scientific reports

Check for updates

OPEN

Enhancing noninvasive pancreatic cystic neoplasm diagnosis with multimodal machine learning

Wei Huang[1,5], Yue Xu[1,5], Zhao Li[2]✉, Jun Li[3], Qing Chen[3], Qiang Huang[4], Yaping Wu[4] & Hongtan Chen[1]✉

Pancreatic cystic neoplasms (PCNs) are a complex group of lesions with a spectrum of malignancy. Accurate differentiation of PCN types is crucial for patient management, as misdiagnosis can result in unnecessary surgeries or treatment delays, affecting the quality of life. The significance of developing a non-invasive, accurate diagnostic model is underscored by the need to improve patient outcomes and reduce the impact of these conditions. We developed a machine learning model capable of accurately identifying different types of PCNs in a non-invasive manner, by using a dataset comprising 449 MRI and 568 CT scans from adult patients, spanning from 2009 to 2022. The study's results indicate that our multimodal machine learning algorithm, which integrates both clinical and imaging data, significantly outperforms single-source data algorithms. Specifically, it demonstrated state-of-the-art performance in classifying PCN types, achieving an average accuracy of 91.2%, precision of 91.7%, sensitivity of 88.9%, and specificity of 96.5%. Remarkably, for patients with mucinous cystic neoplasms (MCNs), regardless of undergoing MRI or CT imaging, the model achieved a 100% prediction accuracy rate. It indicates that our non-invasive multimodal machine learning model offers strong support for the early screening of MCNs, and represents a significant advancement in PCN diagnosis for improving clinical practice and patient outcomes. We also achieved the best results on an additional pancreatic cancer dataset, which further proves the generality of our model.

Keywords Pancreatic cystic neoplasms, Machine learning, Diagnostic classification, Multimodal data, Clinical screening

Pancreatic cystic neoplasms (PCNs) can be histologically categorized into serous cystic neoplasms (SCNs), mucinous cystic neoplasms (MCNs), intraductal papillary mucinous neoplasms (IPMNs), and other rare tumorous cystic lesions[1]. In recent years, due to the increasing availability of medical screenings and the widespread utilization of imaging modalities such as abdominal computed tomography (CT), the incidence of PCNs has exhibited a steady annual increase. PCNs are estimated to be present in 2%–45% of the general population[2]. Within this cohort, IPMNs account for about 21%–41%, MCNs for 10%–45%, and SCNs for 18.3%–39%. Collectively, these three types constitute more than 90% of all PCN instances. It is worth noting that the distribution of PCN types may exhibit slight variations across different geographic regions[2–4].

Nonetheless, there exists a notable disparity in the likelihood of malignancy across various types of PCNs. Specifically, MCNs exhibit a malignancy rate of 10%-39%, whereas IPMNs have the highest potential for malignancy, ranging from 36%-100%[2,5]. Regarding SCNs, they are generally regarded as benign (malignancy rate less than 1%[6]). PCNs are commonly acknowledged as precursors to invasive pancreatic cancer[7]. However, owing to the generally delayed manifestation of symptoms, only a small proportion of patients, approximately 15% to 20%, are able to be diagnosed with PCNs and undergo prompt resection of the lesion[8].

At present, PCNs present two clinical issues: one is the low diagnostic accuracy of PCNs, and the other is the difficulty of risk stratification[9,10]. According to reports, utilizing radiological features to infer the type of PCNs still leads to an incorrect diagnosis in nearly 1/3 of cases[11]. Despite adhering to currently accepted guidelines for PCN treatment[2,10,12], hich may involve surgical resection or other interventions, postoperative pathology often reveals cases of overtreatment or delayed treatment. The European guidelines for pancreatic cystic tumors

[1]Department of Gastroenterology, The First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China. [2] Research Center for Data Hub and Security, Zhejiang Lab, Hangzhou, China. [3]Department of Pathology, The First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China. [4]Department of Imaging, The First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China. [5]Wei Huang and Yue Xu contributed equally to this work. ✉email: lzjoey@gmail.com; chenhongtan@zju.edu.cn

emphasize that the accuracy of relying on a single imaging modality to differentiate PCNs remains relatively low[2]. It is reported that the accuracy of magnetic resonance imaging (MRI) in identifying specific types of PCNs is 40%-95%, and that of Computed Tomography (CT) is 40%-81%[2]. Through early screening and precise diagnostic techniques, malignant PCNs can be promptly addressed, or unnecessary surgical procedures can be averted, ultimately enhancing the prognosis of patients.

The machine learning algorithms not only aids radiologists in their professional tasks but also substantially enhances the accuracy and objectivity in detecting different types of pancreatic cysts on CT scans and MRI. However, existing machine learning approaches predominantly focus on unimodal data analysis, neglecting the synergistic potential of integrating radiomic features with critical clinical biomarkers and patient histories. Dmitriev et al.[13] used machine learning models for diagnosing SCNs, MCNs, IPMNs, and SPNs based on radiomics and clinical features (age and sex). The diagnostic accuracy was 95.9%, 64.3%, 51.7%, and 100%, respectively. DONG et al.[14] selected the radiomics features of CT for model constructing, and the diagnostic accuracy of SCNs, MCNs, and IPMNs was 74.26%, 78.37%, and 68.00%, respectively. For identifying benign and malignant PCNs, Corral et al.[15] had a sensitivity and specificity of 75% and 78%, respectively, for discriminating malignant IPMNs based on MRI. The model constructed by Wang et al.[16] based on CT had the highest accuracy of 90.4% in identifying malignant PCNs. Moreover, as indicated by previous research[17,18], clinically significant variables such as smoking and diabetes history have been found to be associated with the malignant progression of IPMNs, and the Fukuoka guidelines have identified elevated serum CA19.9 levels as a concerning indicator of IPMNs[12]. However, the potential for these clinical features to enhance the precision of diagnosing PCNs has not been extensively investigated.

In this article, we introduce an innovative machine learning model which discriminates effectively among various types of pancreatic cysts by leveraging comprehensive analysis of both imaging features and detailed patient clinical data. It achieves state-of-the-art performance in predicting PCNs with various metrics. Moreover, when applied to patients with MCN undergoing diagnostic CT imaging, the model achieves a perfect prediction accuracy rate of 100%. This highlights the substantial clinical benefits and relevance of our early screening for MCNs.

Method

In this work, we propose a Multimodal Deep Forest (MDF) model for classifying pancreatic cystic neoplasms (PCNs) using both imaging and clinical data. Based on Deep Forest[19], an ensemble method that employs cascaded tree ensembles as an alternative to deep neural networks, MDF excels on medium and small-scale tabular data by automatically learning feature interactions. It handles heterogeneous features with minimal preprocessing, is robust against overfitting, and requires little hyperparameter tuning[20]. Its tree-based structure also enhances interpretability by revealing feature importance and decision pathways[21]. The overall process is illustrated in Fig. 1.

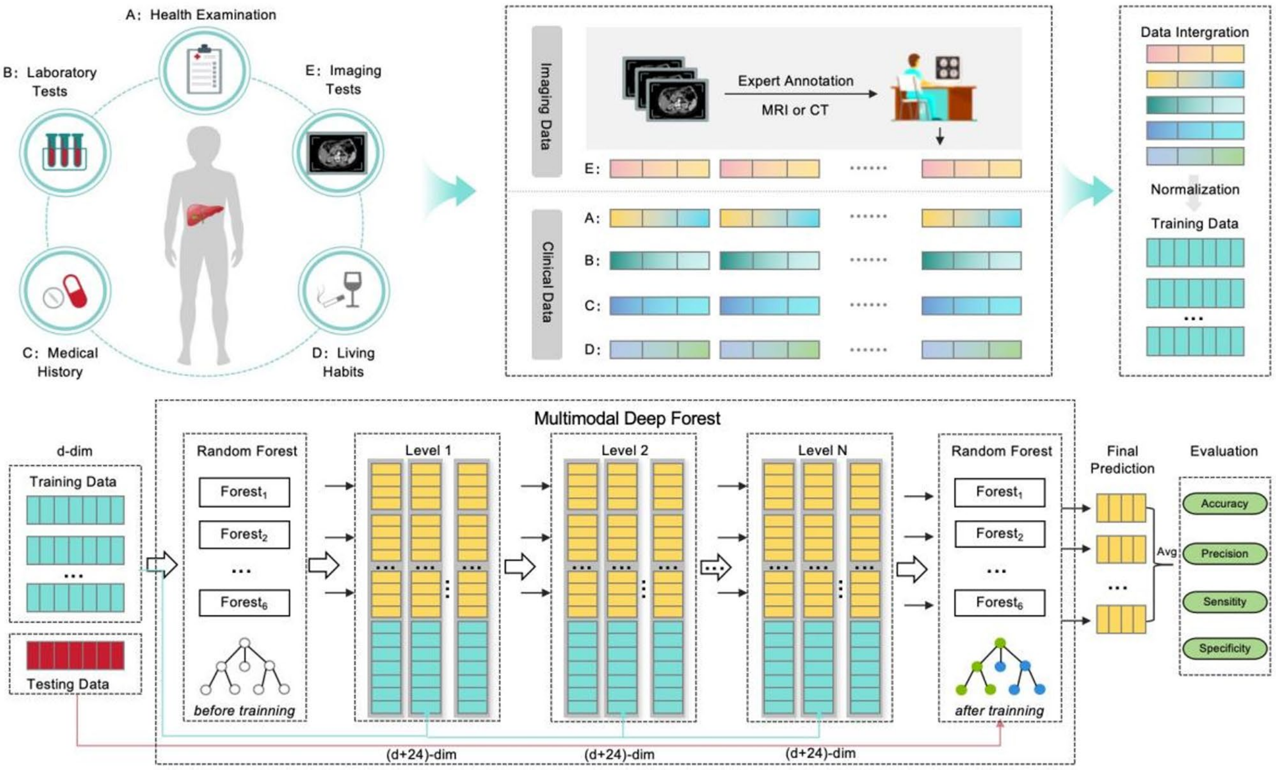

Fig. 1. The proposed multimodal deep forest model framework.

