly developed peripancreatic fat planes, all of which contribute to ambiguous organ boundaries and reduced visual separability from adjacent structures. These observations reflect the real-world complexity of pancreas morphology and imaging variability, emphasizing the need for future refinements that specifically address edge cases and outlier anatomy.

Comparison with a pre-trained public DL model

On the test subsets of the intramural and AbdomenCT-1K datasets, TotalSegmentator achieved DSCs of 0.89 ± 0.05 and 0.89 ± 0.04, respectively. These were significantly lower than the DSCs achieved by our ensemble model (0.94 ± 0.05 and 0.96 ± 0.04, respectively; $p < 0.001$). Similarly, the Hausdorff Distances (HDs) for TotalSegmentator were 2.8 ± 4.4 mm and 3.1 ± 5.4 mm, compared to 1.70 ± 3.04 mm and 2.2 ± 8.8 mm for our model on the intramural and AbdomenCT-1K datasets, respectively ($p < 0.001$). These results demonstrate the superior segmentation accuracy and boundary localization performance of our model across both internal and external datasets.

Discussion

We developed and rigorously validated a fully automated 3D AI model for volumetric pancreas segmentation on standard-of-care CT scans, leveraging the largest and most diverse dataset reported in the literature. This dataset encompassed a broad spectrum of patient demographics, scanner types, and imaging parameters to represent varied clinical settings. In the intramural test subset, our model achieved a mean Dice Similarity Coefficient (DSC) of 0.94 (SD 0.05) and demonstrated high spatial concordance with radiologist-segmented volumes, as evidenced by a concordance correlation coefficient (CCC) of 0.95. Additionally, the model attained a mean Hausdorff Distance (HD) of 1.70 ± 3.04 mm, which indicates strong boundary agreement between AI-predicted and ground truth segmentations. HD provides a crucial complementary metric by quantifying the maximum segmentation error, which is particularly relevant in downstream applications requiring precise

anatomical delineation such as radiomics feature extraction or longitudinal organ volumetry. To further assess its robustness, we validated the model on the international multi-institutional AbdomenCT-1K dataset, where it continued to exhibit superior performance with a mean DSC of 0.96 (SD 0.04), a CCC of 0.98, and an HD of 2.2 ± 8.8 mm. These results underscore the model's ability to provide accurate and reliable pancreas segmentation across heterogeneous imaging environments. In alignment with open science initiatives, we will make the AbdomenCT-1K dataset segmentation outputs publicly available to enable researchers to build upon our work and advance the field of AI-driven pancreatic imaging. This commitment to data sharing is aimed at fostering innovation, facilitating independent validation, and catalyzing new research avenues in pancreatic disease detection and radiomics-based biomarker discovery.

Our decision to focus on morphologically normal pancreas stems from two distinct yet complementary aspects. First, conventional imaging cannot reliably detect PDA at its pre-diagnostic and asymptomatic stage (stage 0). This is because the pancreas appears normal at this pre-diagnostic stage and lacks conspicuous imaging features, even in high-risk individuals, making detection through conventional imaging nearly impossible. This paradigm undermines risk stratification efforts and limits opportunities for early therapeutic intervention as detection at this stage is goal of emerging screening strategies[9–11]. While subtle imaging findings such as pancreatic duct dilation may be present, they are often nonspecific, leading to high false-positive rates and inter-reader variability among radiologists[12,41]. Given the rapid progression of PDA from subclinical disease to metastatic cancer[11], these limitations significantly restrict opportunities for early intervention. To address these diagnostic challenges, radiomic features extracted from normal-appearing pancreas on pre-diagnostic CTs have demonstrated promise in predicting subsequent PDA development[12–16]. However, the clinical translation of these biomarkers depends on the ability to achieve precise, reproducible, and scalable volumetric pancreas segmentation. By leveraging automated segmentation of normal-appearing pancreases, our model provides a scalable solution for extraction of imaging biomarkers that may enhance risk stratification and enable earlier intervention, ultimately addressing the critical limitations of conventional imaging.

Second, beyond oncology, morphometric and radiomic metrics from normal-appearing pancreas have demonstrated utility as imaging biomarkers for various non-neoplastic conditions such as diabetes, obesity, metabolic syndrome, and nonalcoholic fatty liver and pancreas diseases[17–19,42]. Similarly, CT radiomic features extracted from normal pancreas have shown discriminative ability to distinguish between patients with type 2 diabetes and age-matched controls, offering potential for opportunistic detection of undiagnosed diabetes or identifying at-risk individuals[20]. The subtle yet measurable changes in a normal-appearing pancreas that drive these predictive analytics are often imperceptible to human observation. Consequently, the clinical necessity for automated tools for volumetric segmentation of normal-appearing pancreas on cross-sectional imaging is paramount.

