

<https://doi.org/10.1038/s41746-025-01970-y>

Deep learning CT model for stratified diagnosis of pancreatic cystic neoplasms: multicenter development, validation, and real-world clinical impact



Xiaohan Yuan^{1,5}, Chengwei Chen^{1,5}, Zhang Shi^{2,5}, Wenbin Liu¹, Xinyue Zhang¹, Ming Yang¹, Mengmeng Zhu¹, Jieyu Yu¹, Fang Liu¹, Jing Li¹, Yunshuo Zhang³, Hui Jiang³, Bozhu Chen⁴✉, Jianping Lu¹✉, Chengwei Shao¹✉ & Yun Bian¹✉

Pancreatic cystic neoplasms (PCN) are critical precursors for early pancreatic cancer detection, yet current diagnostic methods lack accuracy and consistency. This multicenter study developed and validated an artificial intelligence (AI)-powered CT model (PCN-AI) for improved assessment. Using contrast-enhanced CT images from 1835 patients, PCN-AI extracted 63 quantitative features to classify PCN subtypes through four hierarchical tasks. A multi-reader, multi-case (MRMC) study demonstrated that AI assistance significantly improved radiologists' diagnostic accuracy (AUC: 0.786 to 0.845; $p < 0.05$) and reduced interpretation time by 23.7% (5.28 vs. 4.03 minutes/case). Radiologists accepted AI recommendations in 87.14% of cases. In a prospective real-world cohort, PCN-AI outperformed radiologist double-reading, providing actionable diagnostic benefits to 45.45% of patients (5/11) by correctly identifying missed malignant PCN cases, enabling timely intervention, and simultaneously reducing clinical workload by 39.3%. PCN-AI achieved robust performance across tasks (AUCs: 0.845–0.988), demonstrating its potential to enhance early detection, precision management, and diagnostic efficiency in clinical practice.

Pancreatic cancer, with a five-year survival rate of 11%, ranks as the third and sixth leading cause of cancer-related deaths in the US and China, respectively^{1,2}. Early detection of pancreatic cystic neoplasms (PCN)—a high-risk precursor group—is critical to improving survival and reducing mortality^{3,4}. Current guidelines emphasize imaging-based “high-risk stigmata” (e.g., main pancreatic duct [MPD] dilation ≥ 5 mm) and “worrisome features” (e.g., cyst size ≥ 3 cm) for malignancy risk stratification^{5–9}. However, clinical implementation faces three fundamental limitations:

First, manual quantification shows concerning variability. Inter-observer agreement for mural nodule detection reaches only 62.2% positive predictive value¹⁰, while cyst size thresholds demonstrate paradoxical correlations with malignancy (AUC = 0.52–0.68 across studies)^{11–13}. Second, evolving evidence challenges conventional criteria: MPD dilation thresholds now require anatomical specificity (≥ 9 mm in pancreatic head vs. ≥ 7 mm in

body/tail)¹⁴ and rapid growth rate (≥ 3 mm/year) shows stronger malignancy association than absolute size¹⁴. Third, invasive modalities like endoscopic ultrasound-guided fine needle aspiration (EUS-FNA) suffer from 18–34% false-negative rates and procedure-related complications¹⁵, leaving non-invasive CT/MRI as the frontline modality despite its subjective interpretation.

Artificial intelligence (AI) offers transformative potential for addressing these challenges. While AI has shown promise in binary cyst classification^{16,17}, existing models lack: (1) hierarchical classification of PCN subtypes (e.g., intraductal papillary mucinous neoplasm [IPMN] vs. mucinous cystic neoplasm [MCN]); (2) interpretable feature quantification aligned with clinical criteria; and (3) multicenter validation—gaps that limit clinical adoption^{16,18,19}. For example, most models fail to distinguish serous cystic neoplasms (SCN) from solid pseudopapillary neoplasms (SPN), a critical distinction for management²⁰.

¹Department of Radiology, Changhai Hospital, Shanghai, China. ²Department of Radiology, Zhongshan Hospital, Shanghai, China. ³Department of Pathology, Changhai Hospital, Shanghai, China. ⁴Department of Radiology, Nanjing Drum Tower Hospital, Nanjing, China. ⁵These authors contributed equally: Xiaohan Yuan, Chengwei Chen, Zhang Shi. ✉e-mail: chencbz@qq.com; cjr.lujianping@vip.163.com; chengweishaoch@163.com; bianyun2012@foxmail.com

To address these gaps, we developed PCN-AI, an interpretable CT-based model that: (1) performs hierarchical classification of four PCN subtypes using 63 quantitative imaging biomarkers; (2) aligns with WHO diagnostic criteria by differentiating *precancerous* (IPMN/MCN with low-grade dysplasia) from *malignant* lesions (high-grade dysplasia or invasive carcinoma)²¹; and (3) underwent multicenter validation with real-world implementation. By integrating radiologist workflows and reducing diagnostic variability, PCN-AI aims to enhance early cancer detection and precision management of PCN patients.

Results

Study design

Figure 1 presents a schematic illustration of the overall study design. This study was divided into four stages.

Stage 1: Image Segmentation. Key imaging features were delineated from the training set. A segmentation model was trained to quantify CT characteristics comprehensively.

Stage 2: Classification Tasks: Four tasks were performed using quantitative features from Stage 1. Firstly, Classification Task 1 involved establishing a two-class system to distinguish between mucinous PCN (comprising IPMN and MCN) and non-mucinous PCN (comprising SCN and SPN). Building on the results of Task 1, Classification Task 2 further classified mucinous PCN into precancerous and malignant categories. Subsequently, Classification Task 3 focused on differentiating between IPMN and MCN within the mucinous PCN group. Lastly, Classification Task 4 distinguished between SCN and SPN within the non-mucinous PCN category.

Stage 3: This stage involved a multi-reader, multi-case (MRMC) study conducted in March 2024, with 500 randomly selected patients. Eight readers of varying experience levels participated in a randomized crossover design to evaluate AI's impact on diagnostic accuracy and assess doctors' willingness to use AI.

Stage 4: This stage encompassed a Real-World Data (RWD) Cohort study, with the primary objectives of evaluating the AI's diagnostic accuracy and its impact on radiologists' workload. During this process, AI-assisted diagnosis was compared with traditional double-reader diagnosis (Fig. 2).

Baseline clinical characteristics

We initially recruited 2359 patients with PCN from four hospitals between January 2012 and December 2023. After carefully applying the inclusion and exclusion criteria, we selected 1835 patients (mean age 54 ± 16 years, 1028 females) for inclusion in our study. From this group, 1249 patients formed the primary cohort, enrolled between June 2012 and March 2024. This primary cohort was further divided into two sets: a training set consisting of 999 patients with PCN (mean age 54 ± 16 years, 560 females) and a validation set comprising 250 patients with PCN (mean age 52 ± 16 years, 147 females). Additionally, we established three external test sets: Test 1 included 173 patients with PCN (mean age 54 ± 16 years, 96 females) identified between March 2012 and January 2024. Test 2 comprised 289 patients with PCN (mean age 54 ± 16 years, 172 females) enrolled between January 2017 and December 2021. Test 3 consisted of 124 patients with PCN (mean age 61 ± 13 years, 53 females) identified between May 2014 and June 2024.

The training set included 299 precancerous (29.9%) and 214 malignant (21.4%) cases of IPMN, 72 precancerous (7.2%) and 7 malignant (7.0%) cases of MCN, 212 cases of SCN (21.2%), and 195 cases of SPN (19.5%). The validation set featured 90 precancerous (36.0%) and 39 malignant (15.6%) IPMN cases, 18 precancerous (7.2%) and 1 malignant (0.4%) MCN case, 53 SCN cases (21.2%), and 49 SPN cases (19.6%). Test 1 encompassed 50 precancerous (28.9%) and 40 malignant (23.1%) IPMN cases, 13 precancerous (7.5%) and 1 malignant (0.6%) MCN case, 37 SCN cases (21.3%), and 32 SPN cases (18.4%). Test 2 included 74 precancerous (25.6%) and 30 malignant (10.4%) IPMN cases, 47 precancerous (16.3%) and 1 malignant (0.3%) MCN case, 84 SCN cases (29.0%), and 53 SPN cases (18.3%). Test 3 contained 57 precancerous (46.0%) and 37 malignant (29.8%) IPMN cases,

6 precancerous (4.8%) and 1 malignant (0.8%) MCN case, 10 SCN cases (8.0%), and 13 SPN cases (10.4%). The patient characteristics are presented in Table 1. The patient selection process is shown in Fig. 3.

Segmentation of AI model

The overall segmentation performance of the model was good, with DSCs all above 0.8; details are shown in Supplementary Table 1. Example images of the model's segmentation results are displayed in Supplementary Fig. 1. Based on the segmentation results, 63 quantitative features across four main categories were automatically extracted. The parameters most significant for model construction are shown in Table 2, comprising 14 parameters including patient age, sex, total pancreatic volume, net pancreatic volume, maximum diameter of the MPD in the pancreatic head/body/tail, the number of tumors in the pancreatic head/body/tail, maximum diameter of the tumor, average attenuation of solid tumor component in the largest tumor cross-section in arterial phase and portal phase, solid component volume of the tumor, solid component volume ratio, and maximum attenuation of solid tumor component within the tumor volume in portal phase. The remaining parameters are detailed in Supplementary Table 2. Comparative evaluation of six segmentation architectures revealed Mamba U-Net's superior performance with tumor-only DSC of 0.85 ± 0.02 and all-tissue DSC of 0.90 ± 0.02 , outperforming nnU-Net (0.79 ± 0.05), SwinU-NETR (0.82 ± 0.04), and other baseline methods (Supplementary Table 3, Supplementary Fig. 2).

Performance of AI models in classification task 1

In Classification Task 1, we used the MRMR method to carefully select eight features from clinical characteristics and 63 CT quantitative measurements that are most closely related to distinguishing mucinous from non-mucinous cystic tumors. These eight key features are: sex, total pancreatic volume, solid component volume ratio, number of tumors in the pancreatic head, average attenuation of solid tumor components in the arterial phase on the largest tumor cross-section, and the maximum diameters of the pancreatic ducts in the head, body, and tail of the pancreas. Next, we incorporated these eight features into a regression model for analysis. Both univariate and multivariate analyses revealed that these features are significantly correlated with the distinction between mucinous and non-mucinous cystic tumors ($P < 0.05$) (Fig. 4A, Supplementary Table 4).