Data collection

This study was conducted in accordance with relevant guidelines and regulations and received approval from The First Affiliated Hospital, Zhejiang University School of Medicine [Approval Number: IIT20230145A], which included PCN patients diagnosed with postoperative pathology between January 2009 and May 2022. As SCNs, MCNs, and IPMNs account for more than 90\% of PCNs all three were included in this study, while rare solid pseudopapillary neoplasms and other pancreatic cystic neoplasms were not included. Informed consent was obtained from all subjects and/or their legal guardians prior to participation.

In this work, we initially structured the imaging data of PCNs by manually organizing it into standardized information (including location, size, pancreatic duct diameter, communication between the lesion and pancreatic duct, mural nodules, calcification, border, etc.). Additionally, we incorporated patient clinical information such as health examinations, laboratory tests, medical history, and living habits.

The imaging data of all patients were acquired by certified professionals using CT or MRI enhanced scans, with a slice thickness of 5 mm for both modalities. Subsequently, the acquired imaging material underwent a retrospective review completed by two or more professional radiologists and gastroenterologists each. Following the deliberations of all experts, determinations were made regarding the type, size, pancreatic duct dilation, communication between lesions and the pancreatic duct, etc.

The pathological slides of all lesions resected from patients were evaluated for pathological type by two pathologists with expertise in the field. According to the classification system of the World Health Organization in 2010, SCNs were classified as benign SCNs and serous cystadenocarcinoma, while MCNs and IPMNs were categorized into mild to moderate dysplasia, severe dysplasia, and associated invasive carcinoma based on the degree of epithelial dysplasia[22]. If different degrees of epithelial dysplasia coexisted in a lesion, the highest degree was recorded. However, due to the rarity of malignant cases in SCNs and MCNs, this study mainly focused on conducting further experiments on the benign IPMNs (abbr. IPMN-0) and malignant IPMNs (abbr. IPMN-1).

To address variability arising from different imaging modalities (e.g., MRI-specific features such as T1WI/T2WI and CT-specific features such as calcification patterns), missing values were systematically handled. For instance, MRI-specific features unavailable in CT data were labeled as "unknown," while CT-specific features missing in MRI data were assigned a numerical value of – 1. This approach ensured consistent feature dimensions across all inputs, enabling robust multimodal integration.

Input data preprocessing

The input data is mainly consisted of two dimensions: $\mathrm{x}_1 \in \mathbb{R}^{\mathrm{n}\times \mathrm{d}_1}$ represents the clinical data; $\mathrm{x}_2 \in \mathbb{R}^{\mathrm{n}\times \mathrm{d}_2}$ denotes the imaging data. To mitigate information bias among different features, standardization is performed as $x = \left[\frac{\mathrm{x}_1 - \mu_1}{\sigma_1} \oplus \frac{\mathrm{x}_2 - \mu_2}{\sigma_2}\right]$, $x \in \mathbb{R}^{\mathrm{n}\times(\mathrm{d}_1+\mathrm{d}_2)}$, where $\oplus$ denotes the the concatenation operation.

Model objective

Intermediate tree layer function is defined as follows:

$$f_t\left(x\right) = \left\{ \begin{array}{l} h_1\left(x\right), t = 1, \\ h_t\left(x \oplus f_{t-1}\left(x\right)\right), t > 1, \end{array} \right.$$

where h_t represents the ensemble of forests at layer t, f_t represents the output results of the forest ensemble at layer t. The prediction value of each sample at each layer is $\not{y}_i^t = avg(f_t(x))$. The loss function of our multimodal deep forest is formulated as follows:

$$L = \sum_{t=1}^{N}\sum_{i=1}^{n} l(y_i, \not{y}_i^t) + \sum_{t=1}^{N}\Omega\left(f_t\right),$$

where $l\left(\cdot\right)$ represents the cross-entropy loss function, and

$$\Omega\left(f_t\right) = \gamma k + \frac{1}{2}\lambda\sum_{j=1}^{n}\omega_j^2,$$

where k represents the number of leaf nodes in the tree, and ω represents the value of the leaf nodes.

Model training and testing

For the convenience of data analysis and comparison, we have grouped the three types of PCN categories and the two types of IPNM categories together, thus turning it into a four-category problem. We input the d-dimensional training data into the MDF model, where the model is trained by 6 random forest classifiers through an N-layer cascade network. After training, predictions are made on the test data. Each classifier at each layer outputs four types of predictions, forming a 24-dimensional output result. At the same time, the model connects the original d-dimensional input through a cascading method, turning it into a vector of d + 24 dimensions. This vector then serves as the input data for the classifiers in the next layer until the training process is completed. Finally, the prediction results for the test data are obtained by taking a weighted average of the results from the 6 classifiers. We partitioned the dataset into a training set and a test set. The training set accounted for 70% of the entire dataset, while the test set accounted for 30%. To better evaluate the model's performance, we primarily used five metrics: Accuracy, Precision, Sensitivity, Specificity, and Area Under roc Curve (ROC-AUC) for additional comparisons between different algorithms.

MDF requires moderate memory and benefits from parallel processing, allowing training on medium-sized datasets in minutes to hours-faster than computationally intensive nature of these widely adopted deep learning architectures such as DNNs, CNNs, and Transformers. Once trained, its low memory footprint enables millisecond-level predictions on standard hardware, supporting deployment on low-power devices for real-time clinical applications.

Hyperparameter tuning strategy

The hyperparameter mainly consists of number of layers in the cascade network and number of Decision Trees.

The cascade structure grows layer-by-layer during training. After each layer is added, the model evaluates its performance on a validation set. If adding a new layer improves performance significantly, the cascade continues to grow. If performance stops improving or reaches a plateau, the cascade growth is halted. This process ensures the model adapts to the complexity of the data without overfitting.

Each layer in the cascade consists of multiple random forests and completely-random tree forests. The number of trees in each forest is typically predefined (e.g., 6 random forests) and not optimized during training. However, the ensemble nature of random forests inherently provides robustness, reducing the need for fine-tuning the number of trees.

After each layer, the model's performance is assessed using a validation set. If performance improves, the cascade grows; otherwise, it stops.

In MDF model, the number of layers in the cascade network is optimized dynamically during training based on validation performance, avoiding manual tuning of the number of layers, while the number of decision trees in each forest is typically fixed. This adaptive approach makes the algorithm flexible and data-driven. The ensemble of forests ensures strong generalization without overfitting.

Results

Data characteristics

Our data were presented as mean ± standard deviation, and categorical data were presented as percentages (%). Differences between groups were analyzed using Student's t-test or chi-square test, and a *p*-value less than 0.05 is considered statistically significant. Thus, we obtained two datasets consisting of 449 patients for MRI (Table 1), and 568 patients for CT (Table 2). As shown in tables, the statistically significant features in both the MRI and CT datasets include "age", "gender", "smoking history", "drinking history", "hypertension", "diabetes", "main pancreatic duct", "history of pancreatitis", and "thickness" The data revealed no statistically significant differences in abdominal distension and discomfort observed among the three groups, as demonstrated by the respective *p*-values of 0.628 and 0.355 in the MRI dataset, and 0.170 and 0.284 in the CT dataset.

Moreover, the levels of laboratory blood parameters, including alanine aminotransferase (ALT, 25.23U/L MRI, 25.14U/L CT), aspartate transaminase (AST, 22.70U/L MRI, 22.95U/L CT), alkaline phosphatase (ALP, 88.61U/L MRI, 88.25U/L CT), gamma-glutamyl transferase (Γ-GT, 46.63U/L MRI, 46.00U/L CT), carcinoembryonic antigen (CEA, 7.26 ng/ml MRI, 6.45 ng/ml CT), and cancer antigen 19.9 (CA19.9, 60.61ku/L MRI, 58.00ku/L CT), were significantly elevated in patients with IPMN compared to those with MCN or SCN. Specifically, patients with IPMN exhibited a greater degree of variation in these blood markers, as indicated by the standard deviation (std), signifying less stability in their values. Conversely, patients with SCN demonstrated greater instability in ferritin levels compared to other patient groups.

