ScaleMAI: Accelerating the Development of Trusted Datasets and AI Models

Wenxuan Li[1] Pedro R. A. S. Bassi[1,2,3] Tianyu Lin[1] Yu-Cheng Chou[1] Xinze Zhou[1]
Yucheng Tang[4] Fabian Isensee[5] Kang Wang[6] Qi Chen[1,7] Xiaowei Xu[8] Xiaoxi Chen[9]
Lizhou Wu[10] Qilong Wu[11] Yannick Kirchhoff[5] Maximilian Rokuss[5] Saikat Roy[5]
Yuxuan Zhao[12] Dexin Yu[12] Kai Ding[13] Constantin Ulrich[5] Klaus Maier-Hein[5] Yang Yang[6]
Alan L. Yuille[1] Zongwei Zhou[1,*]

[1]Johns Hopkins University
[2]University of Bologna [3]Italian Institute of Technology [4]NVIDIA [5]DKFZ
[6]University of California, San Francisco [7]University of Chinese Academy of Sciences
[8]Guangdong Provincial People's Hospital [9]University of Illinois Urbana-Champaign
[10]The First Affiliated Hospital of Shandong First Medical University
[11]National University of Singapore [12]Qilu Hospital of Shandong University
[13]Johns Hopkins Medicine

Code, dataset, and models: https://github.com/MrGiovanni/ScaleMAI

Abstract

Building trusted datasets is critical for transparent and responsible Medical AI (MAI) research, but creating even small, high-quality datasets can take years of effort from multidisciplinary teams. This process often delays AI benefits, as human-centric data creation and AI-centric model development are treated as separate, sequential steps. To overcome this, we propose **ScaleMAI***, an agent of AI-integrated data curation and annotation, allowing data quality and AI performance to improve in a self-reinforcing cycle and reducing development time from years to months. We adopt pancreatic tumor detection as an example. First, ScaleMAI progressively creates a dataset of 25,362 CT scans, including per-voxel annotations for benign/malignant tumors and 24 anatomical structures. Second, through progressive human-in-the-loop iterations, ScaleMAI provides Flagship AI Model that can approach the proficiency of expert annotators (30-year experience) in detecting pancreatic tumors. Flagship Model significantly outperforms models developed from smaller, fixed-quality datasets, with substantial gains in tumor detection (**+14%**), segmentation (**+5%**), and classification (**72%**) on three prestigious benchmarks. In summary, ScaleMAI transforms the speed, scale, and reliability of medical dataset creation, paving the way for a variety of impactful, data-driven applications.*

1. Introduction

The pursuit of trusted datasets is critical for developing AI models in medical imaging research [49, 50, 70]. However, data curation and annotation are labor-intensive, requiring close collaboration between medical professionals and technical experts [12, 20, 68, 85]. Poor data practices can introduce problems [24] such as duplication, label noise, biases, and representational disparities arising from limited data sources. These problems compromise AI's robustness in real-world applications. We ask: *To what extent can we automate the arduous process of data curation and annotation for developing trustworthy AI models?*

We present **ScaleMAI**, an AI-integrated data curation and annotation agent that accelerates the development of *AI Trusted Dataset*[1] and *Flagship Model*[2] reducing the data creation time from years to months. Unlike previous methods [8, 11, 28, 39, 48], where human-centric data creation and AI-centric development are completely independent, we combine these two endeavors. It addresses the challenge that as datasets grow exponentially and annotated classes diversify, fully curating and annotating the dataset beforehand becomes increasingly time-consuming. The development of

*Correspondence to Zongwei Zhou (ZZHOU82@JH.EDU)

[1] *AI Trusted Dataset* refers to a large-scale, high-quality, and multi-source dataset that reflect real-world clinical scenarios. The annotation quality of this dataset should match that of expert radiologists (Figure 2).

[2] *Flagship Model* is an AI model optimized alongside the data curation and annotation process. In later iterations, Flagship Model's annotation quality is expected to match or even exceed that of expert annotators and can be reliably applicable to out-of-distribution CT scans (Table 3).

ScaleMAI agent follows three unique **insights**.

- *AI can retrieve patient scans suitable for clinical needs.* By applying large language models (e.g., Llama 3 [22]) to extract information such as pathology findings, contrast enhancement, and patient demographics from radiology reports, we reduced retrieval time per scan from 15 minutes (using keyword search and human review) to 5 seconds, and improved the precision of retrieving CT studies from 89% to 96% (Figure 4).
- *When an AI model struggles to fit a specific data point in the training set, it often signals a labeling error.* This assumes that most labels are correct and errors are infrequent and non-systematic; fitting these errors would increase the overall loss by negatively impacting performance on correct labels. Based on this, we develop the 'test-on-training' approach (§2.2.1) to effectively detect and correct 36% of these labeling errors[3] (Table 1).
- *Criticizing label quality is easier than creating labels.* We developed Label Expert (see §2.2.2) that use large vision-language models (e.g., Qwen2-VL [83]) to identify the best label quality by pair-wise comparing multiple pseudo label candidates predicted by 19 pre-existing AI models [10], efficiently replacing about 50% of problematic labels with better labels (Table 1). Compared to manually checking labels one by one, which takes 5 min/scan, Label Expert reduces this to 5 sec/scan and can tirelessly compare 34.7 million[4] pairs of label candidates.

In this paper, ScaleMAI is applied to the clinical need of pancreatic tumor detection, staging, and planning, contributing public deliverables summarized as follows.

1. **A ScaleMAI Agent** (§2) that accelerates the transformation of specific clinical needs into trusted datasets and AI models. This is the first attempt that integrates large language models, vision-language models, and segmentation models to significantly reduce expert effort in curating and annotating very large medical datasets (25K). While this paper uses pancreatic tumor detection as a demonstration (§4), our ScaleMAI agent can be applicable to address a range of clinical needs such as tumor staging (§C.5) and radiotherapy planning (§C.6).
2. **An AI Trusted Dataset** (§3) comprising 25,362 CT scans with precise per-voxel annotations of benign and malignant pancreatic tumors, along with 24 surrounding structures. Sourced from 112 hospitals, this dataset includes imaging metadata such as patient sex, age, contrast phase, diagnosis, spacing, and scanner details. This dataset enables standard medical imaging tasks—detection, segmentation, and classification—and clinical tasks such as tumor staging and radiotherapy planning.
3. **A Flagship AI Model** (§4), developed through progressive human-in-the-loop iterations and trained with our proposed Data Mix and Data Annealing techniques, can achieve expert-level proficiency in pancreatic tumor detection. Flagship Model significantly outperforms models[5] trained on smaller, fixed-quality datasets, achieving notable gains in tumor detection (+14%), segmentation (+5%), and classification (72%) across three prestigious tumor benchmarks. Moreover, our Flagship Model extends its utility to tumor staging (T1–T4) and radiotherapy planning, where it can perform tumor and multi-organ segmentation on planning CT scans (distinct from the diagnostic CT scans used in training.

Related Work. Current AI development in medical imaging heavily relies on public datasets like TCIA-Pancreas [19] and MSD-Pancreas [8], where clinical needs are predefined by dataset creators. While simplifying the algorithm development, this approach introduces biases and limitations due to task-specific focus and poor data practices [24, 55, 56]. Common issues include low-quality images, noisy labels, inconsistent annotation standards, and partial labeling, all of which undermine AI performance in real-world applications. For instance, MSD-Pancreas focuses only on tumor segmentation, ignoring other key conditions like cysts and pancreatitis, which represent 30–40% of clinical cases [69]. Similarly, TCIA-Pancreas suffers from motion artifacts and inconsistent slice thickness in 20% of cases [75], causing variability. Additionally, datasets from a single hospital often lack diversity in demographics, diseases, and imaging protocols, leading to a 10–15% drop in accuracy when tested on external datasets [57]. Automated data curation is crucial for creating diverse, high-quality datasets, enhancing AI robustness and applicability in medical imaging.

2. ScaleMAI

2.1. Curation: Clinical Needs → Suitable Data

Clinical Needs. We begin by defining clinical requirements C_{clinical} with domain experts, focusing on pancreatic tumor detection, staging, and planning. These clinical objectives guide the translation of target tasks $T = \{T_1, T_2, \ldots, T_n\}$ and corresponding annotated classes, aligning data collection with real-world clinical objectives.

Data Retrieval. Large language models (LLMs) significantly expedite data retrieval as detailed in the §A.1. They reduce search time from approximately 15 minutes per CT scan to just 5 seconds and achieve 96% precision, improv-

[3] This approach is effective for large structures with typical shapes (e.g., liver, pancreas, spleen, aorta) but has limited utility for small, tubular, or complex structures like tumors, colon, or intestine.

[4] # of pairs=3 iterations×19 candidates×24 classes×25,362 CT scans

[5] *Flagship Models* continually enhance performance by increasing data size and quality until they eliminate simple errors (§2.2.2) and approach expert-level inter-annotator variance (§3.2). In contrast, *Conventional Models* rely on static datasets, limiting their performance to the initial data quality and size, with a primary focus on algorithmic optimization rather than dynamic data and label improvement.

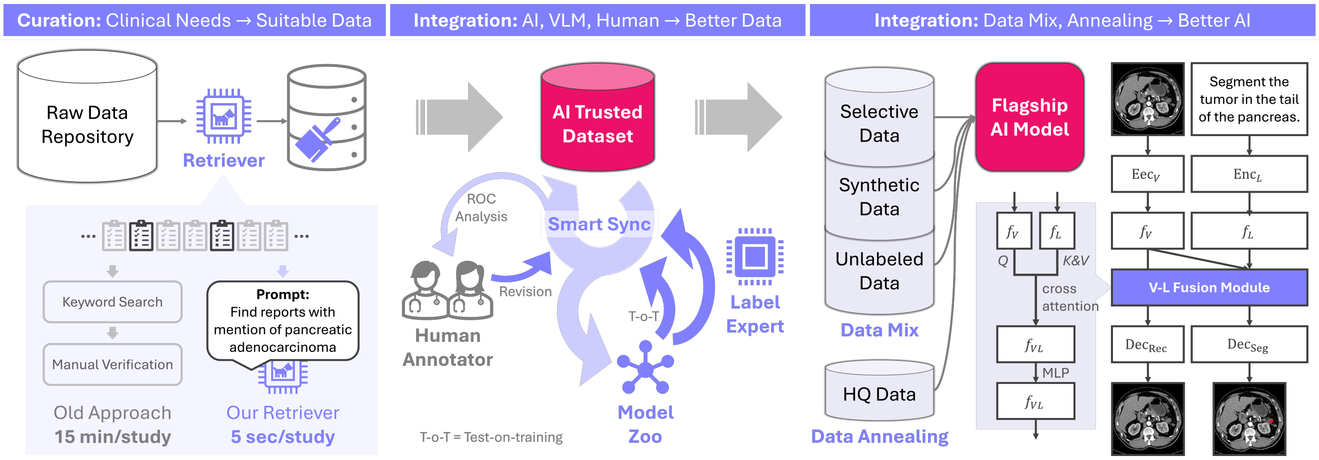


Figure 1. We have developed **ScaleMAI**, an AI agent that accelerates data curation and annotation, reducing development time from years to months. ScaleMAI produces an *AI Trusted Dataset* and a high-performing *Flagship AI Model* through an iterative, self-reinforcing cycle. Key innovations include: (1) Prompting large language models by medical knowledge to retrieve relevant data, perform de-identification, and eliminate duplicates. (2) Introducing a novel ROC analysis to prioritize data for annotation, minimizing manual effort. (3) Maintaining a Model Zoo of state-of-the-art segmentation model architectures (e.g., top-ranking models from Touchstone [10], MSD [7], KiTS [28], etc.) to generate pseudo labels and support data labeling. (4) Devising automated strategies like Test-on-Training (T-o-T) and Label Expert to identify and correct labeling errors, with updates seamlessly integrated into the dataset. (5) Identifying and categorizing data into selective, synthetic, and unlabeled subsets for optimized training. (6) Integrating a vision-language fusion module for learning new classes without full model retraining. (7) Refining models using high-quality, human-annotated data. This iterative cycle concludes when annotation accuracy matches or exceeds human standards, eliminating the need for further manual work.

ing upon the 89% rate obtained with keyword-based systems (Figure 4). For example, simple keyword searches may fail to detect subtle or indirect tumor indicators (e.g., hyperenhancing lesions, duct dilation, or the emergence of suspicious nodules in cystic lesions). LLM-driven retrieval, which we term *Retriever*, can handle these complex queries, including those over longitudinal data. This approach uncovers early, pre-diagnostic scans that radiologists might have initially overlooked, enabling the evaluation of whether AI can detect lesions missed by human experts.

Data Cleaning. We apply a de-identification process $f_{\text{deID}}$to remove patient identifiers and use a vision-language model f_{VLM} to identify corrupted scans. Three-dimensional perceptual hashing $h(x_i) = \text{Hash}(x_i)$ and Approximate Nearest Neighbor search [9] detect duplicates. Standardization in formatting and windowing produces the final, curated dataset $\mathcal{D}_{\text{curated}}$, ready for clinical analysis.

Cold Start of ScaleMAI. Initially, we define the target classes $C = \{C_1, C_2, \ldots, C_{24}\}$ and separate them into C_{public} (with classes already annotated in public datasets) and $C_{\text{new}} = C \setminus C_{\text{public}}$ (with classes requiring new annotations). For each class $C_i \in C_{\text{public}}$, we use 19 pre-existing models which we collected from Touchstone benchmark [10] to generate pseudo labels, creating D_{pseudo}. For each class in $C_j \in C_{\text{new}}$, we annotate a small number of scans (i.e., $N < 50$) and train a new model. We then combine D_{pseudo} and these newly annotated data to train a unified model that produces pseudo labels for all classes in C, forming our initial fully labeled dataset to construct the initial dataset D_0.

method	gallbladder	prostate	rectum	bladder	pancreas	kidney
T-o-T	48.8	52.8	49.4	49.6	14.0	24.6
Label Expert	44.4	46.0	48.1	48.6	74.2	50.1

Table 1. **75% of label errors can be detected and revised by Test-on-Training (T-o-T) and Label Expert.** A total of 51,454 annotations were revised iteratively. We report percentages for the proportion of errors in each class detected and revised by T-o-T and Label Expert. T-o-T is effective for classes that are often absent, while Label Expert excels with organs that have typical shapes. Results of applying T-o-T and Label Expert to revise more anatomical structures are presented in Appendix Table 6.

2.2. Integration: AI, VLM, Human → Better Data

2.2.1. Test-On-Training Detects/Revises 36% Errors

We propose a *Test-On-Training* (T-o-T) strategy to detect and revise annotation errors. By training another model on our dataset of 25,362 CT scans and testing it on the same training set, we can identify discrepancies between the model's predictions and the annotations. Good performance on the training set indicates consistent and accurate annotations, whereas poor performance may suggest noisy or ambiguous annotations that hinder the model's learning

(Table 1). If the Dice Similarity Coefficient (DSC) between the model's prediction and the so-called 'ground truth' in our dataset is zero—meaning that the 'ground truth' has at least one positive pixel but the AI predictions have no overlap—this indicates a potential labeling error. Visual inspection reveals that the current annotations are false positives. In such cases, the T-o-T strategy can also generate annotation candidates that are potentially better than the existing annotations. We replace the erroneous annotations with these new candidates, detecting and revising an average of 35.6% annotation errors for 14 classes (Table 6).

2.2.2. Label Expert Detects/Revises 39% Simple Errors

We propose *Label Expert*, a system that prompts a vision-language model (VLM) with anatomical knowledge to select better labels between two options, detailed in §A.2. The hypothesis is that a majority of label errors are obvious even to non-professionals and can be detected by VLMs trained on diverse and extensive image-text datasets, given their strong performance in various image understanding tasks [5, 41, 47, 54, 71]. Examples of such obvious errors include organ misplacement, abnormal shapes, disconnections, multiple predictions for a single organ, noise artifacts, and label inconsistencies due to poor CT quality. Since pre-existing VLMs are trained on 2D natural images, they cannot directly analyze 3D CT scans. To address this, we project 3D CT scans and labels into 2D images, using a front-view projection. These projections resemble 2D X-rays with overlaid labels in red. The VLM then evaluates these projections with prompts designed to guide its decision-making. We use aorta as an example. The prompt teaches the VLM that '*aorta should appear as a long vertical red line with a curve at the top.*' When comparing two labels, the VLM determines that label #1 matches the description better than label #2. We found that prompt design significantly impacts performance. With carefully designed prompts, the VLM achieved 98% accuracy in selecting the better label. By automating the detection of obvious label errors through 34.7 million pair-wise comparison, Label Expert significantly reduced human review/revision efforts and corrected 39% of annotation errors. In contrast, traditional error detection methods miss around 80% of such obvious label errors, despite being straightforward for humans to identify in under two seconds per scan.

2.2.3. Human-In-The-Loop Tumor/Organ Annotation

ROC Analysis for Pancreatic Tumor Annotation. Annotating per-voxel tumors is time-consuming. Our ROC analysis strategy biases AI predictions toward high sensitivity. Inevitably, this generates more false positives, but removing them is much faster and easier than creating annotations from scratch. False positives in non-tumor CT scans can be automatically excluded using radiology reports, and false positives in tumor CT scans can be erased with a few clicks. Achieving 99% sensitivity with only 0.6 false positives per scan reduces annotation time by up to 92% (see §A.3).