The AI model demonstrated excellent discrimination in the classification Task 1, with an AUC of 0.946 (95% CI: 0.933–0.959) on the training set and 0.942 (95% CI: 0.914–0.970) on the validation set. The AUCs were 0.923 (95% CI: 0.883–0.962), 0.855 (95% CI: 0.813–0.897), and 0.877 (95% CI: 0.811–0.941) for tests 1–3, respectively.

Performance of AI models in classification task 2

In Classification Task 2, we input clinical characteristics along with all 63 CT quantitative features into the XGBoost model to distinguish between precancerous and malignant mucinous tumors. After analyzing the model, we identified the following six CT features as the most closely related to differentiating precancerous from malignant mucinous tumors: pancreatic gland volume, tumor solid component volume, solid component volume ratio, average attenuation of solid tumor components in the portal phase on the largest tumor cross-section, maximum attenuation of solid tumor components volume in the portal phase, and the maximum diameter of the pancreatic duct in the head of the pancreas (Fig. 4B).

In Classification Task 2, the AI model demonstrated impressive discriminatory power. The AUC was 0.976 (95% CI: 0.965–0.987) for the training set and 0.858 (95% CI: 0.777–0.939) for the validation set. For the test sets, the AUCs were 0.826 (95% CI: 0.735–0.917), 0.815 (95% CI: 0.689–0.942), and 0.903 (95% CI: 0.835–0.970) for Test 1, 2, and 3, respectively.

Performance of AI models in classification task 3

In Classification Task 3, we utilized the mRMR method to identify 2 features from clinical data and 63 CT quantitative features that are most relevant for

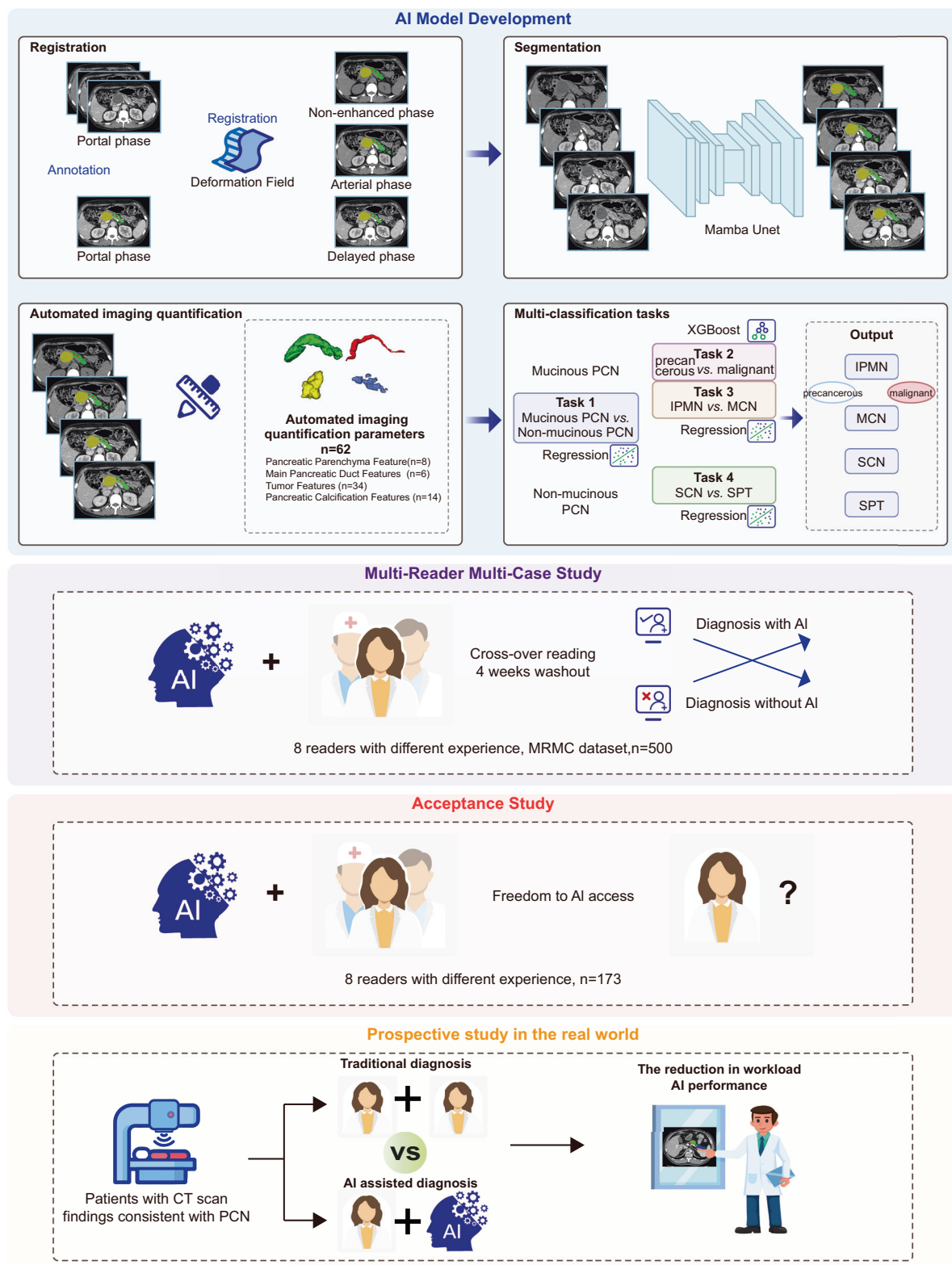


Fig. 1 | Flow chart for development and validation of PCN-AI.

distinguishing IPMN from MCN. The selected features include: age, sex, maximum tumor diameter, the number of tumors in the head of the pancreas, the maximum diameter of the pancreatic duct in the head of the pancreas, and the maximum diameter of the pancreatic duct in the tail of the pancreas. We then input these six features into a regression model for

further analysis. Univariate analysis demonstrated that all six features are significantly associated with differentiating IPMN from MCN ($P < 0.001$). Additionally, multivariate analysis revealed that sex, the number of tumors in the head of the pancreas, the maximum diameter of the pancreatic duct in the head of the pancreas, and the maximum diameter of the pancreatic duct

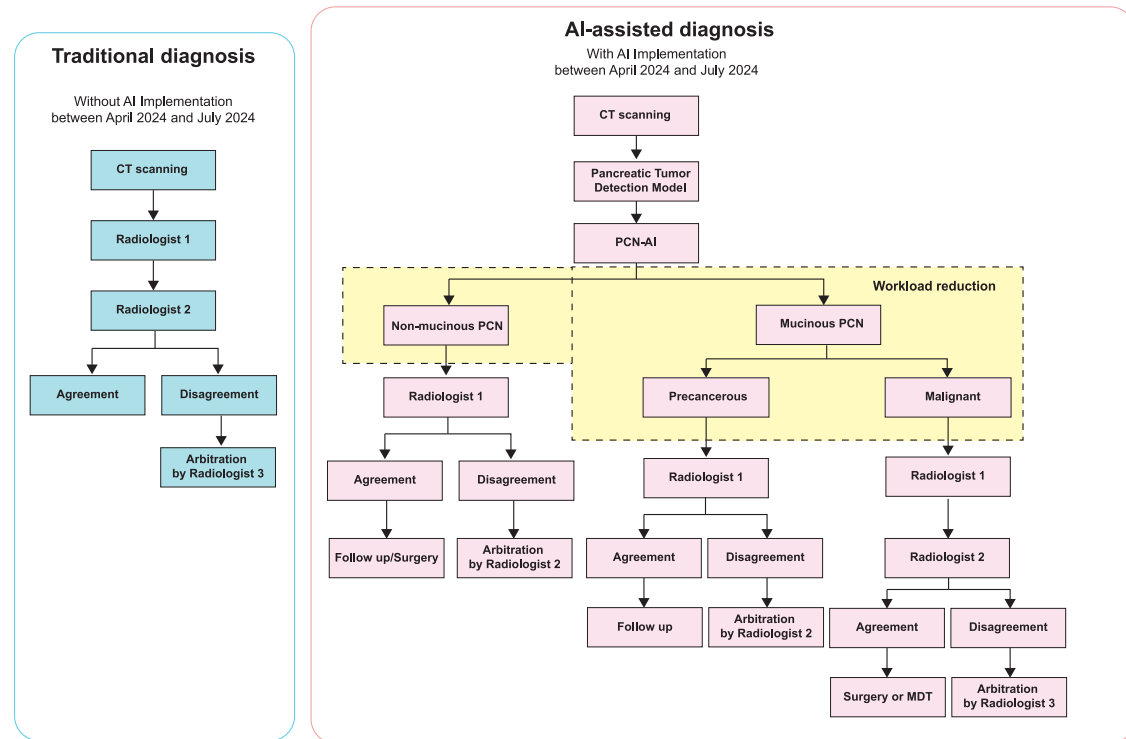


Fig. 2 | Overview of the real-world study. Based on the Pancreatic Tumor Detection Model, PCN-AI was deployed at Hospital 1 for a real-world simulation study.