The model developed in this study outperformed previously reported AI-based pancreas segmentation studies. Prior studies have reported mean DSCs ranging from 0.65 to 0.91[25,26], whereas our model, trained on a significantly larger and more diverse dataset, achieved a mean DSC of 0.94 on the intramural test set and 0.96 on an international multi-institutional public dataset. Furthermore, prior studies have largely focused on anatomical delineation and, therefore, primarily report segmentation accuracy metrics (DSC, HD) without considering the broader clinical implications for early disease detection and biomarker discovery. Conversely, we incorporate clinically relevant evaluations, including Bland–Altman analysis and volume agreement thresholds, to ensure that segmentation errors remain within acceptable clinical limits. Bland–Altman analysis demonstrated limits of agreement between 16.3 and −11.5 cc, which align with the well-documented inter-individual pancreas volume variability[38]. While these limits of agreement (−11.5 cc to +16.3 cc) marginally exceed the pancreas volume alterations reported in some disease states (e.g., 6–12 cc changes following bariatric surgery-induced diabetes remission)[22], it is important to note that these bounds represent 1.96 standard deviations and not typical errors. The average volume difference in our study was only 2.41 cc, indicating high precision in most cases. Given the natural physiological variability in pancreas volume, which can fluctuate by 10–15% due to age, metabolic status, and incidental findings, our model's volume estimations are well within the expected range of biological variation. Importantly, our model is intended to support consistent and reproducible segmentation for downstream tasks—such as radiomics or biomarker development—rather than serving as a standalone diagnostic tool based on absolute volume cutoffs. Hence, the observed limits of agreement remain acceptable and within clinically tolerable boundaries for the intended use of the segmentation model.

Furthermore, previous segmentation pipelines have relied on multi-stage, coarse-to-fine segmentation approaches[26]. However, these methods often require manual intervention or intricate pre-processing, which often introduce additional sources of error. Our study sought to overcome these inefficiencies by employing a single-stage, end-to-end automated 3D nnU-Net model to achieve high performance while maintaining computational efficiency. The enhanced performance and generalizability of this model can be attributed to the robustness of the training dataset and the methodological refinements incorporated into the development pipeline. Our dataset consisted of high-quality, expert-annotated segmentations that spanned a diverse patient population, including variations in scanner types, slice thicknesses, and anatomical presentations. The importance of dataset diversity in improving AI model generalizability has been well-documented in prior studies[26]. Additionally, we leveraged a dynamic augmentation strategy that introduced spatial perturbations, intensity variations, and contrast adaptations to enhance model resilience to real-world variations. Furthermore, for cases where the model exhibited lower performance, a detailed analysis was conducted using the 5th percentile DSC threshold. Segmentation errors were primarily observed at the interfaces between the pancreatic head and the duodenum, as well as between the pancreatic tail and adjacent structures like the jejunum and splenic vessels. These regions are known to be challenging for segmentation, even for expert radiologists, due to their complex anatomy and the subtlety of tissue boundaries. The fact that these specific areas posed difficulties for the model provides

valuable insight into potential areas for further refinement, either through enhanced model training or post-processing techniques.

We chose to develop the volumetric pancreas segmentation model on standard-of-care portal venous phase CT scans due to several key advantages over MRI. While MRI offers superior soft tissue contrast, it suffers from prolonged acquisition times, higher costs, susceptibility to motion artifacts, and the requirement for specialized protocols that limit its widespread use in large-scale screening initiatives[43–45]. Conversely, CT is widely accessible, cost-effective, and already embedded in routine clinical workflows, making it the most practical imaging modality for AI-driven segmentation models[46,47]. Moreover, emerging evidence supports the cost-effectiveness of CT for pancreatic cancer screening in individuals at high-risk for sporadic PDA such as those with new-onset diabetes[48]. Additionally, CT provides isotropic volumetric data with high spatial resolution, which facilitates automated pancreas segmentation with greater precision[47]. The portal venous phase, the most commonly acquired phase in abdominal CT, provides optimal tissue contrast between the pancreas and adjacent structures, making it ideal for accurate and reproducible longitudinal volumetric measurements. Given these advantages, CT remains the optimal imaging modality for developing AI-based pancreas segmentation models, ensuring scalability for population-based applications and enabling the extraction of imaging biomarkers critical for early disease detection and risk stratification.