Comparisons of different data

Quantitative assessment of our proposed method was performed using 20-fold cross-validation. The results of all metrics were averaged from 20 experimental trials. To assess the significance of the differences between the results, t-tests were conducted for all four metrics. Additionally, the ensemble of diverse forests (6 per layer) further mitigated overfitting risks, ensuring reliable generalization despite dataset size constraints, and achieving stable metrics across all trials. The main results are listed as follows:

The impact of multimodal data: Fig. 2A, shows that combining clinical and imaging data significantly improves MDF's performance across all metrics. Specifically, accuracy, precision, sensitivity, and specificity improved by 8.0%, 7.7%, 8.4%, and 2.6%, respectively, compared to using clinical or imaging data alone. These improvements highlight the substantial positive effect of multimodal data on predictive accuracy, particularly in achieving a specificity of 96.5%.

The impact of four categories: Fig. 2B, shows that the MDF's predictive results for MCNs are nearly 100% on all four metrics, with Specificity exceeding 96% for predictions on IPMN1 and SCNs, and Accuracy exceeding 90% for predictions on IPMN-1 (malignant IPMN), IPMN-0 (benign IPMN), and SCNs. The Sensitivity for predictions across the four categories is slightly lower, but overall, it has reached 88.9%.

The impact of MRI and CT data: Fig. 2C,D respectively display the comparative results of the MDF using MRI and CT data, with markers (such as upper triangles, lower triangles) used to compare the superiority or inferiority of MRI data and CT data on the corresponding metrics. An upper triangle indicates that the data is superior to the other data on that metric; a lower triangle indicates that the data is inferior to the other data on that metric; a square marker indicates that the *p*-value for MRI and CT under the current label is > 0.05, meaning there is no statistical difference in the predictive results between the two data types. It can be seen that overall, the MRI data is superior to the CT data on all four metrics for the four-category predictions, with MRI leading in 10 out of 16 dimensions, and the remaining 6 being on par with CT data. Especially in the prediction of MCN, there is no difference between the two, indicating that the MDF model can fully replace MRI in predicting MCN issues.

Feature	IPMN (n = 246)	MCN (n = 109)	SCN (n = 94)	*p*-value
Age (Mean, std.)	63.15 (8.87)	45.81 (12.91)	53.35 (11.35)	0.000
Gender (Male/Female)	146/100	10/99	22/72	0.000
Abdominal Pain (Yes/No)	99/147	34/75	26/67	0.075
Abdominal Distension (Yes/No	39/207	13/96	14/80	0.628
Abdominal Discomfort (Yes/No)	22/224	5/104	8/86	0.355
Weak (Yes/No)	8/238	0/109	1/93	0.100
Appetite Loss (Yes/No)	34/212	16/93	18/76	0.467
Weight Loss (Yes/No)	43/203	11/98	10/84	0.098
Smoking History (Yes/No)	87/159	6/103	12/82	0.000
Drinking History (Yes/No)	63/183	5/104	7/87	0.000
with Hypertension (Yes/No)	102/144	17/92	30/64	0.000
with Diabetes (Yes/No)	34/212	10/99	8/86	0.000
complicated with pancreatitis (Yes/No)	63/183	5/104	7/87	0.000
formerly history of pancreatitis (Yes/No)	19/227	2/107	0/94	0.003
Other tumors (Yes/No)	21/225	3/106	7/87	0.136
Family History of tumors (Yes/No)	45/201	14/95	14/80	0.406
HB (Mean,std.)	135.27 (15.33)	130.93 (13.65)	134.89 (14.45)	–
ALT (Mean,std.)	25.23 (29.54)	18.11 (18.91)	20.40 (23.34)	–
AST (Mean,std.)	22.70 (13.16)	17.93 (5.83)	19.55 (13.07)	–
ALP (Mean,std	88.61 (45.08)	72.33 (25.56)	75.62 (38.84)	–
TB (Mean,std.)	11.92 (14.77)	11.06 (4.96)	12.65 (15.39)	–
DB (Mean,std.)	5.08 (13.08)	4.13 (2.07)	5.52 (15.16)	–
IB (Mean,std.)	6.84 (3.35)	6.92 (3.40)	7.13 (3.11)	–
Γ-GT (Mean,std.)	46.63 (87.54)	26.05 (36.89)	28.06 (48.79)	–
Jaundice (Yes/No)	28/218	13/96	13/81	0.825
CEA (Mean, std.)	7.26 (64.18)	1.76 (1.04)	1.95 (1.11)	–
CA199 (Mean, std.)	60.61 (292.31)	19.09 (28.55)	16.86 (39.41)	–
CA125 (Mean, std.)	14.62 (17.48)	16.50 (11.82)	30.26 (138.79)	–
Ferritin (Mean, std.)	245.35 (244.24)	121.12 (210.14)	410.18 (2414.49)	–
Thickness (Thick/Thin)	68/178	34/60	34/60	0.020
Separation (Yes/No)	210/36	95/14	Aug-86	0.323
Fibrotic scar (Yes/No)	0/246	1/108	7/87	0.000
Mural nodule (Yes/No)	65/181	32/77	15/79	0.064
Pancreatic duct dilation (Yes/No)	208/38	12/97	23/71	0.000
Main pancreatic duct diameter (< 5 mm,5-10 mm, > 10 mm)	100/100/46	104/4/1	1981/11/2	0.000
T1W (low, high, mixed)	240/5/1	102/4/3	1986/4/4	–
T2WI (low, high, mixed)	3/242/1	1/108/0	0/92/2	–

Table 1. MRI dataset.

Comparisons of subgroup

We have conducted the subgroup analysis by age for PCNs as shown in Fig. 3. It reveals distinct performance patterns across age groups: younger cohorts (< 18 and 18–35) achieve near-perfect accuracy (up to 1.0) and specificity (> 0.99) but exhibit notably lower precision and sensitivity (≤ 0.63), suggesting potential overconfidence in negative predictions. Middle-aged groups (35–65) demonstrate the most balanced performance, with precision (0.91), sensitivity (0.89), and ROC-AUC (0.93) peaking, indicating robust generalization. Older adults (65 +) show declines in accuracy (0.83) and specificity (0.92), yet maintain stable ROC-AUC (0.92), highlighting retained discriminative power despite age-related variability. Overall, precision and sensitivity improve with age, while specificity remains high across all groups. It is worth noting that the average age of patients in our data is in their 50 s, with the majority concentrated between 35 and 65, which has caused some bias to the results.

Comparisons of different models

To further verify the effectiveness of the model, we compared it with ten other popular AI models, namely Naive Bayes, Supporting Vector Machine (SVM), Decision Tree, CatBoost, XgBoost, Random Forest, Multilayer Perceptron (MLP) within the sklearn toolbox[23], Deep Neural Network (DNN), Convolutional Neural Networks (CNN)[24], and Transformer[25]. Among them, Naive Bayes is a probabilistic classifier based on Bayes' theorem. SVM is a supervised model for classification and regression. Decision Trees split data using feature conditions. Random Forest combines multiple Decision Trees for better performance, and it's the base classifier of out MDF model. The CatBoost and XgBoost enhance gradient boosted tree models. MLP, a basic neural network

Feature	IPMN(n = 311)	MCN(n = 144)	SCN(n = 113)	*p*-value
Age(Mean, std.)	62.94(9.26)	45.49(12.84)	54.08(11.59)	0.000
Gender(Male/Female)	184/127	14/130	24/89	0.000
Abdominal Pain(Yes/No)	130/181	51/93	33/80	0.049
AbdominalDistension(es/No)	52/259	17/127	23/90	0.17
Abdominal Discomfort(Yes/No)	31/280	8/136	9/104	0.284
Weak(Yes/No)	10/301	2/142	3/110	0.529
Appetite Loss(Yes/No)	42/269	26/118	27/86	0.035
Weight Loss(Yes/No)	48/263	12/132	14/99	0.109
Smoking History(Yes/No)	108/203	8/136	14/99	0.000
Drinking History(Yes/No)	75/236	8/136	10/103	0.000
with Hypertension(Yes/No)	119/192	21/123	33/80	0.000
with Diabetes(Yes/No)	42/269	10/134	7/106	0.000
comorbid pancreatitis(Yes/No)	52/259	10/134	6/107	0.000
formerly history of pancreatitis(Yes/No)	24/287	5/139	0/113	0.000
Other tumors(Yes/No)	20/291	4/140	9/104	0.166
Family History of tumors (Yes/No)	58/253	21/123	16/97	0.400
HB(Mean,std.)	134.84(15.80)	129.91(12.09)	133.50(15.30)	–
ALT(Mean,std.)	25.14(30.48)	17.37(11.88)	19.29(21.90)	–
AST(Mean,std.)	22.95(15.09)	18.41(6.45)	19.27(12.15)	–
ALP(Mean,std.)	88.25(48.38)	69.33(23.62)	74.92(37.19)	–
TB(Mean,std.)	12.09(15.93)	10.93(4.83)	12.21(14.39)	–
DB(Mean,std.)	5.22(14.09)	4.10(2.01)	5.17(14.09)	–
IB(Mean,std.)	6.88(3.47)	6.83(3.41)	7.04(3.19)	–
Γ-GT(Mean,std.)	46.00(87.27)	24.89(33.19)	26.23(45.40)	–
Jaundice(Yes/No)	33/278	16/128	14/99	0.876
CEA (Mean, std.)	6.45(57.69)	1.76(1.09)	1.92(1.11)	–
CA199 (Mean, std.)	58.00(270.24)	23.52(62.73)	15.52(37.08)	–
CA125 (Mean, std.)	14.53(17.63)	16.90(17.31)	28.24(129.35)	–
Ferritin (Mean, std.)	247.61(249.01)	124.55(195.65)	384.63(2251.39)	–
Thickness (Thick/Thin)	44/267	35/108	17/96	0.020
Separation (Yes/No)	276/35	107/37	94/19	0.000
Mural nodule (Yes/No)	110/201	41/103	13/100	0.000
Pancreatic Duct Dilation (Yes/No)	266/45	21/123	35/77	–
Main Pancreatic Duct Length (< 5 cm, 5 cm–10 cm, > 10 cm)	100/100/46	104/4/1	1981/11/2	0.000