Table 1 | Clinical and pathological characteristics of all datasets

Characteristics	Training set (n = 999)	Validation set (n = 250)	Test 1 (n = 173)	Test 2 (n = 289)	Test 3 (n = 124)
Age (years), Median (IQR)	56.0(23.0)	55.0(23.3)	57.0(25.0)	56.0(20.5)	63.0(12.0)
Sex, n(%)					
Female	560(56)	147(58.8)	96(55.4)	172(59.5)	53(42.7)
Male	439(44)	103(41.2)	77(44.6)	117(40.5)	71(57.3)
BMI (kg/m ²), Mean ± SD	23.0 ± 3.3	22.7 ± 3.2	22.7 ± 3.5	23.2 ± 3.7	22.4 ± 3.2
CA19-9 μmol/L, Median (IQR)	9.1(18.4)	9.3(17.0)	7.9(20.5)	10.8(15.4)	13.4(19.6)
CEA (ng/ml), Median (IQR)	2.0(2.0)	1.9(1.8)	1.7(2.2)	0.8(1.4)	2.2(2.1)
Abdominal pain, n(%)	350(35.0)	99(39.6)	58(33.5)	84(29.0)	27(21.7)
Jaundice, n(%)	41(4.1)	9(3.6)	7(4.0)	5(1.7)	4(3.2)
Diabetes, n(%)	145(14.5)	19(7.6)	26(15)	30(10.3)	8(6.4)
Alcohol consumption, n(%)	80(8.0)	22(8.8)	15(8.6)	33(11.4)	7(5.6)
Family history, n(%)	18(1.8)	3(1.2)	1(0.5)	12(4.1)	0(0.0)
Pathological classification, n(%)					
Precancerous IPMN	299(29.9)	90(36.0)	50(28.9)	74(25.6)	57(46.0)
Malignant IPMN	214(21.4)	39(15.6)	40(23.1)	30(10.4)	37(29.8)
Precancerous MCN	72(7.2)	18(7.2)	13(7.5)	47(16.3)	6(4.8)
Malignant MCN	7(0.7)	1(0.4)	1(0.6)	1(0.3)	1(0.8)
SCN	212 (21.2)	53(21.2)	37(21.3)	84(29.0)	10(8.0)
SPN	195 (19.5)	49(19.6)	32(18.4)	53(18.3)	13(10.4)

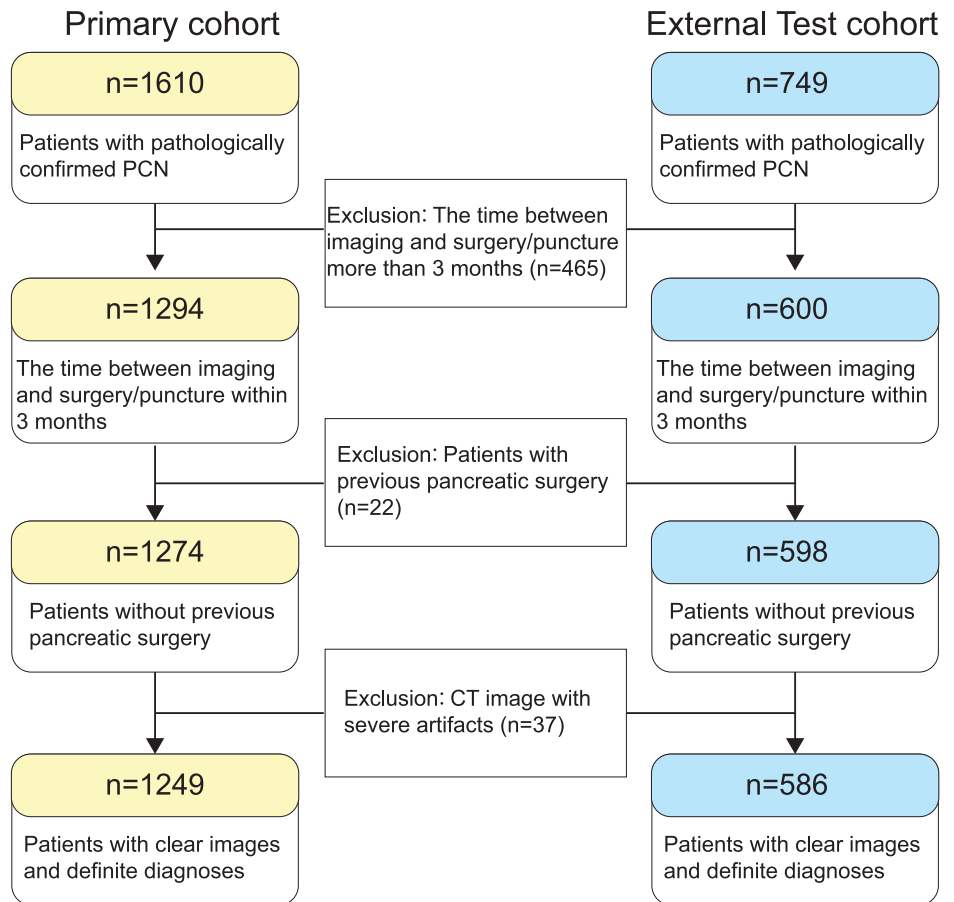
Note: Categorical variables are expressed as the number of cases (%), and continuous variables as the median (interquartile range).

IPMN intraductal papillary mucinous cystic neoplasms, MCN mucinous cystic neoplasms, SCN serous cystic neoplasms, SPN solid pseudopapillary neoplasms.

in the tail of the pancreas are independently significant predictors for distinguishing between IPMN and MCN ($P < 0.001$) (Fig. 4C, Supplementary Table 5).

For Classification Task 3, the AI model maintained good discrimination with an AUC of 0.963 (95% CI: 0.935–0.991) on the training set and an

impressive AUC of 0.971 (95% CI: 0.945–0.996) on the validation set. The AUCs for the test sets were recorded as follows: Test 1 at 0.958 (95% CI: 0.887–1.000), Test 2 at 0.960 (95% CI: 0.919–1.000), and Test 3 with a lower AUC of 0.749 (95% CI: 0.494–1.000). Performance variation in Test 3 reflects extreme class imbalance (1 malignant MCN case, 6 MCN cases total)

Fig. 3 | Flow diagram of the inclusion and exclusion process for the cohort of pancreatic cystic neoplasms (PCN).**Table 2** | CT quantified characteristics in all datasets

Characteristics	Training set (n = 999)	Validation set (n = 250)	test 1 (n = 173)	test 2 (n = 289)	test 3 (n = 124)
Age, years, Median (IQR)	56.0(23.0)	55.0(23.0)	57.0(25.0)	56.0(20.0)	63.0(12.0)
Sex, male, n(%)	439.0(43.9)	103.0(41.2)	77.0(44.5)	117.0(40.5)	71.0(57.3)
Total pancreatic volume, cm ³ , Median (IQR)	80.5(68.6)	80.7(65.9)	79.2(56.7)	70.4(36.7)	57.5(32)
Net pancreatic volume, cm ³ , Median (IQR)	77.2(64.4)	77.1(64.4)	78.2(52.5)	68.8(35.6)	54.8(29.8)
Maximum diameter of the MPD in the pancreatic head, mm, Median (IQR)	4.0(5.5)	4.0(5.1)	4.0(3.8)	2.8(2.5)	2.0(2.0)
Maximum diameter of the MPD in the pancreatic body, mm, Median (IQR)	4.0(6)	4.0(4.6)	4.5(5.5)	2.0(4.5)	2.0(2.0)
Maximum diameter of the MPD in the pancreatic tail, mm, Median (IQR)	2.0(4.9)	2.0(4.5)	2.8(4.9)	2.0(2.8)	2.0(2.0)
The number of patients with tumors in the pancreatic head, n(%)	432.0(43.2)	104.0(41.6)	69(39.9)	111(38.4)	70.0(56.5)
Maximum diameter of the tumor, mm, Median (IQR)	33.4(27.6)	32.1(27)	32.0(26.7)	33.4(30.9)	28.9(19.9)
The average attenuation of solid tumor component in arterial phase on the largest tumor cross-section, HU, Median (IQR)	53.6(14.9)	53.3(14.7)	52.0(14.5)	50.0(12.9)	49.3(11.9)
The average attenuation of solid tumor component in portal phase on the largest tumor cross-section, HU, Median (IQR)	59.3(20.5)	58.3(18.7)	59.4(19.3)	57.2(23.8)	50.2(17.9)
Solid component volume of the tumor, mm ³ , Median (IQR)	4093.2(12314.5)	3397.2(9528.4)	3270.1(9891.8)	4663.6(12682.4)	2404.1(6428.3)
Solid component volume ratio, n(%)	0.5(0.5)	0.5(0.5)	0.5(0.4)	0.4(0.5)	0.3(0.4)
The maximum attenuation of solid tumor component volume in portal phase, HU, Median (IQR)	60.1(19.8)	59.9(18.4)	60.2(17.9)	59.2(23.2)	52.8(16.9)

Note: Categorical variables are expressed as the number of cases (%), and continuous variables as the median (interquartile range).

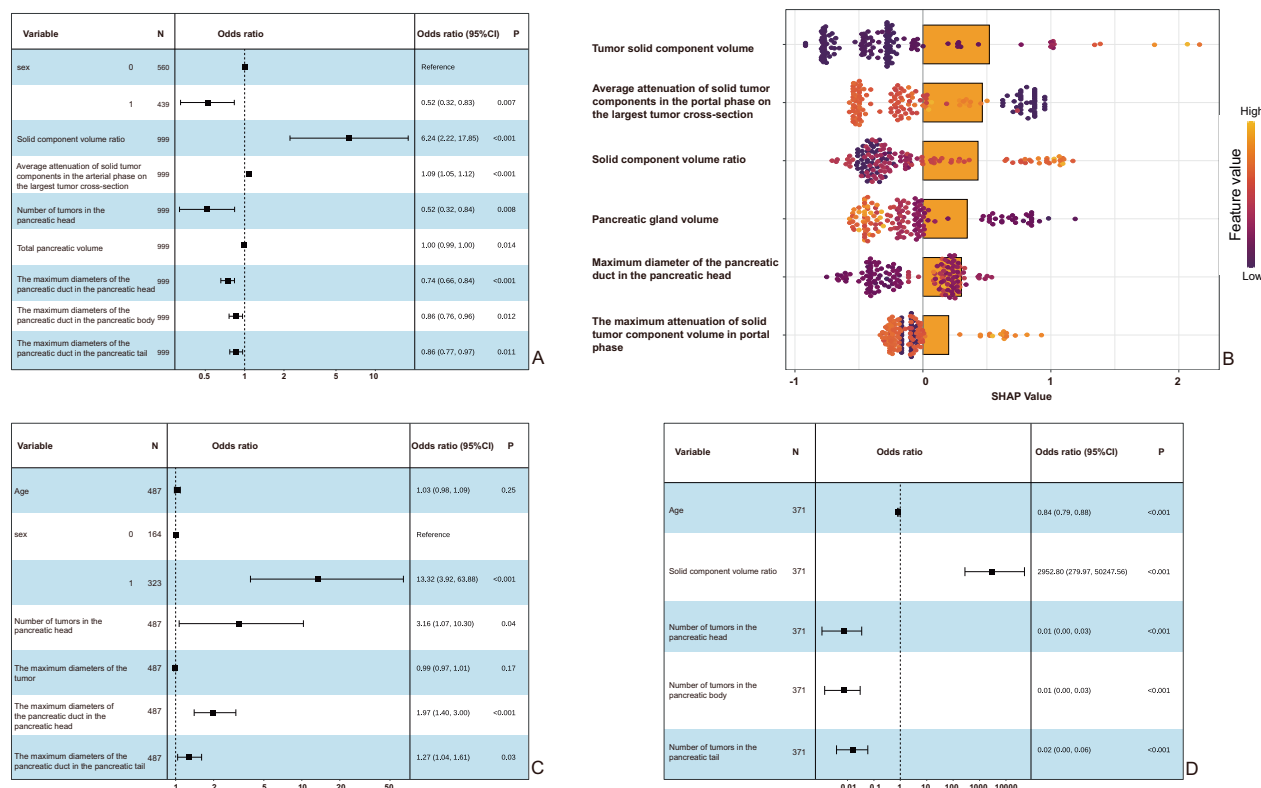


Fig. 4 | Illustrates the feature selection process for 4 tasks across all cohorts. A Forest plot for feature selection in classification task 1. **B** SHAP plot for feature selection in classification task 2. **C** Forest plot for feature selection in classification task 3. **D** Forest plot for feature selection in classification task 4.

rather than overfitting, as demonstrated by stable nested CV performance (± 0.02 standard deviation).

