Abdominal Radiology
https://doi.org/10.1007/s00261-025-05260-9

RESEARCH

Beyond gas bubbles: AI analysis of the "bubble bed" microenvironment improves diagnosis of infected abdominal collections

Yifei Guo[1] · Chengwei Chen[1] · Tiegong Wang[1] · Yixuan Shen[1] · Danqun Zheng[1] · Yilun Zheng[1] · Jieyu Yu[1] · Jing Li[1] · Xu Fang[1] · Fang Liu[1] · Ming Yang[1] · Li Wang[1] · Jianping Lu[1] · Chengwei Shao[1] · Yun Bian[1]

Received: 5 August 2025 / Revised: 29 September 2025 / Accepted: 7 October 2025

Abstract

Background Accurate diagnosis of infected intra-abdominal fluid collections (IAFCs) is challenging, as the conventional "gas bubble sign" on computed tomography (CT) has poor sensitivity. This study aimed to develop and validate a fully automated artificial intelligence (AI) model using non-contrast CT to improve diagnostic accuracy.

Methods In this multicenter retrospective study (July 2011-July 2024), 797 patients with IAFCs confirmed by culture were divided into training ($n=637$), validation ($n=80$), and external test ($n=80$) sets. We developed an AI model, Bubble Bed Based Learning Engine for Abdominal Infection (BUBBLE-AI), based on the novel "bubble bed" concept, which analyzes the inflammatory microenvironment around gas bubbles. The model integrates deep learning and radiomic features, extracted from automated segmentations, with clinical data.

Results The BUBBLE-AI model demonstrated robust and generalizable performance, achieving an area under the curve (AUC) of 0.92 in validation and **0.82** (95% CI: 0.72–0.93) in external testing, significantly outperforming traditional methods ($P<0.001$). The model achieved a total misdiagnosis rate of 24.1%, a substantial improvement over the bubble sign (38.5%) and a clinical-only model (34.3%). Feature analysis confirmed the "bubble bed" microenvironment was the most dominant source of predictive information (35.8% of features), validating our core hypothesis.

Conclusion The BUBBLE-AI provides a fully automated and accurate tool for diagnosing IAFC infections from non-contrast CT. By identifying infection status more reliably, it can guide antimicrobial stewardship, reduce diagnostic errors, and optimize clinical decision-making.

Keywords Artificial intelligence · Computed tomography · Intra-abdominal infection · Machine learning · Radiomics

Abbreviations

AI	Artificial intelligence
AUC	Area under the curve
BMI	Body mass index
CI	Confidence interval
CNN	Convolutional neural network
CRP	C-reactive protein
CSS	Channel-spatial siamese
CT	Computed tomography
DCA	Decision curve analysis
DLR-Score	Deep learning radiomics score
DSC	Dice similarity coefficient
GLDM	Gray level dependence matrix
GLCM	Gray level co-occurrence matrix
GLRLM	Gray level run length matrix
GLSZM	Gray level size zone matrix
IAFC	Intra-abdominal fluid collection
IQR	Interquartile range
LASSO	Least absolute shrinkage and selection operator

Yifei Guo, Chengwei Chen, and Tiegong Wang contributed equally to this work.

Extended author information available on the last page of the article

Published online: 26 December 2025

LDH	Lactate dehydrogenase
MIC	Mamba-in-convolution
NGTDM	Neighboring gray tone difference matrix
NPV	Negative predictive value
OR	Odds ratio
PPV	Positive predictive value
ROC	Receiver operating characteristic
ROI	Region of interest
SAP	Severe acute pancreatitis
SSM	State space model
WBC	White blood cell

Background

Intra-abdominal fluid collections (IAFCs) represent a spectrum of pathological conditions arising from diverse etiologies including acute pancreatitis, postoperative complications, gastrointestinal perforation, and abdominal trauma [1–4]. The clinical significance of these collections is fundamentally determined by their infection status, which directly influences therapeutic strategies, patient outcomes, and healthcare resource utilization. Infected IAFCs are associated with substantial morbidity and mortality, with sepsis rates reaching 30–70% in severe acute pancreatitis and mortality rates approaching 15–20% in complicated cases [3, 5, 6].

The current clinical approach to IAFC management faces significant diagnostic challenges that contribute to suboptimal outcomes. Traditional imaging assessment relies primarily on the morphological "gas bubble sign" on computed tomography (CT), which demonstrates limited diagnostic utility with sensitivity ranging from only 18% to 56% [7]. This poor sensitivity results in delayed diagnosis of infected cases, potentially leading to progression to sepsis and multiple organ failure. Furthermore, the specificity of gas bubbles is compromised by post-interventional artifacts, where iatrogenic gas introduction during drainage procedures creates false-positive findings that confound clinical interpretation [8, 9].

The diagnostic uncertainty surrounding IAFC infection status has led to widespread empirical antibiotic use in clinical practice. Multiple large-scale randomized controlled trials have demonstrated that prophylactic broad-spectrum antibiotics do not significantly reduce infection rates in necrotizing pancreatitis (relative risk 0.81, 95% CI 0.54–1.22) or improve mortality outcomes (relative risk 0.70, 95% CI 0.42–1.17) [10–13]. Despite this evidence, antibiotic overuse remains prevalent, contributing to antimicrobial resistance, healthcare costs, and potential adverse effects [14, 15]. This creates an urgent need for precise diagnostic tools to break the current cycle of empirical treatment and over-medicalization.

Recent advances in artificial intelligence and medical imaging have opened new avenues for addressing these diagnostic challenges. However, existing AI research in this field has been limited by significant methodological constraints. Previous studies have predominantly relied on single-center designs, time-consuming manual region-of-interest delineation, and lack of rigorous external validation, factors that severely limit their clinical applicability and generalizability [16, 17]. Unlike prior radiomics studies requiring manual ROI delineation [18, 19], our approach enables fully automated, real-time analysis, addressing key barriers to clinical adoption.

This study introduces several methodological innovations to overcome these limitations. We developed a comprehensive AI-based diagnostic framework that integrates fully automated image segmentation with multi-dimensional feature analysis. Central to our approach is the introduction of the novel "bubble bed" concept, inspired by the tumor bed concept in oncology, which captures infection-related microenvironmental changes extending beyond discrete gas bubbles. We hypothesized that infected collections create a localized inflammatory microenvironment that alters surrounding tissue density and texture patterns, information that can be captured through advanced image analysis techniques.

The primary objectives of this study were: (1) to develop and validate a fully automated segmentation system for precise delineation of IAFC-related anatomical structures; (2) to construct an integrated prediction model combining imaging and clinical features for infection diagnosis; (3) to compare the diagnostic performance of the AI model against conventional methods including the "gas bubble sign" and clinical assessment; and (4) to evaluate model generalizability across different disease etiologies and external patient populations. The ultimate goal is to provide clinicians with an objective, accurate, and clinically applicable tool to guide therapeutic decision-making and optimize antimicrobial stewardship in patients with intra-abdominal fluid collections.

Methods

Study design and population

This multicenter retrospective cohort study was conducted in accordance with the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) guidelines [20]. The study protocol was approved by the Institutional Review Boards of participating

institutions (Ethics Approval No.: CHEC-Y2024-018), and the requirement for informed consent was waived due to the retrospective nature of the study. The research design flow is illustrated in Fig. 1 and Supplementary Fig. 1.