Table 2. CT dataset. A. MRI Dataset; B. CT Dataset; Abbreviations: -, not applicable; HB, hemoglobin; TB, total bilirubin; DB, direct bilirubin; IB, indirect bilirubin; ALT, alanine aminotransferase; AST, aspartate transaminase; ALP, alkaline phosphatase; Γ-GT, gamma-glutamyl transferase; CEA, carcinoembryonic antigen;CA19.9, cancer antigen 19.9; CA125, cancer antigen 125; IPMN, intraductal papillary mucinous neoplasms; MCN, mucinous cystic neoplasms; SCN,serous cystic neoplasms.

in sklearn, has multiple layers. DNNs, with more hidden layers, model complex relationships. CNNs excel at grid data like images. Transformers handle sequences using self-attention, and becomes the most popular deep learning for various domains.

From Table 3, The Multimodal Deep Forest (MDF) model demonstrates superior performance across all evaluation metrics compared to baseline and state-of-the-art models. It achieves the highest accuracy (91.26% ± 1.70%), precision (91.74% ± 2.50%), sensitivity (88.91% ± 2.02%), specificity (96.53% ± 0.88%), and ROC-AUC (95.84% ± 2.14%), with statistically significant improvements ($p < 0.001$) over all competitors. Notably, MDF outperforms traditional models (e.g., SVM, Naive Bayes) by 75.9% in accuracy and 149% in precision, while surpassing deep learning approaches (e.g., DNN, CNN, Transformer) by 1.8–12.5% in accuracy and 3.36–40.4% in ROC-AUC. Its robust performance (low standard deviations) and balanced metric dominance highlight its efficacy in integrating multimodal data, mitigating overfitting, and generalizing across heterogeneous clinical scenarios, making it a clinically reliable and interpretable tool for complex diagnostic tasks.

Beyond performance metrics, computational efficiency and interpretability were critically evaluated. The Multimodal Deep Forest (MDF) model demonstrated superior classification accuracy compared to baseline methods but required higher computational resources (1.0 s per iteration without GPU acceleration, reduced to 0.39 s with GPU parallelization). In contrast, Random Forest (0.09 s/iteration) and SVM (0.02 s/iteration) exhibited faster per-iteration speeds. While CNNs require 0.94 s per iteration and Transformers exceed 1 s per iteration, these findings highlight the computationally intensive nature of these widely adopted deep

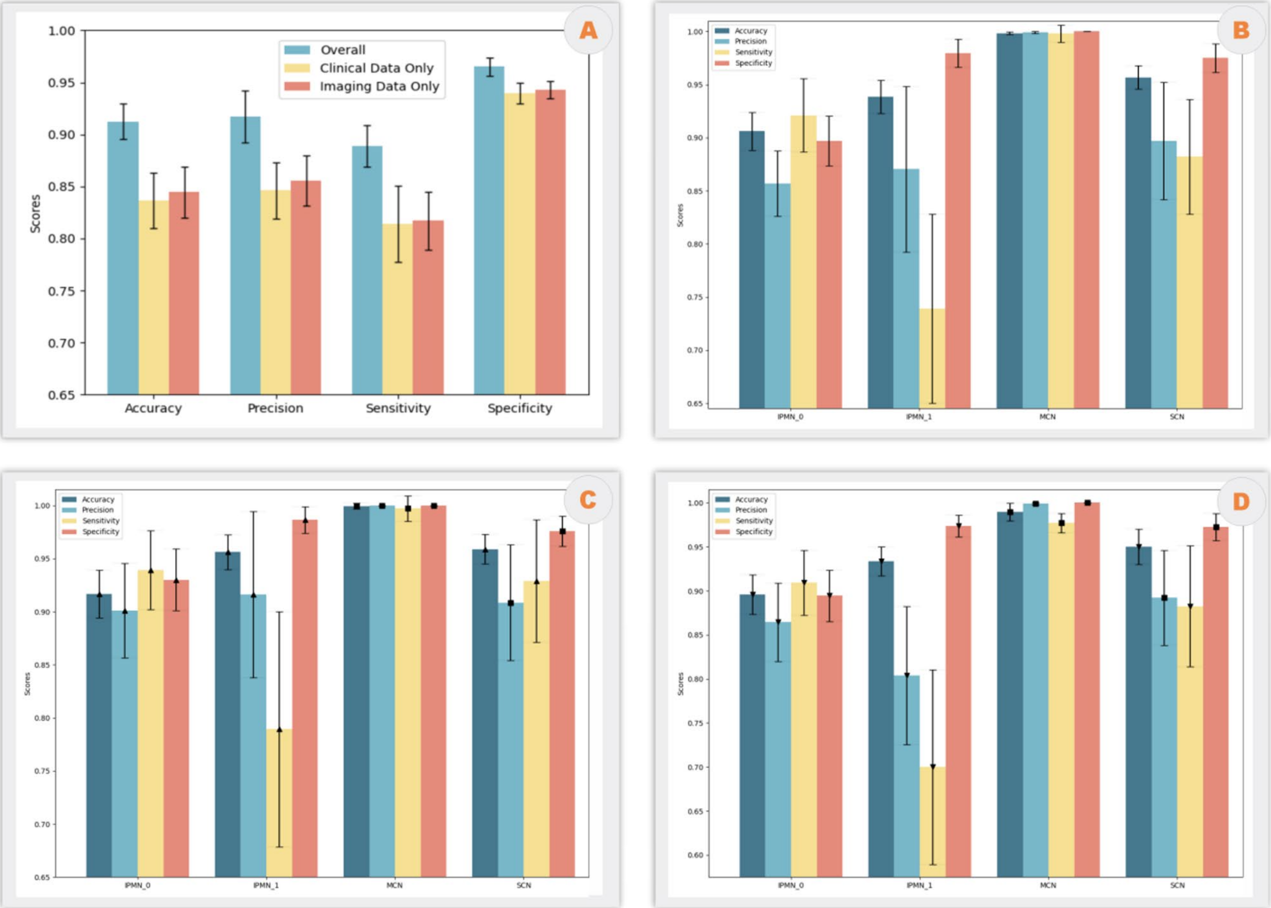

Fig. 2. Comparison of Predictive Results With Different Data: (**A**) The Comparison of Different Modalities of Data; (**B**) The Comparison of Results on Different Categories; (**C,D**) The Comparison of Results on MRI and CT Data, respectively (Note: An upper triangle indicates the win situation between MRI data and CT data; a lower triangle indicates the lose situation ; a square marker indicates the tie situation).

learning architectures. For interpretability, MDF leverages SHAP values to quantify feature importance, akin to Random Forest, while SVM with nonlinear kernels lacks transparency. The MDF's balance of performance and explainability justifies its adoption for complex multimodal tasks, supported by tools like SHAP to map feature contributions.

Results analysis

Compared to the lack of interpretability in AI models such as deep learning, our MDF model based on the random forest can effectively map relationship between features and results, providing convenience for model interpretability. In order to better analyze the key feature factors of model prediction, we employed the SHAP (Shapley Additive explanations) values for interpretation[26], which is a framework for interpreting machine learning model predictions (available at https://github.com/shap/shap).

The Summary Plot in SHAP (Fig. 4) is a visualization tool that displays the average contribution of all features to the model's predictions in the form of a bar chart, with features typically sorted according to the magnitude of their average impact on the model's predictions, aiding in the rapid identification of the most important features for the model's forecasts. From the summary plot, it is found that the top 20 predictive factors include: 7 imaging characteristics, 9 blood indicators, 2 demographic characteristics, and 2 medical histories. This further proves that imaging data and clinical data are equally important for improving the accuracy of predictions. The most important indicators are "complicated with pancreatitis", "main pancreatic duct width", "age", "communicated with the pancreatic duct", "CA19.9", and "CA125". It's worth noting that "age", gender, and medical history also play important roles in our model's predictions, and statistically, these indicators are also highly significant.