Performance of AI models in classification task 4

In Classification Task 4, we employed the MRMR method to carefully select five features from clinical characteristics and 63 CT quantitative features that are most closely related to distinguishing SCN from SPN. These five key features are: age, solid tumor volume ratio, and the number of tumors in the tail of the pancreas, the number of tumors in the body of the pancreas, and the number of tumors in the head of the pancreas. Subsequently, we input these five features into a regression model for analysis. Both univariate and multivariate analyses revealed that these five features are significantly correlated with distinguishing SCN from SPN ($P < 0.001$) (Fig. 4D, Supplementary Table 6).

In Classification Task 4, the AI model demonstrated outstanding discrimination, with an AUC of 0.985 (95% CI: 0.977–0.994) on the training set and 0.982 (95% CI: 0.959–1.000) on the validation set. On test sets 1–3, the AUC was 0.988 (95% CI: 0.970–1.000), 0.945 (95% CI: 0.898–0.992), and 0.981 (95% CI: 0.936–1.000), respectively. The performance of the PCN-AI model is shown in Table 3 and Fig. 5.

Classifier comparison analysis

Comprehensive evaluation revealed XGBoost's superior generalization performance, particularly for Task 2 (malignant detection) with Test 3 AUC of 0.903 compared to MLP (0.635), RF (0.672), and SVM (0.625). XGBoost maintained consistent external validation performance while alternative methods showed significant overfitting (Supplementary Tables 7–10).

The results of MRMC study

In this study, we randomly selected 500 patients from tests 1–3 to compare the performance of AI and radiologists. Among these patients, there were 144 cases of precancerous IPMN (28.8%), 77 cases of malignant IPMN (15.4%), 62 cases of precancerous MCN (12.4%), 1 case of malignant MCN

(0.2%), 120 cases of SCN (24.0%), and 96 cases of SPN (19.2%) confirmed by pathology.

A comparison of the diagnostic AUC between eight readers before and after PCN-AI assistance was conducted. For Task 1, all readers showed an increase in AUC after AI assistance, with statistically significant improvements ($P < 0.05$) for seven out of eight readers. Similar trends were observed for Tasks 2, 3, and 4, with most readers demonstrating significant improvements in AUC following PCN-AI assistance. The P -values were obtained using the DeLong test, indicating the statistical significance of the differences in AUC before and after PCN-AI assistance. The results suggested that PCN-AI assistance has a positive impact on the diagnostic accuracy of the readers. The changes in AUC with and without PCN-AI-assisted diagnosis for radiologists are shown in Supplementary Table 11–15 and Fig. 6A–D.

Without PCN-AI, the 8 radiologists' mean diagnostic time was 5.28 minutes (range: 0.93–20.61). With PCN-AI, it significantly decreased to 4.03 minutes (range: 0.88–16.50) (Without PCN-AI vs With PCN-AI, $P < 0.001$). Detailed comparisons are in Fig. 6E and Supplementary Table 16.

The results of the acceptance of the PCN-AI

Eight readers conducted an acceptance experiment on 173 samples, seamlessly integrating PCN-AI assistance into their diagnostic workflows. In the total of 1384 readings conducted by these eight radiologists, they were willing to rely on PCN-AI assistance to make diagnoses in 1206 cases (87.14%). The acceptance of the AI model is presented in Fig. 6F.

Prospective study in the real world

To assess the real-world performance of PCN-AI, we conducted a prospective study (Stage 4) at Hospital 1. We evaluated a cohort of 3176 patients who underwent pancreatic CT examinations. The Pancreatic Tumor Detection Model categorized 2217 patients (69.8%) as having a normal pancreas or non-neoplastic lesions, while 959 patients (30.2%) were

Table 3 | Classification performance of AI models in all datasets

Performance	Training set	Validation set	Test 1	Test 2	Test 3
Classification Task 1					
Number	<i>n</i> = 999	<i>n</i> = 250	<i>n</i> = 173	<i>n</i> = 289	<i>n</i> = 124
AUC (95%CI)	0.946(0.933,0.959)	0.942(0.914,0.970)	0.923(0.883,0.962)	0.855(0.813,0.897)	0.877(0.811,0.941)
Sensitivity	0.912(371/407)	0.784(80/102)	0.812(56/69)	0.810(111/137)	0.913(21/23)
Specificity	0.823(487/592)	0.960(142/148)	0.865(90/104)	0.717(109/152)	0.753(76/101)
Accuracy	0.859(858/999)	0.888(222/250)	0.844(146/173)	0.761(220/289)	0.782(97/124)
PPV	0.779(371/476)	0.930(80/86)	0.800(56/70)	0.721(111/154)	0.457(21/46)
NPV	0.931(487/523)	0.866(142/164)	0.874(90/103)	0.807(109/135)	0.974(76/78)
Classification Task 2					
Patients	<i>n</i> = 487	<i>n</i> = 142	<i>n</i> = 90	<i>n</i> = 109	<i>n</i> = 76
AUC (95%CI)	0.976(0.965,0.987)	0.858(0.777,0.939)	0.826(0.735,0.917)	0.815(0.689,0.942)	0.903(0.835,0.970)
Sensitivity	0.905(171/189)	0.684(26/38)	0.879(29/33)	0.792(19/24)	0.692(18/26)
Specificity	0.936(279/298)	0.942(98/104)	0.719(41/57)	0.929(79/85)	0.960(48/50)
Accuracy	0.924(450/487)	0.873(124/142)	0.778(70/90)	0.899(98/109)	0.868(66/76)
PPV	0.900(171/190)	0.813(26/32)	0.644(29/45)	0.760(19/25)	0.900(18/20)
NPV	0.939(279/297)	0.891(98/110)	0.911(41/45)	0.941(79/84)	0.857(48/56)
Classification Task 3					
Number	<i>n</i> = 487	<i>n</i> = 142	<i>n</i> = 90	<i>n</i> = 109	<i>n</i> = 76
AUC (95%CI)	0.963(0.935,0.991)	0.971(0.945,0.996)	0.958(0.887,1.000)	0.960(0.919,1.000)	0.749(0.494,1.000)
Sensitivity	0.965(444/460)	0.873(110/126)	0.951(77/81)	0.936(87/93)	0.562(41/73)
Specificity	0.852(23/27)	1.000(16/16)	0.889(8/9)	0.875(14/16)	1.000(3/3)
Accuracy	0.959(467/487)	0.887(126/142)	0.944(85/90)	0.927(101/109)	0.579(44/76)
PPV	0.991(444/448)	1.000(110/110)	0.987(77/78)	0.978(87/89)	1.000(41/41)
NPV	0.590(23/39)	0.500(16/32)	0.667(8/12)	0.700(14/20)	0.086(3/35)
Classification Task 4					
Number	<i>n</i> = 371	<i>n</i> = 80	<i>n</i> = 56	<i>n</i> = 111	<i>n</i> = 21
AUC (95%CI)	0.985(0.977,0.994)	0.982(0.959,1.000)	0.988(0.970,1.000)	0.945(0.898,0.992)	0.981(0.936,1.000)
Sensitivity	0.943(182/193)	0.938(45/48)	0.938(30/32)	0.922(47/51)	0.923(12/13)
Specificity	0.955(170/178)	1.000(32/32)	0.958(23/24)	0.900(54/60)	1.000(8/8)
Accuracy	0.949(352/371)	0.963(77/80)	0.946(53/56)	0.910(101/111)	0.952(20/21)
PPV	0.958(182/190)	1.000(45/45)	0.968(30/31)	0.887(47/53)	1.000(12/12)
NPV	0.939(170/181)	0.914(32/35)	0.920(23/25)	0.931(54/58)	0.889(8/9)

Task 1: Classification of mucinous PCN (IPMN and MCN) versus non-mucinous PCN (SCN and SPN); Task 2: Classification of precancerous versus malignant pancreatic mucinous tumors; Task 3: Differentiation of pancreatic IPMN from MCN; and Task 4: Distinction of pancreatic SPN from SCN.

AUC area under the receiver operating characteristic curve, PPV positive predictive value, NPV negative predictive value.

classified as having neoplastic lesions. The latter group was further subdivided into 628 cases (19.8%) of solid tumors and 331 cases (10.4%) of cystic tumors.

In this study, 331 PCN patients underwent AI-assisted diagnosis, categorized into 147 non-mucinous and 184 mucinous PCNs by the AI. A radiologist reviewed the non-mucinous cases, with 94.6% (139) concordance and 5.4% (8) discordance resolved by another. For mucinous PCNs, the AI identified 120 precancerous and 64 malignant cases; 95% (114) of the precancerous cases matched, with 5% (6) discordance resolved. The malignant cases had 6.3% (4) discordance adjudicated by a third radiologist. This process showcased PCN-AI's efficiency and accuracy, reducing workload by 39.3% through 267 AI-assisted reviews out of 680.

At the end of the study, all patients will be followed up for three months to obtain the final pathological diagnosis results. Pathological confirmation revealed 99 cases of pancreatic cystic neoplasms, comprising 52.5% (52) precancerous IPMN, 11.1% (11) malignant IPMN, 2.0% (2) precancerous MCN, 1.0% (1) malignant MCN, 18.2% (18) SCN, and 15.2% (15) SPN.

