DREAM: ERROR-GUIDED LEARNING FOR ROBUST CANCER DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Segmentation models can surpass radiologists at detecting cancer on computed tomography (CT). However, models that perform well on standard datasets often degrade in real-world deployment, by producing more false negatives (missed tumors) and more false positives (false alarms). These errors concentrate on hard cases. *Hard negatives* are often non-cancer abnormalities that trigger false alarms. *Hard positives* are small early tumors that are often missed. To address these systematic errors, we propose **D**iagnostic **R**obustness by **E**rror **A**nalysis and **M**odeling (DREAM). The framework has three components. (1) *Report-based mining*: LLMs read 440,000 radiology reports to find hard negatives and hard positives. The corresponding CT scans are added to training, so the segmentation model learns from the cases it most often gets wrong. (2) *Report-refined masks*: an algorithm refines AI-generated tumor masks to match the tumor size, count, and location described in the radiology report. This avoids time-consuming manual annotation. (3) A *cascade verifier* combines the tumor segmentation model's output with a radiomics-based classifier of non-cancerous abnormalities to improve the tumor probability estimate and reduce the risk of mistaking other abnormalities for cancer. We evaluated DREAM on the detection of tumors in the pancreas, bladder, and gallbladder. DREAM consistently outperforms 13 state-of-the-art methods and public AI models on internal and external validation. Gains over the nnU-Net baseline reach up to $+6.7$ pp in AUC on tumor detection, and DREAM improves specificity by up to $+40.7$ pp on non-cancer abnormalities. In a reader study, DREAM detected 2.4 times more early pancreatic tumors than the average of three radiologists at matched specificity. These results show that AI errors cluster into systematic, identifiable patterns. Targeted mining, report-refined masks, and abnormality-aware verification are effective strategies to reduce these errors. Code and models will be publicly available.

1 INTRODUCTION

Early detection of cancer in CT scans significantly improves patient survival (Siegel et al., 2025). Yet radiologists miss approximately 40–50% of early pancreatic cancers on CT (Alves et al., 2026; Hoogenboom et al., 2022; Mukherjee et al., 2022), 13.7% of bladder cancer (Trinh et al., 2018), and 44.0% of gallbladder cancer (Akabane et al., 2025). Recent work shows that segmentation models can surpass and assist radiologists at early cancer detection (Bassi et al., 2025c; Cao et al., 2023; Li et al., 2026). Segmentation models can surpass vision-language models (Bassi et al., 2025b) and classifiers (Hooper et al., 2023; Bassi et al., 2026b) at tumor detection. Additionally, they improve explainability: radiologists can inspect segmentation outputs to understand and verify the AI's decision.

However, models that excel on standard test datasets often show higher error rates in real-world clinical use and on less curated external datasets (Xia et al., 2022). These errors are not random. They concentrate on two recurring patterns (Oakden-Rayner et al., 2020): *(i) hard positives*, small early-stage tumors that are difficult to detect, and *(ii) hard negatives*, non-cancer abnormalities (e.g., pancreatitis, bladder stone, fluid collection) that visually resemble cancer. Hard positives and hard negatives are underrepresented in hospital databases. Hard positives are rare because tumors are often only detected when large and symptomatic. Hard negatives are rare because non-cancer abnormalities often have low incidence (e.g., ~ 35 cases per 100,000 population per year for pancreatitis (Danpanichkul et al., 2026)). Few hard positives and hard negatives therefore exist in

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

standard training and testing datasets, which are randomly sampled from hospital databases. As a result, segmentation models train mostly on easy cases. As test datasets also often have few hard cases, benchmarks are insensitive to errors on hard cases. Yet avoiding these errors is critical for real-world performance. Hard positives are the early, curable tumors that screening must catch. Hard negatives are rare, but they are often more common than cancer itself. For example, pancreatic cancer is about 7 times rarer than pancreatitis (Li et al., 2025a). Thus, an AI model that systematically confuses non-cancer abnormalities for cancer will issue more false alarms than true detections in clinical use. Without explicit mining, training, and testing on these hard cases, these systematic errors cannot be measured, let alone solved.

Mining hard cases is an established and effective method in computer vision (Viola & Jones, 2001). However, standard mining techniques involve running the AI model many images to find where it fails. This is unfeasible in the medical domain: mining rare abnormalities (e.g., pancreatitis at ~ 6 cases per 100,000) would require running a segmentation model on tens or hundreds of thousands of medical images. Just downloading these images from hospitals, especially heavy 3D images, could take years due to the slowness of hospital PAC systems. These systems often require manual or rate limited downloading. Therefore, we propose mining hard cases from the text that accompanies the medical images, i.e., radiology and pathology reports. Reports are light, and can be downloaded at scale. After mining from text, the vision model can be deployed on the corresponding images, to identify where it fails and add these cases to the training set.

This paper presents **D**iagnostic **R**obustness by **E**rror **A**nalysis and **M**odeling (DREAM), an end-to-end framework that reduces segmentation errors on hard positives and hard negatives. It has three complementary components. (1) LLM-driven *report-based mining* of hard positives and hard negatives over hospital-scale radiology report datasets. The mined hard cases are added to both training and test sets. (2) *Report-refined masks*, which automatically adjust AI-generated tumor masks to match the tumor descriptions in radiology reports. This allows training on hard positives without manual and expensive mask creation (creating a single tumor mask takes a radiologist up to 30 minutes). (3) A *cascade verifier*, which estimates the final tumor probability by combining the segmentation model's output with the output of a radiomics-based classifier of 14 non-cancer abnormalities. The cascade verifier resolves cases where the segmentation model is uncertain. It reduces confusion between non-cancer abnormalities and cancer.

We mine for hard cases on over 440,000 reports. We evaluate DREAM on pancreatic, bladder, and gallbladder cancers. For each cancer, DREAM yielded 719, 1,948, and 935 hard cases, respectively. We evaluate DREAM on internal and external datasets, spanning 3,719 CT scans from 149 hospitals worldwide. DREAM substantially surpassed previous public AI models (trained on diverse datasets), by +5.9%/+15.2%/+16.9% AUC on detecting pancreas/bladder/gallbladder tumors, on external validation. Comparing to models trained on our own dataset, DREAM improved AUC by +9%/+1.4%/+5.2%. Importantly, DREAM substantially reduced false positive cancer detection on scans with non-cancerous abnormalities in the pancreas/bladder/gallbladder, improving specificity by +17.4%/+8.3%/+8.5%. Ablations demonstrate that the 3 elements of the DREAM pipeline improve AI performance. In a reader study, DREAM detected 2.4 times more small tumors than the average of three radiologists at matched specificity (90.7% sensitivity at 96.8% specificity, Sec. 4). Together, the results show that DREAM improves tumor detection on hard positives (small tumors), on hard negatives (non-cancer abnormalities), and across three organs. Our main contributions are:

1. **Report-based mining of hard negatives at hospital scale.** LLMs read 440,000 radiology reports and identify non-cancer abnormalities. Training on these *hard negative* CT scans suppresses common false alarms. Report-based mining took one week. Similar large-scale image-based mining or active learning could take months.
2. **Hard positive mining with report-refined masks.** We use the same LLM pipeline to find early-stage tumors in reports. We then generate *report-refined tumor masks* that align AI-generated tumor masks with the tumor descriptions in reports. This avoids time-consuming manual mask creation and leverages tumor information already available in reports. Resulting masks are not gold-standard human-made masks, but ablations show they substantially improve AI performance without the need for radiologists' time.
3. **The cascade verifier** analyzes DREAM's segmentation output and a radiomics-based classification of abnormalities. The verifier estimates an abnormality-aware tumor probability. The verifier corrects segmentation errors without retraining the segmentation model.

In summary, errors of medical AI cluster into structured patterns. DREAM addresses these errors by mining hard cases, automatically creating report-refined masks for them, and adding a cascade verifier that double-checks model decisions while accounting for non-cancer abnormalities. This strategy is cheaper, faster, and more effective than collecting random additional training data and asking radiologists to draw tumor masks for it. We demonstrate DREAM on 3 cancer types to show its generality. Missed small tumors and false positives from non-cancer abnormalities are recurring failure modes across cancer-detection tasks (Bassi et al., 2025b). We believe that idea of mining hard cases from text may also be generalizable beyond the medical domain.

Related Work. Tumor mask creation for a single 3D CT scan can take a radiologist up to 30 minutes. Creating 3,125 pancreatic tumor masks required eight radiologists working over five years (Xia et al., 2022). Active learning can accelerate mask creation, but it still requires multiple radiologists to draw or refine masks for months (Li et al., 2025c; Bassi et al., 2025b; Li et al., 2025b). In contrast, our *report-refined masks* (Sec. 2.2) are created automatically. An algorithm refines AI-generated tumor masks to better match the tumor descriptions in radiology reports. Even though these masks are not human made, our ablations show they substantially improve AI performance. **Hard negative mining** reduces errors by prioritizing challenging negative cases in training (Viola & Jones, 2001; Liu et al., 2025; Kamnitsas et al., 2017; Shrivastava et al., 2016). Hard negatives are usually found by testing on images (Xiang et al., 2023; 2024). In medical imaging, this is difficult: downloading hundreds of thousands of images from hospitals can take years. We instead mine hard negatives from radiology reports, which took one week. **Active learning** is an alternative strategy for selecting additional training cases. It prioritizes hard cases. Active learning typically runs models on a large image dataset and selects cases where the model is uncertain or where multiple models disagree (Li et al., 2024; Qu et al., 2023; Bassi et al., 2025c; Chou et al., 2024). It shares the same bottleneck as standard hard negative mining. Finding rare abnormalities would require downloading and running inference on hundreds of thousands of 3D CT scans, which often takes months on hospital PAC systems.