In SHAP, the Beeswarm plot (Fig. 5) is a scatter plot used to display the distribution of SHAP values for each feature. The positioning of blue and red dots relative to a central line provides insights into the impact of features on a model's predictions. When dots are closely clustered around the central line, whether they are blue (indicating low feature values) or red (indicating high feature values), their SHAP values hover near zero. This suggests minimal influence of the feature on the model's predictions. In contrast, when blue and red dots are

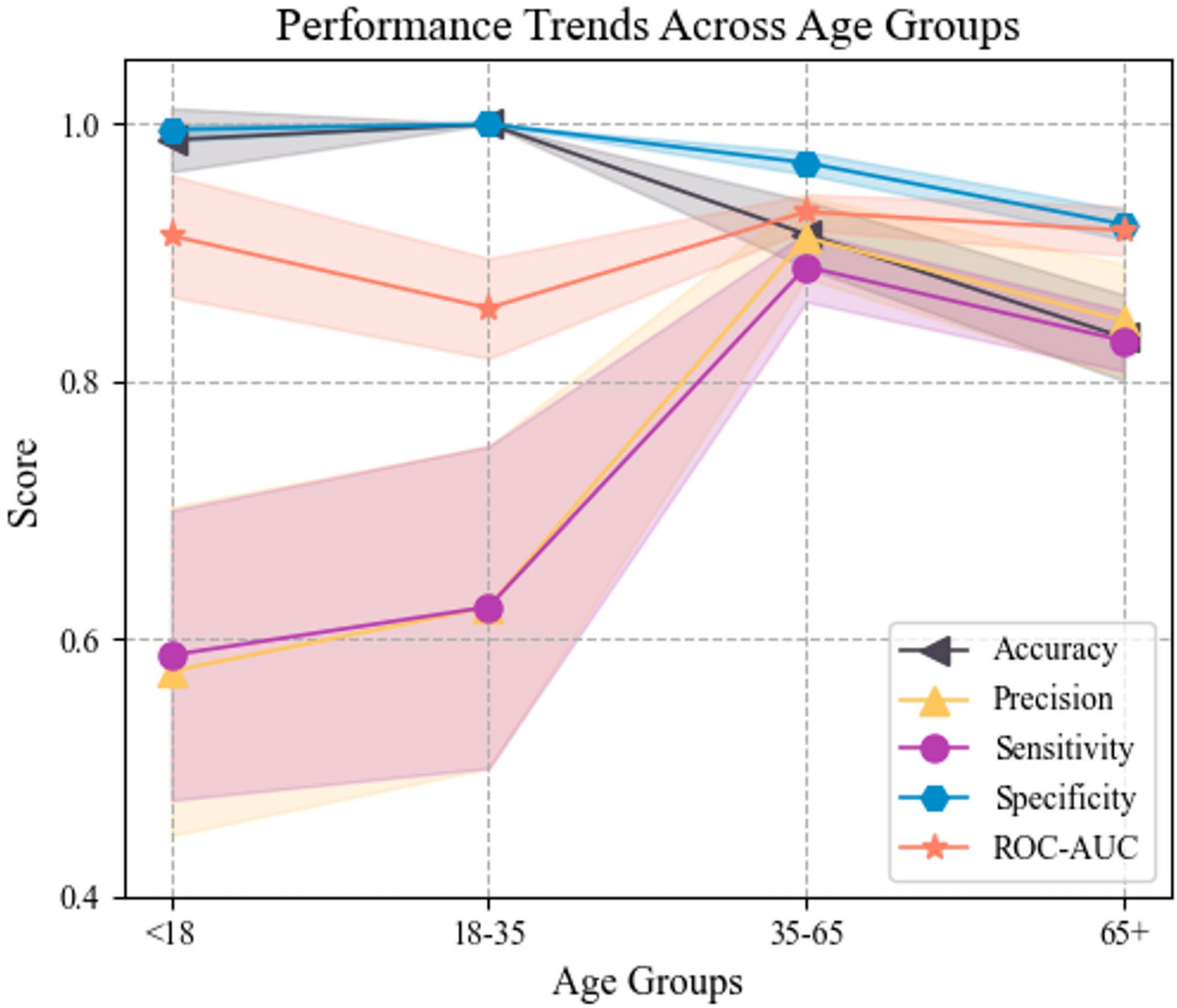

Fig. 3. Performance trends across age groups.

Models	Accuracy	Precision	Sensitivity	Specificity	ROC-AUC
NaiveBayes	0.6203 ± 0.0296	0.6394 ± 0.0237	0.6419 ± 0.0264	0.8741 ± 0.0116	0.7923 ± 0.0897
SVM	0.5188 ± 0.0288	0.3670 ± 0.0346	0.3727 ± 0.0159	0.8049 ± 0.0058	0.6825 ± 0.0927
Decision Tree	0.7878 ± 0.0087	0.7639 ± 0.0189	0.7648 ± 0.0199	0.9256 ± 0.0024	0.8755 ± 0.0522
CatBoost	0.8964 ± 0.0260	0.9070 ± 0.0291	0.8698 ± 0.0206	0.9607 ± 0.0084	0.9272 ± 0.0259
XgBoost	0.8911 ± 0.0246	0.8876 ± 0.0271	0.8681 ± 0.0270	0.9599 ± 0.0088	0.9243 ± 0.0155
Random Forest	0.8680 ± 0.0101	0.8850 ± 0.0245	0.8221 ± 0.0184	0.9493 ± 0.0044	0.9065 ± 0.0432
Multilayer Perceptron	0.7665 ± 0.0298	0.7474 ± 0.0315	0.7334 ± 0.3155	0.7335 ± 0.0370	0.8512 ± 0.0417
DNN	0.8083 ± 0.0141	0.8007 ± 0.0208	0.7851 ± 0.0261	0.9326 ± 0.0054	0.901 ± 0.0153
CNN	0.8135 ± 0.0621	0.8023 ± 0.0981	0.8471 ± 0.0339	0.8382 ± 0.0635	0.8789 ± 0.0313
Transformer	0.8294 ± 0.0509	0.8374 ± 0.0387	0.8146 ± 0.0823	0.8070 ± 0.0143	0.9113 ± 0.0231
Multimodal Deep Forest	**0.9126 ± 0.0170** *	**0.9174 ± 0.0250***	**0.8891 ± 0.0202***	**0.9653 ± 0.0088***	**0.9584 ± 0.0214***
Improvement	**1.8% ~ 75.9%**	**3.35% ~ 149%**	**2.21% ~ 138%**	**0.47% ~ 31.6%**	**3.36% ~ 40.4%**

Table 3. Comparison of different machine learning classifers. A paired *t* test is performed and * indicates a statistical significance $p < .001$ as compared to the best baseline method; ± represents the mean and variance. Significant values are in [bold].

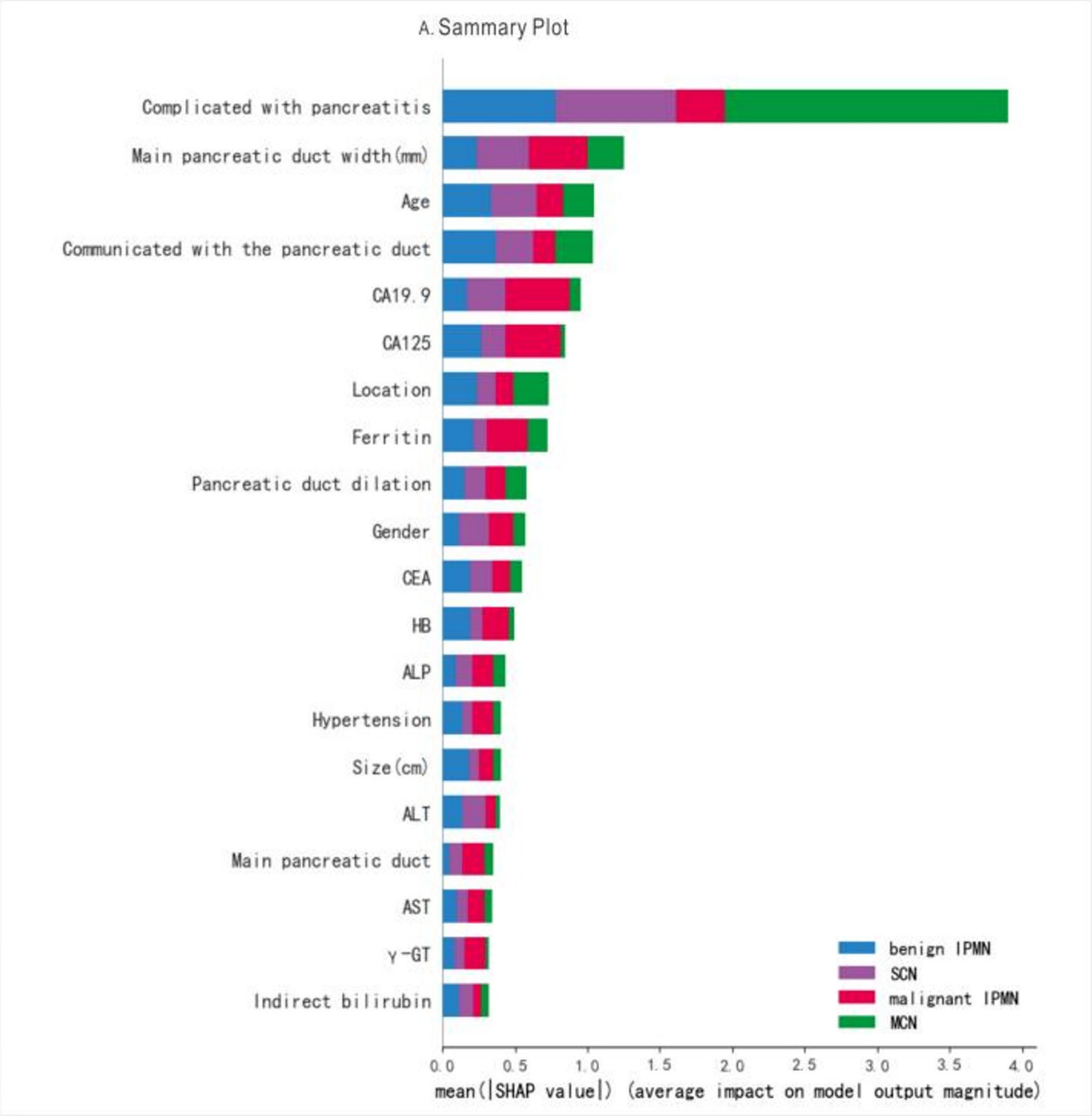

Fig. 4. Interpretability of the MDF Model Results with Summary Plot of PCNs. This figure was generated using SHAP.

widely dispersed away from the central line, it indicates that changes in the feature values significantly affect SHAP values, signifying a strong impact of the feature on the model's predictive outcomes.