SAM-based Tool for Organ/Vessel Annotation. We use 10 strategically placed points from Flagship Model as prompts for SAM [26], enabling highly accurate boundary detection. This approach reduces annotation time by approximately 60% compared to traditional active learning approaches [70] and by over 80% compared to creating organ annotations from scratch. This technique is particularly beneficial for annotating vessels, where Hounsfield Unit (HU) intensity values are often homogeneous, presenting challenges for conventional segmentation methods.

2.3. Integration: Data Mix/Annealing → Better AI

Architecture for Online Continual Learning. Flagship Model is designed as an AI-based segmentation agent that learns incrementally in an online continual learning paradigm. Trained iteratively on the combination of selective [18], synthetic [16, 17, 44, 46, 51], and unlabeled data, the model θ incorporates new class from sequential medical images, where new organ structures or tumor types are iteratively introduced. The model architecture begins by feeding the input image I through a vision encoder $E_v(I)$, which generates high-dimensional image features F_i. In parallel, associated text input T is projected into text features F_t using a language encoder $E_l(T)$. These multimodal features are then concatenated within a fusion module $F_f = [F_i; F_t]$, facilitating comprehensive integration of both visual and textual data. The fused representation F_f is decoded by a segmentation decoder $D(F_f)$, producing the final segmentation output S as a refined prediction of anatomical structures. For online learning, the model learns from sequential tasks iteratively $t \in T = \{1, \dots, k\}$ where each task introduces batches of training data $D_{t,b} = \{X_{t,b}, Y_{t,b}\}$, with $X_{t,b}$ as input data and $Y_{t,b}$ as labels that may include previously unseen classes. This sequential data flow is represented as: $D_{\text{c}} = \{D_{t,b}\}_{b \in B_t, t \in T}$, allowing the model to learn continuously from D_{c} without re-accessing earlier tasks. To mitigate catastrophic forgetting, the full samples $M_t \subset D_t$ from each task are retained, ensuring the model retains key representations over time. The objective function for online learning is thus framed as: $\min_\theta \sum_{t=1}^{k} \sum_{b \in B_t} L(f_\theta, D_{t,b} \cup M_{<t})$, where $M_{<t} = \bigcup_{t'<t} M_{t'}$ represents samples from prior tasks. As new classes appear in the label sets $Y_{t,b}$, the model dynamically adjusts its output space. This architecture and learning strategy enables the model to progressively refine its segmentation capabilities in response to evolving medical data and newly introduced anatomical or pathological classes.

Data Mix consists of three primary data types to enhance the model's training efficiency and robustness, detailed in §A.4. *First*, unlabeled data supports self-supervised representation learning. By tackling the abundance of raw

dataset (year) [source]	# of class	# of CT	# of center
MSD-Pancreas [2022][link]	2	420	1
TCIA-CBCT [2021][link]	-	40	1
TCIA-panNET [2023][link]	-	38	1
PANORAMA [2024][link]	6	3,000	7
PancreaVerse	27	25,362	112

Table 2. PancreaVerse exceeds existing pancreatic tumor datasets in scale and diversity, providing 25,362 CT scans with annotations of 27 classes from 112 hospitals. More comprehensive summary of public tumor/organ datasets is in Appendix Table 8.

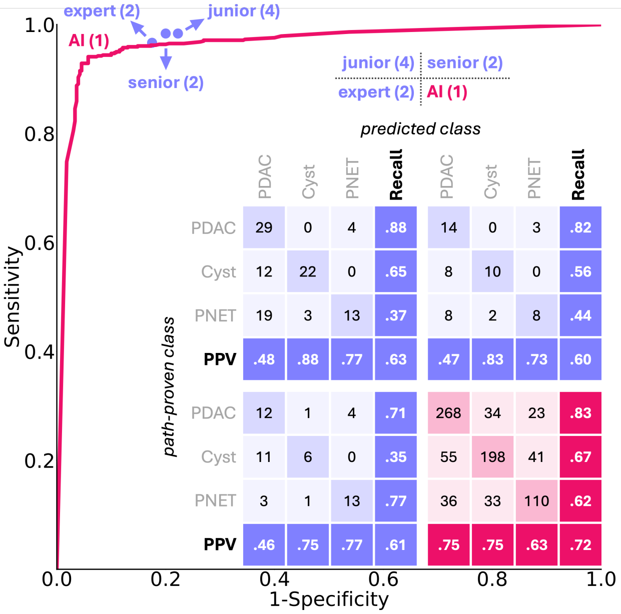
clinical CT scans produced daily—without requiring manual annotations—this approach allows the model to develop rich, generalizable features and reduces the need for annotated data [89, 91]. *Second*, synthetic data introduces variations that may not be present in the original dataset, such as differences in patient demographics, scanner types, contrast phases, or tumor characteristics (e.g., location, shape, texture, size, intensity) [21, 32–34]. This artificial diversity helps the model adapt better to out-of-distribution cases, improving its generalization. *Third*, selective data targets the most challenging regions of CT scans as identified by loss function during training [14, 18, 87, 88, 90]. By prioritizing repeated sampling of these regions, we can avoid the model learning from the non-informative areas such as air, bedding, or irrelevant anatomical regions. This targeted approach ensures that the model focuses on clinically relevant areas, such as the pancreas or abdominal region.

Data Annealing. We introduce data annealing to further fine-tune the model, detailed in §A.4. We identify a gold-standard subset, consisting of voxel-level annotations meticulously created by human experts. This data annealing technique has proven effective in large-scale training efforts in other domains, such as GPT [2] and Llama [22]. However, in the medical field, the lack of gold-standard data and the predominance of silver-standard data have limited its exploration. When releasing the dataset, we will explicitly mark this gold-standard subset to facilitate further research and development in the field.

3. Quality Assessment of AI Trusted Datasets

3.1. AI Trusted Pancreatic Tumor Dataset

Table 2 compares public pancreatic datasets with PancreaVerse in terms of the number of classes, CT scans, and sourcing hospitals. Our PancreaVerse surpasses existing pancreatic tumor datasets in both scale and diversity, containing 25,362 CT scans annotated with 27 classes[6] which are essential for pancreatic tumor diagnosis, collected from 112 hospitals. *We will make our PancreaVerse available.*



	PDAC	Cyst	PNET	Recall	PDAC	Cyst	PNET	Recall
PDAC	29	0	4	.88	14	0	3	.82
Cyst	12	22	0	.65	8	10	0	.56
PNET	19	3	13	.37	8	2	8	.44
PPV	.48	.88	.77	.63	.47	.83	.73	.60
PDAC	12	1	4	.71	268	34	23	.83
Cyst	11	6	0	.35	55	198	41	.67
PNET	3	1	13	.77	36	33	110	.62
PPV	.46	.75	.77	.61	.75	.75	.63	.72

Figure 2. **Flagship Model matches radiologists in tumor detection and surpasses them in classification.** Eight radiologists (4 juniors, 2 seniors, 2 experts) were tasked to detect and classify tumors in 50 cases. Note that radiologists can only make predict based on CT scans without accessing patient medical history, so their performance will be lower than that in the clinical practice. For tumor detection, Flagship Model (pink curve) achieved similar performance to the radiologists (blue points). Junior and senior radiologists achieved high sensitivity (97–100%), but lower specificity (75–80%), while experts showed slightly lower sensitivity (93–100%) with higher specificity (80–85%). For tumor classification (PDAC, Cyst, PNET), Flagship Model achieved 72% accuracy, surpassing junior, senior, and expert radiologists by 9%, 12%, and 11%, respectively. Details of the expanded reader study (13 radiologists) are provided in §B.2.

3.2. Gold Standard vs. Silver Standard Annotation

Flagship Model has achieved human-level annotation quality, allowing us to stop human-in-the-loop dataset refinement since further human annotations add negligible value. To quantify this, we conducted a study where eight radiologists independently annotated pancreatic tumors in 50 unseen CT scans. We consider Flagship Model to have reached human-level performance if its performance exceeds that of all or most radiologists. Such datasets are termed **silver standard**—they match human-level performance but are less precise than **gold standard** annotations, which are pathology-proven but limited in scale. Despite this, silver standard datasets are highly valuable for training AI models that perform strongly on gold standard evaluations (Table 4, Table 5) because they can be scaled more

[6] These classes include aorta, gall bladder, left and right kidneys, liver, pancreas, postcava, spleen, stomach, left and right adrenal glands, bladder, colon, duodenum, left and right femurs, left and right lungs, prostate, superior mesenteric artery, pancreatic duct, celiac artery, common bile duct, veins, and benign and malignant pancreatic tumors.

easily. As model improves, the quality of silver standard annotations is expected to rise, enabling large-scale training while reserving gold standard datasets for evaluation.

3.2.1. Reader Study: Tumor Detection & Classification

Reader study settings. We conducted a multi-reader, multi-case study with eight radiologists (4 juniors, 2 seniors, 2 experts) interpreting 50 contrast-enhanced abdominal CT scans to assess inter-reader variability in pancreatic tumor analysis and compare their performance with Flagship Model. The study focused on pancreatic tumor detection and classification of cysts, pancreatic adenocarcinoma (PDAC), and pancreatic neuroendocrine tumors (PNET). The reader study results show that (1) Flagship Model matched experts in pancreatic tumor detection. (2) Flagship Model outperformed experts in tumor classification.

3.2.2. High Quality Anatomical Structure Annotation

Models trained on PancreaVerse significantly outperform those trained on smaller, manually annotated datasets when evaluated on a high-quality, out-of-distribution dataset, confirming the reliability of PancreaVerse for AI development. Table 3 compared the segmentation performance of AI models trained on different datasets—BTCV, WORD, AbdomenAtlas1.0, and PancreaVerse—and tested on a high-quality, proprietary JHH dataset (N=300). The model trained on our dataset achieves the highest DSC score across most anatomical structures. This demonstrates that the high annotation quality (see §B.1) and diversity of our dataset enable AI models to generalize better, validating its value as a trusted dataset for developing robust AI models.

4. Experimental Results of Flagship AI Model

Baselines. The top-performing Swin UNETR [80] from the MSD leaderboard [8] and top models from the Touchstone benchmark [10] (e.g., MedNext [76] (Top-1), STU-Net-B [35], ResEncL [38], and UniSeg [86]) are considered as baseline. Swin UNETR is built on the MONAI [13] framework, while the others leverage the self-configuring nnU-Net [37], which adapts to diverse datasets.

Metrics. Sensitivity & specificity are used for detection; we only count true positives when tumor predictions intersect with the ground truth. Dice Similarity Coefficient (DSC) & Normalized Surface Distance (NSD) are for segmentation.

Datasets. The MSD-Pancreas dataset [8] was divided into a training set (N=200, 71%) and a test set (N=81, 29%). All 281 CT scans included pancreatic tumors classified as 109 small ($d < 20$mm), 158 medium ($20 < d < 40$mm), and 16 large ($d > 40$mm) tumors, as shown in Appendix Figure 11. The tumor size distribution in the test set is similar to the training set[7]. The baseline models are trained on the MSD-Pancreas training set. PANORAMA [6] and our proprietary JHH datasets were used for external validation and provided rich metadata such as sex, age, scanner type, tumor size, and path-proven methods [10]. The entire PANORAMA dataset (N=1,964)[8] was used for evaluation, comprising 578 positive cases (PDAC) with per-voxel annotations and 1,386 negative cases. The proprietary JHH dataset (N=1,958) was also used for evaluation. This dataset allowed for a detailed analysis across venous and arterial imaging phases. Besides the metadata in PANORAMA, our proprietary JHH dataset provided additional insights by enabling analyses based on contrast methods, tumor size, rough locations, and sub-types. Additional comparisons of dataset attributes for these datasets are provided in §C.1.

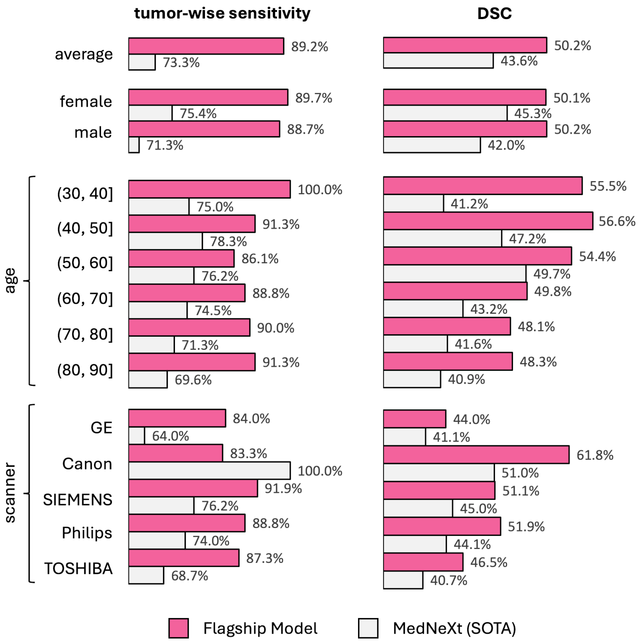


Figure 3. **Flagship Model demonstrates robust generalizability across diverse demographic and technical variations in out-of-distribution evaluation on the PANORAMA dataset.** Flagship Model shows enhanced tumor detection and segmentation across various age groups, genders, and scanner types, consistently outperforming the top-1 performing MedNeXt model [76]—trained on a smaller, fixed-quality dataset [10]. Notably, Flagship Model surpasses MedNeXt in all patient groups except those scanned with Canon scanners. More comprehensive results of Flagship Model vs. MedNeXt are in Appendix Figure 12.

4.1. Tumor Detection (+14% Sensitivity)

Table 4 give an overall performance comparison with all baselines. Our Flagship Model significantly outperforms other baselines on three prestigious benchmarks. Compared to the SOTA method (i.e., Swin UNETR [80]) on

[7] All of the 81 test set cases have PDAC diagnostic labels provided by the PANORAMA dataset.

[8] The 194 MSD-Pancreas cases and 80 cases from National Institute of Health[75] are removed for fair evaluation.

training dataset	# of CTs	annotators	out-of-distribution test on the proprietary JHH dataset				
			spleen	kidneyR	kidneyL	gallbladder	liver
BTCV [45]	47	human	93.6 (75.6–95.4)	43.9 (0.9–89.9)	94.8 (93.0–95.5)	77.1 (30.5–88.3)	95.4 (94.7–95.9)
WORD [59]	120	human	93.0 (89.7–94.3)	95.5 (94.9–95.9)	95.1 (92.8–95.8)	78.5 (51.6–86.5)	94.8 (93.8–95.5)
AbdomenAtlas 1.0 [70]	5,195	human-AI	95.8 (95.1–96.5)	93.2 (91.9–94.4)	92.8 (91.3–93.9)	88.2 (82.0–90.9)	96.4 (95.8–96.9)
PancreaVerse	25,362	human-AI	**96.2 (95.2–96.9)**	**97.7 (97.4–98.0)**	**97.6 (97.3–97.9)**	**88.5 (80.6–92.1)**	**96.7 (96.2–97.2)**
			stomach	aorta	postcava	pancreas	average
BTCV [45]	47	human	92.0 (87.1–94.0)	61.3 (19.6–83.3)	69.1 (36.8–80.6)	74.5 (66.6–79.5)	72.5 (61.8–81.3)
WORD [59]	120	human	90.7 (87.6–92.6)	-	-	75.9 (68.2–80.9)	87.1 (80.8–89.8)
AbdomenAtlas 1.0 [70]	5,195	human-AI	94.7 (93.0–95.5)	90.4 (87.6–91.8)	81.2 (75.1–84.9)	82.9 (78.7–85.9)	89.8 (87.9–91.2)
PancreaVerse	25,362	human-AI	**95.8 (94.3–96.4)**	**91.8 (88.2–94.4)**	**85.8 (82.1–88.8)**	**85.7 (81.8–88.1)**	**92.0 (89.7–93.2)**

Table 3. **AI models trained on PancreaVerse significantly outperform those trained on smaller datasets when evaluating on an out-of-distribution high-quality dataset.** We compare the segmentation performance of AI models trained on BTCV, WORD, AbdomenAtlas1.0, and PancreaVerse, evaluated on an out-of-distribution proprietary JHH dataset (N=300). The model trained on our dataset achieves the highest DSC scores across most anatomical structures—showing better generalization ability and validating the value of our dataset for developing robust AI models. We compare the median and interquartile range (IQR) of the DSC score. In addition, we have further performed an independent two-sample t-test between the best-performed model and others. The performance gain is statistically significant at the $P = 0.05$ level, with highlighting in a pink box.

method	training set	MSD-Pancreas (N=81)	PANORAMA (N=1,964)		proprietary JHH dataset (N=1,958)	
		Sensitivity	Sensitivity	Specificity	Sensitivity	Specificity
Swin UNETR [80]	MSD-Pancreas	81.5 (66/81)	85.5 (497/581)	11.0 (152/1386)	25.2 (824/3426)	16.7 (104/623)
UniSeg [86]	MSD-Pancreas	80.2 (65/81)	77.8 (452/581)	54.8 (760/1386)	23.4 (801/3426)	78.5 (485/623)
ResEncL [38]	MSD-Pancreas	72.8 (59/81)	74.7 (434/581)	84.6 (1173/1386)	23.7 (813/3426)	87.0 (542/623)
STU-Net-Base [35]	MSD-Pancreas	74.1 (60/81)	76.8 (446/581)	84.0 (1164/1386)	23.1 (791/3426)	84.4 (526/623)
MedNeXt [76]	MSD-Pancreas	75.3 (61/81)	73.3 (426/581)	**84.8 (1176/1386)**	24.6 (842/3426)	85.2 (531/623)
Flagship Model	PancreaVerse	**95.1 (77/81)**	**89.2 (518/581)**	73.2 (1015/1386)	**40.2 (1381/3426)**	**88.3 (550/623)**

Table 4. **Flagship Model, with a backbone of ResEncL, achieves the best performance for pancreatic tumor detection.** Note that these are tumor-wise detection results, and the patient-wise results are in Appendix Table 9. The out-of-distribution sensitivity of Flagship Model surpasses the in-distribution sensitivity of existing AI models. Performance is given as sensitivity and specificity. Best-performing results are **bolded** for each dataset. Patient-wise detection results can be found in §C.2.