Our study has limitations. The retrospective nature of this study, while essential for maximizing the sample size required for comprehensive model training, validation, and testing, inherently presents the potential for selection bias. Secondly, the model was developed and validated specifically for the segmentation of morphologically normal-appearing pancreas and does not address advanced stages of diseased pancreas (e.g. chronic pancreatitis). However, this targeted approach was designed to establish a baseline for biomarker identification in individuals without evident pancreatic disease. This strategy aligns with the goal of detecting subtle or visually occult changes that may indicate early disease processes. Another limitation is that the majority of CT scans in our dataset were acquired from a single scanner vendor, which may introduce a potential bias in model generalization to scans from other manufacturers. However, despite this skew, the model demonstrated robust performance across the independent multi-institutional AbdomenCT-1K dataset, which includes CT scans from multiple vendors and imaging protocols. This strong generalizability suggests that the model effectively learned vendor-agnostic pancreas features, reinforcing its potential applicability across diverse clinical settings. Future studies incorporating a more balanced representation of scanner manufacturers could further enhance the model's adaptability.

In conclusion, we developed and rigorously validated a fully automated 3D CNN for volumetric pancreas segmentation using the largest reported CT dataset. The model achieved state-of-the-art segmentation accuracy, demonstrating strong concordance with radiologist-annotated ground truth, as evidenced by high DSC and CCC values. Furthermore, its robust performance across diverse patient demographics, scanner types, and imaging protocols underscores its generalizability. External validation on an international multi-institutional dataset further confirmed its reliability in heterogeneous real-world imaging conditions. By enabling precise and reproducible volumetric pancreas segmentation, this model provides a critical foundation for advancing imaging-based biomarker discovery, facilitating early disease detection, and broadening the scope of quantitative pancreas imaging applications across both endocrine and exocrine diseases.

Data availability

The datasets generated and analyzed during this study are not publicly available due to institutional policies and patient privacy regulations but may be made available from the corresponding author upon reasonable request and subject to approval by the institutional review board. The ground truth pancreas segmentations derived from the AbdomenCT-1K dataset will be publicly shared to facilitate further research and reproducibility efforts. All methodological details, including data preprocessing, model training, and evaluation procedures, are comprehensively described in the Methods section, allowing independent replication of the study findings using non-proprietary tools and publicly available datasets.

References

1. Five-Year Pancreatic Cancer Survival Rate Increases to 12% https://pancan.org/news/five-year-pancreatic-cancer-survival-rate-increases-to-12/.
2. Siegel, R. L., Giaquinto, A. N. & Jemal, A. Cancer statistics, 2024. *CA Cancer J. Clin.* **74**, 12–49. https://doi.org/10.3322/caac.21820 (2024).
3. Holly, E. A., Chaliha, I., Bracci, P. M. & Gautam, M. Signs and symptoms of pancreatic cancer: A population-based case-control study in the San Francisco Bay area. *Clin. Gastroenterol. Hepatol.* **2**, 510–517. https://doi.org/10.1016/s1542-3565(04)00171-5 (2004).
4. Singhi, A. D., Koay, E. J., Chari, S. T. & Maitra, A. Early detection of pancreatic cancer: Opportunities and challenges. *Gastroenterology* **156**, 2024–2040. https://doi.org/10.1053/j.gastro.2019.01.259 (2019).
5. Kenner, B. et al. Artificial intelligence and early detection of pancreatic cancer: 2020 summative review. *Pancreas* **50**, 251–279. https://doi.org/10.1097/MPA.0000000000001762 (2021).
6. Morani, A. C. et al. Hereditary and sporadic pancreatic ductal adenocarcinoma: Current update on genetics and imaging. *Radiol. Imaging Cancer* **2**, e190020. https://doi.org/10.1148/rycan.2020190020 (2020).
7. Chari, S. T. et al. Probability of pancreatic cancer following diabetes: A population-based study. *Gastroenterology* **129**, 504–511. https://doi.org/10.1016/j.gastro.2005.05.007 (2005).
8. Sharma, A. et al. Model to determine risk of pancreatic cancer in patients with new-onset diabetes. *Gastroenterology* **155**, 730–739. https://doi.org/10.1053/j.gastro.2018.05.023 (2018).
9. Overbeek, K. A. et al. Timeline of development of pancreatic cancer and implications for successful early detection in high-risk individuals. *Gastroenterology* **162**, 772–785. https://doi.org/10.1053/j.gastro.2021.10.014 (2022).
10. Chhoda, A. et al. Late-stage pancreatic cancer detected during high-risk individual surveillance: A systematic review and meta-analysis. *Gastroenterology* **162**, 786–798. https://doi.org/10.1053/j.gastro.2021.11.021 (2022).