The study population comprised patients who underwent CT-guided percutaneous drainage of intra-abdominal fluid collections with subsequent microbiological culture analysis between July 2011 and July 2024. Patients were recruited from two tertiary care centers: Hospital 1 ($n=734$) and Hospital 2 ($n=87$). The primary inclusion criteria were: (1) availability of pre-intervention non-contrast CT imaging; (2) CT-guided percutaneous drainage procedure performed; and (3) definitive microbiological culture results available. Exclusion criteria included: (1) absence of pre-intervention CT images; (2) inadequate CT image quality for analysis; (3) incomplete or indeterminate microbiological culture results; and (4) previous drainage procedures within 30 days. After applying the inclusion and exclusion criteria, 797 patients were included in the final study.

Sample size calculation and statistical power

Sample size estimation was performed using PASS 2021 software (NCSS Statistical Software, Kaysville, UT) with the "Test for One ROC Curve" module. A minimum sample size of 664 patients was calculated to provide 90% power at an alpha of 0.05 to distinguish a model with an AUC of 0.85 from a null hypothesis AUC of 0.75, assuming a 75% infection prevalence. The final study cohort of 797 patients exceeded this requirement, ensuring adequate statistical power. A post-hoc analysis confirmed that the study maintained a power of > 0.85 for the achieved external validation AUC of 0.82.

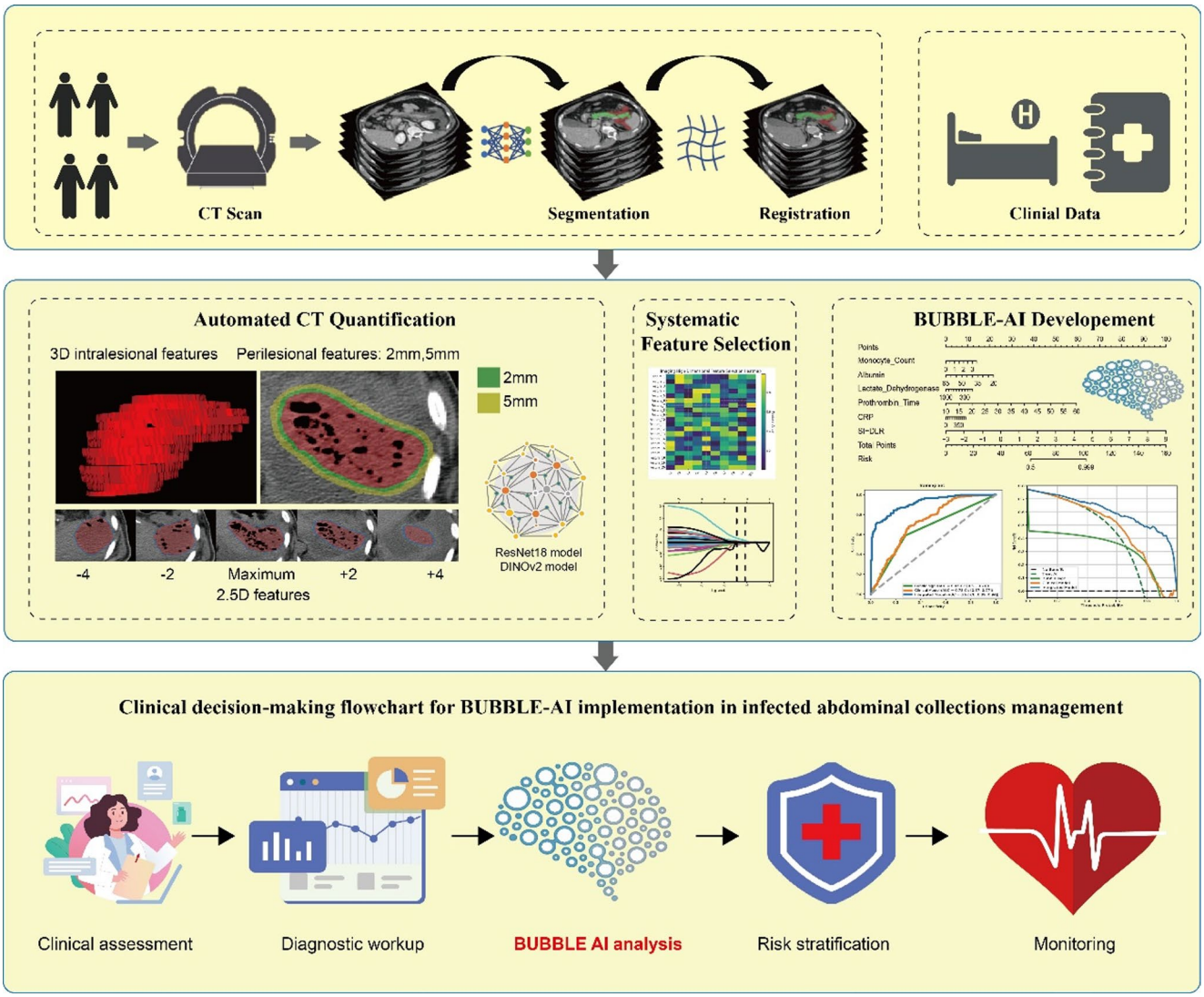


Fig. 1 Study workflow and methodology. Flowchart illustrating the systematic approach for data collection, automated imaging feature extraction, statistical feature selection, dimensionality reduction, and predictive model construction across training, validation, and test cohorts

Dataset partitioning and class balance considerations

Patients from one center were randomly divided into training ($n=637$) and internal validation ($n=80$) sets using stratified sampling to maintain consistent infection rates. Patients from the second center served as an independent external test set ($n=80$) to assess model generalizability. To address the 76.4% infection rate imbalance, sample weighting, inversely proportional to class frequencies, was applied during model training to prevent bias.

CT imaging protocol

CT examinations were performed using standardized protocols across participating centers to ensure image quality consistency. Importantly, the pre-intervention non-contrast CT scan was performed as an integral component of the CT-guided drainage procedure, with imaging acquisition and subsequent needle positioning/sample collection completed as consecutive steps on the same gantry within 30 min. This protocol ensures perfect temporal matching between imaging characteristics and microbiological sampling, eliminating potential confounding from disease progression or interim interventions. Images were acquired on 320-detector row wide-volume CT systems (Aquilion ONE, Toshiba Medical Systems, Japan) and 128-slice spiral CT scanners (Brilliance iCT, Philips Healthcare, Netherlands). The deliberate inclusion of different scanner manufacturers enhances real-world generalizability per FDA guidance on AI validation [21]. Standardized acquisition parameters included: tube voltage 120 kV, tube current 150 *mAs*, detector collimation 100×0.5 mm, reconstruction matrix 350×350, gantry rotation time 0.5 s, slice thickness 0.8 mm, and slice interval 1.0 mm. The scanning range was adjusted to ensure complete coverage of the target lesions and surrounding anatomy.

Reference Standard and Contamination Prevention

The reference standard for infection was established by microbiological culture of fluid obtained via CT-guided percutaneous drainage. performed immediately following the non-contrast CT acquisition (typically within 30 min). This integrated imaging-sampling protocol ensures that the reference standard reflects the infection status at the exact time point of image acquisition. All procedures were performed by experienced interventional radiologists using standardized, strict aseptic techniques to minimize contamination risk. Drainage fluid samples were promptly transported for aerobic, anaerobic, and, when clinically indicated, fungal culture. Infection was defined as any positive growth of pathogenic microorganisms, while cultures were considered sterile if no bacterial or fungal growth was observed after 72 h and 4 weeks, respectively.