MSD leaderboard, Flagship Model improves Sensitivity by **+14%** (95.1% vs. 81.5%) on the MSD-Panreas dataset. On the PANORAMA dataset, our Flagship Model outperforms the top performer, STU-Net [35] by **+12%** in Sensitivity (89.2% vs. 76.8%). Even when compared to the model with the highest Sensitivity (Swin UNETR), our Flagship Model still exceeds **+4%** in Sensitivity and obtains a satisfactory Specificity of 73.2%, significantly surpassing the 11.0% Specificity of Swin UNETR. Moreover, our Flagship Model substantially surpasses other baselines on the proprietary JHH dataset with **+15%** in Sensitivity. Notably, the reduced detection performance of the baselines is due to the constrained and static nature of their training set, i.e., MSD-Pancreas Train, which exhibits an out-of-distribution challenge regarding tumor size when compared to the proprietary JHH Dataset (Appendix Figure 11). In contrast, benefiting from our large and diverse AI-trusted dataset, our Flagship Model exhibits robust performance when encountering out-of-distribution data.

Metadata analysis. We evaluated the generalizability of Flagship Model for tumor detection performance across diverse demographic groups and scanner types in out-of-distribution evaluation on PANORAMA dataset, shown in Figure 3 and on our proprietary dataset shown in Appendix Figure 12. Flagship Model consistently outperforms the top-performing MedNeXt model in tumor detection across different age groups and genders. Additionally, it demonstrates superior performance across various scanner types, except for those manufactured by Canon.

4.2. Tumor Segmentation (+5% DSC)

To assess the model's capability in accurately identifying tumor boundaries, we evaluate the tumor segmentation performance. As shown in Table 5, we present the segmentation performance in DSC and NSD scores. First, on the MSD-Pancreas dataset, our Flagship Model outperforms the SOTA method, i.e., MedNeXt, by **+5%** in DSC and **+3.1%** in NSD. Next, on the PANORAMA dataset, Flagship Model also achieves superior results compared to the SOTA method (STU-Net), with an improvement of **+6.6%** in DSC and **+3.2%** in NSD. Last, Flagship Model considerably surpasses baselines with **+21.5%** in DSC and **+24.1%** in NSD on the proprietary JHH dataset, rendering the model robustness under out-of-distribution evaluation.

		MSD-Pancreas (N=81)		PANORAMA (N=1,964)		proprietary JHH dataset (N=1,958)	
method	training set	DSC	NSD	DSC	NSD	DSC	NSD
Swin UNETR [80]	MSD-Pancreas	50.2 (6.3–68.1)	58.1 (18.1–76.7)	39.4 (9.5–64.3)	31.9 (13.3–52.2)	11.4 (0.0–49.1)	11.2 (0.0–32.5)
UniSeg [86]	MSD-Pancreas	59.4 (5.6–79.1)	69.4 (23.2–89.2)	49.8 (1.1–72.1)	38.7 (6.5–64.2)	12.2 (0.0–56.9)	9.1 (0.0–44.6)
ResEncL [38]	MSD-Pancreas	61.6 (0.0–78.1)	71.1 (0.0–92.1)	54.1 (0.0–72.2)	41.0 (1.8–66.4)	22.6 (0.0–65.6)	13.4 (0.0–52.4)
STU-Net-Base [35]	MSD-Pancreas	62.7 (0.0–77.9)	70.6 (9.1–91.3)	51.8 (0.6–73.1)	41.5 (4.4–67.2)	13.5 (0.0–62.3)	11.2 (0.0–48.4)
MedNeXt [76]	MSD-Pancreas	68.1 (2.7–80.9)	79.4 (18.4–95.9)	54.4 (0.0–74.8)	40.7 (0.4–66.9)	30.6 (0.0–69.9)	20.4 (0.0–60.1)
Flagship Model	PancreaVerse	**63.4 (40.6–77.5)**	**67.7 (45.4–84.5)**	**56.8 (23.7–75.8)**	**44.5 (20.5–66.0)**	**68.6 (34.7–82.9)**	**63.3 (33.2–81.9)**

Table 5. **Benchmarking pancreatic tumor segmentation on MSD-Pancreas, PANORAMA, and a proprietary JHH dataset.** Flagship model tumor segmentation performance is significantly higher than all pre-existing AI models developed on publicly available MSD-Pancreas training set. We compare the median and interquartile range (IQR) of the DSC and NSD scores of the models. For each dataset, we **bold** the best-performing results. In addition, we have further performed an independent two-sample t-test between the best-performed model with others. The performance gain is statistically significant at the $P = 0.05$ level, with highlighting in a pink box. Additional tumor segmentation results for different demographic groups and scanner types are provided in §C.3.

Metadata analysis. We evaluated the out-of-distribution tumor segmentation performance for Flagship Model across demographic groups and scanner types on PANORAMA and our proprietary dataset (Appendix Figure 12). Flagship Model outperformed the top MedNeXt model across ages, genders, and scanners, demonstrating robust performance and potential for broad clinical applicability.

4.3. Tumor Classification (72% Accuracy)

For pancreatic tumor subtype classification (PDAC, Cyst, and PNET), we compared the confusion matrices of Flagship Model with those of radiologists at different experience levels (junior/senior/expert). As shown in Figure 2, radiologists, regardless of experience, achieved modest PPV (Positive Predictive Value) and recall (equivalent to Sensitivity) in classifying the three types of pancreatic tumors. Flagship Model demonstrated superior performance, achieving higher PPV and recall rates across all three tumor types. For instance, while junior radiologists showed slightly higher recall for PDAC (88%), their performance on Cyst and PNET lagged, with recall values of 65% and 37%, respectively. The Flagship Model outperformed radiologists overall, achieving recall rates of 83% for PDAC, 67% for Cyst, and 62% for PNET. Moreover, Flagship Model surpassed the radiologists in overall accuracy, achieving 72% (**+11%**) compared to the average 61% accuracy observed among radiologists. A comprehensive comparison of pancreatic tumor classification between Flagship Model and radiologists of varying experience is presented in §C.4.

5. Conclusion & Clinical Applications

This paper introduces **ScaleMAI**, an AI-integrated data curation and annotation agent that combines iterative, multi-stage processes with AI and human expertise to progressively enhance dataset quality. Using pancreatic tumor detection, segmentation, and classification as case studies, we demonstrate that ScaleMAI significantly reduces data curation and annotation time from years to months by integrating large language models, vision-language models, and human-in-the-loop feedback. The refinement process continues until the dataset achieves human-level performance. ScaleMAI bridges the gap between clinicians, often burdened with labor-intensive annotations, and AI researchers who lack domain-specific expertise. By automating routine tasks, it empowers clinicians to focus on complex cases, accelerating dataset creation, improving model robustness, and enabling scalable AI solutions for clinical practice.

PancreaVerse—an AI trusted dataset created through ScaleMAI—provides 25,362 CT scans (8.5× larger than the largest existing pancreatic tumor dataset) collected from 112 global hospitals with silver-standard annotations verified by eight expert radiologists. AI models trained on PancreaVerse outperform those trained on smaller datasets. This marks an early exploration of Scaling Laws [42] in medical vision, i.e., tumor segmentation plus supervised learning, highlighting the critical need for high-quality annotated data. PancreaVerse will be made public.

Flagship Model—another outcome of ScaleMAI—has the potential for multiple clinical applications. We present preliminary results in Appendix due to the page limit. Firstly, Flagship Model achieves 53% accuracy in pancreatic tumor staging (i.e., T1–T4) by segmenting the pancreas, surrounding vessels, and tumors across 30 CT scans (see details in Appendix §C.5). Precise tumor staging is the key criterion for evaluating resectability. Secondly, Flagship Model excels in stereotactic radiotherapy planning by accurately delineating complex structures such as the duodenum on low-quality planning CTs, significantly reducing manual workload (see details in Appendix §C.6). These capabilities enhance tumor targeting precision, minimize radiation exposure to critical organs, and improve overall treatment outcomes, showcasing the model's transformative potential in clinical practice.

Acknowledgments. This work was supported by the Lustgarten Foundation for Pancreatic Cancer Research and the Patrick J. McGovern Foundation Award.

References

[1] Lorraine Abel, Jakob Wasserthal, Thomas Weikert, Alexander W. Sauter, Ivan Nesic, Marko Obradovic, Shan Yang, Sebastian Manneck, Carl Glessgen, Johanna M. Ospel, Bram Stieltjes, Daniel T. Boll, and Björn Friebe. Automated detection of pancreatic cystic lesions on ct using deep learning. *Diagnostics*, 11(5), 2021. 10

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 5

[3] Hugo J. W. L. Aerts, Emmanuel R. Velazquez, Ralph T. H. Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Carvalho, Johan Bussink, Ronald Monshouwer, Benjamin Haibe-Kains, David Rietveld, Frank Hoebers, Marieke M. Rietbergen, René Leemans, Andre Dekker, John Quackenbush, Robert Gillies, and Philippe Lambin. A whole-body FDG-PET/CT dataset with manually annotated tumor lesions. *Scientific Data*, 9(1):1–9, 2022. 10

[4] Oguz Akin, Paula Elnajjar, Mark Heller, Rosemary Jarosz, Bradley J. Erickson, Sheri Kirk, Yu Lee, Marston W. Linehan, Rajan Gautam, Ramesh Vikram, Kelly M. Garcia, Charles Roche, Emanuele Bonaccio, and Jeffrey Filippini. The Cancer Genome Atlas Kidney Renal Clear Cell Carcinoma Collection (TCGA-KIRC) (Version 3), 2016. [Data set]. 10

[5] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 4, 3

[6] N Alves, M Schuurmans, D Rutkowski, et al. The panorama study protocol: Pancreatic cancer diagnosis-radiologists meet ai. zenodo, 2024. 5, 6, 10, 11

[7] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, Bram van Ginneken, et al. The medical segmentation decathlon. *arXiv preprint arXiv:2106.05735*, 2021. 3

[8] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):1–13, 2022. 1, 2, 5, 6, 10

[9] Sunil Arya, David M Mount, Nathan S Netanyahu, Ruth Silverman, and Angela Y Wu. An optimal algorithm for approximate nearest neighbor searching fixed dimensions. *Journal of the ACM (JACM)*, 45(6):891–923, 1998. 3

[10] Pedro RAS Bassi, Wenxuan Li, Yucheng Tang, Fabian Isensee, Zifu Wang, Jieneng Chen, Yu-Cheng Chou, Yannick Kirchhoff, Maximilian Rokuss, Ziyan Huang, Jin Ye, Junjun He, Tassilo Wald, Constantin Ulrich, Michael Baumgartner, Saikat Roy, Klaus H. Maier-Hein, Paul Jaeger, Yiwen Ye, Yutong Xie, Jianpeng Zhang, Ziyang Chen, Yong Xia, Zhaohu Xing, Lei Zhu, Yousef Sadegheih, Afshin Bozorgpour, Pratibha Kumari, Reza Azad, Dorit Merhof, Pengcheng Shi, Ting Ma, Yuxin Du, Fan Bai, Tiejun Huang, Bo Zhao, Haonan Wang, Xiaomeng Li, Hanxue Gu, Haoyu Dong, Jichen Yang, Maciej A. Mazurowski, Saumya Gupta, Linshan Wu, Jiaxin Zhuang, Hao Chen, Holger Roth, Daguang Xu, Matthew B. Blaschko, Sergio Decherchi, Andrea Cavalli, Alan L. Yuille, and Zongwei Zhou. Touchstone benchmark: Are we on the right way for evaluating ai algorithms for medical segmentation? *Conference on Neural Information Processing Systems*, 2024. 2, 3, 6

[11] Patrick Bilic, Patrick Ferdinand Christ, Eugene Vorontsov, Grzegorz Chlebus, Hao Chen, Qi Dou, Chi-Wing Fu, Xiao Han, Pheng-Ann Heng, Jürgen Hesser, et al. The liver tumor segmentation benchmark (lits). *arXiv preprint arXiv:1901.04056*, 2019. 1, 10

[12] Kai Cao, Yingda Xia, Jiawen Yao, Xu Han, Lukas Lambert, Tingting Zhang, Wei Tang, Gang Jin, Hui Jiang, Xu Fang, et al. Large-scale pancreatic cancer detection via non-contrast ct and deep learning. *Nature medicine*, 29(12): 3033–3043, 2023. 1

[13] M Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, Andriy Myronenko, Can Zhao, Dong Yang, et al. Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*, 2022. 6, 4

[14] Liangyu Chen, Yutong Bai, Siyu Huang, Yongyi Lu, Bihan Wen, Alan L Yuille, and Zongwei Zhou. Making your first choice: To address cold start problem in vision active learning. In *Medical Imaging with Deep Learning*. 2023. 5

[15] Luohai Chen, Wei Wang, Kaizhou Jin, Bing Yuan, Huangying Tan, Jian Sun, Yu Guo, Yanji Luo, Shi-Ting Feng, Xianjun Yu, et al. Special issue "the advance of solid tumor research in china": Prediction of sunitinib efficacy using computed tomography in patients with pancreatic neuroendocrine tumors. *International Journal of Cancer*, 152(1): 90–99, 2023. 5, 10

[16] Qi Chen, Xiaoxi Chen, Haorui Song, Zhiwei Xiong, Alan Yuille, Chen Wei, and Zongwei Zhou. Towards generalizable tumor synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 4

[17] Qi Chen, Yuxiang Lai, Xiaoxi Chen, Qixin Hu, Alan Yuille, and Zongwei Zhou. Analyzing tumors by synthesis. *arXiv preprint arXiv:2409.06035*, 2024. 4

[18] Yu-Cheng Chou, Zongwei Zhou, and Alan Yuille. Embracing massive medical data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 24–35. Springer, 2024. 4, 5

[19] Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, et al. The cancer imaging archive (tcia): maintaining and operating a public information repository. *Journal of digital imaging*, 26(6):1045–1057, 2013. 2

[20] Jeffrey De Fauw, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O'Donoghue, Daniel Visentin, et al. Clinically applicable deep learning for

diagnosis and referral in retinal disease. *Nature medicine*, 24 (9):1342–1350, 2018. 1

[21] Shiyi Du, Xiaosong Wang, Yongyi Lu, Yuyin Zhou, Shaoting Zhang, Alan Yuille, Kang Li, and Zongwei Zhou. Boosting dermatoscopic lesion segmentation via diffusion models with visual and textual prompts. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2024. 5

[22] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 2, 5

[23] Bradley J. Erickson, Sheri Kirk, Yu Lee, Oliver Bathe, Mary Kearns, Cynthia Gerdes, Kristine Rieger-Christ, and Julie Lemmerman. The Cancer Genome Atlas Liver Hepatocellular Carcinoma Collection (TCGA-LIHC) (Version 5), 2016. [Data set]. 10

[24] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2

[25] Hao Guan and Mingxia Liu. Domain adaptation for medical image analysis: a survey. *IEEE Transactions on Biomedical Engineering*, 69(3):1173–1185, 2021. 11

[26] Heng Guo, Jianfeng Zhang, Jiaxing Huang, Tony CW Mok, Dazhou Guo, Ke Yan, Le Lu, Dakai Jin, and Minfeng Xu. Towards a comprehensive, efficient and promptable anatomic structure segmentation model using 3d whole-body ct scans. *arXiv preprint arXiv:2403.15063*, 2024. 4

[27] Xu Han, Jun Hong, Marsha Reyngold, Christopher Crane, John Cuaron, Carla Hajj, Justin Mann, Melissa Zinovoy, Hastings Greer, Ellen Yorke, et al. Deep-learning-based image registration and automatic segmentation of organs-at-risk in cone-beam ct scans from high-dose radiation treatment of pancreatic cancer. *Medical physics*, 48(6):3084–3095, 2021. 5

[28] Nicholas Heller, Niranjan Sathianathen, Arveen Kalapara, Edward Walczak, Keenan Moore, Heather Kaluzniak, Joel Rosenberg, Paul Blake, Zachary Rengel, Makinna Oestreich, et al. The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. *arXiv preprint arXiv:1904.00445*, 2019. 1, 3

[29] Nicholas Heller, Sean McSweeney, Matthew Thomas Peterson, Sarah Peterson, Jack Rickman, Bethany Stai, Resha Tejpaul, Makinna Oestreich, Paul Blake, Joel Rosenberg, et al. An international challenge to use artificial intelligence to define the state-of-the-art in kidney and kidney tumor segmentation in ct imaging., 2020. 10

[30] Craig Holback, Rosemary Jarosz, Fred Prior, David G. Mutch, Priya Bhosale, Kelly Garcia, Yu Lee, Sheri Kirk, Charity A. Sadow, Susan Levine, Elizabeth Sala, Paula Elnajjar, Trevor Morgan, and Bradley J. Erickson. The Cancer Genome Atlas Ovarian Cancer Collection (TCGA-OV) (Version 4), 2016. [Data set]. 10

[31] J Hong, M Reyngold, C Crane, J Cuaron, C Hajj, J Mann, M Zinovoy, E Yorke, E LoCastro, AP Apte, et al. Breath-hold ct and cone-beam ct images with expert manual organ-at-risk segmentations from radiation treatments of locally advanced pancreatic cancer [data set]. the cancer imaging archive. *The Cancer Imaging Archive https://doi. org/10.7937/TCIA. ESHQ-4D90*, 2021. 10

[32] Qixin Hu, Junfei Xiao, Yixiong Chen, Shuwen Sun, Jie-Neng Chen, Alan Yuille, and Zongwei Zhou. Synthetic tumors make ai segment tumors better. *NeurIPS Workshop on Medical Imaging meets NeurIPS*, 2022. 5

[33] Qixin Hu, Yixiong Chen, Junfei Xiao, Shuwen Sun, Jieneng Chen, Alan L Yuille, and Zongwei Zhou. Label-free liver tumor segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7422–7432, 2023.