From Fig. 5, it can be seen that among the most important indicators in the summary plot, the first-ranked is clinical indicator "complicated with pancreatitis" that plays a crucial role in the prediction of all four types of PCNs, especially having an absolute dominant role in the prediction of MCNs. The second-ranked is imaging indicator "main pancreatic duct width" which ranks very high in all four categories and plays a very important role in the prediction of all categories. This also explains the results of Fig. 3, where the effect of using only imaging data is slightly better than that of using only clinical data. The clinical indicators "CA19.9" and "CA125" appear to be more important in the prediction of malignant IPMNs. At the same time, the indicator "age" also has a clear distinguishing effect on the four classifications, which is consistent with the statistical analysis in Tables 1 and 2.

Our study revealed that elevated levels of "complicated with pancreatitis" are associated with an increased likelihood of predicting IPMN-0, IPMN-1, and SCNs, while they decrease the likelihood of MCNs. The presence of acute pancreatitis independently aids in distinguishing between MCNs and IPMNs[27]. Blood tumor markers,

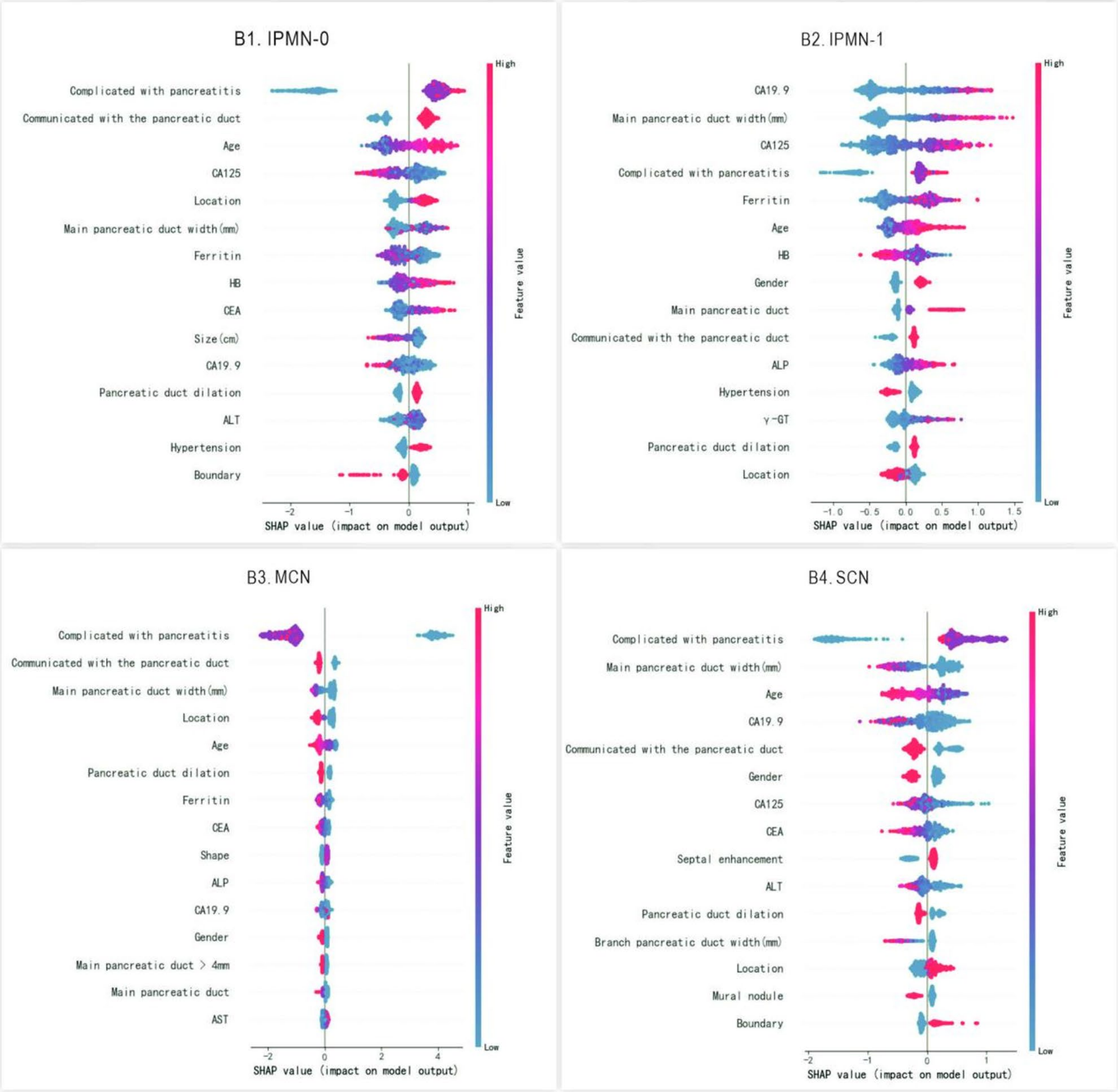

Fig. 5. Interpretability of the MDF Model Results with Beeswarm Plot of PCNs (B1. Beeswarm Plot of IPMN-0; B2. Beeswarm Plot of IPMN-1; B3. Beeswarm Plot of MCN; B4. Beeswarm Plot of SCN. Abbreviations: IPMN-0, benign intraductal papillary mucinous neoplasm; IPMN-1, malignant intraductal papillary mucinous neoplasm; MCN, mucinous cystic neoplasm; SCN,serous cystic neoplasm;HB, hemoglobin; ALT, alanine aminotransferase; AST, aspartate transaminase; ALP, alkaline phosphatase; Γ-GT, gamma-glutamyl transferase; CEA, carcinoembryonic antigen; CA19.9, cancer antigen 19.9; CA125, cancer antigen 125). This figure was generated using SHAP.

including "CA19.9", “CEA”, “CA125”, and “ferritin”, when elevated, increase the predictive likelihood for IPMN-1 and conversely decrease it for IPMN-0, MCNs, and SCNs. A high serum CA199 level is a significant indicator, aligning with high-risk features outlined in current guidelines[2]. The use of combined cyst fluid CEA and CA125 levels may facilitate the differentiation between MCNs and IPMNs[28]. Prior research has indicated that CA125 is effective in predicting invasive IPMN within the CA199-negative subgroup[29]. Additionally, serum ferritin serves as a biomarker for pancreatic cancer[30].

Imaging characteristics, such as "main pancreatic duct width", "lesion communication with the pancreatic duct", “lesion location”, and “size”, along with clinical characteristics including “age”, “gender”, and “ALP”, contribute to the classification diagnosis of PCNs and the differentiation of benign and malignant IPMNs. These findings are consistent with the diagnostic criteria for IPMN[12] and corroborate literature suggesting that lesion location[31], age of onset, gender[32], and ALP[33] are correlated with the type and nature of PCNs.

Models	Accuracy	Precision	Sensitivity	Specificity	ROC-AUC
NaiveBayes	0.9344 ± 0.0639	0.9251 ± 0.0938	0.9382 ± 0.0708	0.9358 ± 0.0656	0.9640 ± 0.0500
SVM	0.6656 ± 0.1157	0.9292 ± 0.2225	0.3268 ± 0.1481	0.8009 ± 0.0679	0.8620 ± 0.0881
Decision Tree	0.8625 ± 0.0852	0.8431 ± 0.1119	0.8755 ± 0.1209	0.8648 ± 0.0869	0.8546 ± 0.0994
CatBoost	0.9250 ± 0.0702	0.9671 ± 0.0852	0.8588 ± 0.1357	0.9298 ± 0.0753	0.9873 ± 0.0259
XgBoost	0.8563 ± 0.0908	0.8908 ± 0.1142	0.8351 ± 0.1059	0.8525 ± 0.0979	0.9600 ± 0.0433
Random Forest	0.9406 ± 0.0608	0.9621 ± 0.0672	0.9060 ± 0.0672	0.9452 ± 0.0598	0.9855 ± 0.0306
Multilayer Perceptron	0.8719 ± 0.0916	0.8976 ± 0.1144	0.8441 ± 0.1580	0.8895 ± 0.0839	0.9375 ± 0.0614
DNN	0.8281 ± 0.0681	0.7926 ± 0.1007	0.8868 ± 0.0991	0.8350 ± 0.0689	0.8849 ± 0.1353
CNN	0.8469 ± 0.0851	0.8111 ± 0.1124	0.9438 ± 0.1108	0.7500 ± 0.1902	0.9906 ± 0.0197
Transformer	0.8812 ± 0.0856	0.9357 ± 0.0872	0.8250 ± 0.1467	0.9375 ± 0.0883	0.9421 ± 0.0642
Multimodal Deep Forest	**0.9563 ± 0.0489** *	**0.9665 ± 0.0589***	**0.9448 ± 0.0716***	**0.9566 ± 0.0494***	**0.9924 ± 0.0149***
Improvement	**1.67% ~ 43.6%**	**0% ~ 21.3%**	**0.1% ~ 189%**	**2% ~ 27.4%**	**0.1% ~ 16.1%**

Table 4. Comparison of Model Performance on Other Pancreatic Disease Dataset. A paired *t* test is performed and * indicates a statistical significance $p < .001$ as compared to the best baseline method; ± represents the mean and variance. Significant values are in [bold].