2 METHODS

2.1 MINING HARD NEGATIVES AND HARD POSITIVES

We mine reports with LLMs because LLMs can capture context and clinical meaning. This is important: radiology reports are variable and describe findings in different ways, making keyword search unreliable. We define hard positives as pancreatic tumors below 2 cm in diameter, the cancer T-stage 1 threshold. We define hard negatives as non-cancer pancreatic abnormalities. Both are extremely rare in hospital archives. Only 1.75% of our $440{,}000$ N-California reports describe hard negatives and 0.05% describe hard positives. Finding them requires searching hundreds of thousands of cases. This is feasible with reports but not with 3D CT scans (most hospital PACS systems are not designed for large-scale 3D image download). Our report-based mining pipeline has three steps, described below. We add the mined hard positives and hard negatives to the training set, then fine-tune the segmentation model on the updated set. This reduces its systematic errors on hard cases.

(I) Fast report-based mining. A small LLM (Llama 3.1 8B AWQ (Dubey et al., 2024)) reads each radiology report and flags cases where the pancreas is described as abnormal. This is a simple task. Small LLMs can run in parallel and quickly process hospital-scale reports. In five days we processed $440{,}000$ abdominal CT reports with 4 NVIDIA A100 GPUs. The selected abnormal cases include both cancer and non-cancer abnormalities, capturing both potential positives and hard negatives.

(II) Temporal report-based refinement. For each patient flagged in Step (I), a large LLM (Llama 3.1 70B AWQ) reads all of the patient's available radiology and pathology reports. From these reports, the LLM determines whether each CT scan contains pancreatic tumors and the tumor sizes. Based on the LLM answer, we label each scan as a hard positive or a hard negative. Hard positives are CT scans with pancreatic tumors smaller than 2 cm. Hard negatives are CT scans (with abnormalities) from patients with no pancreatic cancer at any time point. We exclude from hard negatives any scan without a follow-up scan at least two years later. This excludes abnormalities that the reporting radiologist called non-cancer but that were in fact cancer. This step yields 7,000 hard-negative candidates and 197 hard-positive candidates from the $440{,}000$ reports.

(III) Image-based refinement. We run the tumor segmentation model (trained before mining) on all CT scans selected in Step (II). Only the scans where the model makes mistakes are selected as the

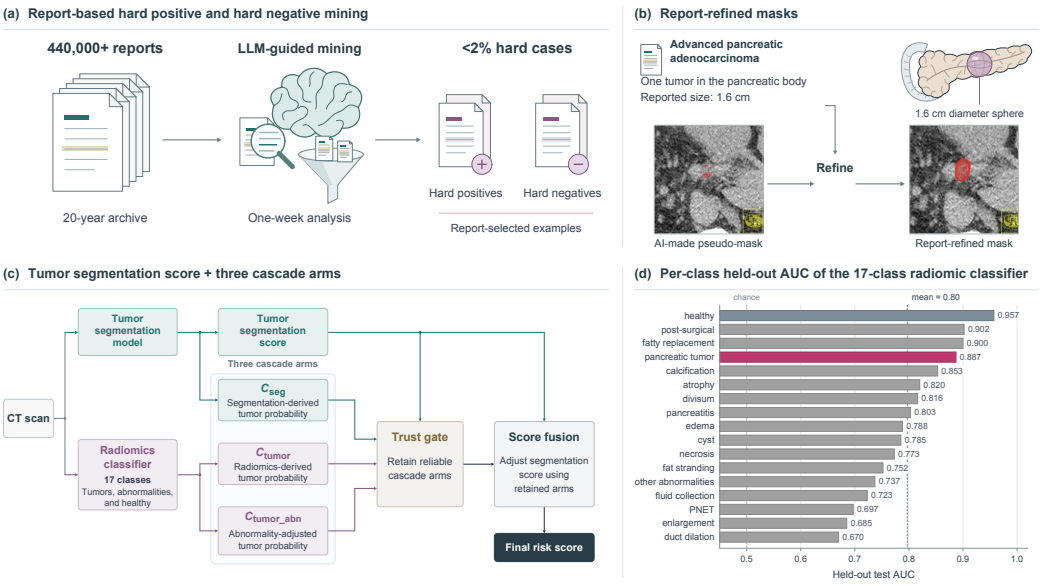


Figure 1: **(a) Report-based mining.** An LLM pipeline read 440,000 radiology reports and selected 374 *hard positives* (small tumors that segmentation models miss) and 3,228 *hard negatives* (non-cancer abnormalities that resemble tumors and trigger false alarms), split 46/189/139 and 673/1,759/796 across pancreas, bladder and gallbladder. Both are rare enough to require a hospital-scale search, which took about one week over reports. **(b) Report-refined masks.** For each hard positive, our algorithm refines an AI-generated tumor mask to match the tumor size, count, and location described in the report, avoiding manual mask creation. **(c) Cascade verifier.** The cascade combines the DREAM tumor segmentation score with three arms (C_{seg}, C_{tumor}, $C_{\text{tumor_abn}}$). The trust gate removes unreliable arms, and score fusion then adjusts the segmentation score without retraining the segmentation model. **(d) Radiomics classifier.** Organ-agnostic and instantiated once per organ; the pancreas is shown here.

final hard positives and hard negatives. The report-based search in the previous steps restricts the cohort to a few thousand scans. This makes downloading and image-based inference feasible. This step yields 673 hard negatives and 46 hard positives that are added to training.

LLM accuracy. Step (II) is where accuracy matters most. It determines whether each scan is labeled as a hard positive or a hard negative. Steps (I) and (III) only narrow the scan pool. We therefore use the largest LLM in Step (II), guided by radiologist-designed prompts that provide clinical knowledge about report interpretation. This combination of LLM and report has been validated by radiologists. It reaches 96% accuracy in detecting tumors from reports, with no cancer patients mislabeled as non-cancer (Bassi et al., 2025b). We also cross-check the LLM output with ICD-10 diagnostic codes (hospital records that flag patients with cancer). Cases where the diagnostic codes disagree with the LLM are excluded. The mining pipeline is robust to occasional LLM errors. Even though LLM accuracy is not 100%, we show that our methods are tolerant to noises (Sec. E).

2.2 Report-refined Masks

After identifying the hard positives, we need segmentation masks to train with them. However, manual mask creation is slow and expensive. CT scans are 3D, and creating a single tumor mask can take a radiologist up to 30 minutes. We therefore propose automatically generating *report-refined masks*. First, the tumor segmentation model (trained before mining) creates soft tumor masks (predicted tumor probability per voxel) for the hard positives. The algorithm below then refines each soft mask into a binary mask that matches the tumor count, diameters, and locations described in the radiology report. Masks that cannot be reliably matched to the report are discarded.

(I) Coarse tumor location. Radiology reports typically specify the pancreatic sub-segment containing the tumor (head, body, or tail). We locate this sub-segment in the CT scan using precomputed pancreas sub-segment masks. These masks are generated with a public segmentation model (Bassi et al., 2025b). Such models are precise (Bassi et al., 2024; 2025b), but we still dilate the sub-segment mask by 1 cm to compensate for small inaccuracies.

(II) Fine tumor location. Inside the selected pancreas sub-segment, we search for the most likely tumor location. We place multiple spheres inside the sub-segment. Each sphere has a diameter equal to the tumor diameter reported by the radiologist (if multiple diameters are reported, we use the largest). The spheres are separated by 1 voxel and cover all positions that overlap the sub-segment. For each sphere, we sum the predicted tumor probabilities inside it. The sphere with the highest probability sum is selected as the most likely tumor location.

(III) Binary tumor delineation. Inside the selected sphere, we convert the predicted tumor probabilities into a binary segmentation mask. We first try a threshold of 50%. If the resulting tumor size matches the diameter in the radiology report (within a 20% margin), we accept this mask. Otherwise, we estimate the expected tumor volume from the reported diameter, compute the number of voxels (K) that correspond to this volume, and select the K voxels with the highest tumor probability inside the sphere. These voxels are set to 1. All other voxels inside the sphere are set to 0. Voxels outside the sphere are always 0.

This algorithm ensures that the resulting segmentation mask approximately matches both the tumor size and location described in the report. It also avoids a limitation of standard thresholding. Even when the model predicts low probabilities, the true tumor location can still have the highest probabilities relative to the rest of the image. A fixed threshold (e.g., 50%) could produce an empty mask, while our report-refined mask can still capture this tumor. If the report describes multiple pancreatic tumors, we repeat the three steps above for each tumor (largest first), while preventing overlap between selected spheres (Step II).

Discarding unreliable masks. If the segmentation model completely fails to detect a tumor, the resulting report-refined mask will be incorrect. Following the principle of auditing training masks before use (Bassi et al., 2025a), we compute a *mask quality score*: the Dice Similarity Coefficient (DSC) between the segmentation model's output (the original soft tumor mask) and the resulting report-refined mask. Low DSC indicates that the segmentation model strongly disagrees with the report-refined mask, suggesting that the mask may be unreliable. For example, if the segmentation model predicts near zero everywhere, the report-refined mask will show a tumor in the wrong place, but the DSC will be near zero, detecting the error. A radiologist reviewed a small subset of 20 masks across different DSC values and selected a threshold for the mask quality score. We excluded all masks below this threshold (20 masks).