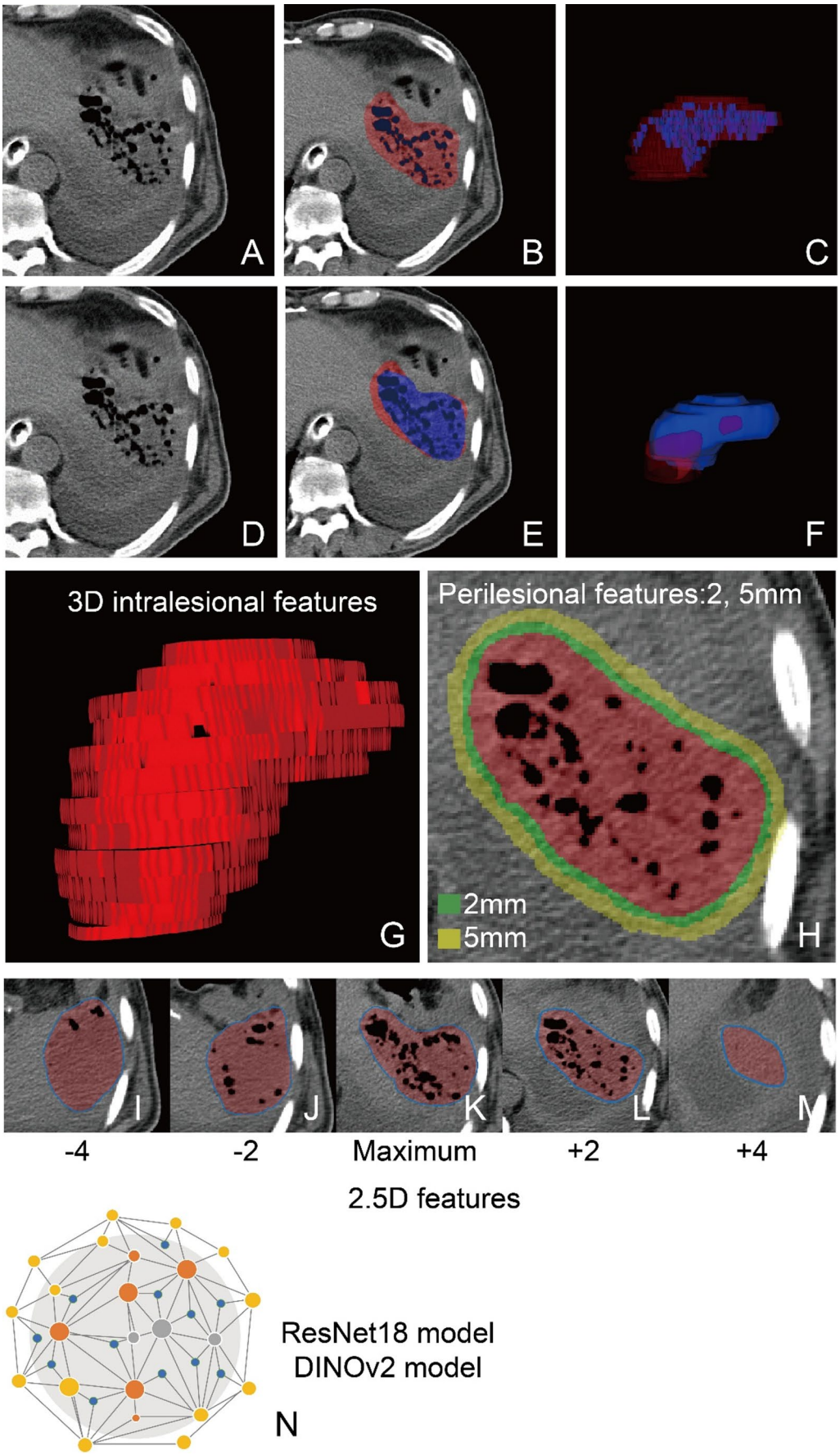
Fig. 2 Schematic Illustration of the Multi-modal Feature Extraction Framework. This figure illustrates the core concepts of lesion segmentation and the different methodologies used for comprehensive feature extraction. **A** A representative axial non-contrast CT image of an intra-abdominal fluid collection with internal gas bubbles. **B**, **C** The standard segmentation of the entire lesion (red overlay in B) and its corresponding 3D rendering (**C**). **D–F** In contrast, the novel segmentation of the "bubble bed" region (blue overlay in E) is shown, which encompasses the gas bubbles and the surrounding hypothesized inflammatory microenvironment, along with its 3D rendering (**F**). The subsequent panels illustrate the four complementary types of features extracted from these regions: **G** Three-dimensional (3D) intralesional radiomic features calculated from the entire lesion volume. **(H)** Perilesional features extracted from concentric 2 mm (green) and 5 mm (yellow) margins around the lesion. **I–M** The 2.5D feature extraction method, which analyzes a stack of five consecutive axial CT slices to capture local spatial context. **N** A conceptual representation of deep learning feature extraction using advanced pre-trained models (e.g., ResNet18, DINOv2)

Clinical data collection

Clinical and laboratory data were systematically extracted from electronic medical records by trained research personnel. A total of 27 clinical variables were collected, including demographic characteristics (age, sex, body mass index), vital signs (pulse rate, body temperature), hematological parameters (white blood cell count, platelet count, red blood cell count, differential counts), biochemical markers (blood glucose, albumin, lactate dehydrogenase, alkaline phosphatase), inflammatory markers (C-reactive protein, procalcitonin), and coagulation parameters (prothrombin time, fibrinogen, thrombin time). All laboratory values represented the most recent results obtained within 24 h prior to the CT examination.

Automated image segmentation and the "bubble bed" concept

To enable a fully automated workflow, we developed a precise segmentation pipeline centered on a novel region of interest (ROI): the "bubble bed" (Fig. 2). This concept is grounded in the hypothesis that gas-producing bacteria create a localized inflammatory microenvironment with subtle tissue alterations that extend beyond the discrete gas bubbles themselves. To segment this and all other key ROIs (fluid, gas, pancreas, drainage tube), we employed the nnMamba architecture, a hybrid CNN-SSM framework designed to capture complex, long-range spatial dependencies in 3D medical images (Supplementary Method 1). The peri-lesional bands (2-mm and 5-mm) were automatically generated by morphological dilation of the segmented


A
B
C
D
E
F
3D intralesional features
G
Perilesional features:2, 5mm
2mm
5mm
H
I
J
K
L
M
-4
-2
Maximum
+2
+4
2.5D features
ResNet18 model
DINOv2 model
N

collection boundaries. Importantly, these bands intentionally encompass heterogeneous adjacent tissues (parenchymal organs, adipose tissue, fascial planes) as our hypothesis posits that infectious inflammation creates detectable microenvironmental changes across all tissue types within the inflammatory sphere of influence. Rather than viewing tissue heterogeneity as contamination, our AI model leverages this as the primary biological signal, learning to distinguish inflammation-affected tissue patterns from normal tissue patterns across different anatomical contexts.

Comprehensive feature extraction framework

Following automated segmentation, a multi-modal feature extraction framework was implemented to comprehensively characterize the imaging phenotype. This involved integrating four complementary feature types: quantitative morphological features, 3D radiomics (PyRadiomics), perilesional features from 2 mm to 5 mm margins, and deep learning features from both local (ResNet18) and global (DINOv2) models. This exhaustive process yielded 47,647 candidate features per patient for model development (Supplementary Method 2).