[34] Qixin Hu, Alan Yuille, and Zongwei Zhou. Synthetic data as validation. *arXiv preprint arXiv:2310.16052*, 2023. 5

[35] Ziyan Huang, Haoyu Wang, Zhongying Deng, Jin Ye, Yanzhou Su, Hui Sun, Junjun He, Yun Gu, Lixu Gu, Shaoting Zhang, et al. Stu-net: Scalable and transferable medical image segmentation models empowered by large-scale supervised pre-training. *arXiv preprint arXiv:2304.06716*, 2023. 6, 7, 8, 12

[36] Tania Idris, Sindhu Somarouthu, Heather Jacene, Ann LaCasce, Elise Ziegler, Steve Pieper, Reza Khajavi, Reuben Dorent, Sonia Pujol, Ron Kikinis, and Gregory Harris. Mediastinal Lymph Node Quantification (LNQ): Segmentation of Heterogeneous CT Data (Version 1), 2024. [Data set]. 10

[37] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021. 6

[38] Fabian Isensee, Tassilo Wald, Constantin Ulrich, Michael Baumgartner, Saikat Roy, Klaus Maier-Hein, and Paul F Jaeger. nnu-net revisited: A call for rigorous validation in 3d medical image segmentation. *arXiv preprint arXiv:2404.09556*, 2024. 6, 7, 8, 12

[39] Wei Ji, Jingjing Li, Cheng Bian, Zongwei Zhou, Jiaying Zhao, Alan L Yuille, and Li Cheng. Multispectral video semantic segmentation: A benchmark dataset and baseline. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1094–1104, 2023. 1

[40] Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *Advances in neural information processing systems*, 35:36722–36732, 2022. 10

[41] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 4, 3

[42] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for

neural language models. *arXiv preprint arXiv:2001.08361*, 2020. 8

[43] Sheri Kirk, Yu Lee, Francesco R. Lucchesi, Natassia D. Aredes, Neringa Gruszauskas, James Catto, Kelly Garcia, Rosemary Jarosz, Vaibhav Duddalwar, Bindu Varghese, Kristine Rieger-Christ, and Julie Lemmerman. The Cancer Genome Atlas Urothelial Bladder Carcinoma Collection (TCGA-BLCA) (Version 8), 2016. [Data set]. 10

[44] Yuxiang Lai, Xiaoxi Chen, Angtian Wang, Alan Yuille, and Zongwei Zhou. From pixel to cancer: Cellular automata in computed tomography. *arXiv preprint arXiv:2403.06459*, 2024. 4

[45] Bennett Landman, Zhoubing Xu, J Igelsias, Martin Styner, T Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, page 12, 2015. 7, 10

[46] Bowen Li, Yu-Cheng Chou, Shuwen Sun, Hualin Qiao, Alan Yuille, and Zongwei Zhou. Early detection and localization of pancreatic cancer by label-free tumor synthesis. *MICCAI Workshop on Big Task Small Data, 1001-AI*, 2023. 4

[47] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 4, 3

[48] Jianning Li, Antonio Pepe, Christina Gsaxner, Gijs Luijten, Yuan Jin, Narmada Ambigapathy, Enrico Nasca, Naida Solak, Gian Marco Melito, Afaque R Memon, et al. Medshapenet–a large-scale dataset of 3d medical shapes for computer vision. *BIOMEDIZINISCHE TECHNIK*, pages 1–20, 2024. 1

[49] Wenxuan Li, Chongyu Qu, Xiaoxi Chen, Pedro RAS Bassi, Yijia Shi, Yuxiang Lai, Qian Yu, Huimin Xue, Yixiong Chen, Xiaorui Lin, et al. Abdomenatlas: A large-scale, detailed-annotated, & multi-center dataset for efficient transfer learning and open algorithmic benchmarking. *Medical Image Analysis*, page 103285, 2024. 1

[50] Wenxuan Li, Alan Yuille, and Zongwei Zhou. How well do supervised models transfer to 3d image segmentation? In *International Conference on Learning Representations*, 2024. 1

[51] Xinran Li, Yi Shuai, Chen Liu, Qi Chen, Qilong Wu, Pengfei Guo, Dong Yang, Can Zhao, Pedro RAS Bassi, Daguang Xu, et al. Text-driven tumor synthesis. *arXiv preprint arXiv:2412.18589*, 2024. 4

[52] Marston Linehan, R Gautam, S Kirk, Y Lee, C Roche, E Bonaccio, et al. The cancer genome atlas cervical kidney renal papillary cell carcinoma collection (tcga-kirp), version 4. *The Cancer Imaging Archive*, 2016. 10

[53] MW Linehan, R Gautam, CA Sadow, and S Levine. Radiology data from the cancer genome atlas kidney chromophobe [tcga-kich] collection. *The Cancer Imaging Archive*, 2016. 10

[54] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 4, 3

[55] Jie Liu, Yixiao Zhang, Jie-Neng Chen, Junfei Xiao, Yongyi Lu, Bennett A Landman, Yixuan Yuan, Alan Yuille, Yucheng Tang, and Zongwei Zhou. Clip-driven universal model for organ segmentation and tumor detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21152–21164, 2023. 2

[56] Jie Liu, Yixiao Zhang, Kang Wang, Mehmet Can Yavuz, Xiaoxi Chen, Yixuan Yuan, Haoliang Li, Yang Yang, Alan Yuille, Yucheng Tang, et al. Universal and extensible language-vision models for organ segmentation and tumor detection from abdominal computed tomography. *Medical Image Analysis*, page 103226, 2024. 2

[57] Weixuan Liu, Bairui Zhang, Tao Liu, Juntao Jiang, and Yong Liu. Artificial intelligence in pancreatic image analysis: A review. *Sensors*, 24(14), 2024. 2

[58] FR Lucchesi and ND Aredes. The cancer genome atlas stomach adenocarcinoma collection (tcga-stad). *The Cancer Imaging Archive*, 2016. 10

[59] Xiangde Luo, Wenjun Liao, Jianghong Xiao, Tao Song, Xiaofan Zhang, Kang Li, Guotai Wang, and Shaoting Zhang. Word: Revisiting organs segmentation in the whole abdominal region. *arXiv preprint arXiv:2111.02403*, 2021. 7, 10

[60] Jun Ma, Yao Zhang, Song Gu, Cheng Zhu, Cheng Ge, Yichi Zhang, Xingle An, Congcong Wang, Qiyuan Wang, Xin Liu, et al. Abdomenct-1k: Is abdominal organ segmentation a solved problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 10

[61] Jun Ma, Yao Zhang, Song Gu, Cheng Ge, Ershuai Wang, Qin Zhou, Ziyan Huang, Pengju Lyu, Jian He, and Bo Wang. Automatic organ and pan-cancer segmentation in abdomen ct: the flare 2023 challenge. *arXiv preprint arXiv:2408.12534*, 2024. 10

[62] Cynthia McCollough, Bo Chen, David R. Holmes III, Xiaoli Duan, Zhi Yu, Lifeng Yu, Shuai Leng, and Joel Fletcher. Low Dose CT Image and Projection Data (LDCT-and-Projection-data) (Version 6), 2020. [Data set]. 10

[63] Ahmed W. Moawad, Diana Fuentes, Ahmed Morshid, Ahmed M. Khalaf, Mohamed M. Elmohr, Ahmed Abusaif, John D. Hazle, Ahmed O. Kaseb, Mohamed Hassan, Amir Mahvash, Jan Szklaruk, Ali Qayyom, and Khaled Elsayes. Multimodality annotated HCC cases with and without advanced imaging segmentation, 2021. [Data set]. 10

[64] National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC). The Clinical Proteomic Tumor Analysis Consortium Pancreatic Ductal Adenocarcinoma Collection (CPTAC-PDA) (Version 14), 2018. [Data set]. 10

[65] National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC). The Clinical Proteomic Tumor Analysis Consortium Clear Cell Renal Cell Carcinoma Collection (CPTAC-CCRCC) (Version 13), 2018. [Data set]. 10

[66] National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC). The Clinical Proteomic Tumor Analysis Consortium Sarcomas Collection (CPTAC-SAR) (Version 10), 2019. [Data set]. 10

[67] National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC). The Clinical Proteomic Tumor

Analysis Consortium Uterine Corpus Endometrial Carcinoma Collection (CPTAC-UCEC) (Version 12), 2019. [Data set]. 10

[68] NLST. Reduced lung-cancer mortality with low-dose computed tomographic screening. *New England Journal of Medicine*, 365(5):395–409, 2011. 1

[69] Wungki Park, Akhil Chawla, and Eileen M. O'Reilly. Pancreatic cancer: A review. *JAMA*, 326(9):851–862, 2021. 2

[70] Chongyu Qu, Tiezheng Zhang, Hualin Qiao, Jie Liu, Yucheng Tang, Alan Yuille, and Zongwei Zhou. Abdomenatlas-8k: Annotating 8,000 abdominal ct volumes for multi-organ segmentation in three weeks. In *Conference on Neural Information Processing Systems*, 2023. 1, 4, 7

[71] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4, 3

[72] Blaine Rister, Darvin Yi, Kaushik Shivakumar, Tomomi Nobashi, and Daniel L Rubin. Ct-org, a new dataset for multiple organ segmentation in computed tomography. *Scientific Data*, 7(1):1–9, 2020. 10

[73] C Roche, E Bonaccio, and J Filippini. The cancer genome atlas sarcoma collection (tcga-sarc)(version 3)[data set]. *The Cancer Imaging Archive*, 2016. 10

[74] Holger Roth, Le Lu, Ari Seff, Kevin M. Cherry, Joanne Hoffman, Shun Wang, Jiamin Liu, Evrim Turkbey, and Ronald M. Summers. A new 2.5 D representation for lymph node detection in CT, 2015. [Data set]. 10

[75] Holger R Roth, Le Lu, Amal Farag, Hoo-Chang Shin, Jiamin Liu, Evrim B Turkbey, and Ronald M Summers. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 556–564. Springer, 2015. 2, 6, 10

[76] Saikat Roy, Gregor Koehler, Constantin Ulrich, Michael Baumgartner, Jens Petersen, Fabian Isensee, Paul F Jaeger, and Klaus H Maier-Hein. Mednext: transformer-driven scaling of convnets for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 405–415. Springer, 2023. 6, 7, 8, 12

[77] Jeffrey D Rudie, Hui-Ming Lin, Robyn L Ball, Sabeena Jalal, Luciano M Prevedello, Savvas Nicolaou, Brett S Marinelli, Adam E Flanders, Kirti Magudia, George Shih, et al. The rsna abdominal traumatic injury ct (ratic) dataset. *Radiology: Artificial Intelligence*, 6(6):e240101, 2024. 10

[78] Hoo-Chang Shin, Holger R Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald M Summers. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging*, 35(5):1285–1298, 2016. 11

[79] Kevin Smith, Kirk Clark, William Bennett, Thomas Nolan, Jason Kirby, Michael Wolfsberger, Jonathan Moulton, Bruce Vendt, and John Freymann. Data From CT COLONOGRAPHY, 2015. [Data set]. 10

[80] Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20730–20740, 2022. 6, 7, 8, 12

[81] Tao Tong and Ming Li. Abdominal or pelvic enhanced CT images within 10 days before surgery of 230 patients with stage II colorectal cancer (StageII-Colorectal-CT), 2022. [Dataset]. 10

[82] Vanya V Valindria, Nick Pawlowski, Martin Rajchl, Ioannis Lavdas, Eric O Aboagye, Andrea G Rockall, Daniel Rueckert, and Ben Glocker. Multi-modal learning from unpaired images: Application to multi-organ segmentation in ct and mri. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 547–556. IEEE, 2018. 10

[83] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 2, 3

[84] Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck, Alexander W Sauter, Tobias Heye, Daniel T Boll, Joshy Cyriac, Shan Yang, et al. Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5), 2023. 10

[85] Yingda Xia, Qihang Yu, Linda Chu, Satomi Kawamoto, Seyoun Park, Fengze Liu, Jieneng Chen, Zhuotun Zhu, Bowen Li, Zongwei Zhou, et al. The felix project: Deep networks to detect pancreatic neoplasms. *medRxiv*, 2022. 1

[86] Yiwen Ye, Yutong Xie, Jianpeng Zhang, Ziyang Chen, and Yong Xia. Uniseg: A prompt-driven universal segmentation model as well as a strong representation learner. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 508–518. Springer, 2023. 6, 7, 8, 12

[87] Zongwei Zhou, Jae Shin, Lei Zhang, Suryakanth Gurudu, Michael Gotway, and Jianming Liang. Fine-tuning convolutional neural networks for biomedical image analysis: actively and incrementally. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7340–7351, 2017. 5

[88] Zongwei Zhou, Jae Shin, Ruibin Feng, R Todd Hurst, Christopher B Kendall, and Jianming Liang. Integrating active learning and transfer learning for carotid intima-media thickness video interpretation. *Journal of Digital Imaging*, 32(2):290–299, 2019. 5

[89] Zongwei Zhou, Vatsal Sodha, Md Mahfuzur Rahman Siddiquee, Ruibin Feng, Nima Tajbakhsh, Michael B Gotway, and Jianming Liang. Models genesis: Generic autodidactic models for 3d medical image analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 384–393. Springer, 2019. 5

[90] Zongwei Zhou, Jae Y Shin, Suryakanth R Gurudu, Michael B Gotway, and Jianming Liang. Active, continual

fine tuning of convolutional neural networks for reducing annotation efforts. *Medical Image Analysis*, 71:101997, 2021. 5

[91] Zongwei Zhou, Vatsal Sodha, Jiaxuan Pang, Michael B Gotway, and Jianming Liang. Models genesis. *Medical Image Analysis*, 67:101840, 2021. 5

Appendix

Table of Contents

A. Technical Details of ScaleMAI

A.1. Retriever: 180× Faster with 96% Accuracy

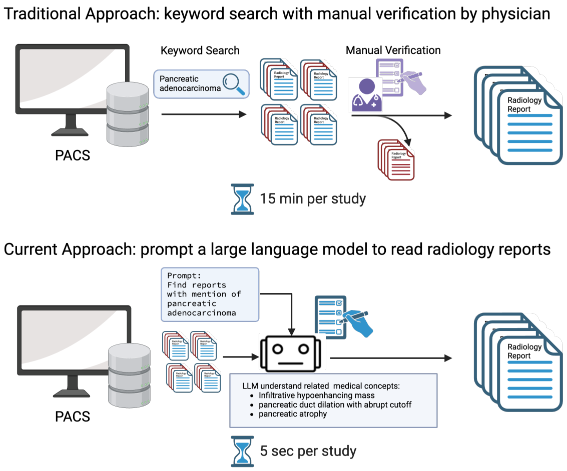

Traditional Approach: keyword search with manual verification by physician
Keyword Search
Pancreatic adenocarcinoma
Manual Verification
Radiology Report
PACS
15 min per study
Current Approach: prompt a large language model to read radiology reports
Prompt: Find reports with mention of pancreatic adenocarcinoma
LLM understand related medical concepts:
• Infiltrative hypoenhancing mass
• pancreatic duct dilation with abrupt cutoff
• pancreatic atrophy
Radiology Report
PACS
5 sec per study

Figure 4. The use of large language models (LLMs) for CT retrieval offers three key advantages. ***First***, it reduces retrieval time from 15 minutes per scan to just 5 seconds, outperforming traditional keyword searches followed by manual verification, while also enabling 24/7 operation. ***Second***, with respect to a search system in a real-world hospital, LLM-based searches improve precision from 89% to 96%. Specifically, using the keyword search system with terms like pancreatic cancer, pancreatic tumor, pancreatic adenocarcinoma, PDAC, and PNET, several relevant reports were missed but could be successfully identified by our LLM-based search, such as: (1) '*Redemonstrated 3 mm hyperenhancing lesion at the tail of the pancreas*'—keywords cannot differentiate between benign and malignant lesions; the term 'hyperenhancing' suggests suspicion for PNET and could be understood by LLMs. (2) '*Diffuse moderate dilatation of the pancreatic duct measuring up to 7 mm in diameter in the neck with moderate to severe dilatation in the pancreatic head up to 14 mm*'—no pancreatic mass is explicitly mentioned, but duct dilation is highly suspicious for a tumor. (3) '*A 1.0 cm cystic lesion in the pancreatic tail, now with a nodular component that is new from prior*'—the presence of a new nodule in a cystic lesion is suspicious for malignancy. ***Third***, LLMs enable complex searches, especially across longitudinal scans, that cannot be easily achieved with simple keyword combinations. For example, in early pancreatic tumor detection, the goal is to retrieve the earliest scan where a lesion was detected. Ideally, this would be the scan from the earliest date for each patient. However, the reports for these scans may not always indicate a tumor, as the lesion might have been undetectable by the radiologist at the time. These early, pre-diagnostic scans are particularly valuable because they provide opportunities to evaluate whether AI can identify tumors that were missed by human observers. These capabilities make our LLM-based approach, termed *Retriever*, a powerful tool for efficiently and accurately retrieving clinically relevant CT scans.