To evaluate the AI-assisted group's potential as a screening tool, we assessed its diagnostic efficacy in Tasks 1 and 2. In Task 1, the AI-assisted

group outperformed radiologist double-reading, achieving a higher AUC (0.947 [95% CI: 0.903–0.991] vs. 0.909 [95% CI: 0.853–0.966], $p = 0.225$). In Task 2, the AI-assisted group again excelled in diagnosing malignant lesions, with a significantly higher AUC (0.819 [95% CI: 0.723–0.916] vs. 0.600 [95% CI: 0.476–0.724], $p = 0.008$). With PCN-AI assistance, radiologists correctly identified 81.8% (9/11) of malignant cases, compared to 36.4% (4/11) with double-reading alone. This PCN-AI-assisted diagnosis provided additional benefits to 45.45% (5/11) of patients over radiologist double-reading. Detailed performance is shown in Supplementary Table 17. Representative case images are shown in Supplementary Fig. 3–6.

Discussion

The widespread adoption of advanced imaging modalities, coupled with increasing life expectancy, has led to a rising incidence of pancreatic cystic neoplasms (PCNs)²². Accurate risk stratification of these lesions is critical, given their heterogeneous malignant potential—ranging from indolent serous cystic neoplasms (SCNs) to aggressive intraductal papillary mucinous neoplasms (IPMNs) with invasive carcinoma²². To address this challenge, we developed PCN-AI, a comprehensive artificial intelligence model

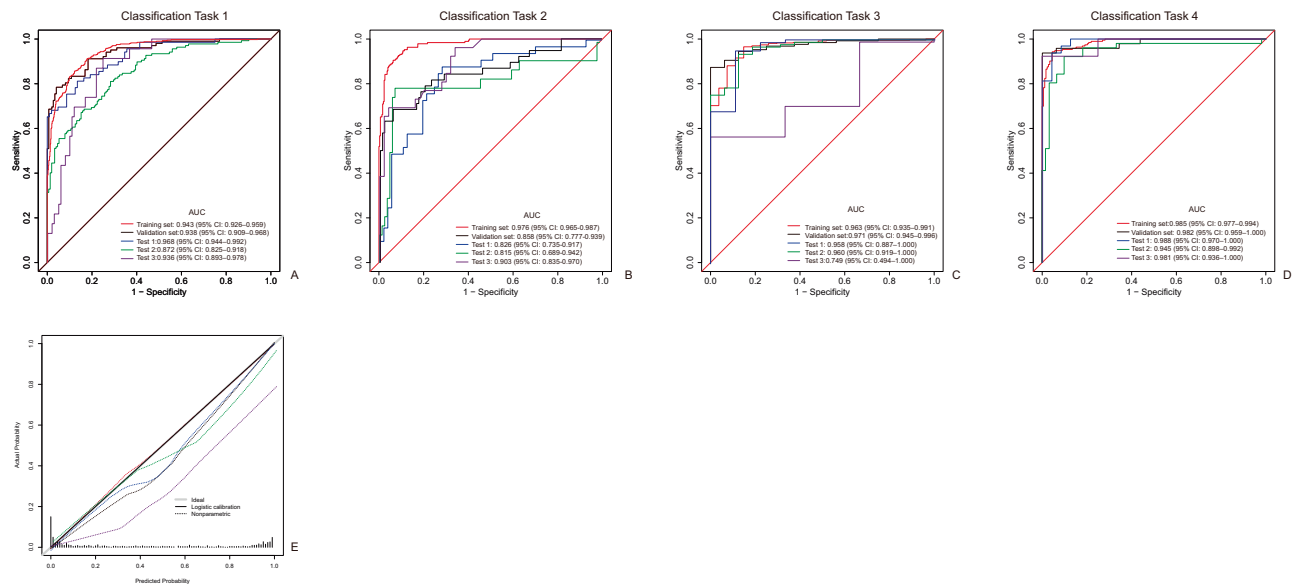


Fig. 5 | Performance of the diagnostic models in 4 tasks across all cohorts. The figure shows ROC curves for: (A) classification of mucinous PCN (IPMN and MCN) versus non-mucinous PCN (SCN and SPN); (B) classification of precancerous versus malignant pancreatic mucinous tumors; (C) differentiation of pancreatic IPMN from MCN; (D) distinction of pancreatic SPN from SCN; (E) presents the

calibration plot for model prediction probabilities. ROC receiver operating characteristic, PCN pancreatic cystic neoplasms, IPMN intraductal papillary mucinous cystic neoplasms, MCN mucinous cystic neoplasms, SCN serous cystic neoplasms, SPN solid pseudopapillary neoplasms.

that leverages 63 quantitative CT features to hierarchically classify PCNs across four clinically actionable tasks. Trained on a multicenter cohort of 1835 patients, PCN-AI first distinguishes mucinous (IPMN/MCN) from non-mucinous (SCN/SPN) lesions—a pivotal step in identifying high-risk subgroups requiring intensive surveillance. Within mucinous subtypes, the model further stratifies lesions into *precancerous* (low-grade dysplasia) and *malignant* (high-grade dysplasia or invasive carcinoma) categories, enabling tailored management while relegating non-mucinous cases to routine follow-up. In a multi-reader, multi-case (MRMC) study, PCN-AI significantly enhanced radiologists' diagnostic accuracy, elevating the average AUC from 0.814 to 0.880 ($p < 0.05$) across all tasks, while reducing interpretation time by 23.7% (5.28 vs. 4.03 minutes per case). Radiologists endorsed AI assistance in 87.14% of cases, reflecting its seamless integration into clinical workflows. Prospective validation further underscored its clinical value: PCN-AI-assisted diagnoses identified 81.8% (9/11) of malignant PCNs, compared to 36.4% (4/11) with conventional double-reading—a 45.45% improvement in actionable detection. This translated to a 39.3% reduction in clinician workload through optimized triage, demonstrating its dual role in enhancing diagnostic precision and operational efficiency.

The PCN-AI model synergizes radiologists' diagnostic expertise with interpretable AI-driven decision-making, a cornerstone of personalized oncology²³. By translating subjective imaging assessments into quantitative radiomic features (e.g., solid component volume ratio, portal-phase attenuation), our framework enables transparent risk stratification aligned with clinical workflows²⁴. In Task 1, PCN-AI achieved 89.9% accuracy in distinguishing non-mucinous PCNs, with solid-predominant lesions (e.g., SPNs) showing significantly higher solid component volumes compared to SCNs ($p < 0.001$). Task 2 further identified that malignant mucinous PCNs exhibited larger solid components (solid component volume ratio: 0.56 ± 0.29 vs. 0.32 ± 0.19 in precancerous lesions) with elevated portal-phase attenuation (ΔCT : $+11.8$ HU), corroborating the "high-risk stigmata" criteria defined by international guidelines^{25,26}. Unlike manual measurements, which suffer from inter-observer variability ($\kappa = 0.45$ – 0.62)^{25,27}, PCN-AI's automated segmentation reduced measurement discrepancies by 72% (DSC: 0.85 vs. 0.24 for manual delineation), ensuring reproducible quantification. The XGBoost model effectively assigned clinically significant weights to imaging biomarkers. SHAP summary plots revealed that an

increased solid component volume ratio and pancreatic duct dilation were positively correlated with malignancy classification, whereas higher portal-phase attenuation values exhibited protective effects against malignancy. This allowed for an objective risk stratification that surpassed guideline-based methods, with accuracy rates ranging from 77.8% to 89.9% compared to 53.0% to 59.0%^{28–30}. Notably, Tasks 3–4 revealed demographic and anatomic insights: mucinous cystic neoplasms (MCNs) predominantly affected middle-aged women (mean age: 54 years) and localized to the pancreatic tail (82.6%), whereas branch-duct IPMN (BD-IPMN) correlated with pancreatic head involvement (76.3%) and MPD dilation (≥ 5 mm, 68.9%)—findings consistent with established epidemiological patterns⁸. These interpretable outputs not only validate PCN-AI's biological plausibility but also empower radiologists to refine diagnostic thresholds in ambiguous cases.

PCN-AI distinguishes itself in PCN classification through rigorous technical validation and clinical generalizability—a critical advancement over prior AI models constrained by limited training data ($n < 500$ in 80% of studies) and single-center designs^{31,32}. Developed on multi-institutional cohorts (4 centers, $n = 1,835$), our model demonstrated robust diagnostic performance across all tasks (AUC: 0.749–0.988), with consistent results in external validation cohorts ($\Delta AUC < 0.05$ between internal/external tests). Notably, PCN-AI maintained high accuracy (AUC: 0.877–0.981) even when applied to CT scanners with 5 mm slice thickness—a common yet technically challenging scenario in routine practice where thinner slices (1–2 mm) are often unavailable. While Task 3 (IPMN vs. MCN differentiation) showed a marginally lower AUC (0.749), this reflects the inherent class imbalance in Test 3 (IPMN: MCN = 7:1) rather than model limitations, emphasizing the need for balanced datasets in rare disease AI³³. Unlike existing tools restricted to binary classification (e.g., precancerous vs. malignant)^{31,32}, PCN-AI's cascading architecture enables hierarchical differentiation of four PCN subtypes—mucinous (IPMN/MCN), non-mucinous (SCN/SPN), precancerous, and malignant—with diagnostic accuracy comparable to senior radiologists (mean AUC difference: 0.02–0.05). This multi-class capability addresses a critical clinical gap, as misclassification of SPN (a lesion with malignant potential) as SCN could delay life-saving interventions^{33,34}. By bridging technical innovation with clinical pragmatism, PCN-AI sets a new benchmark for translatable AI in pancreatic oncology.