Feature selection and model development

This high-dimensional feature set was first preprocessed, then distilled using a three-stage selection pipeline (univariate screening, correlation analysis, and LASSO regression) to identify the most informative predictors while preventing overfitting. Subsequently, two models were constructed for comparison: a baseline model using only clinical data, and the integrated **BUBBLE-AI** (Bubble Bed Based Learning Engine for Abdominal Infection) model. The BUBBLE-AI combines clinical parameters with a composite imaging biomarker, the Deep Learning Radiomics Score (DLR-Score), which is calculated from the final weighted set of selected imaging features.

Statistical analysis and model evaluation

Statistical comparisons for baseline characteristics were performed using the Student's t-test or Mann-Whitney U test for continuous variables and the chi-square or Fisher's exact test for categorical variables. The predictive models' performance was comprehensively evaluated using the area under the receiver operating characteristic curve (AUC)—with DeLong's test for model comparisons—along with sensitivity, specificity, accuracy, PPV, and NPV. Decision curve analysis was used to assess clinical utility. The underlying segmentation model's accuracy was specifically quantified using the Dice Similarity Coefficient (DSC).

All statistical analyses were conducted using Python (v3.8.20) and R (v4.0.2), with a two-sided P-value < 0.05 considered significant.

Data and code availability

To ensure reproducibility, the model code is publicly available at [https://github.com/CHANGHAI-AILab/AP_Task2]. While the clinical dataset cannot be publicly shared due to patient privacy regulations, it may be available to qualified researchers upon reasonable request and appropriate data use agreements.

The nnMamba segmentation model and prediction model code are available at [GitHub repository: https://github.com/institution/iafc-infection-prediction]. Data are available through corresponding authors. Computer codes are available as online inference codes on GitHub (https://github.com/CHANGHAI-AILab/AP_Task2).

IRB approval statement

This study was approved by the Institutional Review Board (IRB Approval No. CHEC-Y2024-018; Approval Date: January 25, 2024) with a waiver of written informed consent. The ethics committee confirmed the study's compliance with the Declaration of Helsinki and other applicable international ethical standards for human subject research.

Results

Study population characteristics and baseline balance

The final study cohort comprised 797 patients with intra-abdominal fluid collections who met all inclusion criteria, with disease etiologies primarily being acute pancreatitis (52.6%) and postoperative complications (33.1%) (Fig. 3). The overall infection rate was 76.4%. This rate was balanced across the training (77.4%), validation (72.5%), and test (70.0%) sets, confirming the success of stratified sampling and the comparability of the datasets (Table 1).

Within the training set, infected patients presented with a distinct and significant systemic inflammatory profile, characterized by elevated white blood cell counts and C-reactive protein concentrations, alongside lower albumin and lymphocyte percentages (all $P \leq 0.001$). Furthermore, microbiological analysis revealed a diverse and clinically expected spectrum of pathogens, most commonly *Escherichia coli* and *Klebsiella pneumoniae*. The overall microbial composition remained consistent across the different disease

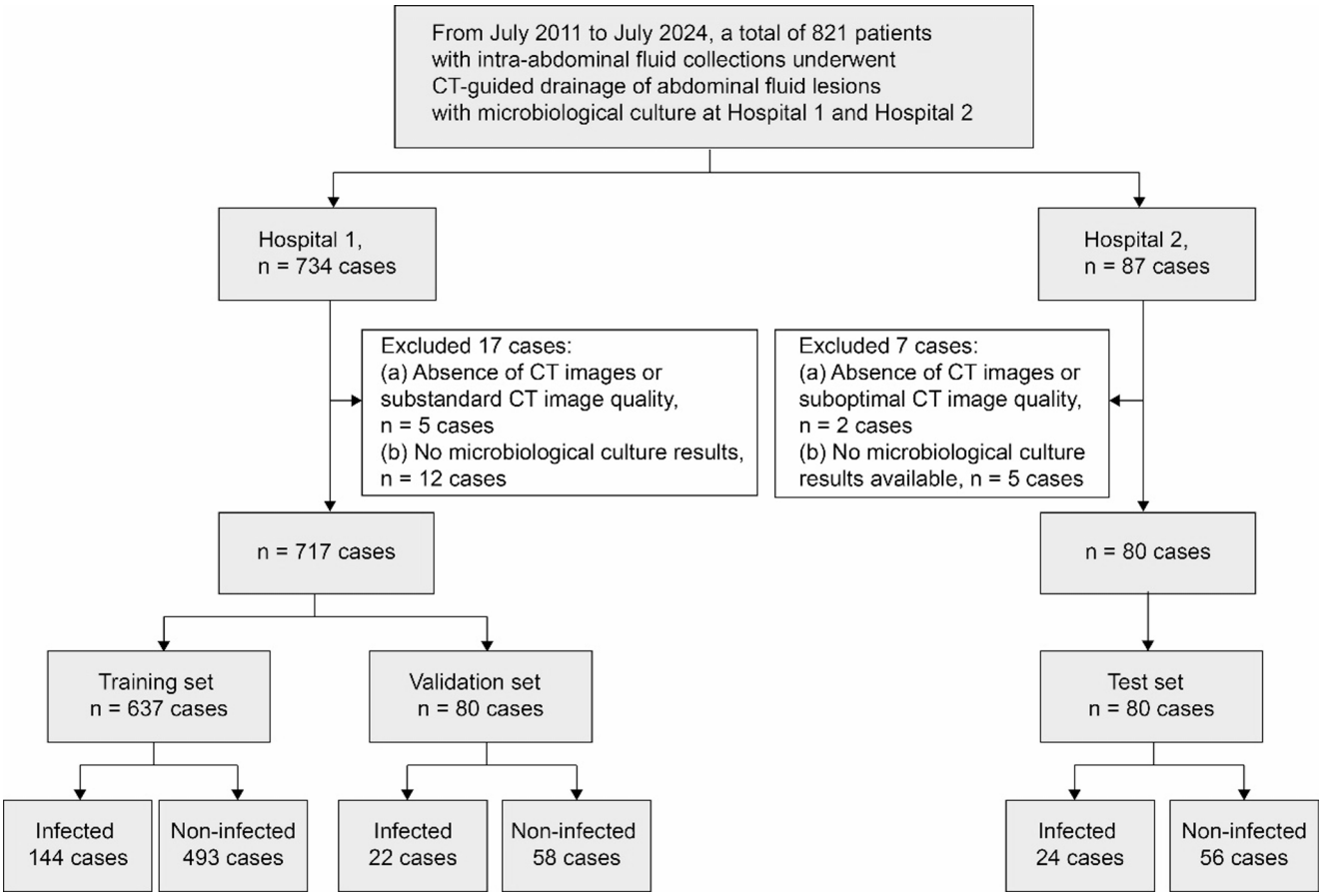


Fig. 3 Patient Enrollment Flowchart. This flowchart details the patient selection process for the multicenter study, including all inclusion and exclusion criteria applied. It illustrates the final partitioning of the cohort into a training set ($n=637$), an internal validation set ($n=80$), and an independent external test set ($n=80$)

etiologies, supporting the generalizability of the infection prediction model **(Supplementary Fig. 2)**.