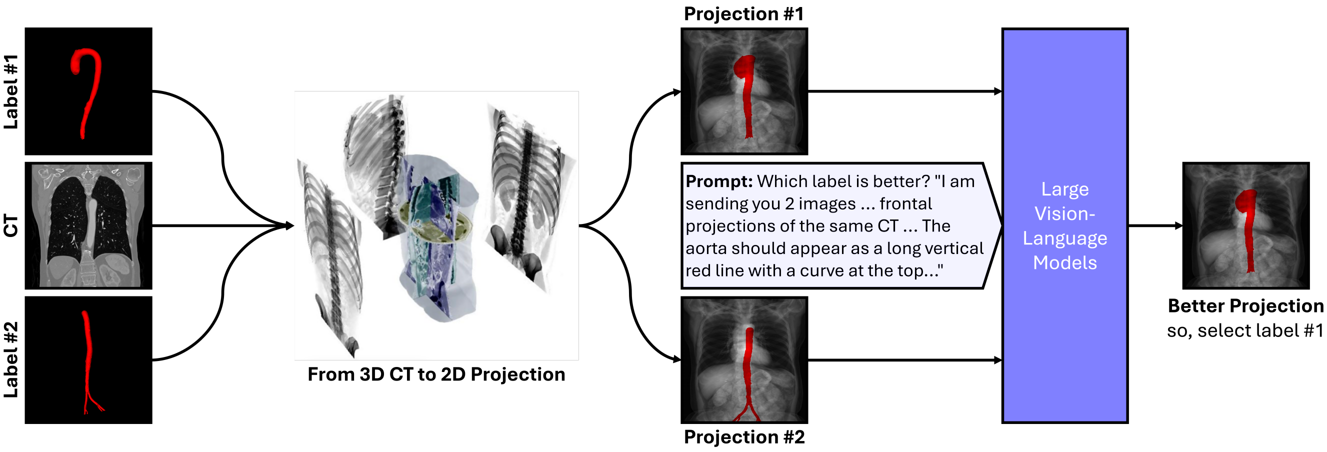
LLMs can semantically interpret language, consider context, and understand multiple ways a radiologist may describe a finding (e.g., a tumor may be referred to as a mass, neoplasm, lesion, growth, etc). With the assistance of radiologists, we have developed prompts that guide *Retriever* in interpreting reports, resulting in three main aspects:

1. **Step-wise approach.** To encourage step-by-step reasoning, we ask the LLM a sequence of questions, progressing from general to specific. For instance, to identify small pancreatic tumors, we first ask if the reports mention pancreatic tumors, then whether the tumor is malignant, and finally, the tumor's size. This structured approach builds a hierarchical database, starting with broad categories (e.g., pancreatic tumor) and narrowing down to specific details (e.g., malignant pancreatic tumor $\leq$ 2 cm). This hierarchy accelerates future searches; for example, to find large malignant pancreatic tumors, one only needs to search within reports already categorized as containing malignant pancreatic tumors.
2. **Medical Guidance.** With radiologist's support, we included in our prompts medical information and rules for report interpretation. For example, we provide the LLM names of pancreatic tumors that are benign, malignant or possibly both, and we explain other findings that may be confused with tumors, such as pancreatitis. Furthermore, we also explain how common expressions used by radiologists should be interpreted. E.g., 'unremarkable' indicates tumor absence, hypo- or hyperattenuation may indicate tumors, etc.
3. **Answer Template and Justification.** To easily process the LLM answer, we ask it to fill out templates, like: 'substitute __ by 'yes', 'no' or 'uncertain': liver tumor presence=__; kidney tumor presence=__; pancreas tumor presence=__'. For explainability, we also request the LLM a justification for its answers.

A.2. Label Expert: 39% Simple Errors Detected and Revised

method	spleen	kidney	gallbladder	liver	stomach	aorta	pancreas	prostate
Test-On-Training	28.0	24.6	48.8	8.3	38.4	10.0	14.0	52.8
Label Expert	26.4	50.1	44.4	32.3	30.8	33.0	74.2	46.0
method	duodenum	femur	esophagus	lung	bladder	rectum	average	
Test-On-Training	40.8	48.7	24.6	47.4	49.6	49.4	35.6	
Label Expert	37.7	48.0	27.1	35.3	48.6	48.1	39.3	

Table 6. **75% of label errors can be detected and revised by Test-on-Training (T-o-T) and Label Expert.** T-o-T replaces the current label when the pseudo label has no intersection with it (DSC = 0), while Label Expert replaces the current label when vision-language models (VLMs) judge the pseudo label to be superior. For this study, we used Qwen2-VL [83] as a demonstration. A total of 51,454 annotations were revised iteratively using the ScaleMAI agent. We report the percentage of label errors detected and revised by T-o-T and Label Expert, respectively. These two strategies detect different types of errors: T-o-T is effective for structures often absent in abdominal CT scans, such as the lung, prostate, rectum, and bladder. The current labels may generate false positives for these structures, whereas pseudo labels avoid these errors, leading to no overlap between them. Label Expert significantly improves label quality for structures with fixed and typical shapes, such as the pancreas and kidneys. VLMs excel at identifying shape errors when prompted with descriptive anatomical shapes or a few examples as in-context learning. In our study, the pancreas, for example, is prompted as '*a soft, elongated, comma-shaped organ nestled in the upper abdomen, extending horizontally from the duodenum to the spleen, with variable contours.*'



Prompt: Which label is better? "I am sending you 2 images ... frontal projections of the same CT ... The aorta should appear as a long vertical red line with a curve at the top..."

Figure 5. We propose *Label Expert*, a system that prompts a vision-language model (VLM) with anatomical knowledge to select better labels between two options. The hypothesis is that a majority of label errors are obvious even to non-professionals and can be detected by VLMs trained on diverse and extensive image-text datasets, given their strong performance in various image understanding tasks [5, 41, 47, 54, 71]. Examples of such obvious errors include organ misplacement, abnormal shapes, disconnections, multiple predictions for a single organ, noise artifacts, and label inconsistencies due to poor CT quality. Since pre-existing VLMs are trained on 2D natural images, they cannot directly analyze 3D CT scans. To address this, we project 3D CT scans and labels into 2D images, using a front-view projection. These projections resemble 2D X-rays with overlaid labels in red. The VLM then evaluates these projections with prompts designed to guide its decision-making. We use aorta as an example. The prompt teaches the VLM that '*aorta should appear as a long vertical red line with a curve at the top.*' When comparing two labels, the VLM determines that label #1 matches the description better than label #2. We found that prompt design significantly impacts performance. With carefully designed prompts, the VLM achieved 98% accuracy in selecting the better label. By automating the detection of obvious label errors through 34.7 million pair-wise comparison, Label Expert significantly reduced human review/revision efforts and corrected 39% of annotation errors. In contrast, traditional error detection methods miss around 80% of such obvious label errors, despite being straightforward for humans to identify in under two seconds per scan.

A.3. ROC Analysis for Tumor Annotation

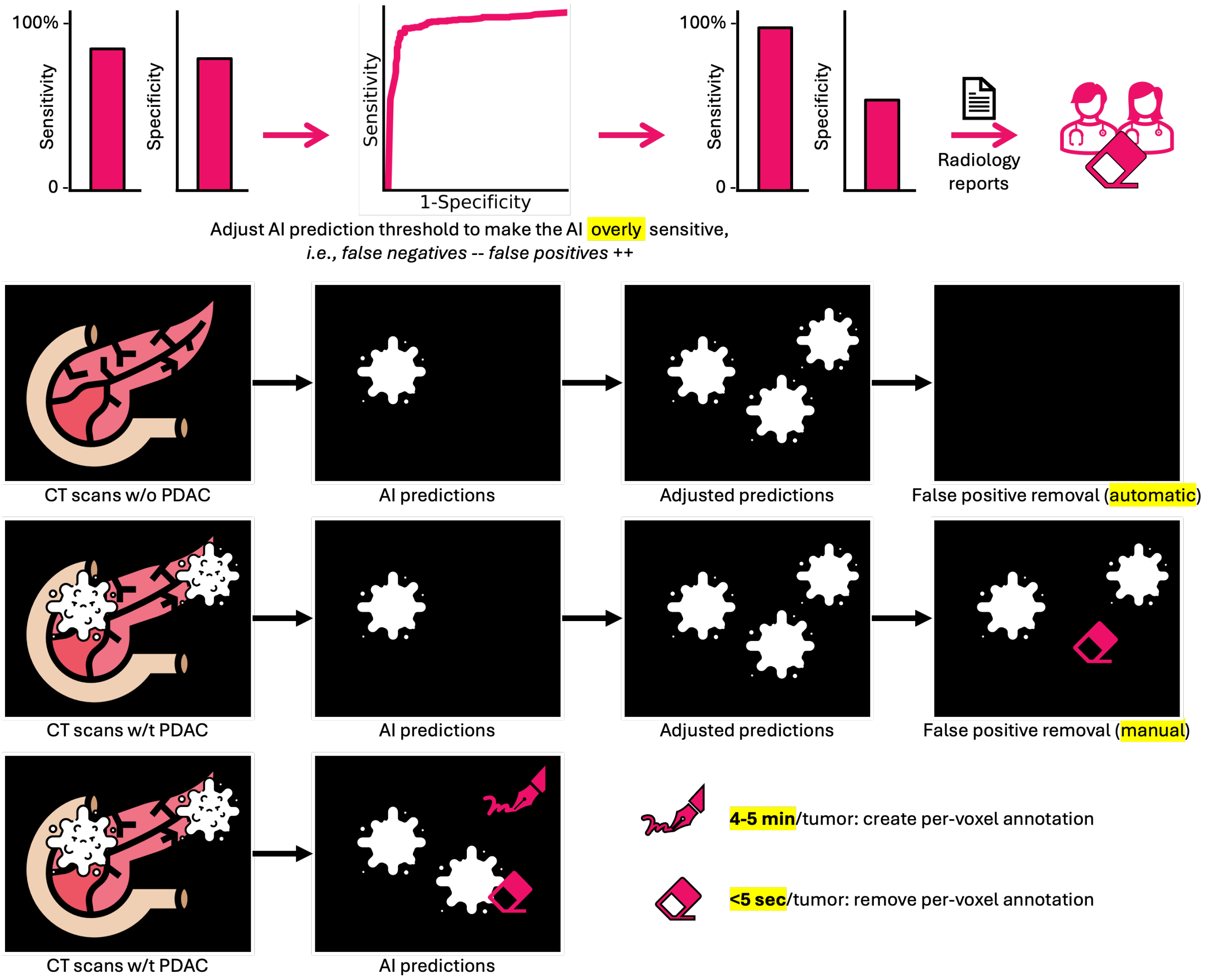

100%
0
Sensitivity
Specificity
Sensitivity
1-Specificity
Adjust AI prediction threshold to make the AI overly sensitive,
i.e., false negatives -- false positives ++
100%
0
Sensitivity
Specificity
Radiology reports
CT scans w/o PDAC
AI predictions
Adjusted predictions
False positive removal (automatic)
CT scans w/t PDAC
AI predictions
Adjusted predictions
False positive removal (manual)
CT scans w/t PDAC
AI predictions
4-5 min/tumor: create per-voxel annotation
<5 sec/tumor: remove per-voxel annotation

Figure 6. **ROC Analysis for Pancreatic Tumor Annotation.** We propose an efficient strategy, called ROC analysis, to assist radiologists in annotating tumors within a large-scale dataset (e.g., over 25,000 CT scans in our study). During the iterative data curation and annotation process facilitated by the ScaleMAI agent, AI model performance improves as data quality increases. In turn, stronger AI models generate more accurate pseudo labels with high sensitivity and specificity, significantly reducing radiologists' workload. Our observations show that removing AI false positives is much faster than creating per-voxel annotations for false negatives (missed tumors). Removing a false positive takes less than five seconds, whereas creating per-voxel annotations for a missed tumor can take 4–5 minutes. This insight motivates us to analyze the AI model's receiver operating characteristic (ROC) curve, which allows us to adjust the prediction threshold to prioritize sensitivity over specificity. To minimize radiologists' workload, we aim for nearly perfect sensitivity while maintaining acceptable specificity. By intentionally biasing the model towards high sensitivity, the AI minimizes missed tumors but inevitably introduces more false positives. Since handling false positives is simpler, this trade-off optimizes efficiency: (1) ***False positives in non-tumor CT scans*** can be automatically removed by cross-referencing radiology reports, which are typically available in clinical repositories (as illustrated in the second line in the Figure). (2) ***False positives in tumor CT scans*** can be efficiently removed using open-source annotation tools [13]. These tools enable radiologists to erase false positives with a few clicks, leveraging the AI's highly sensitive per-voxel predictions. In our study, we achieved 99% sensitivity for pancreatic tumor detection with only 0.6 false positives per scan. This means radiologists only have to remove just one false positive for every two CT scans (as illustrated in the third line in the Figure). Compared to creating per-voxel annotations from scratch, our ROC analysis approach reduces annotation time by up to 92%, significantly streamlining the workflow.

A.4. Data Mix and Data Annealing

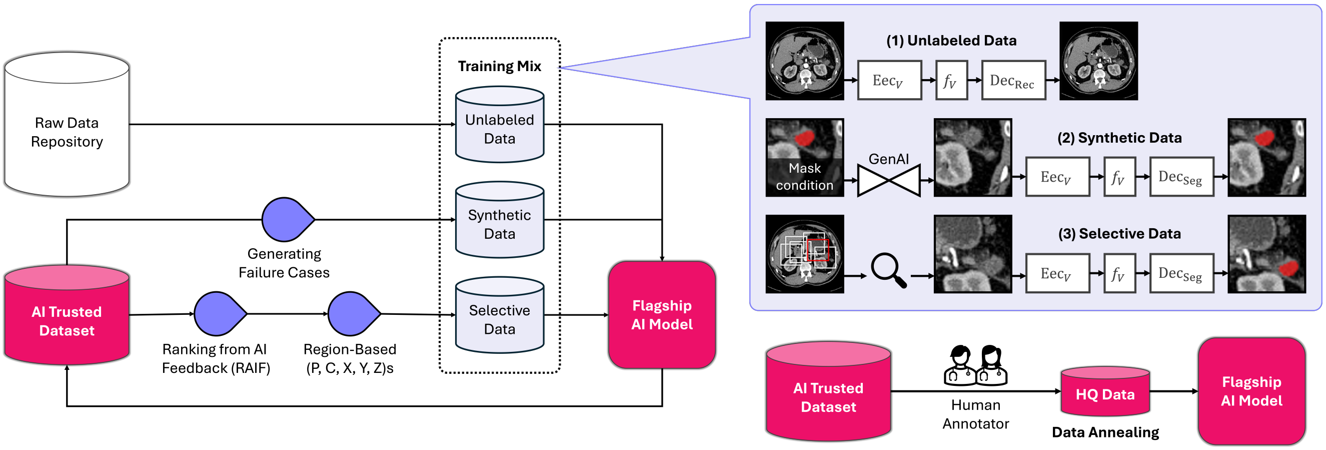


Figure 7. **Training Flagship Model with Data Mix and Data Annealing.** To optimize the training of Flagship Model, we incorporate a combination of data mix and data annealing strategies. The data mix consists of three primary types: ***First***, unlabeled data is utilized for self-supervised representation learning. This approach leverages the vast quantities of raw clinical data generated daily, requiring no manual annotation. The learned representations effectively regularize the model, enabling faster and more efficient learning of segmentation tasks with reduced reliance on annotated data. This methodology, supported by extensive literature, demonstrates the potential to exploit unlabeled clinical data for robust model training. ***Second***, synthetic data is employed to generate a diverse array of scans. These include variations across demographics, scanner types, and contrast enhancements, as well as tumors with differing locations, shapes, textures, sizes, and intensities that are not fully represented in the training set. This diversity enhances the model's robustness, particularly when encountering out-of-distribution test cases. ***Third***, selective data focuses on the most challenging regions of CT scans that confuse the model during training, as identified by the loss function. By prioritizing repeated sampling of these regions, the model learns more efficiently, avoiding the inefficiencies of processing non-informative areas such as air, bedding, or irrelevant anatomical regions. This targeted approach ensures that the model focuses on clinically relevant areas, such as the pancreas or abdominal region. ***Finally***, once the model is trained on data mix, we introduce data annealing to further fine-tune the model. We identify a gold-standard subset, consisting of voxel-level annotations meticulously created by human experts. This data annealing technique has proven effective in large-scale training efforts in other domains, such as GPT and Llama. However, in the medical field, the lack of gold-standard data and the predominance of silver-standard data have limited its exploration. When releasing the dataset, we will explicitly mark this gold-standard subset to facilitate further research and development in the field.