Report-refined masks act as a silver standard. They are less precise than radiologist-drawn masks, but they exploit tumor information and human knowledge in radiology reports to improve upon simple pseudo-labels. Compared with radiologist-drawn masks, report-refined masks require little radiologist effort (at most selecting a filtering threshold). Compared with standard pseudo-labels (e.g., applying a 0.5 threshold to AI-generated tumor masks), they incorporate clinical knowledge from reports. This creates a better training signal and improves tumor detection and segmentation (Sec. E).

The segmentation model's output can be converted into a tumor segmentation score. For example, the diameter of the largest segmented tumor (after binarization) can serve as a confidence score for tumor detection. However, this segmentation score ignores complementary signals that the segmentation model produces. DREAM segments not only the tumor but also the pancreatic duct and the pancreas itself. This provides valuable signals for tumor detection, such as pancreatic-duct dilation, pancreas atrophy (common signs of early cancer, even when tumors are barely visible), and the relative distribution of pancreas, duct, and tumor volumes. The segmentation score also ignores the presence of non-cancer pancreatic abnormalities. We therefore propose a lightweight cascade verifier that adjusts the tumor segmentation score using three cascade arms. These arms analyze the complementary signals while the segmentation network is kept frozen.

Three cascade arms estimate three tumor probabilities ($\mathbf{C}_{\text{seg}}$, $\mathbf{C}_{\text{tumor}}$, and $\mathbf{C}_{\text{tumor_abn}}$) from two signal sources (Fig. 1a): DREAM's multi-class segmentation mask, and a separately trained radiomics classifier. The classifier predicts tumor, non-cancer abnormalities, healthy, and other abnormalities from radiomics features extracted from the organ region (Appendix C). All three cascade arms output

tumor probabilities, but compute them from different inputs. (i) $\mathbf{C}_{\text{seg}}$ is the tumor probability based on 26 summary features of DREAM's segmentation mask (per-class volumes, component counts, top logits, and derived ratios such as tumor-to-organ volume; full list in Appendix C). These summary features are passed to an ensemble of five LightGBM (Ke et al., 2017) binary classifiers (different random seeds) combined by a logistic-regression meta-learner. The meta-learner estimates $\mathbf{C}_{\text{seg}}$. (ii) $\mathbf{C}_{\text{tumor}}$ is the tumor probability based on radiomics features. (iii) $\mathbf{C}_{\text{tumor_abn}}$ is the tumor probability based on radiomics features and on the probabilities of different abnormalities. To create it, we feed the full output of the radiomics classifier (all classes, including tumor and abnormalities) into a second LightGBM ensemble with the same configuration. The intuition is that the joint pattern across all different classes is more discriminative than the tumor probability $\mathbf{C}_{\text{tumor}}$ alone. For example, simultaneously high tumor probability, high duct dilation, and low pancreatitis is more consistent with pancreatic cancer. Every normalization in the cascade verifier is an inductive empirical CDF estimated on a calibration split ($n = 543$), ensuring to be held out from the test set. Three kinds of quantity are normalized this way: (i) the member probabilities inside $\mathbf{C}_{\text{seg}}$ and $\mathbf{C}_{\text{tumor_abn}}$, (ii) the arm scores $\mathbf{C}_k$, and (iii) the segmentation score D.

Trust Gate. Classifiers may not generalize across hospitals as well as segmentation models (Hooper et al., 2023). The Trust Gate therefore activates each cascade arm only when its score ranking aligns well with the tumor segmentation score ranking on a held-out calibration split of our development data ($n = 543$, stratified by tumor status and disjoint from all three test cohorts). We measure this alignment by the Spearman rank correlation (Spearman, 1904): $w_k = \mathbf{1}[\rho(C_k, D) > 0.20]$ for $k \in \{\text{seg}, \text{tumor}, \text{tumor_abn}\}$, where D is the tumor segmentation score. The combined cascade score C is the rank-normalized average of the active arms, or C_{seg} alone if no arm passes the gate:

$$C = \begin{cases} \frac{1}{\sum_k w_k} \sum_k w_k \, C_k & \sum_k w_k > 0, \\ C_{\text{seg}} & \text{otherwise.} \end{cases} \tag{1}$$

Score fusion. To produce the final tumor score used for tumor detection, we balance the original segmentation score D and the combined cascade score C using a score fusion formula with three components:

$$S_{\text{final}} = D + \gamma \, (C - 0.5) \cdot \max\big(0, \, 1 - 2|D - C|\big), \tag{2}$$

where $\gamma = 0.05$ is the maximum nudge size (chosen by training-cohort cross-validation). The term $(C - 0.5)$ sets the direction. It is positive when the cascade considers the case tumor-like and negative otherwise. The factor $\max(0, 1 - 2|D - C|)$ measures agreement between D and C. It equals 1 when D and C are identical, 0 when they are opposite, and varies smoothly between. When D and C roughly agree, the cascade nudges D up or down by a small amount. When they disagree strongly, the agreement factor zeros out the adjustment and $S_{\text{final}} = D$. This compensates for cases where the segmentation model generalizes to external hospitals better than the classifiers, which can otherwise create unreasonable disagreements.

3 Datasets and Training

Training. For pancreatic cancer we first train on **E-Coast** ($N = 3{,}151$, private), CT scans from a US east-coast hospital with radiologist-drawn pancreatic masks. Of these, 2,101 contain a pancreatic tumor. For the bladder and gallbladder we instead start from **Merlin Plus** (Bassi et al., 2026a), a public dataset of 1,153 radiologist-drawn tumor masks across nine organs, including 127 bladder and 97 gallbladder tumors. We then mine hard cases from **N-California Reports** ($N = 440{,}000$, private) and download only the CT scans the mining selects: 46/189/139 hard positives and 673/1,759/796 hard negatives for the pancreas/bladder/gallbladder respectively. Dataset details are in Appendix G.

Testing. We evaluate the pancreas on three cohorts: **N-California** ($N = 1{,}046$, internal; 94 small tumors, 952 non-cancer), **PanTS** ($N = 901$, public external; 151 tumors, 750 non-cancer), and **Euro** ($N = 976$, external; 722 tumors, 254 non-cancer). DREAM never trains on PanTS or Euro. A fourth pancreatic cohort, **Pre-diagnostic** ($N = 33$, internal), holds gold standard labeled scans where radiologists missed tumors (scans acquired before diagnosis). For bladder and gallbladder we test internally on N-California ($40/44$ positives/negatives for bladder and $13/30$ for gallbladder), and test externally on Euro (bladder $171/227$ and gallbladder $28/243$). Details are in Appendix G.

Architectures and training. DREAM is an nnU-Net ResEncL (Isensee et al., 2021) followed by the cascade verifier (Sec. C). We pretraining on each organ's mask dataset (E-Coast for pancreas, Merlin Plus for the others), then fine-tune on the mined hard cases with their report-refined masks, retaining the original pretraining data to avoid catastrophic forgetting. We compare against two kinds of baseline. The first is public AI models trained on diverse datasets and evaluated as released, which measures what a practitioner could download today. The second is models trained on our own data, which isolates the contribution of the method from the contribution of the data. We train an nnU-Net on the datasets DREAM is pretrained on (E-Coast for the pancreas and Merlin Plus for the bladder and gallbladder). Fine-tuning and preprocessing details are in Appendices D and F.

4 RESULTS

DREAM reduces false positives and false negatives. Across multiple test sets, including external validation (PanTS and Euro), DREAM outperforms multiple state-of-the-art baselines (Table 1). The baselines include public segmentation models (MedFormer, Merlin, PANORAMA, DiffTumor, Universal Model) and nnU-Net models we trained on PanTS, Merlin Plus, and on E-Coast. DREAM fine-tunes the nnU-Net Baselines (nnU-Net trained on E-Coast for pancreas, Merlin Plus for others) on the mined hard cases and their report-refined masks, then applies the cascade verifier. It surpasses the nnU-Net Baseline for all tumor types. The public models we compare against differ by organ. Several segmentation models are trained specifically for pancreatic tumors (Li et al., 2025c; Gao et al., 2022; Chen et al., 2024; Alves et al., 2024), but no comparable organ-specific model exists for the bladder or the gallbladder. For those two organs we therefore compare against universal lesion segmentation models (de Grauw et al., 2025), pan-abdominal tumor models from the FLARE challenges (Ma et al., 2024b), and text-promptable or vision-language models that can be queried for any tumor (Blankemeier et al., 2026; Rokuss et al., 2026). The combined sensitivity and specificity gains over the nnU-Net Baselines confirm that report-based mining, report-refined masks, and the cascade verifier together reduce both false negatives and positives. Appendix I shows qualitative examples of both error types: a pancreatitis case the baseline calls a tumor and DREAM does not, and a 9 mm tumor the baseline misses and DREAM detects.

DREAM improves the detection of early and pre-diagnostic tumors. DREAM raises small tumor sensitivity by up to $+22.4$ pp over the nnU-Net Baseline for prediagnostic pancreatic tumors (Table 5; pre-diagnostics were unavailable in our test hospitals for other tumor types, which are rarer). Mining hard positives therefore improves detection on the cases that matter most clinically: prediagnostic scans where the tumor was missed by radiologists. That is because tumors missed by radiologists are often the smallest ones, similar to our hard positives. Detecting small pancreatic tumors ($\leq$2 cm in diameter) is critical because they are often still operable and removable. Early detection raises the chance of survival more than tenfold (Kaur et al., 2012). On those small tumors DREAM raises sensitivity from 64.6% to 83.0% across all pancreas, bladder and gallbladder tumor and on PanTS from 59.7% to 82.1% at a matched 84.0% specificity, a gain of 22.4 pp (Tab. 5). Figure 3 shows a 9 mm tumor missed by the nnU-Net Baseline but detected after fine-tuning with DREAM. Its radiology report describes the tumor as borderline operable, indicating that detection could be life-saving. In summary, by mining pathology-confirmed small tumors, we improved the detection of cancers missed in clinical practice.