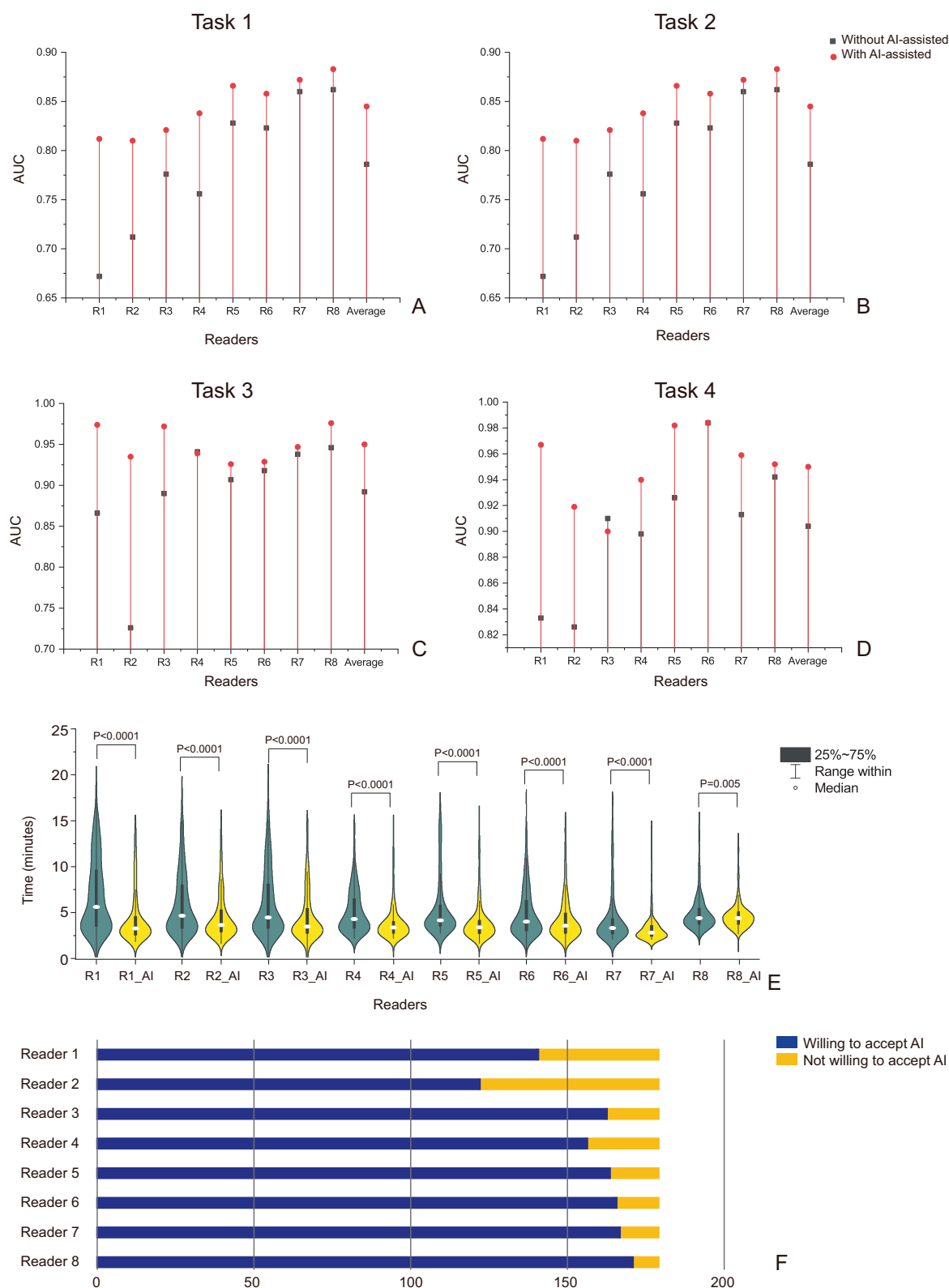


Fig. 6 | The comparison of Multi-Reader Multi-Case study between with AI and without AI assisted. **A–D** Performance of nine readers without and with the assistance of PCN-AI. **A** Task 1: ROC curves for Classification of mucinous vs non-mucinous pancreatic cystic neoplasms (PCN). **B** Task 2: ROC curves for distinguishing precancerous from malignant pancreatic mucinous tumors. **C** Task 3: ROC curves for differentiating pancreatic intraductal papillary mucinous neoplasms (IPMN) from mucinous cystic neoplasms (MCN). **D** Task 4: ROC curves for

distinguishing pancreatic solid pseudopapillary neoplasms (SPN) from serous cystic neoplasms (SCN). The curves represent the AI model performance. Grey dots indicate the performance of 8 readers without AI assistance, while red dots show the performance of the same 8 readers with AI assistance. **E** Changes in diagnosis time for the eight readers without and with PCN-AI assistance. **F** Willingness of the eight readers to use PCN-AI-assisted diagnosis.

While AI holds promise for augmenting diagnostic workflows, its tangible impact on radiologist performance and adoption in complex oncologic imaging remains poorly characterized^{16,18,19,35,36}. Our study addresses this gap by quantifying PCN-AI's clinical utility across four high-stakes classification tasks. AI assistance elevated radiologists' AUC by 0.046–0.158 ($p < 0.05$), most notably in differentiating malignant mucinous PCNs (AUC: 0.675 → 0.776)—a critical determinant of surgical intervention. The model's ability to reduce false negatives was evidenced by a 24.8% increase in sensitivity (0.597 → 0.745) and 17.9% improvement in PPV (0.557 → 0.657), directly addressing a key limitation of guideline-based approaches that miss up to 63.6% of malignancies in double-reader workflows³⁴. These gains, however, revealed task-dependent variations: while SCN/SPN differentiation saw near-perfect AI-radiologist concordance (AUC: 0.904 → 0.950), IPMN/MCN classification showed more modest improvements (Δ AUC: +0.058), likely reflecting the subtler imaging distinctions between these mucinous subtypes. Such heterogeneity underscores the need for lesion-specific AI optimization, particularly for rare entities where training data scarcity compounds diagnostic uncertainty¹⁹. Nevertheless, the high radiologist acceptance rate (87.14%) and 23.7% reduction in interpretation time demonstrate PCN-AI's potential to overcome operational hurdles in high-volume practices.

Our study revealed that AI's impact on diagnostic performance varied significantly by radiologist experience. Junior radiologists (≤ 5 years' experience) exhibited the most pronounced improvements, particularly in identifying malignant mucinous PCNs (AUC: 0.62 → 0.81, Δ +0.19), outperforming senior counterparts (Δ +0.07; $p = 0.003$)—a finding consistent with prior studies showing AI's potential to mitigate expertise gaps in early-career practitioners³⁷. This disparity likely stems from AI's ability to codify tacit knowledge (e.g., subtle attenuation patterns in portal-phase CT) that typically requires years of subspecialty training. Across all experience levels, AI assistance reduced diagnostic time by 23.7% (5.28 → 4.03 minutes/case) while maintaining accuracy (κ agreement with pathology: 0.82 vs. 0.79 unassisted), demonstrating its dual role as both an educational tool and workflow optimizer. These results underscore AI's transformative potential in democratizing access to high-quality pancreatic imaging expertise, particularly in resource-constrained settings.

To assess PCN-AI's translational value, we implemented the model prospectively in a high-volume pancreatic cancer screening program. By automating triage—flagging 63.2% of low-risk PCNs for routine follow-up versus 28.4% in manual workflows—PCN-AI reduced redundant radiologist double-reads by 39.3%, equivalent to 11.2 hours/month saved per clinician. Over a 3-month follow-up, the model maintained robust diagnostic accuracy (AUC: 0.947 for Task 1; 0.819 for Task 2), with 45.45% (5/11) of malignant cases correctly reclassified from initial false-negative double-reads. For example, In Task 2, a 52-year-old female patient was initially misdiagnosed with a precancerous IPMN but PCN-AI's exhaustive analysis of critical indicators led to a revised diagnosis of malignant IPMN with a 0.67928 probability. This was confirmed by multidisciplinary consultation and surgery. Such cases highlight AI's capacity to reinforce guideline adherence while personalizing risk stratification—a critical advance for precision oncology.

While PCN-AI represents a significant advance in pancreatic cyst characterization, several limitations warrant consideration: (1) Anatomic Contextualization: The current model focuses exclusively on intralesional features (e.g., solid components, MPD dilation) and does not incorporate peripancreatic findings such as lymphadenopathy or vascular invasion. In one instance, PCN-AI misclassified a malignant IPMN as precancerous due to unaccounted peripancreatic lymph node enlargement (short-axis: 12 mm)—a feature subsequently flagged by radiologists. Future iterations will integrate PET/CT or MRI data to address this gap. (2) Recall Bias in MRMC Design: Despite randomization and washout periods, radiologists' prior exposure to cases may have influenced their unassisted interpretations, potentially inflating baseline diagnostic performance. Blinded re-reading of a subset ($n = 50$) showed moderate agreement ($\kappa = 0.54$) between initial and washout evaluations, suggesting residual bias. (3) Spectrum Bias: The

exclusion of pseudocysts ($n = 32$) and cystic neuroendocrine tumors ($n = 19$) limits generalizability to all pancreatic cystic lesions. These entities, while clinically distinct, often mimic PCN on CT—a challenge requiring expanded training data. (4) Translational Validation: Although our prospective study demonstrated workflow benefits (39.3% workload reduction), multicenter randomized trials are needed to quantify PCN-AI's impact on long-term outcomes (e.g., survival, cost-effectiveness) across diverse healthcare settings.

PCN-AI establishes a new paradigm for AI-driven pancreatic cyst management, combining interpretable feature quantification with hierarchical risk stratification. Trained on multi-institutional data ($n = 1835$), the model matches subspecialty radiologists in diagnostic accuracy (AUC: 0.845–0.988) while reducing interpretation time by 23.7%—a critical advantage in high-volume practices. Its ability to correct 45.45% of missed malignancies in double-reading workflows underscores AI's role in safeguarding diagnostic quality, particularly for early-career radiologists (AUC improvement: Δ +0.19). By translating subjective imaging criteria into reproducible biomarkers (e.g., solid component attenuation, MPD dilation), PCN-AI bridges the gap between guideline-based thresholds and personalized risk assessment. Future work will focus on: (1) multi-modal integration (EUS/MRI) to address anatomic contextualization limits; (2) international multi-center trials to validate survival benefits; and (3) embedding AI-derived biomarkers into evolving classification systems (e.g., WHO 2023). As precision oncology advances, tools like PCN-AI will be indispensable for transforming pancreatic cancer from a lethal enigma to a preventable disease.

Methods

Study design

This study fundamentally adhered to the STARD³⁸ and TRIPOD³⁹ guidelines, involving retrospective data collection from four hospitals (January 2012–December 2023). Hospital 1 served as the primary cohort (training/validation sets), while Hospitals 2–4 formed independent test sets 1–3. The study was approved by Hospital 1 Ethics Committee (CHEC-Y2024-020). Four stages were implemented: Stage 1 (image segmentation and feature extraction), Stage 2 (four classification tasks), Stage 3 (multi-reader, multi-case study), and Stage 4 (real-world prospective validation, April–October 2024, Clinical Trial: ChiCTR2500096592). Independent IRB approvals for each participating institution were also obtained: Hospital 2: Ethics Committee of Shanghai 411 Hospital (Approval No. AH411-2024-11). Hospital 3: Institutional Review Board of Drum Tower Hospital, the Affiliated Hospital of Nanjing University (Approval No. 2024-081-02). Hospital 4: Ethics Committee of Zhongshan Hospital, Fudan University (Approval No. B2024-250R).