Automated segmentation performance and technical validation

The technical foundation of our analysis was validated by the excellent performance of the automated nnMamba segmentation model, which showed robust agreement with expert annotations (Supplementary Fig. 3). The model achieved high Dice Similarity Coefficients (DSC) across all key structures, including gas regions (0.95 ± 0.12), drainage tubes (0.88 ± 0.08), pancreatic parenchyma (0.83 ± 0.12), and the more morphologically complex fluid collections (0.76 ± 0.11). This high-fidelity segmentation was crucial, as it enabled the automated extraction of 19 quantitative morphological features that revealed profound differences between infected and non-infected groups (Table 2). Specifically, infected collections exhibited substantially higher bubble volumes (median 537.4 vs. 0.0 mm^3), increased bubble counts (6.0 vs. 0.0), and vastly larger "bubble bed" volumes (15,178.7 vs. 0.0 mm^3) (all $P<0.001$). These findings highlight the superiority of our quantitative approach, moving beyond the traditional binary assessment of gas presence to capture biologically relevant signatures of infection.

Feature selection results and model interpretability

From an initial, high-dimensional set of 47,647 candidate features, a systematic three-stage selection pipeline distilled a final, robust set of 53 optimal predictors. This rigorous process, culminating in a LASSO regression model that balanced predictive power with model parsimony ($\lambda=0.0155$), effectively isolated the most significant imaging biomarkers of infection (Supplementary Fig. 4).

Analysis of this final feature set offered two crucial insights into the biological basis of the model's predictions. First, it validated our core hypothesis regarding the diagnostic importance of the perilesional microenvironment: features derived from the "bubble bed" regions were the most dominant anatomical source of information, contributing

Table 1 Baseline characteristics of patients with peritoneal effusion

Characteristics	Training Set ($n=637$)			Validation Set ($n=80$)			Test Set ($n=80$)		
	Non-infected $n=144$	Infected $n=493$	*P*-value	Non-infected $n=22$	Infected $n=58$	*P*-value	Non-infected $n=24$	Infected $n=56$	*P*-value
Age (years), M (IQR)	51.00 (13.00–86.00)	54.00 (12.00–87.00)	0.022	49.50 (26.00–71.00)	60.00 (14.00–83.00)	0.066	51.00 (30.00–77.00)	51.50 (17.00–78.00)	0.722
Gender [n (%)]			0.444			0.486			0.082
Female	45 (31.25%)	171 (34.69%)		7 (31.82%)	14 (24.14%)		5 (20.83%)	23 (41.07%)	
Male	99 (68.75%)	322 (65.31%)		15 (68.18%)	44 (75.86%)		19 (79.17%)	33 (58.93%)	
Body mass index (kg/m²), M (IQR)	23.47 (13.59–33.80)	22.77 (13.33–36.33)	0.173	23.44 (17.58–29.41)	23.03 (17.22–36.05)	0.745	22.85 (17.30–32.53.30.53)	22.86 (16.41–30.22)	0.808
Pulse rate (beats/min), M (IQR)	87.50 (58.00–134.00.00.00)	90.00 (40.00–152.00.00.00)	0.034	90.00 (74.00–125.00.00.00)	90.00 (70.00–142.00.00.00)	0.601	84.00 (72.00–130.00.00.00)	94.50 (74.00–138.00.00.00)	0.027
Body temperature (°C), M (IQR)	37.00 (36.10–39.80)	37.40 (36.00–40.00)	0.008	37.05 (36.40–38.60)	37.20 (36.00–39.70.00.70)	0.152	37.50 (36.40–39.30)	37.40 (36.00–39.20.00.20)	0.498
White blood cell count (×10⁹/L), M (IQR)	8.84 (2.37–30.62)	10.69 (2.27–46.85)	<0.001	11.73 (3.46–21.63)	10.39 (3.20–39.53.20.53)	0.923	7.79 (3.21–18.79)	10.02 (3.20–39.61.20.61)	0.067
Lymphocyte percentage (%), M (IQR)	12.70 (3.10–47.00)	10.50 (1.20–45.30)	<0.001	10.65 (1.90–32.30)	10.65 (3.40–30.30)	0.149	10.10 (4.00–51.40.00.40)	11.50 (1.80–22.70)	0.092
Neutrophil percentage (%), M (IQR)	76.40 (44.10–92.00)	81.10 (42.70–96.30)	<0.001	79.75 (50.30–97.10)	80.35 (50.50–93.80)	0.207	78.30 (37.50–95.30)	81.90 (63.90–94.50)	0.006
Eosinophil percentage (%), M (IQR)	1.00 (0.10–18.40)	0.60 (0.10–17.50)	<0.001	0.85 (0.10–4.40)	0.70 (0.10–17.50)	0.701	0.75 (0.10–4.60)	0.55 (0.10–4.80)	0.246
Basophil percentage (%), M (IQR)	0.30 (0.10–1.80)	0.20 (0.10–1.10)	0.001	0.25 (0.10–1.50)	0.30 (0.10–1.00.10.00)	0.095	0.25 (0.10–1.10)	0.20 (0.10–1.00.10.00)	0.430
Neutrophil count (×10⁹/L), M (IQR)	6.81 (1.09–26.64)	8.45 (1.32–41.02)	<0.001	9.16 (1.99–17.26)	8.28 (1.67–35.04)	0.696	5.43 (2.05–16.57)	8.19 (2.05–37.39)	0.040
Eosinophil count (×10⁹/L), M (IQR)	0.08 (0.01–1.45)	0.05 (0.01–1.23)	0.001	0.10 (0.01–0.72)	0.06 (0.01–1.95)	0.978	0.06 (0.01–0.31)	0.07 (0.01–0.46)	0.739
Monocyte count (×10⁹/L), M (IQR)	0.64 (0.22–2.12)	0.76 (0.08–3.49)	0.003	0.76 (0.12–3.71)	0.60 (0.09–2.33)	0.166	0.57 (0.21–2.98)	0.58 (0.05–1.48)	0.452
Albumin (g/L), M (IQR)	36.00 (20.20–48.00)	33.00 (20.00–61.00)	<0.001	33.50 (26.00–40.00)	33.50 (21.00–49.00)	0.720	35.50 (29.00–61.00)	34.00 (26.00–44.00)	0.064
Lactate dehydrogenase (U/L), M (IQR)	210.00 (101.00–819.00.00.00)	198.00 (76.00–960.00.00.00)	<0.001	252.00 (130.00–664.00.00.00)	196.50 (84.00–782.00.00.00)	0.197	201.00 (84.00–564.00.00.00)	218.50 (105.00–774.00.00.00)	0.908
Prothrombin time (s), M (IQR)	14.25 (11.50–25.10)	15.00 (11.40–55.10)	0.005	15.00 (12.50–17.90)	15.20 (12.60–25.90)	0.435	14.90 (12.50–25.10)	15.20 (13.00–21.50.00.50)	0.827
C-reactive protein (mg/L), M (IQR)	61.63 (0.50–520.00)	89.10 (2.39–520.00)	0.001	89.35 (5.88–239.00)	105.00 (5.46–274.00)	0.208	73.25 (2.39–424.00)	98.97 (6.99–382.00)	0.198

Table 1 (continued)