DREAM generalizes to external test sets (Tab. 1). On Euro, DREAM reaches 0.971 pancreatic tumor AUC despite training only on US CT scans, and significantly surpasses the nnU-Net Baseline by 3 pp. On PanTS (Li et al., 2025c), which draws on 145 external institutions, DREAM raises pancreatic AUC from 0.824 to 0.915, a gain of 10 pp, and significantly outperforms public segmentation models trained on PanTS even though DREAM never trains on it. The same recipe transfers to two further organs. Bladder detection AUC rises from 80.7 to 82.1 externally and from 75.3 to 81.5 internally, and gallbladder detection AUC rises from 84.3 to 89.5 and from 85.9 to 92.6. Together these results show that DREAM improves out-of-distribution generalization, not just in-distribution accuracy.

Ablations. We performed comprehensive ablations on pancreatic tumors (Appendix E). We progressively removed each component of DREAM to isolate its contribution (Tab. 3). Removing the cascade verifier drops AUC by 0.008 on PanTS and 0.045 on N-California. Second, removing the report-refined masks (replacing them with AI-generated masks thresholded at 0.5) drops AUC by another 0.045 on PanTS and 0.012 on N-California, and lowers PanTS segmentation DSC from 0.60

Table 1: **Tumor detection across three organs.** All entries are percentage points: sensitivity (Se), specificity (Sp), AUC, and DSC, the last reported only on PanTS, the one test set with per-voxel tumor masks. Left: pancreas. Right: bladder and gallbladder. Public models are evaluated as released. DREAM and nnU-Net use operating points frozen on a validation split; public checkpoints, which have none. No organ-specific public model exists for the bladder or gallbladder, so those comparators are universal lesion segmentation models (de Grauw et al., 2025), pan-abdominal FLARE models, and text-promptable or vision-language models (Blankemeier et al., 2026; Rokuss et al., 2026); the pancreatic comparators additionally include MedFormer (Gao et al., 2022), DiffTumor (Chen et al., 2024), Universal Model (Liu et al., 2023) and PANORAMA (Alves et al., 2024), with nnU-Net (Isensee et al., 2021), SwinUNETR (Hatamizadeh et al., 2022), MedNeXt (Roy et al., 2023) and U-Mamba (Ma et al., 2024a). On the pancreas **Euro test set** ($N = 976$) the nnU-Net baseline attains Sen/Spec/AUC $= 89.6/89.6/0.940$ and DREAM reaches $91.9/93.5/0.971$. The second table is ladder and gallbladder tumor detection. **Blue** marks statistical significance.

	PanTS				N-Cal.		
Model	Se	Sp	AUC	DSC	Se	Sp	AUC
Zero-shot							
nnU-Net (PanTS)	73	93	87	46	90	67	88
MedFormer	75	90	90	53	75	89	87
ULS23	46	65	55	16	49	56	52
ULS 400	20	86	52	3	23	73	48
ULS 901	18	88	51	3	23	74	49
DiffTumor	76	77	83	37	71	78	79
Universal	97	16	77	34	88	9	57
PANORAMA	52	99	85	27	54	99	91
Merlin	49	78	65	n/a	60	91	86
Trained on Same Dataset							
DREAM	**87**	84	**92**	**60**	**96**	92	**97**
nnU-Net	77	84	82	0.43	88	89	91
SwinUNETR	64	76	74	42	54	90	81
MedNeXt	72	89	85	52	55	95	87
U-Mamba	76	87	87	55	75	94	87

	Bladder						Gallbladder					
	Euro			N-Cal.			Euro			N-Cal.		
Model	Se	Sp	AUC	Se	Sp	AUC	Se	Sp	AUC	Se	Sp	AUC
ULS 400	57	58	56	85	29	53	85	25	52	92	46	66
ULS 901	66	53	57	67	38	49	82	24	44	92	46	65
FLARE23-A	74	32	52	85	25	51	70	80	70	53	83	70
FLARE23-B	83	20	47	95	20	54	64	77	69	76	66	73
FLARE24	8	94	49	15	90	45	92	30	63	30	100	65
FLARE25	50	64	56	7	95	43	64	75	70	69	86	75
Merlin	68	50	58	99	6	50	69	80	72	50	83	61
VoxTell	55	66	62	30	75	40	57	84	72	76	90	89
DREAM	**74**	**81**	**82**	**85**	**75**	**82**	**71**	**91**	**90**	**85**	**93**	**93**
nnU-Net	68	84	81	83	66	75	61	88	84	85	87	86

to 0.40. Third, removing the mined hard cases (replacing them with the same number of randomly sampled positives and negatives from N-California) drops AUC by an additional 0.033 on PanTS and 0.010 on N-California. Fourth, replacing the report-based mining with *image-based mining*, which retrains on the deployed model's confirmed errors instead of reading reports, lowers specificity to 91.7% on PanTS and 91.3% on N-California, against 96.0% and 96.2% for DREAM. Fifth, perturbing every reported tumor diameter by up to $\pm 25\%$ leaves AUC statistically unchanged on both cohorts ($p \geq 0.73$), so report-refined masks tolerate errors in reported tumor sizes. Retaining 50% more report-refined masks moves small-tumor AUC by 0.002, so performance is insensitive to the mask-quality threshold. Therefore, DREAM is robust to hyper-parameter choices and report errors, and each DREAM component raises detection AUC. Detailed ablation results are in Appendix E.

DREAM substantially reduces false positives on hard negatives. Table 2 reports specificity on 656 cases with different pancreatic abnormalities collected from N-California (sensitivity is in Tab. 1). These patients were not used in training. DREAM improves specificity over the nnU-Net Baseline by $+17.4$ pp on average. The largest gains (up to $+40.7$ pp) appear on abnormalities known to mimic cancer, such as fluid collection, fat stranding, pancreatitis, and duct dilatation. The same pattern holds in the other two organs, where DREAM raises specificity from 50.0% to 58.3% on 48 bladder observations and from 65.0% to 73.5% on 117 gallbladder observations.

DREAM may surpass human performance. To understand whether DREAM has potential to help radiologists in early cancer detection, the first step is to test whether DREAM can surpass human performance. We performed a small reader study. We asked three general radiologists (two residents with 2 years of experience and one board-certified radiologist with 5 years of experience) to detect tumors in our N-California test dataset. For feasibility, we randomly selected 43 CT scans with small tumors and 31 non-cancer patients. Detailed results are shown in Fig. 4. The average performance of the radiologists was 37% sensitivity at 97% specificity. In contrast, DREAM achieved 90.7%

Table 2: **Specificity on non-cancer abnormalities.** Hard negative cases with abnormalities but no tumor; these are hard because abnormalities can mimic tumors on CT. Values are percentages and Δ is the change in percentage points. Boxed and blue marks a statistically significant improvement by DREAM (Fisher's exact test, $p < 0.05$).

Abnormality	n	nnU-Net	DREAM	Δ	Abnormality	n	nnU-Net	DREAM	Δ
Atrophy	272	68.8	86.4	+17.6	Fluid collection	27	7.4	48.1	+40.7
Calcification	99	42.4	57.6	+15.2	Necrosis	5	0.0	20.0	+20.0
Enlargement	22	63.6	72.7	+9.1	Pancreas divisum	42	90.5	85.7	−4.8
Fat stranding	41	22.0	43.9	+21.9	Duct dilatation	69	46.4	69.6	+23.2
Fatty replacement	65	72.3	90.8	+18.5	Pancreatitis	70	27.1	47.1	+20.0
Pancreas overall ($n = 656$): nnU-Net $= 57.0$ **DREAM $= 74.4$** $\Delta = +17.4$									
Diverticulum	18	61.1	72.2	+11.1	Wall thickening	14	21.4	35.7	+14.3
Stone/calcification	16	62.5	62.5	0.0					
Bladder overall ($n = 48$): nnU-Net $= 50.0$ **DREAM $= 58.3$** $\Delta = +8.3$									
Adenomyomatosis	14	78.6	85.7	+7.1	Distension	26	57.7	65.4	+7.7
Gallstones	42	83.3	90.5	+7.1	Sludge	22	50.0	59.1	+9.1
Wall thickening	13	30.8	46.2	+15.4					
Gallbladder overall ($n = 117$): nnU-Net $= 65.0$ **DREAM $= 73.5$** $\Delta = +8.5$									

sensitivity at the same 96.8% specificity. For the same number of false positives, DREAM therefore achieved 2.4-fold higher sensitivity for early tumors than the radiologist average (+53.7 pp).

5 Conclusion

Tumor detection models often fail systematically on *hard negatives* (non-cancer abnormalities that resemble cancer) and *hard positives* (small tumors that are difficult to see). Both are rare in standard training and test sets. Tumors are often diagnosed only when large, and non-cancer abnormalities are uncommon in the general population. In clinical use, however, even rare non-cancer abnormalities outnumber pancreatic tumors (e.g., pancreatitis is $\sim 7\times$ more common). Models that confuse these abnormalities for cancer can therefore produce more false positives than true detections. AI models that miss small tumors also miss exactly the cancers still curable.