Sample size calculations employed Richard D. Riley's methodology with 6–8 features per task and rMPSE criterion of 0.06, yielding minimum requirements of 580, 390, 290, and 440 cases for Tasks 1–4. MRMC study calculations using 'MRMCsamplesize' determined requirements of 124, 144, 215, and 124 cases ($\alpha = 0.05$, $\beta = 0.20$), with post-hoc power analyses exceeding 0.90.

Clinical information

Clinical data were collected within three months of CT scanning, encompassing demographics (sex, age, BMI), symptoms (abdominal pain, jaundice), medical history (diabetes, family history), and biomarkers (CEA, CA19-9). Inclusion criteria: pathologically confirmed IPMN/MCN/SCN/SPN with contrast-enhanced CT within 3 months of surgery/biopsy.

Inclusion and exclusion criteria

In the retrospective study, the inclusion and exclusion criteria for consecutive patients from different centers were consistent with those of the main cohort. The inclusion criteria for our research were as follows: (a) patients who were confirmed with a pathological diagnosis of PCN, (b) patients who underwent contrast-enhanced multiphase CT within 3 months before surgical resection or puncture. The exclusion criteria for

our research were as follows: (a) patients who were not confirmed with a pathological diagnosis of pancreatic IPMN/MCN/SCN/SPN, (b) patients for whom the time between imaging and surgery/puncture was more than 3 months, and (c) patients for whom the CT image had severe artifacts.

Pathology evaluation

All pathological results were diagnosed by two pathologists with 10 and 15 years of experience in pancreatic pathology. If there was a discrepancy between their findings, a pathologist with 20 years of experience made the final decision. The results were obtained from either biopsy pathology or radical surgical pathology, with radical surgical pathology considered definitive when both were available.

The pathological assessment results were neatly categorized into precancerous IPMN (mild dysplasia), malignant IPMN (severe dysplasia and invasive carcinoma), precancerous MCN (mild dysplasia), malignant MCN (severe dysplasia and invasive carcinoma), SCN, and SPN⁴⁰.

CT scan and analysis

Standardized contrast-enhanced CT protocols used non-ionic iodine contrast (355 mg I/mL, 1.5 mL/kg) with arterial (20–25 s), portal (60–70 s), and delayed (110–130 s) phases. CT Scanners and Scanning Parameters for Each Institution are shown in Supplementary Table 18. IPMN subtypes were classified by a senior radiologist: BD-IPMN (cystic lesions >5 mm connecting to main duct), MD-IPMN (main duct dilation >5 mm), and MT-IPMN (combined features)⁶.

All CT volumes underwent end-to-end 3D processing rather than 2D slice-based analysis. Image preprocessing involved six steps: z-score normalization, soft tissue windowing (level 50, width 350), resampling to $1 \times 1 \times 1 \text{ mm}^3$, DEEDS algorithm registration, ROI cropping to $128 \times 128 \times 128$ voxels, and data augmentation (geometric transformations: rotation $\pm 15^\circ$, scaling 0.8–1.2 $\times$; intensity transformations: Gaussian noise $\sigma = 0.02$, brightness/contrast $\pm 30\%$; advanced techniques: CutMix $\alpha = 1.0$, MixUp $\lambda = 0.2$).

AI Model Development

Image annotation. To train the segmentation model, two experienced radiologists, with 15 and 20 years of expertise respectively, used ITK-SNAP software (Version 3.8.0, available at <http://www.itksnap.org>, 2020) to annotate 500 training set cases. They delineated key structures, including the pancreatic parenchyma, pancreatic duct, and tumors, with final approval from the senior radiologist. These 500 cases were used to train the segmentation model, which was then optimized through an iterative process. The model's performance was evaluated using the Dice Similarity Coefficient (DSC). If the DSC fell below 0.80, the senior radiologist would correct the annotations, and the model was retrained on the revised dataset. This cycle of correction and retraining continued until the model's performance reached a DSC of 0.80 or greater.

Segmentation Architecture. Mamba U-Net was selected after comparing nnU-Net, SwinUNETR, UNETR, nnSAM, and SegResNet, achieving superior DSC (0.85 ± 0.02). The architecture features: encoder with $3 \times 3 \times 3$ convolutions, batch normalization, ReLU activation, and $2 \times 2 \times 2$ max pooling; bottleneck with dilated convolutions (rate = 2) and deformable convolutions (offset groups = 4); decoder with attention-gated skip connections (reduction ratio=8) and deep supervision at layers 2–4 (weights [0.3, 0.2, 0.1]). Training used ADAM optimizer, batch size 16, learning rate 1×10^{-4} , weight decay 0.0005, and combined Dice/cross-entropy loss.

Feature Extraction. Automated segmentation generated 63 quantitative features across four categories: pancreatic parenchyma (8), pancreatic duct (6), calcifications (14), and tumors (35). Volume quantification counted non-zero voxels with physical volume calculation using spatial resolution. Area calculation analyzed 2D slices with pixel counting. Diameter measurement employed maximum inscribed circle algorithms.

CT values were measured across four phases using 35 HU threshold for solid/cystic differentiation. Count analysis used connected component analysis for calcification enumeration. All quantification parameters are shown in Supplementary Table 19.

Classification models. Four sequential tasks addressed clinical needs (Fig. 1 and Supplementary Fig. 7): Task 1 (mucinous vs. non-mucinous) used logistic regression with 8 mRMR-selected features; Task 2 (precancerous vs. malignant mucinous) employed XGBoost with 6 key features; Task 3 (IPMN vs. MCN) used logistic regression with 6 features; Task 4 (SCN vs. SPN) employed logistic regression with 5 features. Nested cross-validation (outer 5-fold, inner 3-fold) prevented data leakage, with temporal separation maintaining training (2012–2020) and test (2021–2023) isolation.

Multi-reader multi-case (MRMC) study

Between March and May 2024, we conducted an MRMC study involving 500 randomly selected patients from both internal and external test datasets. The study followed a randomized crossover design. Eight independent readers from Hospital 1 participated, comprising two residents with three to five years of experience (R1 and R2), four attending radiologists with six to ten years (R3–R6), and two senior radiologists with fifteen to twenty years (R7–R8). None were involved in patient recruitment or image labeling.

Prior to the diagnostic process, all readers completed a two-session training course on evaluating pancreatic cystic neoplasms, followed by a practical exercise of writing 50 diagnostic reports. During the study, specialized workstations were used to read de-identified pancreatic CT scans. To avoid bias, readers outlined test cases both with and without AI assistance, and were randomly assigned to start with either assisted or unassisted segmentation. A four-week washout period was implemented between sessions, during which readers continued their clinical duties. In the second session, cases were presented in a random order, and readers had no access to clinical information, other readers' results, or pathological findings.

The readers completed four tasks: (1) determining if a tumor was mucinous or non-mucinous; (2) if it was mucinous, diagnosing it as precancerous or malignant; (3) identifying it as either IPMN or MCN; and (4) if it was non-mucinous, differentiating between SCN and SPN. All assessments were completed within a week, with the time taken for each case recorded.

Acceptance of the PCN-AI

In June 2024, we evaluated the clinicians' acceptance of PCN-AI using a dataset comprising 173 patients from Test 2. The eight readers, identical to those in the MRMC study, concurrently recorded their diagnoses on designated local diagnostic computers. Although the PCN-AI results were not automatically displayed, the readers had the flexibility to view them by clicking on a specially designed software interface whenever needed. This methodology allowed us to assess the clinicians' reception of the AI model effectively.

Prospective study in the real world

Study Phase 4 (April 2024–July 2024) aims to integrate PCN-AI into our center's Pancreatic Tumor Detection Model with tumor and cyst-solid discrimination. Deployed at Hospital 1, it will enroll patients for CT exams and follow-up until October 2024 to gather pathological results.

During this stage, subjects undergoing pancreatic-enhanced CT scans will follow the traditional protocol: two radiologists will review images independently, and a senior radiologist will arbitrate disagreements to determine lesion type (mucinous/non-mucinous PCN). Mucinous PCNs will undergo further diagnosis to distinguish precancerous from malignant lesions.

Concurrently, CT images will be imported into an AI-assisted system for preliminary classification by our "Pancreatic Tumor Detection Model" (Supplementary Fig. 8). This model employs a two-stage approach. Stage 1 uses a pre-trained Mamba U-Net segmentation model, trained on a diverse

dataset of 3700 cases (1000 normal, 1000 inflammatory, and 1700 neoplastic), to accurately detect tumorous lesions. If a lesion is detected, Stage 2 uses Mamba U-Net and DenseNet121 model to classify it as either cystic or solid. This classification model was trained on 1000 balanced samples (500 cystic, 500 solid).

The cystic lesions (identified as PCN) will then be subdivided. The PCN-AI algorithm will further diagnose PCN, with diagnostic rules as follows: for non-mucinous lesions, patient enrollment in follow-up will be based on the PCN-AI and radiologist agreement; disagreements will be resolved by a higher-level radiologist. For precancerous lesions, the same process applies. For malignant PCN, confirmation by two radiologists is required; if they disagree, a higher-level radiologist will decide (Fig. 2).

At the end of the study, all patients will be followed up for three months to obtain the final pathological diagnosis results, which will serve as crucial evidence for the study evaluation.

Statistical analysis

Dichotomous data were presented as absolute numbers with corresponding relative percentages, while continuous data were summarized using means (SD) or medians with interquartile ranges (IQR). The Dice similarity coefficient (DSC) was used to assess segmentation consistency between observers and AI models. It ranges from 0 to 1, where 0 indicates no overlap between manually delineated and AI-generated ROIs, and 1 represents complete overlap⁴¹. The diagnostic performance of the AI model was evaluated using the Area under the Curve (AUC) of the receiver operating characteristic (ROC), along with sensitivity, specificity, accuracy, positive predictive value (PPV), and negative predictive value (NPV). The DeLong test was used to compare the diagnostic AUC performance of readers with and without AI assistance. The paired samples t-test was used to assess the difference in diagnosis time between readers with and without AI model assistance.

Statistical significance was determined by a p-value less than 0.05 for two-tailed tests. All analyses were conducted using SPSS (Statistics v 26.0, IBM Corp, Armonk, NY, USA) and R language (R version 4.6.0, R Foundation for Statistical Computing, USA).