Characteristics	Training Set ($n=637$)			Validation Set ($n=80$)			Test Set ($n=80$)		
	Non-infected $n=144$	Infected $n=493$	*P*-value	Non-infected $n=22$	Infected $n=58$	*P*-value	Non-infected $n=24$	Infected $n=56$	*P*-value
Platelet count ($\times10^9$/L), M (IQR)	263.50 (54.00–701.00.00.00)	277.00 (15.00–931.00.00.00)	0.381	273.50 (74.00–1208.00.00.00)	297.00 (90.00–651.00.00.00)	0.995	275.00 (139.00–764.00.00.00)	315.00 (51.00–762.00.00.00)	0.581
Red blood cell count ($\times10^{12}$/L), M (IQR)	3.25 (2.09–5.06)	3.28 (1.67–5.41)	0.693	3.05 (2.18–4.37)	3.27 (2.16–5.45)	0.286	3.21 (2.63–4.92)	3.15 (1.98–4.70)	0.081
Monocyte percentage (%), M (IQR)	7.00 (2.60–29.30)	7.20 (0.70–22.20)	0.816	8.00 (0.90–17.20)	6.55 (0.80–14.70)	0.013	7.25 (2.10–22.80)	6.00 (0.30–12.30)	0.016
Blood glucose (mmol/L), M (IQR)	6.40 (2.30–19.90)	6.90 (1.70–28.20)	0.071	6.60 (4.60–14.50)	6.40 (3.50–16.10)	0.872	7.15 (4.40–18.80)	7.65 (2.30–20.30)	0.399
Lymphocyte count ($\times10^9$/L), M (IQR)	1.12 (0.41–3.95)	1.06 (0.17–7.49)	0.622	1.12 (0.26–4.09)	1.17 (0.30–2.74)	0.420	0.96 (0.37–4.35)	1.02 (0.30–2.44)	0.971
Basophil count ($\times10^9$/L), M (IQR)	0.02 (0.01–0.12)	0.02 (0.01–0.32)	0.760	0.03 (0.01–0.21)	0.03 (0.01–0.11)	0.060	0.02 (0.01–0.06)	0.03 (0.01–0.10)	0.125
Alkaline phosphatase (U/L), M (IQR)	95.00 (42.00–3662.00.00.00)	111.00 (32.00–2361.00.00.00)	0.701	136.00 (64.00–2352.00.00.00)	103.50 (37.00–563.00.00.00)	0.152	84.00 (49.00–395.00.00.00)	104.50 (49.00–400.00.00.00)	0.568
Procalcitonin (ng/mL), M (IQR)	0.17 (0.02–18.30)	0.26 (0.01–47.11)	0.048	0.20 (0.03–27.10)	0.26 (0.02–38.49)	0.588	0.22 (0.01–2.70)	0.24 (0.02–15.08)	0.169
Fibrinogen (g/L), M (IQR)	4.01 (1.62–9.44)	4.55 (1.18–11.96)	0.119	4.13 (1.98–9.52)	4.63 (1.82–8.99)	0.903	4.43 (1.89–8.33)	4.40 (1.60–7.52)	0.455
Thrombin time (s), M (IQR)	15.60 (13.10–22.10)	16.00 (13.00–24.90.00.90)	0.112	16.25 (13.40–20.90)	16.15 (13.10–20.40)	0.671	15.30 (13.10–20.70)	16.30 (13.70–20.40)	0.070
Disease [n (%)]			0.492			0.882			0.125
Acute pancreatitis	80 (55.55)	248 (50.30)		13 (59.09)	31 (53.45)		11 (45.83)	36 (64.29)	
Postoperative fluid collection	46 (31.94)	169 (34.28)		7 (31.82)	20 (34.48)		7 (29.17)	15 (26.79)	
Other	18 (12.5)	76 (15.42)		2 (9.09)	7 (12.07)		6 (25)	5 (8.93)	
Bubbles [n (%)]			< 0.001			0.069			0.18
Absent	111 (77.08)	291 (59.03)		17 (77.27)	30 (51.72)		19 (79.17)	34 (60.71)	
Present	33 (22.92)	202 (40.97)		5 (22.73)	28 (48.28)		5 (20.83)	22 (39.29)	
Drainage tube [n (%)]			< 0.001			0.24			0.063
Absent	112 (77.78)	299 (60.65)		16 (72.73)	32 (55.17)		20 (83.33)	33 (58.93)	
Present	32 (22.22)	194 (39.35)		6 (27.27)	26 (44.83)		4 (16.67)	23 (41.07)	

M, median; IQR, interquartile range; $P<0.05$ was considered statistically significant

the largest portion of predictors (35.8%, $n=19$). Second, it highlighted the superiority of advanced image analysis techniques, as deep learning features were overwhelmingly the most prevalent type, with those extracted by the DINOv2 model alone comprising 60.4% ($n=32$) of the final feature set.

This conclusion was powerfully exemplified by the single most influential feature in the model: a wavelet-based texture feature (wavelet.LLL_glcm_Imc1) extracted

Table 2 Quantitative features based on fully automated CT segmentation

Characteristics	Training Set			Validation Set			Test Set		
	Non-infected	Infected	*P*-value	Non-infected	Infected	*P*-value	Non-infected	Infected	*P*-value
Volume of fluid region [(mm^3), M (Range)]	503685.99(7795.75–4449950.43.75.43)	368633.00(7636.16–6957158.35.16.35)	0.006	396176.77(76113.28–4705150.33.28.33)	322978.57(6418.57–5902688.46.57.46)	0.346	550627.07(123287.29–2622565.60.29.60)	511338.17(59708.28–4450041.27.28.27)	0.589
Maximum 3D diameter of fluid region [(mm), M (Range)]	634.11(62.21–1459.35.21.35)	647.24(125.39–1462.95.39.95)	0.313	592.40(120.73–1274.20.73.20)	631.66(135.08–1193.96.08.96)	0.639	634.69(307.39–1331.34.39.34)	681.19(221.88–1033.06.88.06)	0.846
Maximum CT value of fluid region [(HU), M (Range)]	102.00(41.00–1122.00.00.00)	128.00(27.00–2424.00.00.00)	0.004	90.50(53.00–696.00.00.00)	149.50(52.00–1009.00.00.00)	0.876	97.50(52.00–505.00.00.00)	148.00(52.00–1156.00.00.00)	0.037
Minimum CT value of fluid region [(HU), M (Range)]	−99.00(−1000.00–4.00)	−99.00(−1024.00–30.00)	0.018	−103.50(−997.00–55.00)	−99.00(−968.00–49.00)	0.258	−99.00(−976.00–50.00)	−99.00(−994.00–56.00)	0.602
Mean CT value of fluid region [(HU), M (Range)]	14.34(−5.45-54.67)	14.99(−33.78-124.76)	0.892	15.35(−2.34-28.07)	14.62(−4.39-38.03)	0.474	14.02(−0.42-25.69)	16.90(−21.12-39.33)	0.440
Bubble volume [(mm^3), M (Range)]	0.00(0.00–349695.28.00.28)	537.40(0.00–522571.35.00.35)	<0.001	0.00(0.00–253191.95.00.95)	103.14(0.00–410989.68.00.68)	0.013	0.00(0.00–175561.49.00.49)	257.94(0.00–144686.81.00.81)	0.005
Number of bubbles [(count), M (Range)]	0.00(0.00–103.00.00.00)	6.00(0.00–772.00.00.00)	<0.001	0.00(0.00–642.00.00.00)	1.00(0.00–438.00.00.00)	0.018	0.00(0.00–90.00)	5.00(0.00–172.00.00.00)	0.004
Mean bubble volume [(mm^3), M (Range)]	0.00(0.00–13449.82.00.82)	28.76(0.00–22706.09.00.09)	<0.001	0.00(0.00–394.38.00.38)	18.14(0.00–2138.08.00.08)	0.01	0.00(0.00–7315.06.00.06)	47.31(0.00–2720.46.00.46)	0.002
Maximum bubble volume [(mm^3), M (Range)]	0.00(0.00–349634.34.00.34)	175.24(0.00–503480.51.00.51)	<0.001	0.00(0.00–250256.54.00.54)	48.28(0.00–350775.05.00.05)	0.015	0.00(0.00–171487.57.00.57)	198.86(0.00–135128.70.00.70)	0.005
Minimum bubble volume [(mm^3), M (Range)]	0.00(0.00–460.82.00.82)	1.58(0.00–2985.76.00.76)	<0.001	0.00(0.00–4.77.00.77)	0.39(0.00–149.54.00.54)	0.044	0.00(0.00–3.76.00.76)	1.91(0.00–237.55.00.55)	0.001
Minimum distance between bubbles [(mm), M (Range)]	0.00(0.00–17.64.00.64)	1.41(0.00–250.24.00.24)	<0.001	0.00(0.00–3.52.00.52)	0.00(0.00–37.13.00.13)	0.032	0.00(0.00–5.69.00.69)	1.41(0.00–58.83.00.83)	0.003