DREAM addresses these critical failure modes with three components. *Report-based mining* searches over 440,000 radiology reports to find hard negatives and hard positives, and adds them to training. *Report-refined masks* automatically generate tumor masks for the mined hard positives by refining AI-generated masks to match the report descriptions. This avoids manual mask creation, which is time-consuming and expensive. The *cascade verifier* combines the tumor segmentation output, the organ and duct segmentations, and abnormality classifications to estimate a more accurate tumor probability. Each DREAM component contributes a measurable gain (Sec. E). Together, they reduce errors on hard cases. DREAM raises sensitivity on cancer under 2 cm by 22.4 pp at matched specificity, and raises specificity on non-cancer abnormalities by 17.4 pp on average and by up to 40.7 pp, surpassing 13 previous state-of-the-art models on internal and external validation.

Limitations. DREAM surpassed three radiologists in one small, enriched reader study of 43 small tumors and 31 non-cancer cases, with neither prior imaging nor clinical context. We chose that setting to simulate incidental detection of small, asymptomatic pancreatic tumors on CT scans ordered for other reasons, following previous pancreatic cancer reader studiesCao et al. (2023). We therefore claim only that DREAM surpassed radiologists in this specific setting, and large-scale studies are needed for further clinical validation. The LLM inside the mining pipeline reaches 96% cancer-identification accuracy, not 100%, and report-refined masks are approximations filtered by a mask-quality score. Our results and ablations show that imperfect LLMs and masks still lead to substantial improvement in cancer detection, without the need for manual radiologist effort.

AI USE STATEMENT

Generative AI is a component of the method in this work, and we state its role first. We used Llama 3.1 8B and Llama 3.1 70B (Dubey et al., 2024) to read radiology and pathology reports and to label each scan as a hard positive or a hard negative, as described in Sec. 2.1. These models select training cases and supply the tumor sizes, counts and locations that drive report-refined mask generation. They therefore clean and reformat dataset content, support qualitative analysis of clinical text, and provide a critical ingredient of our supervision. We report their accuracy and its limits in Sec. 2.1 and in the Limitations. In addition, we also used AI to help polish writing.

ETHICS STATEMENT

This study analyzes retrospective, de-identified CT scans and their associated radiology and pathology reports from participating hospitals. Each contributing site approved the study under its own institutional review board, with a waiver of informed consent for the secondary use of de-identified clinical data. All scans and reports were de-identified at the source before any analysis, and no protected health information entered our pipeline or is reproduced in this paper.

The reader study involved three radiologists reading de-identified scans in a research setting. It measured model and reader performance, not patient care, and no clinical decision was made or withheld on the basis of any output reported here.

The private datasets used here cannot be redistributed under their governing data use agreements. We release code and trained models so that the method can be reproduced on public data, and we evaluate on the public PanTS benchmark for that purpose.

REPRODUCIBILITY STATEMENT

Code and models will be publicly available. Our mining pipeline is described in Sec. 2.1, the report-refinement algorithm in Sec. 2.2, and the cascade verifier in Sec. C, with its features, training pools and hyperparameters in Appendix C. Fine-tuning settings are in Appendix D and the preprocessing and normalization pipeline in Appendix F. Cohort definitions, inclusion rules, positive and negative counts. The public PanTS test split lets any reader reproduce our external comparison without access to our private cohorts.

REFERENCES

Nabila Abraham and Naimul Mefraz Khan. A novel focal tversky loss function with improved attention u-net for lesion segmentation. In *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*, pp. 683–687. IEEE, 2019.

M. Akabane, J. Kawashima, A. Altaf, S. Woldesenbet, F. Aucejo, M. Kitago, Y. Imaoka, A. Ruzzenente, I. Endo, H. P. Marques, V. Lam, S. K. Maithel, T. Hugh, N. Bhimani, and T. M. Pawlik. Preoperative diagnostic failure in gallbladder cancer: Influence of tumor location and size on imaging precision. *European Journal of Surgical Oncology*, 51(7):109979, July 2025. doi: 10.1016/j.ejso.2025.109979.

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

Natalia Alves, Megan Schuurmans, Dawid Rutkowski, Anindo Saha, Pierpaolo Vendittelli, Nancy Obuchowski, Marjolein H Liedenbaum, Ingfrid S Haldorsen, Anders Molven, Derya Yakar, et al. Artificial intelligence and radiologists in pancreatic cancer detection using standard of care ct scans (panorama): an international, paired, non-inferiority, confirmatory, observational study. *The Lancet Oncology*, 27(1):116–124, 2026.

Pedro R. A. S. Bassi, Wenxuan Li, Szymon Płotka, Ruby Honjol, Jakub Prządo, Xinze Zhou, Kang Wang, Yang Yang, Malte Jensen, Akshay S. Chaudhari, Curtis P. Langlotz, Alan L. Yuille, and Zongwei Zhou. Merlin Plus: A Large-Scale, Multi-cancer, Image-Mask-Report Dataset . In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2026*, volume LNCS 16895. Springer Nature Switzerland, September 2026a.

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? *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 37:15184–15201, 2024. URL `https://github.com/MrGiovanni/Touchstone`.

Pedro RAS Bassi, Qilong Wu, Wenxuan Li, Sergio Decherchi, Andrea Cavalli, Alan Yuille, and Zongwei Zhou. Label critic: Design data before models. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5. IEEE, 2025a. URL `https://github.com/PedroRASB/LabelCritic`.

Pedro RAS Bassi, Mehmet Can Yavuz, Ibrahim Ethem Hamamci, Sezgin Er, Xiaoxi Chen, Wenxuan Li, Bjoern Menze, Sergio Decherchi, Andrea Cavalli, Kang Wang, Yang Yang, Alan Yuille, and Zongwei Zhou. Radgpt: Constructing 3d image-text tumor datasets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23720–23730, 2025b. URL `https://github.com/MrGiovanni/RadGPT`.

Pedro RAS Bassi, Xinze Zhou, Wenxuan Li, Szymon Płotka, Jieneng Chen, Qi Chen, Zheren Zhu, Jakub Prządo, Ibrahim E Hamamci, Sezgin Er, Xiaoxi Chen, Mehmet Can Yavuz, Yu-Cheng Chou, Tianyu Lin, Kang Wang, Yucheng Tang, Jaroslaw B Cwikla, Sergio Decherchi, Andrea Cavalli, Yang Yang, Alan L Yuille, and Zongwei Zhou. Scaling artificial intelligence for multi-tumor early detection with more reports, fewer masks. *arXiv preprint arXiv:2510.14803*, 2025c. URL `https://github.com/MrGiovanni/R-Super`.

Pedro RAS Bassi, Wenxuan Li, Jakob Wasserthal, Jieneng Chen, Xinze Zhou, Zheren Zhu, Chuntung Zhuang, Sergio Decherchi, Andrea Cavalli, Kang Wang, et al. Report supervision. *Medical Image Analysis*, pp. 104257, 2026b.

Louis Blankemeier, Ashwin Kumar, Joseph Paul Cohen, Jiaming Liu, Longchao Liu, Dave Van Veen, Syed Jamal Safdar Gardezi, Hongkun Yu, Magdalini Paschali, Zhihong Chen, Jean-Benoit Delbrouck, Eduardo Reis, Robbie Holland, Cesar Truyts, Christian Bluethgen, Yufu Wu, Long Lian, Malte Engmann Kjeldskov Jensen, Sophie Ostmeier, Maya Varma, Jeya Maria Jose Valanarasu, Zhongnan Fang, Zepeng Huo, Zaid Nabulsi, Diego Ardila, Wei-Hung Weng, Edson Amaro Junior, Neera Ahuja, Jason Fries, Nigam H. Shah, Greg Zaharchuk, Marc Willis, Adam Yala, Andrew Johnston, Robert D. Boutin, Andrew Wentland, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, and Akshay S. Chaudhari. Merlin: a computed tomography vision-language foundation model and dataset. *Nature*, 2026. doi: 10.1038/s41586-026-10181-8. URL https://doi.org/10.1038/s41586-026-10181-8.

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.

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 (CVPR)*, pp. 11147–11158, 2024. URL https://github.com/MrGiovanni/DiffTumor.

Yu-Cheng Chou, Zongwei Zhou, and Alan Yuille. Embracing massive medical data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 24–35. Springer, 2024. URL https://github.com/MrGiovanni/OnlineLearning.

Pojsakorn Danpanichkul, Yanfang Pang, {Luis Antonio} Diaz, Disatorn Dejvajara, {Ethan Kai Jun} Tham, Siwanart Kongarin, Benjamin Koh, Thanathip Suenghataiphorn, Noppachai Siranart, Pooja Rangan, Kanak Das, Kanokphong Suparan, {Juan Pablo} Arab, {Michael B.} Wallace, {Jorge D.} Machicado, and Karn Wijarnpreecha. Global, regional, and national disparities in the burden of acute pancreatitis and alcohol-related pancreatitis from 2000 to 2021. *Mayo Clinic proceedings*, 101(7):1151–1162, July 2026. ISSN 0025-6196. doi: 10.1016/j.mayocp.2025.06.026. Publisher Copyright: © 2025 Mayo Foundation for Medical Education and Research. Published by Elsevier Inc. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

MJJ de Grauw, E Th Scholten, Ewoud J Smit, Matthieu JCM Rutten, Mathias Prokop, Bram van Ginneken, and Alessa Hering. The uls23 challenge: A baseline model and benchmark dataset for 3d universal lesion segmentation in computed tomography. *Medical image analysis*, 102:103525, 2025.

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.

Yunhe Gao, Mu Zhou, Di Liu, Zhennan Yan, Shaoting Zhang, and Dimitris N Metaxas. A data-scalable transformer for medical image segmentation: architecture, model efficiency, and benchmark. *arXiv preprint arXiv:2203.00131*, 2022.

Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images, 2022. URL https://arxiv.org/abs/2201.01266.