B. Quality Assessment of AI Trusted Datasets

B.1. Annotation Standard for Structures Adjacent to Pancreatic Tumors

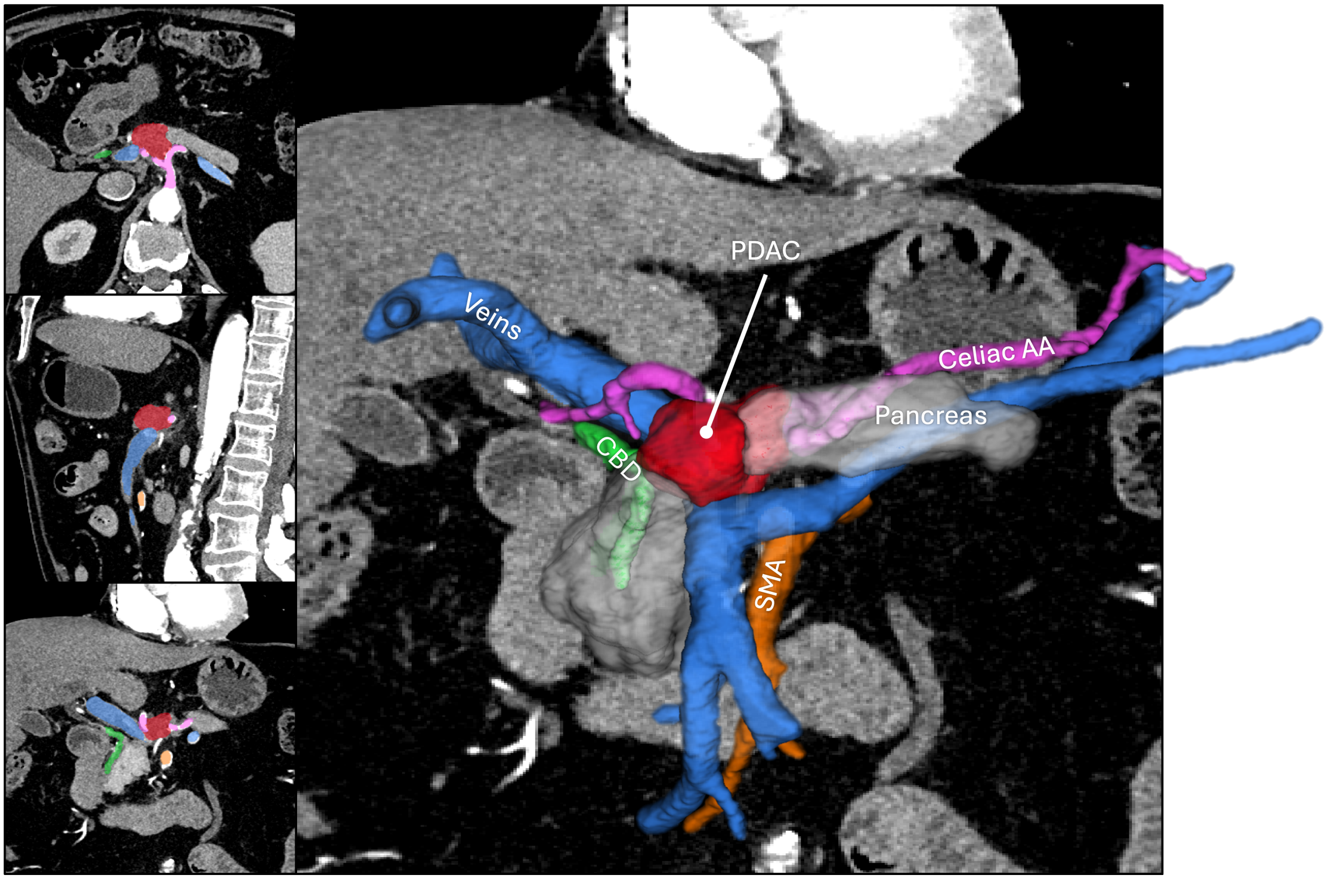
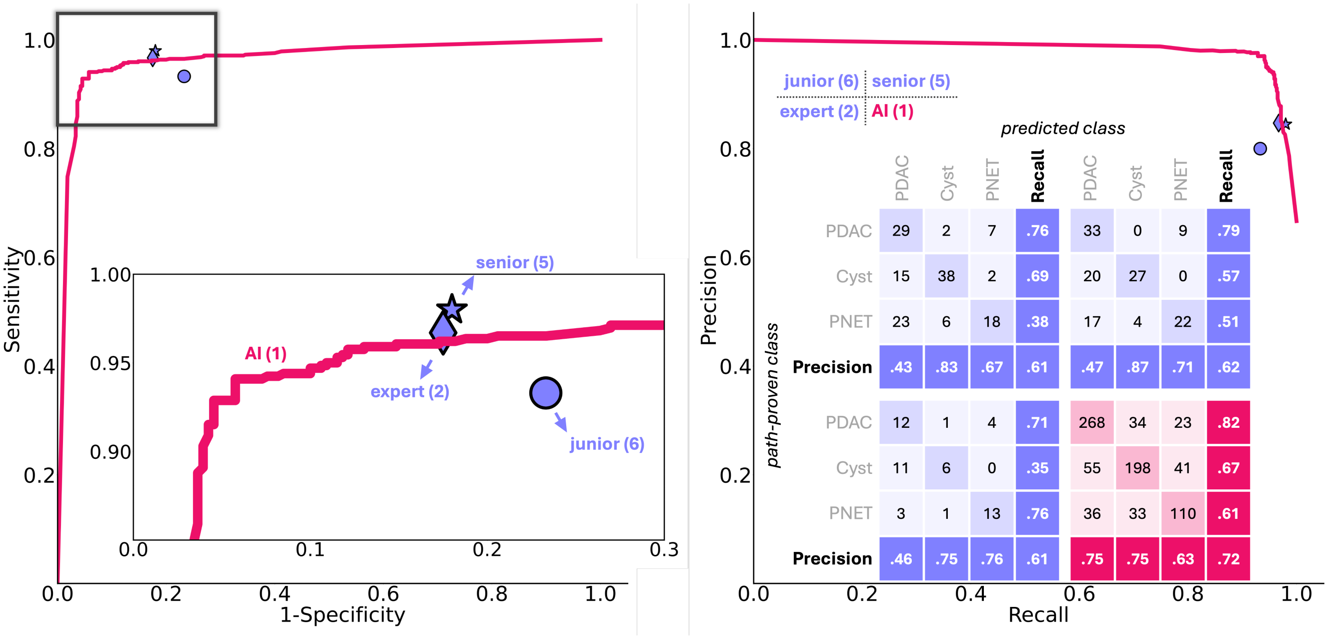


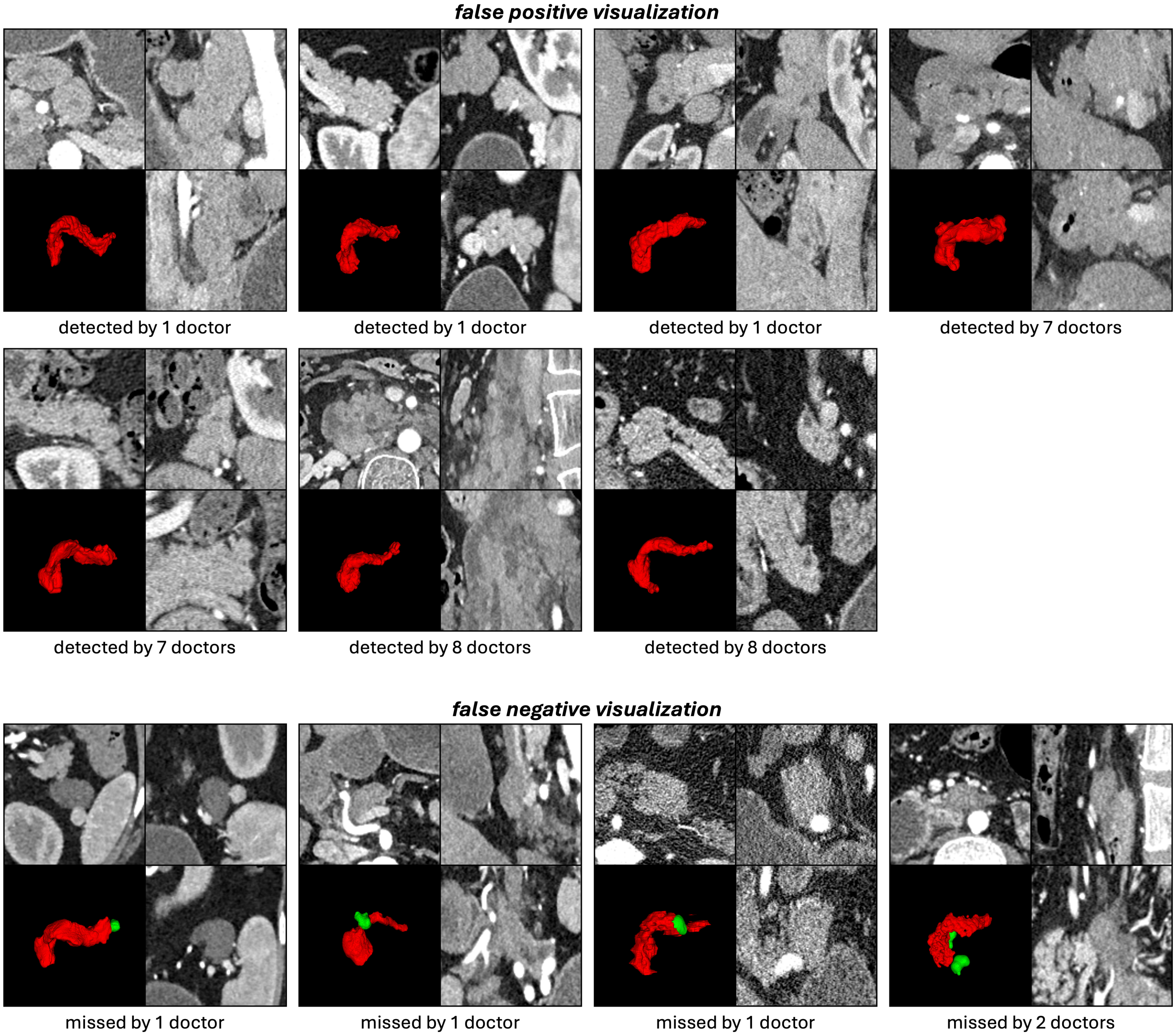
Figure 8. The annotated areas of all tubular structures include both the tube wall and the lumen but exclude surrounding tissues, such as organs, mesentery, and adipose tissue. The pancreatic duct is identified as a low-attenuation tubular structure within the pancreas and should be marked from the tail to the ampulla of Vater. The common bile duct (CBD) appears as a low-attenuation tubular structure and should be annotated from the confluence of the common hepatic duct and bile duct to the ampulla of Vater. The superior mesenteric artery (SMA) is highlighted as a bright structure originating from the aorta. Trace the SMA from its origin to the point where it branches. The celiac artery is a short vessel that can be identified branching from the aorta into the left gastric, splenic, and common hepatic arteries. Annotate from its origin to where it branches. Veins include the portal vein, splenic vein and superior mesenteric vein. The portal vein is a bright, enhanced vessel formed by the confluence of the superior mesenteric and splenic veins. It should be traced from this confluence to its entry into the liver. The splenic vein runs behind the pancreas and should be marked from its origin at the spleen to where it merges with the superior mesenteric vein. The superior mesenteric vein (SMV) merges with the splenic vein to form the portal vein. The mask encompasses the area from the branches of these veins to the confluence of the splenic vein. The masks of different structures are displayed in various colors. The three images on the left row demonstrate the masks on axial, sagittal, and coronal planes. On the right, the mask effect is illustrated with a 3D rendering on the coronal plane.

B.2. Reader Study: Tumor Detection & Classification



	PDAC	Cyst	PNET	Recall	PDAC	Cyst	PNET	Recall
PDAC	29	2	7	.76	33	0	9	.79
Cyst	15	38	2	.69	20	27	0	.57
PNET	23	6	18	.38	17	4	22	.51
Precision	.43	.83	.67	.61	.47	.87	.71	.62
PDAC	12	1	4	.71	268	34	23	.82
Cyst	11	6	0	.35	55	198	41	.67
PNET	3	1	13	.76	36	33	110	.61
Precision	.46	.75	.76	.61	.75	.75	.63	.72

Figure 9. **Flagship Model matches senior and expert radiologists in tumor detection and surpasses them in tumor classification accuracy.** We conducted an extensive multi-institution, multi-reader study comparing Flagship Model with radiologists with varying levels of experience. Thirteen board-certified radiologists were participated, including 6 juniors (<8 years of experience), 5 seniors (8–15 years), and 2 experts (>15 years). Each radiologist independently reviewed contrast-enhanced abdominal CT scans in the venous and arterial phases from 50 patients (100 CT scans), representing a broad spectrum of pancreatic conditions, including normal cases and tumors of three common subtypes: cysts, pancreatic adenocarcinoma (PDAC), and pancreatic neuroendocrine tumors (PNET). Radiologists were blinded to the proportion of normal and tumor cases and tasked with detecting and localizing tumors using 3D Slicer by marking any point within the tumor. They also classified tumors into the specified subtypes without access to patient medical history or symptom information. Flagship Model was evaluated under identical conditions on a larger cohort of 982 patients (1,964 CT scans). Performance was assessed using Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves for tumor detection and confusion matrices for classification. For tumor detection, Flagship Model (pink curve) achieved performance comparable to expert (blue diamond) and senior radiologists (blue star), outperforming junior radiologists (blue circle). In tumor classification of PDAC, cysts, and PNET, Flagship Model achieved 72% accuracy, exceeding junior, senior, and expert radiologists by 11%, 10%, and 11%, respectively.



false positive visualization

detected by 1 doctor

detected by 1 doctor

detected by 1 doctor

detected by 7 doctors

detected by 7 doctors

detected by 8 doctors

detected by 8 doctors

false negative visualization

missed by 1 doctor

missed by 1 doctor

missed by 1 doctor

missed by 2 doctors

Figure 10. **Visualization of false positives and negatives predicted by radiologists.** In the false positive cases, the radiologists noticed slight irregularities in the pancreas tissue texture. However, these cases lacked two key reliable warning signs that typically indicate pancreatic tumor: abnormal widening of the main pancreatic duct and localized tissue shrinkage. The false negative cases demonstrated more subtle findings. One case showed a tumor growing outward from the tail end of the pancreas—a location that is often difficult for human readers if not examined thoroughly. In two other cases, while no obvious tumors were visible, there were areas where the pancreas tissue had become unusually thin, which often signals an underlying tumor in that location.

		Sensitivity, %				Specificity, %
career stage	reader	all-size	small (<20mm)	medium (20–40mm)	large (>40mm)	normal
junior (<8 years)	Reader 1	96.7 (29/30)	100 (10/10)	90.0 (9/10)	100 (10/10)	75.0 (15/20)
	Reader 2	100 (30/30)	100 (10/10)	100 (10/10)	100 (10/10)	75.0 (15/20)
	Reader 3	100 (30/30)	100 (10/10)	100 (10/10)	100 (10/10)	80.0 (16/20)
	Reader 4	96.7 (29/30)	100 (10/10)	90.0 (9/10)	100 (10/10)	80.0 (16/20)
	Reader 5	76.7 (23/30)	90.0 (9/10)	80.0 (8/10)	60.0 (6/10)	85.0 (17/20)
	Reader 6	90.0 (27/30)	100 (10/10)	90.0 (9/10)	80.0 (8/10)	65.0 (13/20)
	average	93.3 (168/180)	98.3 (59/60)	91.7 (55/60)	90.0 (54/60)	76.7 (92/120)
senior (8–15 years)	Reader 7	96.7 (29/30)	100 (10/10)	90.0 (9/10)	100 (10/10)	80.0 (16/20)
	Reader 8	100 (30/30)	100 (10/10)	100 (10/10)	100 (10/10)	80.0 (16/20)
	Reader 9	96.7 (29/30)	100 (10/10)	90.0 (9/10)	100 (10/10)	80.0 (16/20)
	Reader 10	100 (30/30)	100 (10/10)	100 (10/10)	100 (10/10)	80.0 (16/20)
	Reader 11	96.7 (29/30)	100 (10/10)	90.0 (9/10)	100 (10/10)	90.0 (18/20)
	average	98.0 (147/150)	100 (50/50)	94.0 (47/50)	100 (50/50)	82.0 (82/100)
expert (>15 years)	Reader 12	93.3 (28/30)	100 (10/10)	80.0 (8/10)	100 (10/10)	80.0 (16/20)
	Reader 13	100 (30/30)	100 (10/10)	100 (10/10)	100 (10/10)	85.0 (17/20)
	average	96.7 (58/60)	100 (20/20)	90.0 (18/20)	100 (20/20)	82.5 (33/40)
	Flagship Model	94.1 (640/680)	85.2 (468/549)	100 (105/105)	100 (10/10)	83.8 (253/302)

Table 7. **Flagship Model matches the pancreatic tumor detection performance of senior and expert radiologists, outperforming junior radiologists.** We present the sensitivity (%) and specificity (%) of pancreatic tumor detection across different tumor sizes—small (<20 mm), medium (20–40 mm), large (>40 mm), and all sizes—for radiologists at various career stages and Flagship Model. 13 radiologists of varying experience levels each evaluated 50 patients individually, while Flagship Model was tested on 982 patients. For all tumor sizes, all radiologist groups demonstrated high sensitivity, but only senior and expert radiologists achieved good specificity. Flagship Model surpassed all radiologist groups in specificity and had higher sensitivity than junior radiologists, approaching the performance of senior and expert radiologists. Notably, Flagship Model attained 100% sensitivity for medium and large tumors, suggesting it could assist all radiologists in detecting medium-sized tumors and help junior radiologists with large tumor detection. These findings indicate that Flagship Model performs at a level comparable to experienced radiologists, highlighting its potential as a reliable tool in clinical practice.