Computational infrastructure: Ubuntu 22.04, 128GB RAM, Intel i7-14700K, NVIDIA RTX 4060Ti. Mamba U-Net provided 30% parameter reduction (7.4M vs. 23.4M) with FP16/INT8 quantization for deployment flexibility. All code and models are publicly available at <https://github.com/CHANGHAI-AILab/PCN> with fixed random seeds (42), ensuring reproducibility.

Data availability

The datasets generated and analyzed during the current study are available from the corresponding author upon reasonable request. All data supporting the findings of this study are included within the manuscript and its supplementary information files. All source code and pre-trained models used in this study are publicly available at: <https://github.com/CHANGHAI-AILab/PCN>, ensuring full reproducibility of our results.

Received: 21 February 2025; Accepted: 20 August 2025;

Published online: 13 October 2025

References

1. Stoffel, E. M., Brand, R. E. & Goggins, M. Pancreatic cancer: changing epidemiology and new approaches to risk assessment, early detection, and prevention. *Gastroenterology* **164**, 752–765 (2023).
2. Commission, G. O.o.N. H. Standard for diagnosis and treatment of pancreatic cancer (2022 edition). *J. Clin. Hepatol.* **38**, 1006–1015 (2022).
3. Peters, M. L. B. et al. Progression to pancreatic ductal adenocarcinoma from pancreatic intraepithelial neoplasia: Results of a simulation model. *Pancreatology* **18**, 928–934 (2018).
4. Shin, E. J. & Canto, M. I. Pancreatic cancer screening. *Gastroenterol. Clin. North Am.* **41**, 143–157 (2012).
5. Vege, S. S. et al. American gastroenterological association institute guideline on the diagnosis and management of asymptomatic neoplastic pancreatic cysts. *Gastroenterology* **148**, 819–822 (2015).
6. Tanaka, M. et al. Revisions of international consensus Fukuoka guidelines for the management of IPMN of the pancreas. *Pancreatology* **17**, 738–753 (2017).
7. Megibow, A. J. et al. Management of incidental pancreatic cysts: a white paper of the acr incidental findings committee. *J. Am. Coll. Radiol.* **14**, 911–923 (2017).
8. European Study Group on Cystic Tumours of the, P. European evidence-based guidelines on pancreatic cystic neoplasms. *Gut* **67**, 789–804 (2018).
9. Elta, G. H., Enestvedt, B. K., Sauer, B. G. & Lennon, A. M. ACG clinical guideline: diagnosis and management of pancreatic cysts. *Am. J. Gastroenterol.* **113**, 464–479 (2018).
10. Marchegiani, G. et al. Systematic review, meta-analysis, and a high-volume center experience supporting the new role of mural nodules proposed by the updated 2017 international guidelines on IPMN of the pancreas. *Surgery* **163**, 1272–1279 (2018).
11. Pozzi Mucelli, R. M. et al. Branch-duct intraductal papillary mucinous neoplasm (IPMN): Are cyst volumetry and other novel imaging features able to improve malignancy prediction compared to well-established resection criteria?. *Eur. Radiol.* **32**, 5144–5155 (2022).
12. Jang, J. Y. et al. Validation of international consensus guidelines for the resection of branch duct-type intraductal papillary mucinous neoplasms. *Br. J. Surg.* **101**, 686–692 (2014).
13. El Chafic, A. et al. Does cyst growth predict malignancy in branch duct intraductal papillary mucinous neoplasms? Results of a large multicenter experience. *Dig. Liver Dis.* **50**, 961–968 (2018).
14. Crippa, S. et al. Main duct thresholds for malignancy are different in intraductal papillary mucinous neoplasms of the pancreatic head and body-tail. *Clin. Gastroenterol. Hepatol.* **20**, 390–399 e397 (2022).
15. Suzuki, R., Thosani, N., Annangi, S., Guha, S. & Bhutani, M. S. Diagnostic yield of EUS-FNA-based cytology distinguishing malignant and benign IPMNs: a systematic review and meta-analysis. *Pancreatology* **14**, 380–384 (2014).
16. Liang, W. et al. Classification prediction of pancreatic cystic neoplasms based on radiomics deep learning models. *BMC Cancer* **22**, 1237 (2022).
17. Yao, L. et al. A review of deep learning and radiomics approaches for pancreatic cancer diagnosis from medical imaging. *Curr. Opin. Gastroenterol.* **39**, 436–447 (2023).
18. Schulz, D. et al. Accurate prediction of histological grading of intraductal papillary mucinous neoplasia using deep learning. *Endoscopy*, <https://doi.org/10.1055/a-1971-1274> (2022).
19. Corral, J. E. et al. Deep learning to classify intraductal papillary mucinous neoplasms using magnetic resonance imaging. *Pancreas* **48**, 805–810 (2019).
20. Tjoa, E. & Guan, C. A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. *IEEE Trans. Neural Netw. Learn. Syst.* **32**, 4793–4813 (2021).
21. J. A. M. et al. *WHO Classification of the Digestive System*. 5th edn, 303–305 (IARC Press, 2019).
22. Miller, F. H. et al. Pancreatic cystic lesions and malignancy: assessment, guidelines, and the field defect. *Radiographics* **42**, 87–105 (2022).
23. Wood, M. L. Variability and standardization of Quantitative Imaging. *Invest Radiol.* **55**, 617–618 (2020).
24. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215 (2019).
25. Zhong, N. et al. Histologic and imaging features of mural nodules in mucinous pancreatic cysts. *Clin. Gastroenterol. Hepatol.* **10**, 192–198, 198 e191–192 (2012).

26. Choi, S. Y. et al. Diagnostic performance and imaging features for predicting the malignant potential of intraductal papillary mucinous neoplasm of the pancreas: a comparison of EUS, contrast-enhanced CT and MRI. *Abdom. Radiol. (NY)* **42**, 1449–1458 (2017).
27. Lee, J. E. et al. Determining malignant potential of intraductal papillary mucinous neoplasm of the pancreas: CT versus MRI by using revised 2017 international consensus guidelines. *Radiology* **293**, 134–143 (2019).
28. Uribarri-Gonzalez, L. et al. Agreement among magnetic resonance imaging/magnetic resonance cholangiopancreatography (MRI-MRCP) and endoscopic ultrasound (EUS) in the evaluation of morphological features of branch duct intraductal papillary mucinous neoplasm (BD-IPMN). *Pancreatology* **18**, 170–175 (2018).
29. Olar, M. P. et al. Clinical utility of the contrast-enhanced endoscopic ultrasound guided fine needle aspiration in the diagnosis of pancreatic cyst. *Diagnostics (Basel)* **12**, <https://doi.org/10.3390/diagnostics12092209> (2022).
30. Lekkerkerker, S. J. et al. Comparing 3 guidelines on the management of surgically removed pancreatic cysts with regard to pathological outcome. *Gastrointest. Endosc.* **85**, 1025–1031 (2017).
31. Wang, X. et al. A deep learning algorithm to improve readers' interpretation and speed of pancreatic cystic lesions on dual-phase enhanced CT. *Abdom. Radiol. (NY)* **47**, 2135–2147 (2022).
32. Jan, Z., El Assadi, F., Abd-Alrazaq, A. & Jithesh, P. V. Artificial intelligence for the prediction and early diagnosis of pancreatic cancer: scoping review. *J. Med. Internet Res.* **25**, e44248 (2023).
33. Scholten, L., van Huijgevoort, N. C. M., van Hooft, J. E., Besselink, M. G. & Del Chiaro, M. Pancreatic cystic neoplasms: different types, different management, new guidelines. *Visc. Med.* **34**, 173–177 (2018).
34. Liu, Q. et al. Long-term survival, quality of life, and molecular features of the patients with solid pseudopapillary neoplasm of the pancreas: a retrospective study of 454 cases. *Ann. Surg.* **278**, 1009–1017 (2023).
35. Hu, B. et al. A deep-learning model for intracranial aneurysm detection on CT angiography images in China: a stepwise, multicentre, early-stage clinical validation study. *Lancet Digit Health* **6**, e261–e271 (2024).
36. Cao, K. et al. Large-scale pancreatic cancer detection via non-contrast CT and deep learning. *Nat. Med.* **29**, 3033–3043 (2023).
37. Lee, J. H. et al. Improving the performance of radiologists using artificial intelligence-based detection support software for mammography: a multi-reader study. *Korean J. Radiol.* **23**, 505–516 (2022).
38. Cohen, J. F. et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open* **6**, e012799 (2016).
39. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ* **350**, g7594 (2015).
40. Nagtegaal, I. D. et al. The 2019 WHO classification of tumours of the digestive system. *Histopathology* **76**, 182–188 (2020).
41. Taha, A. A. & Hanbury, A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Med. Imaging* **15**, 29 (2015).
- 62402501, 82202125, 82572215 and 82202145), The Natural Science Foundation of Shanghai Science and Technology Innovation Action Plan (21ZR1478500, 21Y11910300), Clinical Research Plan of SHDC (SHDC2022CRD028), Shanghai Municipal Health Commission Seed Program for Research and Translation of Medical New Technologies Project (2024ZZ1015), and Plan for Promoting Scientific Research Paradigm Reform and Enhancing Disciplinary Advancement through Artificial Intelligence (2025SHJW1717), 2025 Special Project for Clinical Research from the Shanghai Municipal Health Commission (202540148) and Funding for Clinical Trials from the Affiliated Drum Tower Hospital, Medical School of Nanjing University (grant number 2022-YXZX-YX-07).

Author contributions

Xiaohan Yuan, Chengwei Chen, and Zhang Shi designed the study and developed the artificial intelligence models. Wenbin Liu, Xinyue Zhang and Ming Yang collected and processed the data. Yunshuo Zhang and Hui Jiang verified the pathological data. Mengmeng Zhu, Jieyu Yu, Fang Liu, and Jing Li performed the statistical analyses. Xiaohan Yuan, Chengwei Chen, and Zhang Shi wrote the main manuscript text. Xiaohan Yuan and Yun Bian prepared all the figures. Bozhu Chen, Jianping Lu, Chengwei Shao and Yun Bian provided critical revisions. All authors reviewed and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41746-025-01970-y>.

Correspondence and requests for materials should be addressed to Bozhu Chen, Jianping Lu, Chengwei Shao or Yun Bian.

Reprints and permissions information is available at <http://www.nature.com/reprints>

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

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025

Acknowledgements

This work was supported in part by the National Science Foundation for Scientists of China (81871352, 82171915, 82171930, 82271972 82371955,