Sanne A Hoogenboom, Megan ML Engels, Anthony V Chuprin, Jeanin E van Hooft, Jordan D LeGout, Michael B Wallace, and Candice W Bolan. Prevalence, features, and explanations of missed and misinterpreted pancreatic cancer on imaging: a matched case–control study. *Abdominal Radiology*, 47(12):4160–4172, 2022.

Sarah Hooper, Mayee Chen, Khaled Saab, Kush Bhatia, Curtis Langlotz, and Christopher Ré. A case for reframing automated medical image classification as segmentation. *Advances in Neural Information Processing Systems*, 36:55415–55441, 2023.

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.

Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert, and Ben Glocker. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical image analysis*, 36:61–78, 2017.

Sukhwinder Kaur, Michael J Baine, Maneesh Jain, Aaron R Sasson, and Surinder K Batra. Early diagnosis of pancreatic cancer: challenges and new developments. *Biomarkers in medicine*, 6(5): 597–612, 2012.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

Tianyu Li, Chen Lin, and Weibin Wang. Global, regional, and national burden of pancreatic cancer from 1990 to 2021, its attributable risk factors, and projections to 2050: a systematic analysis of the global burden of disease study 2021. *BMC cancer*, 25(1):189, 2025a.

Wenxuan Li, Chongyu Qu, Xiaoxi Chen, Pedro RAS Bassi, Yijia Shi, Yuxiang Lai, Qian Yu, Huimin Xue, Yixiong Chen, Xiaorui Lin, Yutong Tang, Yining Cao, Haoqi Han, Zheyuan Zhang, Jiawei Liu, Tiezheng Zhang, Yujiu Ma, Jincheng Wang, Guang Zhang, Alan Yuille, and Zongwei Zhou. Abdomenatlas: A large-scale, detailed-annotated, & multi-center dataset for efficient transfer learning and open algorithmic benchmarking. *Medical Image Analysis*, pp. 103285, 2024. URL https://github.com/MrGiovanni/AbdomenAtlas.

Wenxuan Li, Pedro RAS Bassi, Tianyu Lin, Yu-Cheng Chou, Xinze Zhou, Yucheng Tang, Fabian Isensee, Kang Wang, Qi Chen, Xiaowei Xu, Jin Ye, Zheren Zhu, Sergio Decherchi, Andrea Cavalli, Alan L Yuille, and Zongwei Zhou. Scalemai: Accelerating the development of trusted datasets and ai models. *arXiv preprint arXiv:2501.03410*, 2025b. URL https://github.com/MrGiovanni/ScaleMAI.

Wenxuan Li, Xinze Zhou, Qi Chen, Tianyu Lin, Pedro RAS Bassi, Xiaoxi Chen, Chen Ye, Zheren Zhu, Kai Ding, Heng Li, Kang Wang, Yang Yang, Yucheng Tang, Daguang Xu, Alan L Yuille, and Zongwei Zhou. Pants: The pancreatic tumor segmentation dataset. In *Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2025c. URL https://github.com/MrGiovanni/PanTS.

Wenxuan Li, Pedro R. A. S. Bassi, Lizhou Wu, Xinze Zhou, Yuxuan Zhao, Qi Chen, Szymon Plotka, Tianyu Lin, Zheren Zhu, Marisa Martin, Justin Caskey, Shanshan Jiang, Xiaoxi Chen, Jarosław B. Ćwikła, Artur Sankowski, Yaping Wu, Sergio Decherchi, Andrea Cavalli, Chandana Lall, Cristian Tomasetti, Yaxing Guo, Xuan Yu, Yuqing Cai, Hualin Qiao, Jie Bao, Chenhan Hu, Ximing Wang, Arkadiusz Sitek, Kai Ding, Heng Li, Meiyun Wang, Dexin Yu, Guang Zhang, Yang Yang, Kang Wang, Alan L. Yuille, and Zongwei Zhou. Early and prediagnostic detection of pancreatic cancer from computed tomography. *arXiv preprint arXiv:2601.22134*, 2026. URL https://github.com/BodyMaps/ePAI.

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*, pp. 21152–21164, 2023. URL https://github.com/ljwztc/CLIP-Driven-Universal-Model.

Junqi Liu, Dongli He, Wenxuan Li, Ningyu Wang, Alan L Yuille, and Zongwei Zhou. Shapekit. In *International Workshop on Shape in Medical Imaging*, pp. 44–58. Springer, 2025. URL https://github.com/BodyMaps/ShapeKit.

Jun Ma, Feifei Li, and Bo Wang. U-mamba: Enhancing long-range dependency for biomedical image segmentation, 2024a. URL https://arxiv.org/abs/2401.04722.

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, 2024b. URL https://arxiv.org/abs/2408.12534.

Sovanlal Mukherjee, Anurima Patra, Hala Khasawneh, Panagiotis Korfiatis, Naveen Rajamohan, Garima Suman, Shounak Majumder, Ananya Panda, Matthew P Johnson, Nicholas B Larson, et al. Radiomics-based machine-learning models can detect pancreatic cancer on prediagnostic computed tomography scans at a substantial lead time before clinical diagnosis. *Gastroenterology*, 163(5):1435–1446, 2022.

Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM conference on health, inference, and learning*, pp. 151–159, 2020.

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 (NeurIPS) Datasets and Benchmarks Track*, volume 21, 2023. URL https://github.com/MrGiovanni/AbdomenAtlas.

Maximilian Rokuss, Moritz Langenberg, Yannick Kirchhoff, Fabian Isensee, Benjamin Hamm, Constantin Ulrich, Sebastian Regnery, Lukas Bauer, Efthimios Katsigiannopulos, Tobias Norajitra, and Klaus Maier-Hein. VoxTell: Free-text promptable universal 3D medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37538–37557, 2026.

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*, pp. 405–415. Springer, 2023.

Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 761–769, 2016.

Rebecca L. Siegel, Tyler B. Kratzer, Angela N. Giaquinto, Hyuna Sung, and Ahmedin Jemal. Cancer statistics, 2025. *CA: A Cancer Journal for Clinicians*, 75(1):10–45, 2025. doi: https://doi.org/10.3322/caac.21871. URL https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.3322/caac.21871.

Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

Tony W. Trinh, Daniel I. Glazer, Cheryl A. Sadow, V. Anik Sahni, Nina L. Geller, and Stuart G. Silverman. Bladder cancer diagnosis with CT urography: test characteristics and reasons for false-positive and false-negative results. *Abdominal Radiology*, 43(3):663–671, March 2018. doi: 10.1007/s00261-017-1249-6.

Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, volume 1, pp. I–I. Ieee, 2001.

Yingda Xia, Qihang Yu, Linda Chu, Satomi Kawamoto, Seyoun Park, Fengze Liu, Jieneng Chen, Zhuotun Zhu, Bowen Li, Zongwei Zhou, Yongyi Lu, Yan Wang, Wei Shen, Lingxi Xie, Yuyin Zhou, Christopher Wolfgang, Ammar Javed, Daniel Fadaei Fouladi, Shahab Shayesteh, Jefferson Graves, Alejandra Blanco, Eva S. Zinreich, Benedict Kinny-Köster, Kenneth Kinzler, Ralph H. Hruban, Bert Vogelstein, Alan L. Yuille, and Elliot K. Fishman. The felix project: Deep networks to detect pancreatic neoplasms. *medRxiv*, 2022.

Tiange Xiang, Yixiao Zhang, Yongyi Lu, Alan L Yuille, Chaoyi Zhang, Weidong Cai, and Zongwei Zhou. Squid: Deep feature in-painting for unsupervised anomaly detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23890–23901, 2023. URL https://github.com/tiangexiang/SQUID.

Tiange Xiang, Yixiao Zhang, Yongyi Lu, Alan Yuille, Chaoyi Zhang, Weidong Cai, and Zongwei Zhou. Exploiting structural consistency of chest anatomy for unsupervised anomaly detection in radiography images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. URL https://github.com/MrGiovanni/SimSID.

APPENDIX

TABLE OF CONTENTS

A Pancreas abnormalities mined as hard negatives .....
B Computational Resources .....
C Cascade verifier: training data, features, and hyperparameter .....
D Fine-tuning Details .....
E Ablation Study .....F CT normalization across sources .....
G Dataset details .....
H Detection stratified by tumor size .....I Qualitative examples .....

A PANCREAS ABNORMALITIES MINED AS HARD NEGATIVES

Out of 440,000 N-California radiology reports, our report-based mining pipeline (Sec. 2.1, Step II) flagged ∼7,000 patients with non-cancer pancreatic abnormalities. Fig. 2 breaks these candidates down by abnormality category. The distribution is long-tailed and motivates report-based mining. The rarest categories (necrosis, fluid collection, fat stranding) are precisely the ones most underrepresented in standard training datasets and most likely to drive false positives at deployment time.