Our research also discovered that a decrease in the “HB” value is linked to an increased possibility of predicting malignant IPMN. Currently, no studies have reported on the role of “HB” in differentiating the benign and malignant nature of IPMNs. Furthermore, serum indicators such as "Γ-GT", “ALT”, and “AST” have shown value in predicting the type and nature of PCNs, although they have not been extensively documented in the literature.

Jaundice, while considered an indicative sign of malignancy in established guidelines, was found to be of relatively minor importance in differentiating among the three categories of PCNs and in discerning the malignancy of IPMNs in this study. This could be due to the increased availability of CT scans and heightened health awareness, leading to earlier detection and management of PCNs before jaundice becomes evident.

These findings align with clinical guidelines (e.g., Fukuoka criteria[12]) and provide actionable insights for risk stratification.

Generalizability to pancreatic cancer

To further evaluate the generalizability of the proposed MDF model, we collected clinical and imaging data from an additional cohort of 40 patients diagnosed with mass-forming autoimmune pancreatitis (benign) and pancreatic cancer (malignant). This cohort consisted of 65 males and 14 females, with ages ranging from 46 to 84 years. The detailed data are listed below.

Demographic Information: Gender, age;

Symptoms and Medical History: Abdominal pain, jaundice, weight loss, history of diabetes, alcohol consumption, smoking;

Physical Examination and Physiological Indicators: Height, weight, BMI;

Laboratory Tests: ALT (U/L), AST (U/L), ALP (U/L), GGT (U/L), TBIL (μ mol/L), DBIL (μ mol/L), albumin (g/L), globulin (g/L), fasting blood glucose (mmol/L), CEA (ng/mL), CA199 (U/mL), CA125 (U/mL), ferritin (ng/mL), WBC (10E9/L), Hgb (g/L), PLT (10E9/L), INR, fibrinogen (g/L), D-dimer (μ mol/L), IgG4 (g/L).

Imaging Structural Descriptions: Gallbladder wall thickening, bile duct wall thickening, abdominal lymph node enlargement, collateral circulation (e.g., portal vein branches), thrombosis formation;

Vessel Involvement: Involvement of the hepatic artery, abdominal aorta, splenic artery, splenic vein, superior mesenteric vein (SMV), superior mesenteric artery (SMA), portal vein;

Lesion Characteristics (CT/MRI): Location, size (mm), boundary (clear/unclear), lesion plain CT value, arterial phase CT value, portal venous phase CT value, corresponding control values, and differences between lesion and control values;

Pancreaticobiliary System: Type of pancreatic duct dilatation (no/uniform/segmental), pancreatic duct diameter (mm), common bile duct diameter (mm), pancreatic morphology (normal/atrophic/enlarged).

Table 4 compares the performance of the MDF algorithm with other popular AI models on this dataset. The Multimodal Deep Forest (MDF) model achieves state-of-the-art performance across all metrics, outperforming both classical and deep learning models. With the highest accuracy (95.63% ± 4.89%), precision (96.65% ± 5.89%), sensitivity (94.48% ± 7.16%), specificity (95.66% ± 4.94%), and ROC-AUC (99.24% ± 1.49%), MDF demonstrates statistically significant improvements ($p < 0.001$) over all baselines. Notably, it surpasses top competitors like Random Forest (accuracy: + 1.6%), CatBoost (sensitivity: + 9.9%), and CNN (ROC-AUC: + 0.18%), while addressing critical limitations of other models—e.g., SVM’s poor sensitivity (improvement: + 189%) and DNN’s unstable specificity (improvement: + 14.5%). MDF’s robust performance (low standard deviations) and balanced metric dominance highlight its ability to integrate multimodal data effectively, mitigate overfitting, and generalize across diverse clinical scenarios, solidifying its role as a highly reliable, interpretable, and clinically actionable diagnostic tool.

Clinical discussion

Clinical guidelines

As far as its contribution to clinical decision-making, we list them for each category in details by analyzing both SHAP Summary plot and Beeswarm plot as follows:

Benign IPMN: several factors enhance its predictability as a benign condition. Specifically, patients with a history of pancreatitis, those whose lesions communicate more extensively with the pancreatic duct, older individuals, and those exhibiting smaller CA125 values are more likely to be classified as having benign IPMN. Additionally, such predictions are stronger when the tumors are located towards the head or tail of the pancreas, present a smaller size, and show elevated levels of HB and CEA.

Malignant IPMN: higher CA19.9 values, wider main pancreatic ducts, increased CA125 levels, a history of pancreatitis, and advanced age are significant predictors for malignancy. These factors collectively contribute to a more accurate identification of malignant IPMN.

MCN: the clinical application highlights that patients with pancreatitis whose lesions exhibit greater communication with the pancreatic duct are more likely to be diagnosed with MCN. This factor is crucial in distinguishing MCN from other pancreatic cystic neoplasms.

SCN: the prediction leans towards younger patients with a history of pancreatitis and wider main pancreatic ducts. Furthermore, SCN is more likely when there is increased communication between the lesion and the pancreatic duct, coupled with lower CA19.9 values. These characteristics help in differentiating SCN from other types of pancreatic cystic lesions.

Clinical deployment

The MDF model is suitable for real-time clinical deployment with limited resources such as edge-devices for following aspects:

Real-Time Prediction: MDF's fast inference time makes it well-suited for real-time clinical deployment. In clinical settings, where quick and accurate predictions are crucial (e.g., for diagnostic support or treatment planning), MDF can provide timely results without significant delays. The ability to parallelize inference further enhances its suitability for real-time applications. Multiple predictions can be made simultaneously, which is beneficial in scenarios where multiple patients need to be assessed concurrently.

Resource Efficiency: The relatively low computational requirements during inference mean that MDF can be deployed on a wide range of hardware, including edge devices and low-power servers. This flexibility is important in clinical environments where resources may be limited. The model's efficiency also reduces the need for high-end GPUs or specialized hardware, making it more accessible and cost-effective for clinical deployment.

Scalability: MDF can handle varying loads efficiently. In a clinical setting, the number of patients requiring predictions can fluctuate, and MDF's ability to scale horizontally (by adding more processing units) ensures that it can maintain performance under different loads.

Robustness and Reliability: The model's robustness to noise and missing data is another advantage in clinical deployment. Medical data often contain missing values or noisy measurements, and MDF's ability to handle such data without significant performance degradation is beneficial in real-world clinical scenarios.

The computational requirements of MDF make it a suitable candidate for real-time clinical deployment. Its fast inference time, low memory footprint during prediction, and ability to run on a variety of hardware platforms ensure that it can provide timely and accurate predictions in clinical settings. Additionally, its robustness and scalability further enhance its practical relevance for real-time applications.

Limitation

The process of manually processing image information is not automated but relies on radiologist intervention and operation. This process typically involves professionals carefully reviewing and analyzing images to ensure the accuracy and completeness of the information. Since subjective judgment and interpretation may introduce errors, this process may not be as precise and error-free as an automated system. However, it provides a depth of insight and detailed analysis that artificial intelligence cannot currently replace.

Meanwhile, our study is further constrained by the retrospective and single-center nature of the data. Selection and information biases may limit the generalizability of findings to populations with diverse demographics or healthcare protocols. To address this, future work must prioritize multi-center collaborations and external validation across heterogeneous cohorts, ensuring broader applicability. Also, integrating the model into clinical workflows requires addressing data privacy, bias, and transparency while ensuring rigorous validation, seamless integration, and continuous monitoring.

Conclusion

In summary, the study presents a significant advancement in the non-invasive detection of Pancreatic Cystic Neoplasms (PCNs) through the multimodal machine learning model by the integration of comprehensive imaging features and detailed clinical data. The proposed Multimodal Deep Forest (MDF) model consistently outperforms competing approaches, achieving the highest accuracy (95.63%), precision (96.65%), sensitivity (94.48%), specificity (95.66%), and ROC-AUC (99.24%) with minimal variability, demonstrating statistically significant improvements over classical and deep learning models. Its balanced, robust performance across metrics, coupled with superior generalization and interpretability, ensures reliable integration of multimodal data while mitigating overfitting, making it a clinically actionable tool for precise, transparent diagnostics in diverse real-world scenarios.