Table 2 (continued)

Characteristics	Training Set			Validation Set			Test Set		
	Non-infected	Infected	*P*-value	Non-infected	Infected	*P*-value	Non-infected	Infected	*P*-value
Minimum bounding sphere diameter of bubbles [(mm), M (Range)]	0.00(0.00–403.01.00.01)	83.76(0.00–839.34.00.34)	<0.001	0.00(0.00–313.37.00.37)	0.00(0.00–495.80.00.80)	0.029	0.00(0.00–487.63.00.63)	82.64(0.00–549.43.00.43)	0.029
Bubble bed volume [(mm³), M (Range)]	0.00(0.00–1030338.27.00.27)	15178.67(0.00–5156613.51.00.51)	<0.001	0.00(0.00–567569.73.00.73)	357.79(0.00–1825522.44.00.44)	0.018	0.00(0.00–530064.47.00.47)	8821.53(0.00–1803457.58.00.58)	0.008
Maximum 3D diameter of bubble bed [(mm), M (Range)]	0.00(0.00–205.84.00.84)	49.69(0.00–314.55.00.55)	<0.001	0.00(0.00–177.90.00.90)	7.24(0.00–279.04.00.04)	0.020	0.00(0.00–283.09.00.09)	54.08(0.00–247.87.00.87)	0.024
Maximum CT value of bubble bed [(HU), M (Range)]	0.00(0.00–1122.00.00.00)	67.00(0.00–1838.00.00.00)	<0.001	0.00(0.00–578.00.00.00)	24.00(0.00–711.00.00.00)	0.016	0.00(0.00–184.00.00.00)	62.00(0.00–1156.00.00.00)	0.002
Minimum CT value of bubble bed [(HU), M (Range)]	0.00(−1024.00-0.00)	−612.00(−1111.00-0.00)	<0.001	0.00(−1024.00-0.00)	−343.00(−1037.00-0.00)	0.012	0.00(−1023.00-0.00)	−689.50(−1057.00-0.00)	0.002
Mean CT value of bubble bed [(HU), M (Range)]	0.00(−720.96-29.95)	0.00(−734.85-41.22)	0.033	0.00(−389.48-35.21)	0.00(−265.70-25.29)	0.845	0.00(−265.87-13.15)	0.00(−189.95-29.33)	0.276
Contact area between drainage tube and fluid collection [(mm²), M (Range)]	0.00(0.00–16922.30.00.30)	0.00(0.00–75990.26.00.26)	<0.001	0.00(0.00–24381.34.00.34)	0.00(0.00–36769.53.00.53)	0.133	0.00(0.00–13267.61.00.61)	0.00(0.00–31257.98.00.98)	0.096
Contact diameter between drainage tube and fluid collection [(mm), M (Range)]	0.00(0.00–142.36.00.36)	0.00(0.00–264.49.00.49)	<0.001	0.00(0.00–98.22.00.22)	0.00(0.00–265.55.00.55)	0.121	0.00(0.00–137.47.00.47)	0.00(0.00–224.50.00.50)	0.096

M, median; IQR, interquartile range; $P<0.05$ was considered statistically significant

from the 2 cm region surrounding the gas bubbles (coefficient = 0.847). Its top ranking confirms that the textural heterogeneity of the perilesional tissue is a more powerful predictor than the mere presence of gas itself. This key finding, along with other highly-weighted deep learning features from the "bubble bed", grounds the model's predictions in an interpretable, pathophysiological foundation, moving it beyond a simple "black box" (Supplementary Fig. 5).

Clinical model development and performance

To establish a baseline for comparison, a model was developed using only clinical data. From 15 initial parameters with significant univariate associations with infection (Table 3), multivariate logistic regression identified five independent predictors: lactate dehydrogenase (OR 0.994, $P<0.001$), albumin (OR 0.926, $P=0.001$), C-reactive protein (OR 1.004, $P=0.019$), monocyte count (OR 2.709, $P=0.024$), and prothrombin time (OR 1.130, $P=0.055$) (Table 4). While this clinical model achieved a moderate AUC of 0.73 (95% CI: 0.67–0.78) in the training set, its performance showed poor generalizability, with the AUC declining sharply to 0.59 in the validation set and 0.57 in the external test set (Table 5 and Fig. 4). This substantial performance drop-off underscores the inherent limitations of relying on systemic clinical parameters alone and powerfully justifies the necessity of incorporating advanced imaging features for a robust diagnostic solution.

BUBBLE-AI performance and clinical impact

The final BUBBLE-AI model (Table 4), which integrates the DLR-Score with clinical parameters, demonstrated superior and robust diagnostic performance across all datasets, validating the significant value of our AI-based imaging analysis (Table 5; Fig. 4). In both the training and internal validation sets, the model achieved an excellent and stable AUC of 0.92, with high specificity (95% and 91%, respectively). Crucially, the model's performance showed strong generalizability on the independent external test set, achieving an AUC of 0.82 (95% CI: 0.72–0.93) with a sensitivity of 66% and a specificity of 79%. The visualizations of the two models as nomograms are presented in Supplementary Fig. 6.

The clinical significance of this robust external validation performance is substantial. The high specificity, for instance, directly translates to improved antimicrobial stewardship; when applied to the external test cohort, BUBBLE-AI