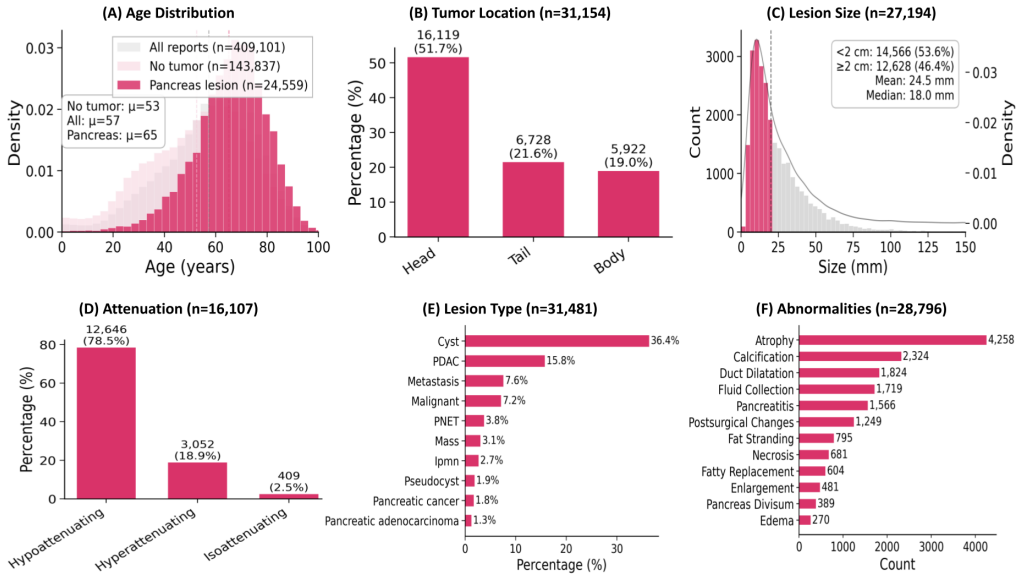


Figure 2: **Non-cancer abnormalities selected by our LLM as hard negatives.** Abnormalities in the pancreas can look like cancer and confound tumor detection models. These abnormalities are often rare and underrepresented in training datasets. DREAM used LLMs to read 440,000 radiology reports and found 7,000 non-cancer pancreatic abnormalities (hard negatives). Adding these abnormalities to the training set teaches the segmentation model that they are not cancer.

B COMPUTATIONAL RESOURCES

All experiments were carried out on two internal compute nodes. The primary training node is a Dell PowerEdge XE8640 server equipped with **NVIDIA H100** GPUs (80 GB HBM3 each). The secondary node has 8×**NVIDIA TITAN RTX** GPUs (24 GB each). All code is released for reproducibility.

C CASCADE VERIFIER: TRAINING DATA, FEATURES, AND HYPERPARAMETERS

Multi-class radiomics classifier. We train the classifier on $n = 26{,}711$ CT scans from N-California, with two further splits (validation $n = 5{,}739$, test $n = 5{,}729$) that share no patients with each other or with the training set. Per-case labels for the 17 classes (pancreatitis, atrophy, fluid collection, calcification, fat stranding, duct dilation, fatty replacement, enlargement, necrosis, divisum, edema, post-surgical, pancreatic tumor, PNET, cyst, healthy, and other abnormalities) are extracted from the corresponding radiology reports using the same hospital-scale LLM pipeline of Sec. 2.1. For each case we extract 132 pyradiomics features (shape, first-order intensity, and texture descriptors) from the pancreas region. We then train one binary LightGBM classifier (Ke et al., 2017) per class. Hyperparameters are tuned on the validation split.

Cascade arm C_{seg}. For each case we extract 26 summary features from DREAM's multi-class voxel mask (pancreas, duct, PDAC, cyst, PNET): per-class volume, top-two voxel sum, connected-component count, and largest-component logit. We add six derived features: $\log$-volume, volume$\times$logit, tumor-to-pancreas volume ratio, duct-dilation indicator, high-logit indicator, and duct$\times$tumor interaction. We train an ensemble of five LightGBM models with different random seeds, combined with a logistic-regression meta-learner. Training uses $n = 2{,}239$ cases: $1{,}500$ tumor-positive scans and 739 strictly clean-healthy negatives. "Clean-healthy" means negative across the tumor label and 14 other pancreatic abnormality labels. The strict clean-healthy filter is necessary. Without it, duct-volume features become *negatively* correlated with the tumor label on training data ($\text{AUC} = 0.31$) because chronic pancreatitis and pancreatic tumor share many imaging signatures. Restricting to clean-healthy negatives flips this to a clinically correct positive correlation ($\text{AUC} = 0.78$).

Cascade arm C_{tumor}. We use the pancreatic-tumor binary head of the classifier above, frozen and without any further fine-tuning. Each case's C_{tumor} is the head's predicted tumor probability.

Cascade arm $C_{\text{tumor_abn}}$. We use the 17-dimensional probability vector emitted by the radiomics classifier as a feature representation. We train a second ensemble of five LightGBM models with different random seeds, combined with a logistic-regression meta-learner, to predict the tumor label. Training pool, label filtering, and hyperparameters match the C_{seg} setup. The intuition is that the relative pattern across all 17 classes (e.g., simultaneously high tumor probability, high duct dilation, and low pancreatitis) carries information beyond any single head's probability.

D FINE-TUNING DETAILS

The nnU-Net segmentation model in DREAM is first trained with its automatically defined training parameters. During fine-tuning on hard negatives and hard positives, we freeze the U-Net encoder and train only the decoder. We use a batch size of 16, a learning rate of 1×10^{-4}, and a weight decay of 1×10^{-5}. Training runs for 100 epochs. We select the best checkpoint based on validation performance, using a randomly sampled 20% hold-out split from the training data. We optimize the model with cross-entropy loss, DSC loss, and Focal Tversky loss (Abraham & Khan, 2019).

E ABLATION STUDY

We ablate the three components of DREAM to isolate each one's contribution: (i) hard positive and hard negative mining, (ii) report-refined tumor masks, and (iii) the cascade verifier. We remove each component in turn and compare the resulting variants against the full DREAM (Tab. 3). The *no report-refine.* variant keeps mining as in DREAM but replaces the report-refined masks with naive pseudo-labels thresholded at 0.5. The *no mining* variant additionally replaces the mined hard cases with the same number of randomly sampled positives and negatives from N-California. Beyond the removal ladder, three further arms probe the design: *image-based mining* substitutes the mining strategy, and two *report noise* arms perturb the reported tumor diameters.

Every component raises detection AUC. Training-time mining and report-refined masks together account for the largest improvement on PanTS, lifting AUC from 0.829 to 0.907 ($+0.078$); the

Table 3: **Component ablations and a mining-strategy comparison on the PanTS and N-Cal. test sets.** We remove each DREAM component in turn: (i) report-based mining, (ii) report-refined masks, and (iii) the cascade verifier. The *image-based mining* row replaces report-based mining with a baseline that retrains on the deployed model's errors, confirmed through follow-up and pathology, holding every other factor fixed: initialization, architecture, optimizer, loss, schedule, mask creation, and fine-tuning set size. It sits beside the no-cascade row because that is the configuration it matches. The two *report noise* rows perturb every reported tumor diameter by up to $\pm 10\%$ or $\pm 25\%$ and re-run the full pipeline; neither change is significant ($p \geq 0.73$). The lower block sweeps the mask-quality threshold; it runs on a different cohort and operating point, both named in the block header, so its rows are not comparable to the rows above. Sensitivity and specificity in the upper blocks are F1-optimal and values are percentages; dashes mark quantities the source run did not report. **Blue** denotes statistical significance ($p < 0.05$) by DeLong's paired test for AUC. No significance test exists for the mining contrast.

	PanTS				N-Cal.		
Model	Sen	Spec	AUC	DSC	Sen	Spec	AUC
DREAM, no cascade, no report-ref., no mining	73.5	84.1	0.829	0.42	85.9	85.0	0.906
DREAM, no cascade, no report-ref.	74.2	90.5	0.862	0.40	85.9	86.7	0.916
DREAM, no cascade, image-based mining	75.5	91.7	0.895	–	83.8	91.3	0.922
DREAM, no cascade	75.5	96.0	0.907	0.60	80.9	96.2	0.928
+ report noise $\pm 10\%$	–	–	0.911	–	–	–	0.927
+ report noise $\pm 25\%$	–	–	0.912	–	–	–	0.927
DREAM	75.5	96.0	0.915	0.60	84.0	95.9	**0.973**
Mask-quality threshold. N-Cal. small-tumor subset, mask-donor patients excluded; no cascade verifier; sensitivity compared at a matched 95.4% specificity.							
retain 70 masks ($+50\%$)	–	–	–	–	87.2	95.4	0.926
retain 46 masks (as shipped)	–	–	–	–	85.1	95.4	0.928
retain 23 masks (-50%)	–	–	–	–	80.9	95.4	0.893
no mined hard positives	–	–	–	–	77.7	95.4	0.877

same ordering holds on N-Cal ($0.906 \rightarrow 0.928$, $+0.022$). Report-refined masks also carry the segmentation gain, raising PanTS DSC from 0.40 to 0.60. The cascade verifier adds 0.045 AUC on N-Cal ($p = 0.0009$) and 0.009 on Euro ($p = 0.0009$); on PanTS it adds 0.008, which does not reach significance ($p = 0.196$), and leaves the PanTS operating point unchanged.

Report-based mining suppresses more false alarms than image-based mining. Hard cases can also come from the model itself: deploy the initial model, confirm its true errors through follow-up and pathology, and retrain on those. We implemented this *image-based mining* baseline over the $6{,}294$ N-California scans we could download, holding every other factor fixed. Compared at the same configuration, report-based mining raises specificity by 4.3 pp on PanTS and 4.9 pp on N-California and raises AUC from 0.895 to 0.907 and from 0.922 to 0.928. Image-based mining falls further behind on individual abnormalities, reaching 59.4% specificity on duct dilatation against 69.6% for DREAM, and 77.2% on atrophy against 86.4%.

Report-refined masks tolerate approximate reported sizes. Radiology reports state tumor size approximately, so we measured how much that approximation costs. We perturbed every reported tumor diameter by up to $\pm 10\%$ or $\pm 25\%$, regenerated the report-refined masks, fine-tuned DREAM again, and re-evaluated detection without the cascade verifier. AUC did not move: N-California 0.928 against 0.927 at both noise levels ($p = 0.98$, $p = 0.96$), PanTS 0.907 against 0.911 and 0.912 ($p = 0.75$, $p = 0.73$). Two design choices explain the tolerance. DREAM confirms tumor existence and malignancy against pathology reports, not radiology reports alone, and it leaves an AI-generated mask unrefined when its diameter already falls within $\pm 20\%$ of the reported diameter.