Meanwhile, its diagnostic precision not only exceeds CT scan-based accuracy as outlined in European guidelines but also approaches MR imaging standards. A key highlight of this research is the model's exceptional performance in predicting MCN with a perfect accuracy rate of 100%, irrespective of whether MRI or CT imaging data was used. This underscores the model's robustness and its potential to serve as a reliable early screening tool for MCN, offering substantial clinical value.

Data availability

Considering that the data still requires de-identification processing to prevent privacy leaks and other issues, a download link is not currently available. If needed, you can request it by contacting the first author (Wei Huang with 1520159@zju.edu.cn or Yue Xu with xuyue1996@zju.edu.cn). Once the data is processed, we will make it public.

Received: 29 October 2024; Accepted: 6 May 2025
Published online: 12 May 2025

References

1. van Huijgevoort, N. C. M., del Chiaro, M., Wolfgang, C. L., van Hooft, J. E. & Besselink, M. G. Diagnosis and management of pancreatic cystic neoplasms: current evidence and guidelines. *Nat. Rev. Gastroenterol Hepatol.* **16**(11), 676–689. https://doi.org/10.1038/s41575-019-0195-x (2019).
2. Del Chiaro, M. et al. European evidence-based guidelines on pancreatic cystic neoplasms. *Gut* **67**(5), 789–804. https://doi.org/10.1136/gutjnl-2018-316027 (2018).
3. Yoon, W. J. & Brugge, W. R. Pancreatic cystic neoplasms: Diagnosis and management. *Gastroenterol. Clin. N. Am.* **41**(1), 103–118. https://doi.org/10.1016/j.gtc.2011.12.016 (2012).
4. Brugge, W. R. L. G., Sahani, D., Fernandez-del, C. C. & Warshaw, A. L. Cystic neoplasms of the pancreas. *N. Engl. J. Med.* **351**(16), 1218–1226. https://doi.org/10.1056/NEJMra031623 (2004).
5. Tanaka, M. et al. International consensus guidelines 2012 for the management of IPMN and MCN of the pancreas. *Pancreatology* **12**(3), 183–197. https://doi.org/10.1016/j.pan.2012.04.004 (2012).
6. Chandwani, R. & Allen, P. J. Cystic neoplasms of the pancreas. *Annu. Rev. Med.* **67**, 45–57. https://doi.org/10.1146/annurev-med-051914-022011 (2016).
7. Patra, K. C., Bardeesy, N. & Mizukami, Y. Diversity of precursor lesions for pancreatic cancer: The genetics and biology of intraductal papillary mucinous neoplasm. *Clin. Transl. Gastroenterol.* **8**(4), e86. https://doi.org/10.1038/ctg.2017.3 (2017).
8. Howlader, N. SEER cancer statistics review, 1975–2011 2011 [Available from: http://seer.cancer.gov/csr/1975_2011/.]
9. Lekkerkerker, S. J. et al. Comparing 3 guidelines on the management of surgically removed pancreatic cysts with regard to pathological outcome. *Gastrointest. Endosc.* **85**(5), 1025–1031. https://doi.org/10.1016/j.gie.2016.09.027 (2017).
10. Singhi, A. D. et al. American Gastroenterological Association guidelines are inaccurate in detecting pancreatic cysts with advanced neoplasia: A clinicopathologic study of 225 patients with supporting molecular data. *Gastrointest. Endosc.* **83**(6), 1107–1117. https://doi.org/10.1016/j.gie.2015.12.009 (2016).
11. Salvia, R. et al. Pancreatic resections for cystic neoplasms: From the surgeon's presumption to the pathologist's reality. *Surgery.* **152**(3 Suppl 1), S135–S142. https://doi.org/10.1016/j.surg.2012.05.019 (2012).
12. Tanaka, M. et al. Revisions of international consensus Fukuoka guidelines for the management of IPMN of the pancreas. *Pancreatology* **17**(5), 738–753. https://doi.org/10.1016/j.pan.2017.07.007 (2017).
13. Dmitriev, K. et al. Classification of pancreatic cysts in computed tomography images using a random forest and convolutional neural network ensemble. *Med. Image Comput. Comput. Assist. Interv.* **10435**, 150–158. https://doi.org/10.1007/978-3-319-66179-7_18 (2017).
14. Dong, Z. et al. Differential diagnosis of pancreatic cystic neoplasms through a radiomics-assisted system. *Front. Oncol.* **12**, 941744. https://doi.org/10.3389/fonc.2022.941744 (2022).
15. Corral, J. E. et al. Deep learning to classify intraductal papillary mucinous neoplasms using magnetic resonance imaging. *Pancreas* **48**(6), 805–810. https://doi.org/10.1097/MPA.0000000000001327 (2019).
16. 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**(6), 2135–2147. https://doi.org/10.1007/s00261-022-03479-4 (2022).
17. Carr, R. A. et al. Smoking and IPMN malignant progression. *Am. J. Surg.* **213**(3), 494–497. https://doi.org/10.1016/j.amjsurg.2016.10.033 (2017).
18. Pergolini, I. et al. Diabetes and weight loss are associated with malignancies in patients with intraductal papillary mucinous neoplasms. *Clin. Gastroenterol. Hepatol.* **19**(1), 171–179. https://doi.org/10.1016/j.cgh.2020.04.090 (2021).
19 Zhou, Z. H. & Feng, J. Deep forest. *Natl. Sci. Rev.* **6**(1), 74–86. https://doi.org/10.1093/nsr/nwy108 (2019).
20. Yasukawa, S. & Yanagisawa, A. Pathological classification of tumors of the pancreas. *Nihon Rinsho* **73**(Suppl 3), 42–44 (2015).
21. Pedregosa, F. et al. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
22. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions, 31st Annual Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, DEC 04–09 (2017).
23. Crippa, S. et al. Mucin-producing neoplasms of the pancreas: An analysis of distinguishing clinical and epidemiologic characteristics. *Clin. Gastroenterol. Hepatol.* **8**(2), 213–219. https://doi.org/10.1016/j.cgh.2009.10.001 (2010).
24. Nagashio, Y. et al. Combination of cyst fluid CEA and CA 125 is an accurate diagnostic tool for differentiating mucinous cystic neoplasms from intraductal papillary mucinous neoplasms. *Pancreatology* **14**(6), 503–509. https://doi.org/10.1016/j.pan.2014.09.011 (2014).
25. Qian, Y. et al. Carbohydrate antigen 125 supplements carbohydrate antigen 19–9 for the prediction of invasive intraductal papillary mucinous neoplasms of the pancreas. *World J. Surg. Oncol.* **20**(1), 310. https://doi.org/10.1186/s12957-022-02720-0 (2022).
26. Ramirez-Carmona, W. et al. Are serum ferritin levels a reliable cancer biomarker? A systematic review and meta-analysis. *Nutr. Cancer* **74**(6), 1917–1926. https://doi.org/10.1080/01635581.2021.1982996 (2022).
27. Kerlakian, S. et al. Cyst location and presence of high grade dysplasia or invasive cancer in intraductal papillary mucinous neoplasms of the pancreas: a seven institution study from the central pancreas consortium. *HPB* **21**(4), 482–488. https://doi.org/10.1016/j.hpb.2018.09.018 (2019).
28. 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**(4), 464–479. https://doi.org/10.1038/ajg.2018.14 (2018).
29. Roch, A. M. et al. The natural history of main duct-involved, mixed-type intraductal papillary mucinous neoplasm: Parameters predictive of progression. *Ann. Surg.* **260**(4), 680–688. https://doi.org/10.1097/SLA.0000000000000927 (2014).
30. Zhou, Z. & Feng, J. Deep forest: Towards an alternative to deep neural networks, IJCAI'17. https://doi.org/10.24963/ijcai.2017/497.
31. Feng, J. & Zhou, Z. H. Autoencoder by forest. *Proc. AAAI Conf. Artif. Intell.* https://doi.org/10.1609/aaai.v32i1.11732 (2018).
32. Rehman, A., Mahmood, T. & Saba, T. Robust kidney carcinoma prognosis and characterization using Swin-ViT and DeepLabV3+ with multi-model transfer learning. *Appl. Soft Comput.* **170**, 112518. https://doi.org/10.1016/j.asoc.2024.112518 (2025).
33. Mahmood, T. et al. Harnessing the power of radiomics and deep learning for improved breast cancer diagnosis with multiparametric breast mammography. *Expert Syst. Appl.* **249**, 123747. https://doi.org/10.1016/j.eswa.2024.123747 (2024).

Author contributions

W. H. finished clinical data collection,diagnosed PCN according to imaging, and manuscript writing; Y. X. finished clinical data collection and manuscript writing; Z. L. organized the technical system of this paper, which encompasses data analysis, algorithm model design, and experimental comparison; J. L. and Q. C. finished evaluating the type of PCN; Q. H. and Y. W. finished diagnosed PCN according to imaging; H. C. finished conception study and manually diagnosed PCN according to imaging.

Declarations

Competing interests

The authors declare no competing interests.

Ethics approval and consent to participate

This study was conducted in accordance with relevant guidelines and regulations and received approval from The First Affiliated Hospital, Zhejiang University School of Medicine [Approval Number: IIT20230145A]. Informed consent was obtained from all subjects and/or their legal guardians prior to participation.

Additional information

Correspondence and requests for materials should be addressed to Z.L. or H.C.

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

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