B.3. PancreaVerse vs. Public Tumor/Organ Datasets

dataset [year] [source]	# of CTs	# of classes	# of hospitals	annotation standard	main clinical need
organ datasets					
1. CHAOS [2018] [link]	40	1	1	gold	liver segmentation
2. BTCV [2015] [link]	50	13	1	gold	abdomen segmentation
3. Pancreas-CT [2015] [link]	82	1	1	gold	pancreas segmentation
4. CT-ORG [2020] [link]	140	6	8	silver	abdomen segmentation
5. WORD [2021] [link]	170	16	1	gold	abdomen segmentation
6. AMOS22 [2022] [link]	500	15	2	silver	abdomen segmentation
7. AbdomenCT-1K [2021] [link]	1,112	4	12	silver	abdomen segmentation
8. TotalSegmentator [2023] [link]	1,228	117	1	silver	anatomic structures segmentation
9. Trauma Detect. [2024] [link]	4,274	6	23	silver	traumatic abdominal injuries
tumor datasets					
10. TCGA-SARC [2016] [link]	5	1	1	-	sarcomas cancer analysis
11. TCGA-KICH [2016] [link]	15	1	1	-	kidney chromophobe analysis
12. TCGA-KIRP [2016] [link]	33	1	1	-	kidney renal papillary cell carcinoma analysis
13. CTpred-Sunitinib-panNET [2023] [link]	38	1	1	-	pancreas cancer classification
14. Pancreatic-CT-CBCT-SEG [2021] [link]	40	7	1	gold	pancreatic cancer segmentation
15. TCGA-STAD [2016] [link]	46	1	1	-	stomach adenocarcinoma analysis
16. MSD Spleen [2022] [link]	61	1	1	gold	spleen segmentation
17. CPTAC-SAR [2019] [link]	88	1	1	-	sarcomas cancer analysis
18. MSD Lung [2022] [link]	96	1	1	gold	lung tumor segmentation
19. TCGA-LIHC [2016] [link]	97	1	1	-	liver hepatocellular carcinoma analysis
20. HCC-TACE-Seg [2021] [link]	105	1	1	gold	liver tumor segmentation
21. TCGA-BLCA [2016] [link]	120	1	1	-	bladder endothelial carcinoma analysis
22. TCGA-OV [2016] [link]	143	1	1	-	ovarian serous cystadenocarcinoma analysis
23. CPTAC-PDA [2018] [link]	168	2	1	gold	pancreatic ductal adenocarcinoma
24. CT Lymph Nodes [2015] [link]	176	2	1	gold	lymph nodes segmentation
25. MSD Colon [2022] [link]	190	1	1	gold	colon tumor segmentation
26. MSD Liver [2022] [link]	201	2	7	gold	liver tumor segmentation
27. LiTS [2019] [link]	201	2	7	gold	liver tumor segmentation
28. PCL[2021] [link]	221	3	1	silver	pancreatic cystic lesions segmentation
29. StageII-Colorectal-CT [2022] [link]	230	1	1	-	colorectal cancer analysis
30. CPTAC-UCEC [2019] [link]	250	1	1	-	corpus endometrial carcinoma analysis
31. CPTAC-CCRCC [2018] [link]	262	1	1	-	clear cell carcinoma analysis
32. TCGA-KIRC [2016] [link]	267	1	1	-	kidney renal clear cell carcinoma analysis
33. TCIA-LDCT [2020] [link]	299	1	1	-	various tumor (Head/Chest/Abdomen) analysis
34. MSD Pancreas [2022] [link]	420	2	1	gold	pancreas tumor segmentation
35. MSD Hepatic Vessels [2022] [link]	443	2	1	gold	liver vessels segmentation
36. Med-Lymph-Node-SEG [2024] [link]	513	1	3	gold	lymph node segmentation
37. KiTS23 [2020] [link]	599	3	2	gold	kidney tumor segmentation
38. TCIAColon [2015] [link]	825	1	1	-	colon cancer analysis
39. autoPET [2022] [link]	1,014	1	2	gold	tumor lesions segmentation
40. PANORAMA [2024] [link]	3,000	6	7	silver	pancreatic tumor diagnosis
41. FLARE23 [2024] [link]	4,650	14	50	silver	abdomen and pan-cancer segmentation
PancreaVerse	**25,362**	**24**	**112**	**silver**	**tumor diagnosis, staging & planning**

Table 8. **Comparison of PancreaVerse with public datasets.** An AI Trusted Dataset is defined as large-scale, high-quality, and multi-source, reflecting real-world clinical scenarios and tailored to clinical needs. By these criteria, PancreaVerse demonstrates significant advancements: **Scale:** 25,362 CT scans, making it 8.5 × larger than the largest existing datasets for pancreatic tumor detection. **Quality:** Silver-standard tumor annotations, validated by a group of eight expert radiologists. **Diversity:** CT scans sourced from 112 global hospitals, offering 3 × more diversity than current datasets. *We will make our PancreaVerse publicly available.*

C. Experimental Results of Flagship AI Model

C.1. Benchmarking Flagship AI Model

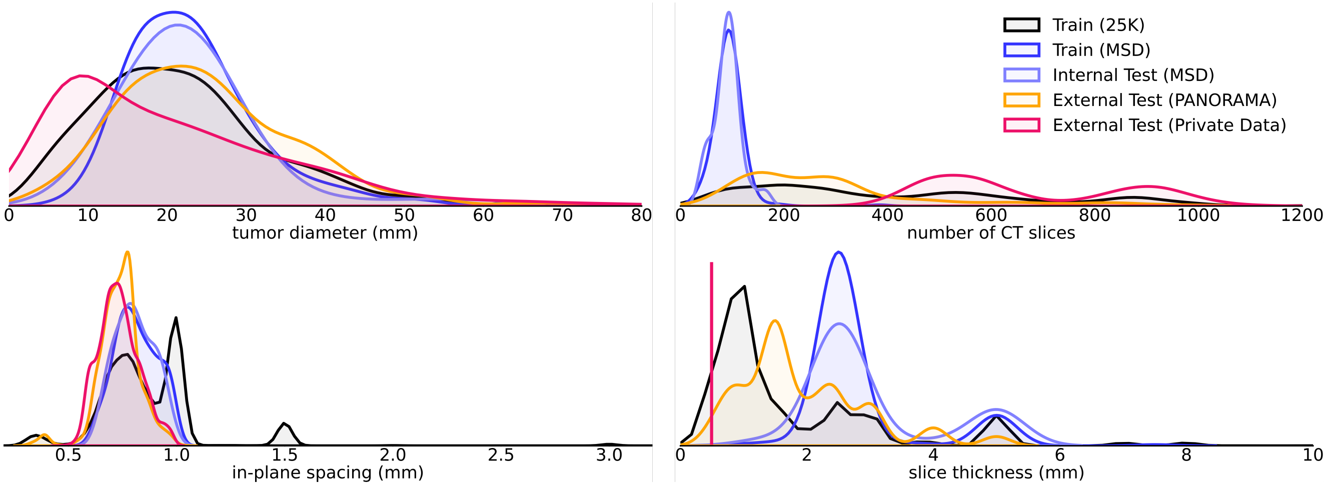
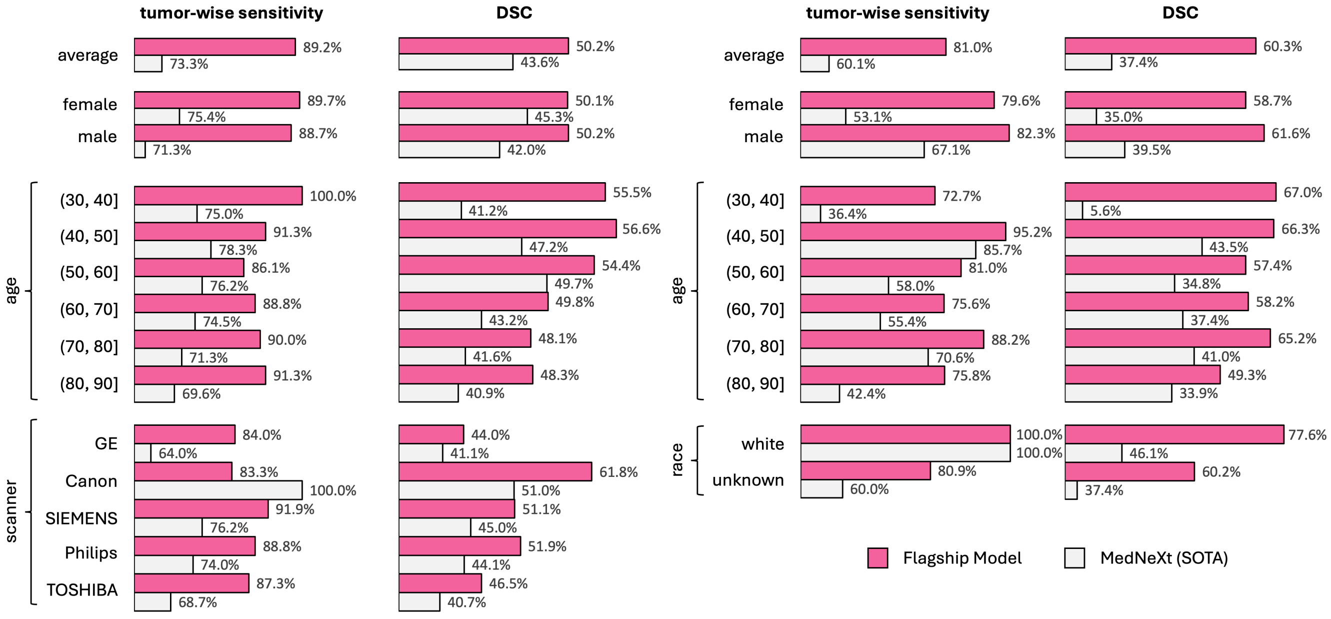


Figure 11. **Dataset attributes.** Our study used five datasets: two for training AI models and three for testing them. Before the creation of PancreaVerse (25K cases), the publicly available MSD-Pancreas dataset was the *only* resource for training pancreatic tumor segmentation models. Therefore, all baseline AI models in this study were trained and tested on MSD-Pancreas. Since the training and test sets in MSD-Pancreas were randomly sampled from the same dataset, their CT scans exhibit high similarity. This inherent similarity is expected to enhance AI performance on such tests. Existing literature suggests that AI is vulnerable when applied to CT scans from datasets with differing attributes [25, 78], such as variations in tumor diameter, the number of CT slices, in-plane spacing, and slice thickness. To evaluate the robustness of the baseline models, we conducted external validation using two additional datasets sourced from hospitals distinct from those contributing to MSD-Pancreas (Memorial Sloan Kettering Cancer Center, USA). The first external dataset, PANORAMA, was sourced from five hospitals across three countries, including Dutch, Sweden and Norway [6]. The second dataset, a proprietary JHH dataset, was gathered from hospitals in a country distinct from MSD-Pancreas. As seen, the distributions of tumor diameter, number of CT slices, in-plane spacing, and slice thickness, in the MSD-Pancreas test set are similar to those of its training set. In contrast, the PANORAMA dataset differs in the number of CT slices and slice thickness, while the proprietary JHH dataset diverges significantly across all four attributes. The test results in Table 5 reveal that baseline models perform best on the MSD-Pancreas test set, followed by PANORAMA, with the proprietary JHH dataset yielding the lowest performance. These findings validate the hypothesis that discrepancies between training and test data significantly impact AI performance and robustness. This motivated us to create PancreaVerse that offers a substantially larger training set (25,362 annotated CT scans vs. MSD-Pancreas's 200) with greater diversity in key attributes as illustrated by the black curve. This diversity enhances the generalizability of models trained on it. As evidenced in Table 5, models trained on our trusted dataset outperform all baseline models by at least 5% DSC in the internal test and by 7% and 22% in the two external tests, i.e., PANORAMA and proprietary JHH datasets, respectively.

C.2. Pancreatic Tumor Detection (+14% Sensitivity)

		MSD-Pancreas (N=81)	PANORAMA (N=1,964)		proprietary JHH dataset (N=1,958)	
method	training set	Sensitivity	Sensitivity	Specificity	Sensitivity	Specificity
Swin UNETR [80]	MSD-Pancreas	97.5 (79/81)	**98.4 (569/578)**	11.0 (152/1386)	88.1 (1176/1335)	16.7 (104/623)
UniSeg [86]	MSD-Pancreas	92.6 (75/81)	89.4 (517/578)	54.8 (760/1386)	69.1 (922/1335)	78.5 (485/623)
ResEncL [38]	MSD-Pancreas	76.5 (62/81)	83.0 (480/578)	84.6 (1173/1386)	66.1 (883/1335)	87.0 (542/623)
STU-Net-Base [35]	MSD-Pancreas	81.5 (66/81)	82.9 (479/578)	84.0 (1164/1386)	65.6 (876/1335)	84.4 (526/623)
MedNeXt [76]	MSD-Pancreas	79.0 (64/81)	81.1 (469/578)	**84.8 (1176/1386)**	68.8 (919/1335)	85.2 (531/623)
Flagship Model	PancreaVerse	**98.8 (80/81)**	92.0 (532/578)	73.2 (1015/1386)	**90.9 (1214/1335)**	**88.3 (550/623)**

Table 9. **Flagship Model, with a backbone of ResEncL, achieves the best performance for pancreatic tumor detection.** Note that these are patient-wise detection results. The out-of-distribution sensitivity of Flagship Model surpasses the in-distribution sensitivity of existing AI models. Although Swin UNETR achieves top-1 for sensitivity in the PANORAMA dataset, its low specificity denotes its suboptimal performance. Performance is given as sensitivity and specificity. Best-performing results are **bolded** for each dataset.



(a) PANORAMA

(b) private dataset

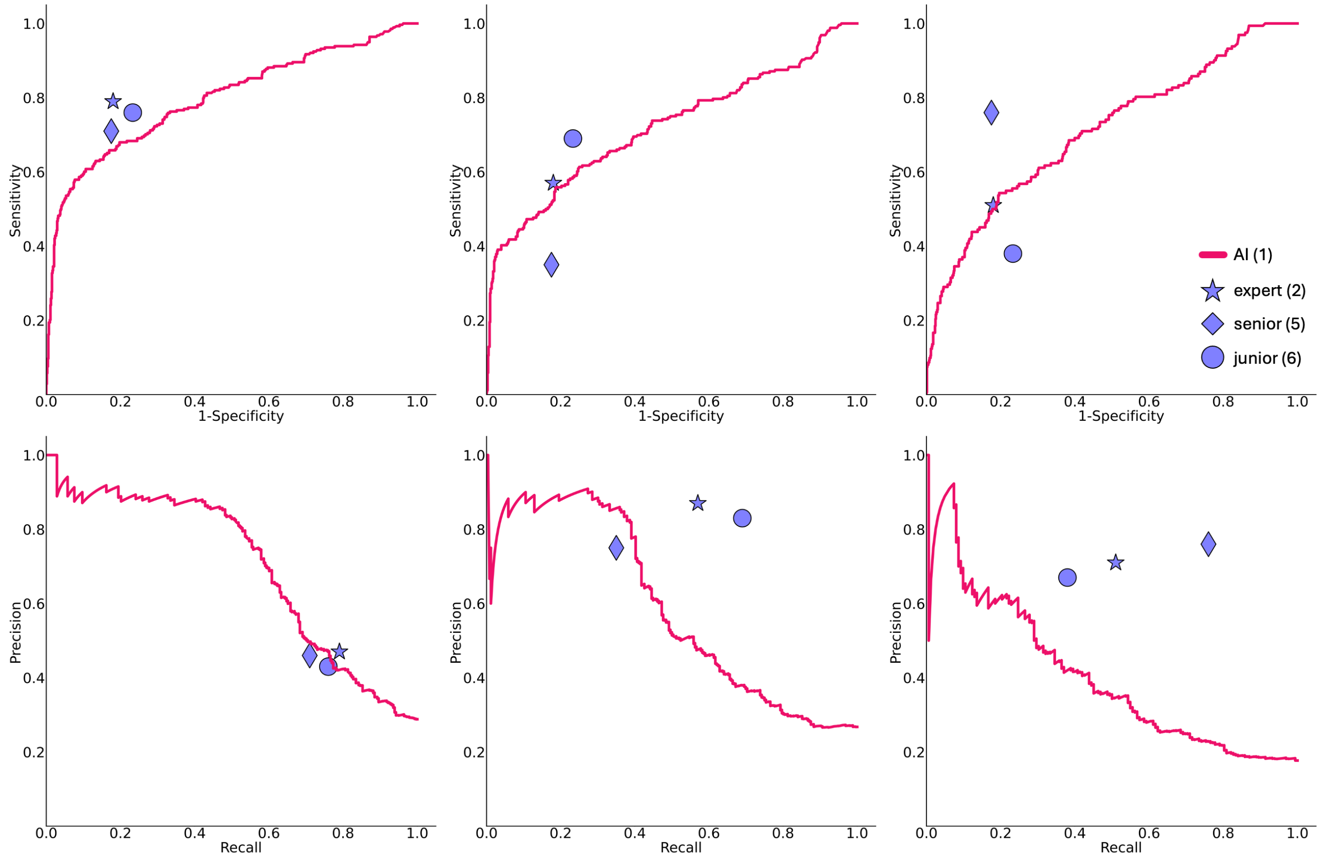
Figure 12. **Flagship Model demonstrates robust generalizability across diverse demographic and technical variations in out-of-distribution evaluations.** Flagship Model shows enhanced tumor detection and segmentation across various age groups, sex, scanner types, and race consistently outperforming the top-1 performing MedNeXt model—trained on a smaller, fixed-quality dataset. Notably, in the PANORAMA dataset, Flagship Model surpasses MedNeXt in all patient groups except those scanned with Canon scanners. In the proprietary JHH dataset, Flagship Model surpasses MedNeXt in all patient groups except both of them achieved 100% accuracy when detecting tumors in the white patient group.