11. Yu, J., Blackford, A. L., Dal Molin, M., Wolfgang, C. L. & Goggins, M. Time to progression of pancreatic ductal adenocarcinoma from low-to-high tumour stages. *Gut* **64**, 1783–1789. https://doi.org/10.1136/gutjnl-2014-308653 (2015).

12. Mukherjee, S. et al. Radiomics-based machine-learning models can detect pancreatic cancer on prediagnostic computed tomography scans at a substantial lead time before clinical diagnosis. *Gastroenterology* **163**, 1435–1446. https://doi.org/10.1053/j.gastro.2022.06.066 (2022).

13. Mukherjee, S. et al. 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. *Abdom. Radiol. (NY)* **49**, 964–974. https://doi.org/10.1007/s00261-023-04127-1 (2024).

14. Chen, P. T. et al. Radiomic features at CT can distinguish pancreatic cancer from noncancerous pancreas. *Radiol. Imaging Cancer* **3**, e210010. https://doi.org/10.1148/rycan.2021210010 (2021).

15. Qureshi, T. A. et al. Predicting pancreatic ductal adenocarcinoma using artificial intelligence analysis of pre-diagnostic computed tomography images. *Cancer Biomark.* **33**, 211–217. https://doi.org/10.3233/CBM-210273 (2022).

16. Javed, S. et al. Risk prediction of pancreatic cancer using AI analysis of pancreatic subregions in computed tomography images. *Front. Oncol.* **12**, 1007990. https://doi.org/10.3389/fonc.2022.1007990 (2022).

17. DeSouza, S. V. et al. Pancreas volume in health and disease: A systematic review and meta-analysis. *Expert. Rev. Gastroenterol. Hepatol.* **12**, 757–766. https://doi.org/10.1080/17474124.2018.1496015 (2018).

18. Lim, S. et al. Differences in pancreatic volume, fat content, and fat density measured by multidetector-row computed tomography according to the duration of diabetes. *Acta Diabetol.* **51**, 739–748. https://doi.org/10.1007/s00592-014-0581-3 (2014).

19. Lohr, J. M., Panic, N., Vujasinovic, M. & Verbeke, C. S. The ageing pancreas: A systematic review of the evidence and analysis of the consequences. *J. Intern. Med.* **283**, 446–460. https://doi.org/10.1111/joim.12745 (2018).

20. Wright, D. E. et al. 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**, 3806–3816. https://doi.org/10.1007/s00261-022-03668-1 (2022).

21. Lautenbach, A. et al. Adaptive changes in pancreas post Roux-en-Y gastric bypass induced weight loss. *Diabetes Metab. Res. Rev.* **34**, e3025. https://doi.org/10.1002/dmrr.3025 (2018).

22. Al-Mrabeh, A. et al. 2-year remission of type 2 diabetes and pancreas morphology: A post-hoc analysis of the DiRECT open-label, cluster-randomised trial. *Lancet Diabetes Endocrinol.* **8**, 939–948. https://doi.org/10.1016/S2213-8587(20)30303-X (2020).

23. Shinagare, A. B., Steele, E., Braschi-Amirfarzan, M., Tirumani, S. H. & Ramaiya, N. H. Sunitinib-associated pancreatic atrophy in patients with gastrointestinal stromal tumor: A toxicity with prognostic implications detected at imaging. *Radiology* **281**, 140–149. https://doi.org/10.1148/radiol.2016152547 (2016).

24. Zhou, Y. et al. A Fixed-Point Model for pancreas segmentation in abdominal CT Scans. 693–701 (2017). https://arxiv.org/abs/1612.08230.

25. Moglia, A., Cavicchioli, M., Mainardi, L. & Cerveri, P. Deep Learning for Pancreas Segmentation: a Systematic Review. https://arxiv.org/abs/2407.16313. (2024).

26. Panda, A. 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.* **48**, 2468–2481. https://doi.org/10.1002/mp.14782 (2021).

27. Panda, A. 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. *AJR Am. J. Roentgenol.* **217**, 730–740. https://doi.org/10.2214/AJR.20.24567 (2021).

28. MIRC CTP http://mircwiki.rsna.org/index.php?title=CTP-The_RSNA_Clinical_Trial_Processor.

29. Khasawneh, H. et al. 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**, 841–847. https://doi.org/10.1097/RCT.0000000000001374 (2022).

30. Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J. & Maier-Hein, K. H. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **18**, 203–211. https://doi.org/10.1038/s41592-020-01008-z (2021).

31. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation, 234–241 (2015).https://arxiv.org/abs/1505.04597.

32. Wasserthal, J. et al. TotalSegmentator: Robust segmentation of 104 anatomic structures in CT images. *Radiol. Artif. Intell.* **5**, e230024. https://doi.org/10.1148/ryai.230024 (2023).

33. Ma, J. et al. AbdomenCT-1K: Is abdominal organ segmentation a solved problem?. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 6695–6714. https://doi.org/10.1109/TPAMI.2021.3100536 (2022).

34. Bilic, P. et al. The liver tumor segmentation benchmark (LiTS). *Med. Image Anal.* **84**, 102680. https://doi.org/10.1016/j.media.2022.102680 (2023).

35. Heller, N. et al. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: Results of the KiTS19 challenge. *Med. Image Anal.* **67**, 101821. https://doi.org/10.1016/j.media.2020.101821 (2021).

36. Antonelli, M. et al. The Medical Segmentation Decathlon. *Nat. Commun.* **13**, 4128. https://doi.org/10.1038/s41467-022-30695-9 (2022).

37. Clark, K. et al. The cancer imaging archive (TCIA): Maintaining and operating a public information repository. *J. Digit Imaging* **26**, 1045–1057. https://doi.org/10.1007/s10278-013-9622-7 (2013).

38. Li, L. et al. Normal pancreatic volume assessment using abdominal computed tomography volumetry. *Medicine* **100**, e27096. https://doi.org/10.1097/MD.0000000000027096 (2021).

39. Lu, J. et al. Association between pancreatic atrophy and loss of insulin secretory capacity in patients with Type 2 diabetes mellitus. *J. Diabetes Res.* **2019**, 6371231. https://doi.org/10.1155/2019/6371231 (2019).

40. Suman, G. et al. Quality gaps in public pancreas imaging datasets: Implications & challenges for AI applications. *Pancreatology* **21**, 1001–1008. https://doi.org/10.1016/j.pan.2021.03.016 (2021).

41. Toshima, F. 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. *AJR Am. J. Roentgenol.* https://doi.org/10.2214/AJR.21.26014 (2021).

42. Petrov, M. S. Harnessing analytic morphomics for early detection of pancreatic cancer. *Pancreas* **47**, 1051–1054. https://doi.org/10.1097/MPA.0000000000001155 (2018).

43. Antony, A. et al. AI-Driven insights in pancreatic cancer imaging: From pre-diagnostic detection to prognostication. *Abdom. Radiol. (NY)* https://doi.org/10.1007/s00261-024-04775-x (2024).

44. Huang, C. et al. Advancements in early detection of pancreatic cancer: The role of artificial intelligence and novel imaging techniques. *Abdom. Radiol. (NY)* https://doi.org/10.1007/s00261-024-04644-7 (2024).

45. Huang, C. et al. Imaging for early detection of pancreatic ductal adenocarcinoma: Updates and challenges in the implementation of screening and surveillance programs. *AJR Am. J. Roentgenol.* **223**, e2431151. https://doi.org/10.2214/AJR.24.31151 (2024).

46. Inoue, A. et al. Diagnostic performance in low- and high-contrast tasks of an image-based denoising algorithm applied to radiation dose-reduced multiphase abdominal CT examinations. *AJR Am. J. Roentgenol.* **220**, 73–85. https://doi.org/10.2214/AJR.22.27806 (2023).

47. Almeida, R. R. et al. Advances in pancreatic CT imaging. *AJR Am. J. Roentgenol.* **211**, 52–66. https://doi.org/10.2214/AJR.17.18665 (2018).

48. Schwartz, N. R. M. et al. Potential cost-effectiveness of risk-based pancreatic cancer screening in patients with new-onset diabetes. *J. Natl. Compr. Canc. Netw.* https://doi.org/10.6004/jnccn.2020.7798 (2021).

Author contributions

SM developed the software, performed statistical analyses, and generated figures. SM and AG structured and drafted much of the manuscript. AA, NP, KT, and AK curated the dataset and performed pancreas segmentations. AA, MN, and MM conducted failure mode analyses. AG provided manuscript revisions, finalization, and overall leadership and oversight. All authors contributed to the study design, reviewed the manuscript, and approved the final version for submission.

Funding

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to A.H.G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.