Performance is insensitive to the mask-quality threshold. DREAM fine-tunes on the 46 report-refined masks with the highest mask quality score. We moved that threshold to retain 70 and 23 masks, and separately trained a variant with no mined hard positives at all. We evaluated every variant on small-tumor detection in N-California, without the cascade verifier, after excluding each patient who supplied one of the 70 masks, and compared sensitivity at a matched 95.4% specificity. Retaining 50% more masks leaves detection essentially unchanged: AUC 0.926 against 0.928, sensitivity 87.2%

Table 4: **The DREAM recipe applied to three other backbones.** Each backbone is trained on E-Coast, then fine-tuned with DREAM's mined hard cases and report-refined masks. Values are detection AUC in percentage points. Operating points are not comparable across arms, so we report AUC only.

	PanTS			N-Cal.
Backbone	base	+ DREAM	Δ	+ DREAM
SwinUNETR Hatamizadeh et al. (2022)	69.6	78.7	+9.1	83.5
MedNeXt Roy et al. (2023)	82.4	88.9	+6.5	90.5
U-Mamba Ma et al. (2024a)	83.5	89.9	+6.4	90.8

against 85.1%. Halving the masks costs 0.035 AUC, and training without mined hard positives costs 0.051.

F CT NORMALIZATION ACROSS SOURCES

Normalization governs whether a segmentation model transfers between hospitals, so we describe ours in full. We adopt the nnU-Net V2 preprocessing pipeline unchanged, which derives its normalization parameters from an analysis of the training set rather than from hand-set constants. We chose this pipeline for two reasons. Public benchmarks, including Touchstone (Bassi et al., 2024), report that segmentation models built on nnU-Net V2 generalize across hospitals better than the alternatives. The pipeline is also public, documented, and widely used, which makes our preprocessing reproducible.

The pipeline applies three operations. It reorients every scan to a standard RAS orientation. It then clips CT intensities to the 0.5th and 99.5th percentiles of the foreground intensities of the training set, which for our data span -1000 to 371 HU. It standardizes the clipped values using the foreground mean and standard deviation of that same training set. Finally, it resamples every scan to the median resolution of the training set, which for our data is $3.0 \times 0.71875 \times 0.71875$ mm (slice thickness $\times$ in-plane). We apply the identical transform at training and at inference, including on all external cohorts.

Together these operations remove differences in orientation, intensity range, and voxel spacing between acquisition sites. Resolution standardization matters most. Models that skip it inherit the scanner's slice thickness as a confounder, and Touchstone reports that such models transfer poorly between hospitals.

G DATASET DETAILS

E-Coast (training, $N = 3{,}440$, private). CT scans from a major US east-coast hospital, including 2101 pancreatic tumor cases of which 11.8% carry tumors under 2 cm. Eight radiologists drew the tumor masks over five years, which is the annotation cost DREAM is designed to avoid repeating.

N-California Reports (training, $N = 440{,}000$, private). Twenty-seven years of abdominal CT reports from northern California hospitals. The archive carries no tumor masks, which reflects ordinary clinical data. We mine it for hard cases (Sec. 2.1) and download only the CT scans the mining selects.

N-California Test (pancreas, $N = 1{,}046$, internal). Drawn from the same hospitals as the report archive: 952 randomly selected non-cancer scans and 94 pancreatic tumors under 2 cm. No patient in this cohort contributes a mined training case.

PanTS (pancreas, $N = 901$, public, external). The largest public pancreatic tumor CT dataset (Li et al., 2025c), drawn from 145 institutions in 12 countries. We use the official test split: 151 tumor cases and 750 non-cancer scans, with no duplicates and no case in both rosters. DREAM never trains on PanTS, and its fine-tuning rosters share no case with this cohort. Two properties of the ground truth bound what we can claim from it. The annotation is a single binary pancreatic-lesion channel with no subtype labels, so detection here covers any pancreatic lesion rather than adenocarcinoma

alone. Tumor size is the sphere-equivalent diameter of the largest connected component of that mask, not a measured diameter. On the 26 cases where an independent reported size also exists, half fall in a different size stratum under the two definitions.

Euro (pancreas, $N = 976$, private, external). CT scans with radiology and pathology reports and no tumor masks, from one hospital in Türkiye and several in Poland: 722 pancreatic tumor cases and 254 non-cancer cases.

Pre-diagnostic (pancreas, $N = 33$, internal). Gold standard scans acquired before the patient's pancreatic cancer was diagnosed, each paired to the later scan that confirmed it, a median of 563 days later (range 94–1,116). A scan enters the positive set only when a pathology report confirms malignancy, and enters the negative set only when a tumor-free follow-up at least two years later confirms the absence of cancer. A lesion the original radiologist overlooked therefore cannot enter our data as a negative.

Bladder and gallbladder cohorts. We evaluate each organ on one internal and one external cohort. In N-California the bladder cohort holds 40 tumors and 44 non-cancer scans, and the gallbladder cohort holds 13 tumors and 30 non-cancer scans. In Euro the bladder cohort holds 171 tumors and 227 non-cancer scans, and the gallbladder cohort holds 28 tumors and 243 non-cancer scans. The Euro organ cohorts come from a different source than the pancreatic Euro set, and 199 of those scans carry an assessment for both organs, so the two are not independent. Gallbladder cancer is rare, which is why these are the smallest cohorts in the paper.

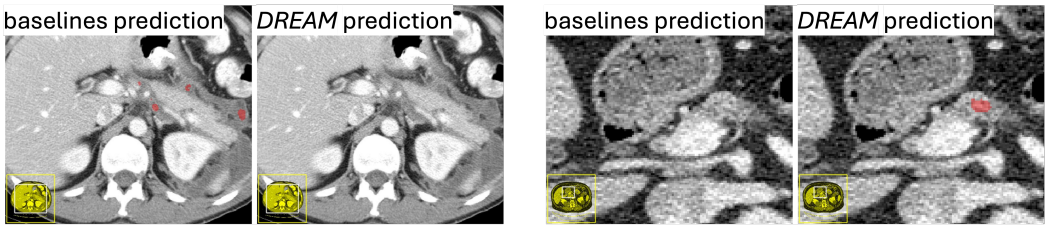
H Detection stratified by tumor size

Table 5: **Detection stratified by tumor size.** In N-California the $\leq$2 cm arms are scored on 1,051 and 993 scans (99 and 94 tumors) and the $>$2 cm arms on a 50-tumor subsample against the same non-cancer scans. PanTS splits into 67 tumors $\leq$2 cm and 84 above 2 cm against 750 non-cancer scans, and both PanTS arms are compared at a matched 84.0% specificity. PanTS tumor size is the sphere-equivalent diameter of the largest connected component of the ground-truth lesion mask. **Blue** marks DREAM's improvement.

	nnU-Net Isensee et al. (2021)			DREAM		
Cohort and tumor size	Sen	Spec	AUC	Sen	Spec	AUC
N-California, $\leq$2 cm	64.6	94.4	0.887	**83.0**	96.3	**0.931**
N-California, $>$2 cm	95.8	88.9	0.968	**97.9**	93.2	**0.983**
PanTS, $\leq$2 cm	59.7	84.0	0.759	**82.1**	84.0	**0.875**
PanTS, $>$2 cm	91.7	84.0	0.877	91.7	84.0	**0.947**

I Qualitative examples

J ROC Curve and Reader Studies



(a) *DREAM* reduced false positives (b) *DREAM* reduced false negatives

Figure 3: **DREAM reduces false positives and false negatives. (a)** The nnU-Net Baseline (left) misclassifies pancreatitis with fluid collection as a tumor. DREAM (right), trained with mined hard negatives, correctly predicts no tumor. **(b)** The nnU-Net Baseline (left) misses a 9 mm borderline-resectable tumor that DREAM (right) detects after training on mined hard positives with report-refined masks.

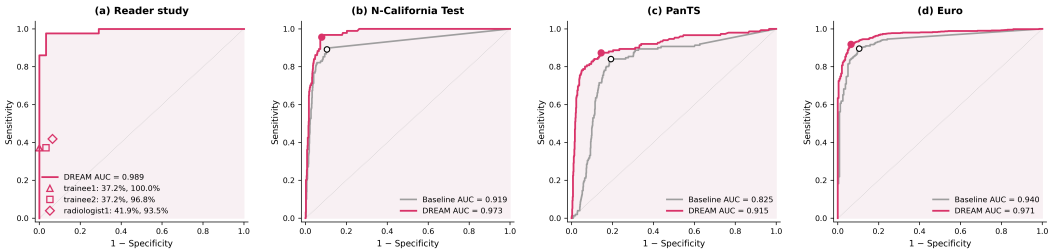


Figure 4: **DREAM surpasses state-of-the-art models and radiologists.** Plots show ROC curves for pancreatic cancer detection. **(a)** Reader study on 74 N-California scans: DREAM reaches 90.7% sensitivity at 96.8% specificity, well above three radiologists (37–42% sensitivity at 94–100% specificity). **(b–d)** DREAM improves AUC over the nnU-Net Baseline (from which DREAM was fine-tuned) on N-California ($0.919 \rightarrow 0.973$), PanTS ($0.824 \rightarrow 0.915$), and Euro ($0.940 \rightarrow 0.971$) In particular, DREAM performs much better than the baseline on PanTS test set under high specificity.