C.3. Pancreatic Tumor Segmentation (+5% DSC)

	Median DSC (IQR) on PANORAMA, %				Median DSC (IQR) on JHH, %			
group	Flagship Model	MedNeXt (SOTA)	difference	p-value	Flagship Model	MedNeXt (SOTA)	difference	p-value
all test samples	58.1 (24.4–76.2)	54.4 (0.0–74.8)	0.1 (-11.2–20.1)	<0.001	70.4 (44.4–81.3)	38.8 (0.0–67.7)	15.2 (2.7–40.8)	<0.001
sex								
female	57.5 (25.0–77.0)	55.9 (0.6–76.3)	0.0 (-12.7–18.7)	0.072	68.8 (39.8–79.8)	31.4 (0.0–68.0)	14.0 (1.5–41.2)	<0.001
male	58.9 (23.9–74.8)	53.4 (0.0–72.9)	0.7 (-9.8–21.0)	0.002	71.4 (49.2–82.5)	45.7 (0.3–67.5)	15.5 (3.5–40.1)	<0.001
age								
30–40	66.6 (40.4–81.7)	44.0 (12.9–72.3)	4.9 (-1.4–20.6)	0.608	70.5 (63.6–73.2)	0.1 (0.0–12.5)	62.8 (54.6–69.9)	<0.001
40–50	66.9 (52.9–82.3)	57.4 (8.8–74.2)	0.6 (-2.8–15.8)	0.325	74.8 (65.0–84.0)	54.0 (16.3–68.9)	17.7 (7.6–58.2)	0.012
50–60	63.5 (31.1–80.2)	60.0 (16.9–77.9)	0.0 (-14.6–19.2)	0.299	65.4 (42.1–78.6)	36.5 (0.0–62.7)	16.0 (1.1–42.4)	<0.001
60–70	57.5 (26.2–75.9)	54.1 (0.0–71.9)	0.0 (-11.7–20.8)	0.038	68.0 (38.0–81.5)	41.1 (0.0–66.8)	13.8 (3.1–35.7)	<0.001
70–80	55.8 (21.6–74.5)	50.1 (0.0–74.4)	0.0 (-11.1–19.2)	0.039	73.9 (57.0–83.1)	47.8 (0.7–71.9)	16.1 (3.8–40.6)	<0.001
80–90	55.4 (24.2–70.7)	44.2 (0.0–75.4)	0.4 (-6.3–25.0)	0.263	57.3 (9.3–78.7)	31.7 (0.0–66.8)	8.2 (0.0–18.7)	<0.001
scanner								
GE	47.1 (20.4–69.4)	54.5 (0.0–74.2)	0.0 (-4.1–10.0)	0.754	-	-	-	-
Canon	70.2 (54.5–84.1)	54.0 (27.3–76.0)	13.4 (1.9–16.5)	0.582	-	-	-	-
SIEMENS	57.7 (26.4–76.9)	57.3 (0.8–75.7)	0.5 (-12.9–21.9)	0.072	-	-	-	-
Philips	59.8 (30.0–75.9)	55.4 (0.0–72.0)	0.2 (-11.6–22.4)	0.006	-	-	-	-
TOSHIBA	57.3 (16.1–75.4)	44.1 (0.0–75.5)	0.0 (-8.3–17.6)	0.157	-	-	-	-
race								
white	-	-	-	-	77.6 (76.6–78.5)	46.1 (43.0–49.3)	31.4 (29.3–33.6)	0.102
unknown	-	-	-	-	70.3 (44.0–81.3)	38.6 (0.0–67.8)	15.0 (2.5–40.9)	<0.001

Table 10. **Flagship Model demonstrates robust generalizability across various demographic groups and scanner types in tumor segmentation.** We compare the median DSC and interquartile range (IQR) of Flagship Model and public SOTA MedNeXt model on the PANORAMA and proprietary JHH datasets for sex, age, scanner type, and race. Notably, Flagship Model consistently achieves higher median DSC with statistically significant improvements (p-value < 0.001 in most cases). For example, in the age group 70–80, Flagship Model achieved a median DSC of 73.9% compared to MedNeXt's 47.8%, a difference of 16.1% (p-value < 0.001).

C.4. PDAC, Cyst, and PNET Classification (72% Accuracy)



(a) PDAC

(b) Cyst

(c) PNET

Figure 13. **Our Flagship Model can approach radiologists' performance in pancreatic tumor classification.** Classifying pancreatic tumor types (Cyst, PDAC, PNET) directly from CT scans is challenging due to the subtle and overlapping visual features among these tumors. Key characteristics such as shape, size, and enhancement patterns can vary significantly within the same tumor type and often mimic those of other types. Additionally, CT scans lack the biological and molecular context provided by patient symptoms, medical history, follow-up imaging, or biopsy results, which are crucial for accurate diagnosis. Furthermore, variations in imaging protocols and scanner settings across institutions add complexity, making it difficult for both radiologists and AI models to achieve high accuracy. Using our annotated dataset of tumor types, this study marks the first time that: (1) AI performance is evaluated on a publicly available dataset, enabling reproducibility. (2) Radiologists are tested on the same dataset, allowing others to benchmark their performance. (3) AI is directly compared with radiologists across different career stages.

C.5. Tumor Staging (53% Accuracy)

Pancreatic tumor staging is essential for determining surgical options and is based on assessing tumor size, location, involvement of nearby vessels, and the presence of metastasis. The tumor's relationship to major vessels—such as the celiac trunk, superior mesenteric artery (SMA), portal vein (PV), splenic vein (SV), and superior mesenteric vein (SMV)—is a key criterion for evaluating resectability, as shown in Figure 14. Accurate assessment requires detailed annotations of the pancreas and surrounding vascular structures. Existing AI models are limited by sparsely annotated datasets, hindering effective segmentation of these critical areas. Leveraging PancreaVerse with comprehensive vascular annotations, our Flagship Model effectively segments both the pancreas and surrounding vessels, enabling tumor staging estimation by measuring tumor size and its contact with related vessels (e.g., SMA).

We first dilated the pancreatic tumor mask to ensure overlap with adjacent vessel masks. We then aligned the vessels along the x-axis using Principal Component Analysis (PCA) and calculated the percentage of overlap between the tumor and vessel borders. The maximum overlap along the vessel was multiplied by 360 to obtain an angle measurement. Tumor staging was determined as follows: if the maximum contact between the pancreatic ductal adenocarcinoma (PDAC) and key vessels—the superior mesenteric artery/vein (SMA/SMV), and celiac artery (CA)—reached 180 degrees or more, the tumor was classified as T4. Otherwise, staging was based on tumor size thresholds: T1 (<20 mm), T2 (20–40 mm), and T3 (>40 mm). Evaluated on 30 CT scans with staging metadata, we achieved an accuracy of 53% for classify tumor stage (T1, T2, T3, T4), demonstrating the model's potential to aid pancreatic tumor staging in clinical applications.

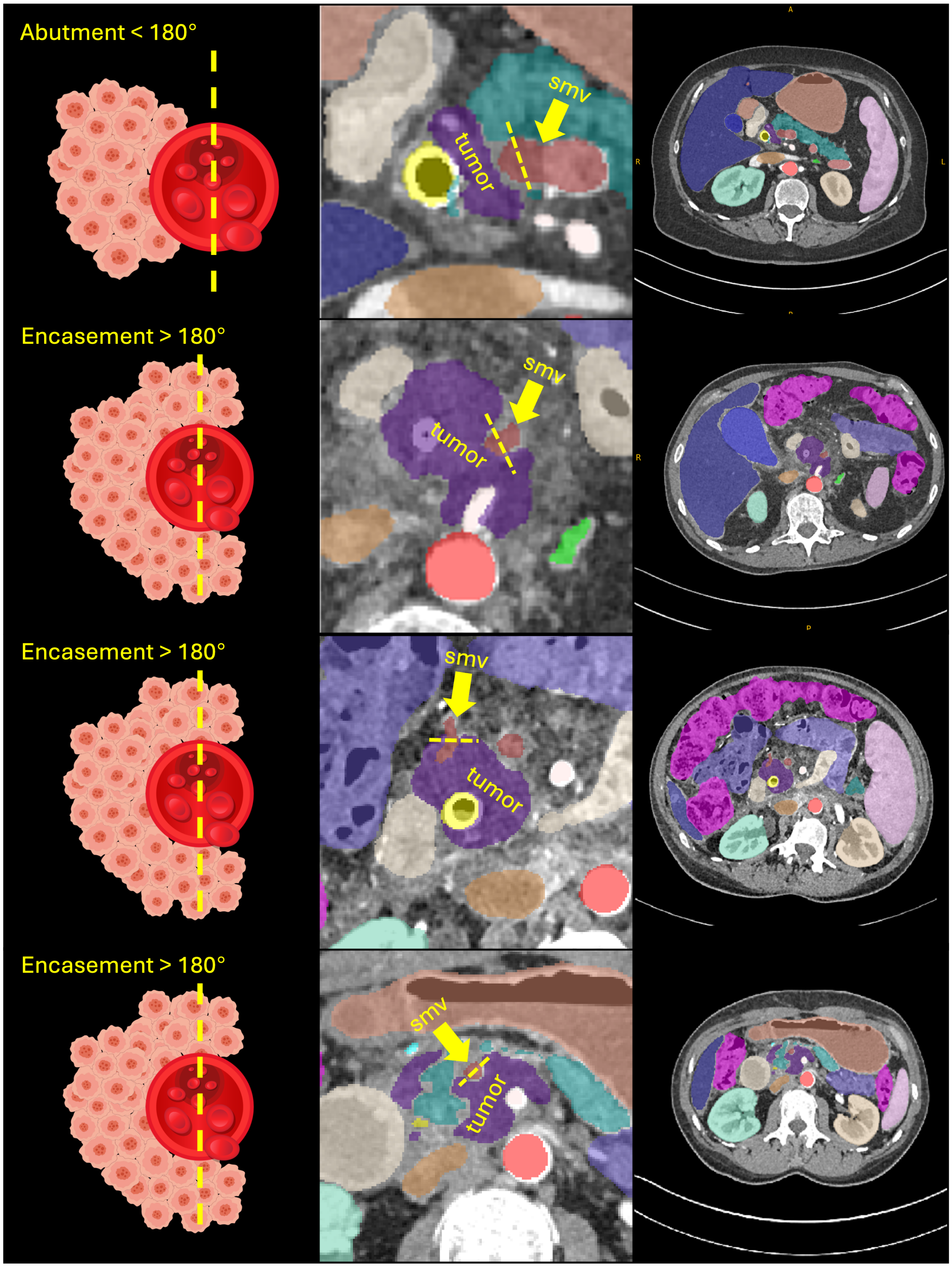


Figure 14. **We calculated pancreatic tumor stages based on tumor size and contact with surrounding vessels, all segmented by Flagship Model.** In the top example, the tumor is in contact with the superior mesenteric vein (SMV) but does not encase it by 180°, classifying it as resectable. In the bottom example, the tumor encircles more than 180° of SMV, making it unresectable. Our model enables precise segmentation of vital structures, such as SMV and tumors, which are crucial for evaluating tumor staging and resectability. This reduces the manual annotation workload of radiologists and aids in accurate staging assessments.

C.6. Radiotherapy Planning

In pancreatic tumor stereotactic body radiotherapy, treatment planning CT scans are acquired prior to the start of treatment. Unlike diagnostic CTs, which focus on disease detection and characterization, planning CTs are specifically used for beam optimization and precise radiation dose calculation. The radiotherapy treatment planning process begins with an accurate delineation of the target tumor and surrounding normal tissues, such as the bowel, stomach, duodenum, kidneys, and spinal cord. However, manual annotation by radiologists can be time-consuming and challenging, particularly for complex structures like the duodenum due to their intricate nature. The integration of Flagship Model can significantly reduce the manual workload and improve delineation accuracy. For instance, as shown in Figure 15, while radiologists may miss parts of the duodenum during annotation, Flagship Model can accurately identify and segment the entire structure. This comprehensive segmentation enables more precise treatment planning, reducing radiation exposure to critical structures and minimizing potential collateral damage. Following confirmation of these segmentation, the plans are refined to meet target objectives and adhere to normal tissue constraints before being delivered on a linear accelerator. This use of AI-driven multi-organ segmentation enhances treatment accuracy, decreases the risk of damage to essential organs, and supports better overall patient outcomes.

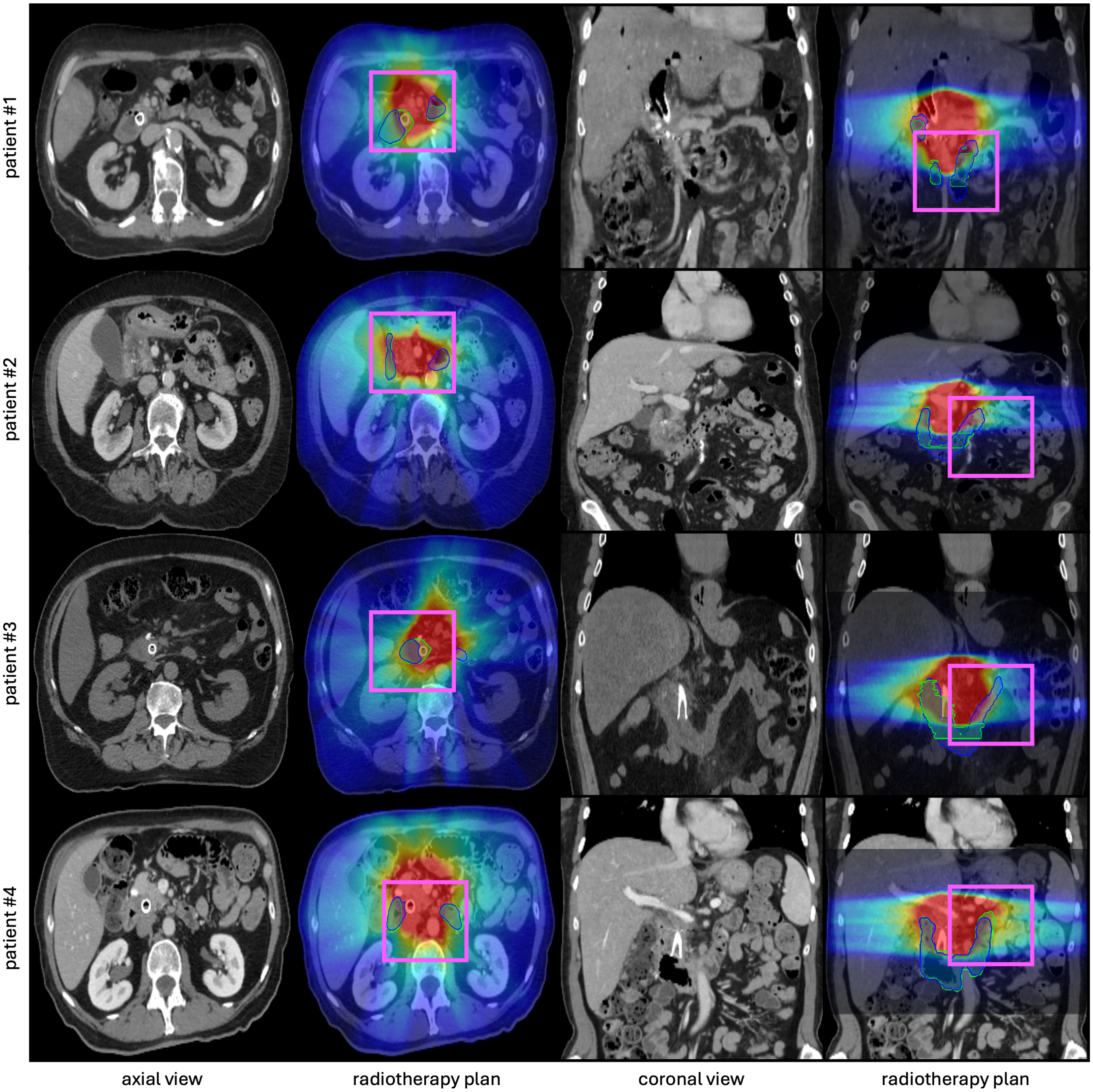


Figure 15. **Improved tumor radiotherapy planning with Flagship Model's superior delineation accuracy over radiologists.** An example of the duodenum, a critical dose-limiting organ due to its proximity to the target, delineated by our Flagship Model (blue) and manually by radiologists (green). The AI model provides precise segmentation of anatomical structures near the tumor, which radiologists miss. This reduces the time required for manual annotation and helps protect vital structures from excessive radiation, minimizing the risk of collateral damage.