Table 3 Clinical data univariate and multivariate logistic regression analysis

Characteristics	Univariate analysis				Multivariate analysis			
	coefficient	Odds ratio	95% CI	*P* value	coefficient	Odds ratio	95% CI	*P* value
Thrombin time	0.088	1.092	(0.982, 1.221)	0.113				
Fibrinogen	0.091	1.095	(0.978, 1.229)	0.119				
Procalcitonin	0.099	1.104	(1.015, 1.249)	0.058				
Alkaline phosphatase	0	1	(0.999, 1.001)	0.702				
Basophil count	1.155	3.174	(0.003, 9027.222)	0.76				
Lymphocyte count	−0.068	0.934	(0.715, 1.237)	0.621				
Blood glucose	0.053	1.054	(0.997, 1.119)	0.072				
Monocyte percentage	−0.007	0.993	(0.94, 1.052)	0.816				
Red blood cell count	−0.059	0.942	(0.703, 1.267)	0.692				
Platelet count	0.001	1.001	(0.999, 1.002)	0.38				
C-reactive protein	0.004	1.004	(1.002, 1.007)	0.001	0.003	1.004	(1.001, 1.007)	0.019
Prothrombin time	0.188	1.207	(1.08, 1.36)	0.001	0.123	1.13	(1.006, 1.288)	0.055
Lactate dehydrogenase	−0.003	0.997	(0.995, 0.999)	*P*<0.001	−0.006	0.994	(0.992, 0.996)	*P*<0.001
Albumin	−0.108	0.897	(0.862, 0.933)	*P*<0.001	−0.077	0.926	(0.885, 0.967)	0.001
Monocyte count	0.746	2.108	(1.304, 3.532)	0.003	0.996	2.709	(1.157, 6.554)	0.024
Eosinophil count	−1.943	0.143	(0.039, 0.494)	0.002	−0.066	0.936	(0.082, 10.837)	0.957
Neutrophil count	0.078	1.081	(1.034, 1.134)	0.001	−0.121	0.886	(0.764, 1.021)	0.103
Basophil percentage	−1.33	0.264	(0.113, 0.609)	0.002	0.338	1.402	(0.474, 4.217)	0.543
Eosinophil percentage	−0.251	0.778	(0.686, 0.872)	*P*<0.001	−0.103	0.902	(0.737, 1.118)	0.303
Neutrophil percentage	0.047	1.048	(1.03, 1.068)	*P*<0.001	0.066	1.068	(0.988, 1.156)	0.102
Lymphocyte percentage	−0.057	0.944	(0.923, 0.966)	*P*<0.001	0.029	1.03	(0.948, 1.119)	0.488
White blood cell count	0.07	1.073	(1.03, 1.12)	0.001	0.09	1.094	(0.967, 1.246)	0.167
Body temperature	0.34	1.404	(1.095, 1.82)	0.009	0.002	1.002	(0.724, 1.397)	0.993
Pulse rate	0.014	1.014	(1.001, 1.027)	0.034	0.01	1.01	(0.995, 1.026)	0.195
Body mass index	−0.036	0.965	(0.916, 1.016)	0.173				
Age	0.014	1.014	(1.002, 1.027)	0.023	0.007	1.007	(0.993, 1.021)	0.315
Gender	−0.156	0.856	(0.571, 1.268)	0.444				

Table 4 The multivariable logistic regression of clinical and BUBBLE-AI

Variable	Clinical Model			BUBBLE-AI		
	Coefficient	OR (95%CI)	*P* value	Coefficient	OR (95%CI)	*P* value
Intercept	2.514	12.359 (1.002, 149.341)	0.049	−0.826	0.438 (0.023, 7.960)	0.580
Monocyte count	0.911	2.488 (1.477, 4.350)	0.001	0.684	1.981 (1.003, 4.054)	0.054
Albumin	−0.098	0.907 (0.869, 0.944)	< 0.001	−0.082	0.921 (0.873, 0.970)	0.002
Lactate dehydrogenase	−0.005	0.995 (0.993, 0.997)	< 0.001	−0.002	0.998 (0.996, 1.000)	0.056
Prothrombin time	0.143	1.154 (1.026, 1.310)	0.023	0.203	1.225 (1.072, 1.412)	0.004
C reactive protein	0.004	1.004 (1.002, 1.007)	0.002	0.003	1.003 (1.000, 1.006)	0.092
SI-DLR				1.421	4.142 (3.138, 5.686)	< 0.001

Clinical model = 2.514 + 0.911×Monocyte count-0.098×Albumin-0.005×Lactate dehydrogenase + 0. 143×Prothrombin.time + 0.004×C reactive protein

BUBBLE-AI = −0.826 + 0.684×Monocyte count-0.082×Albumin-0.002×Lactate dehydrogenase + 0. 203×Prothrombin.time + 0.003×C reactive protein + 1.421×SI-DLR

BUBBLE-AI: Bubble Bed Based Learning Engine for Abdominal Infection

Table 5 Performance comparison of predictive models across Training, Validation, and test datasets

Models	Bubble sign	Clinical Model	BUBBLE-AI
Training Set(*n* = 637)			
AUC	0.66(0.61, 0.70)	0.73(0.67, 0.78)	0.92(0.89, 0.94)
Sensitivity	0.59(291/493)	0.68(337/493)	0.72(353/493)
Specificity	0.72(104/144)	0.69(100/144)	0.95(137/144)
Accuracy	0.62(395/637)	0.69(437/637)	0.77(490/637)
Positive Predictive Value (PPV)	0.88(291/331)	0.88(337/381)	0.98(353/360)
Negative Predictive Value (NPV)	0.34(104/306)	0.39(100/256)	0.49(137/277)
Validation Set(*n* = 80)			
AUC	0.55(0.42, 0.67)	0.59(0.44, 0.74)	0.92(0.85, 0.98)
Sensitivity	0.50(29/58)	0.60(35/58)	0.72(42/58)
Specificity	0.59(13/22)	0.59(13/22)	0.91(20/22)
Accuracy	0.53(42/80)	0.60(48/80)	0.78(62/80)
Positive Predictive Value (PPV)	0.76(29/38)	0.80(35/44)	0.95(42/44)
Negative Predictive Value (NPV)	0.31(13/42)	0.36(13/36)	0.56(20/36)
Test Set(*n* = 80)			
AUC	0.69(0.58, 0.80)	0.57(0.42, 0.72)	0.82(0.72, 0.93)
Sensitivity	0.62(35/56)	0.54(30/56)	0.66(37/56)
Specificity	0.75(18/24)	0.54(13/24)	0.79(19/24)
Accuracy	0.66(53/80)	0.54(43/80)	0.70(56/80)
Positive Predictive Value (PPV)	0.85(35/41)	0.73(30/41)	0.88(37/42)
Negative Predictive Value (NPV)	0.46(18/39)	0.33(13/39)	0.50(19/38)

AUC: area under the curve; CI: confidence interval; PPV: positive predictive value; NPV: negative predictive value

BUBBLE-AI: Bubble Bed Based Learning Engine for Abdominal Infection

reduced false-positive diagnoses compared to the traditional "gas bubble sign" (5 vs. 6 cases), a 16.7% relative reduction in unnecessary antibiotic therapy for this group. Furthermore, at a clinically relevant 30% decision threshold, the model would reduce unnecessary interventions by 38% while maintaining acceptable sensitivity for infection detection. Representative cases illustrating the model's diagnostic capabilities in these scenarios are presented in Fig. 5.

Comparative analysis and statistical validation

A systematic head-to-head comparison revealed the stark superiority of the BUBBLE-AI model over traditional diagnostic approaches. The conventional "gas bubble sign" demonstrated limited and inconsistent utility, with AUC values of 0.66, 0.55, and 0.69 across the training, validation, and test sets, respectively. Similarly, the model based solely on clinical parameters showed poor generalizability, with its performance collapsing from an AUC of 0.73 in training to 0.57 on external validation. In sharp contrast, the BUBBLE-AI's superiority was statistically significant across all datasets ($P < 0.001$ vs. both comparators, DeLong's test). More importantly, it achieved a clinically optimal performance profile, balancing a high specificity (> 79%)—which minimizes false-positive diagnoses and unnecessary antibiotic use—with a reasonable sensitivity (66–72%) for detecting true infections.

Disease-specific performance analysis

To define the model's optimal application scope, a disease-specific subgroup analysis on the external test set revealed both its robust capabilities and critical limitations **(Supplementary Fig. 7)**. The BUBBLE-AI demonstrated excellent and highly reliable performance for the two predominant etiologies, achieving an AUC of 0.90 (95% CI: 0.77–0.99) in patients with acute pancreatitis and 0.89 (95% CI:

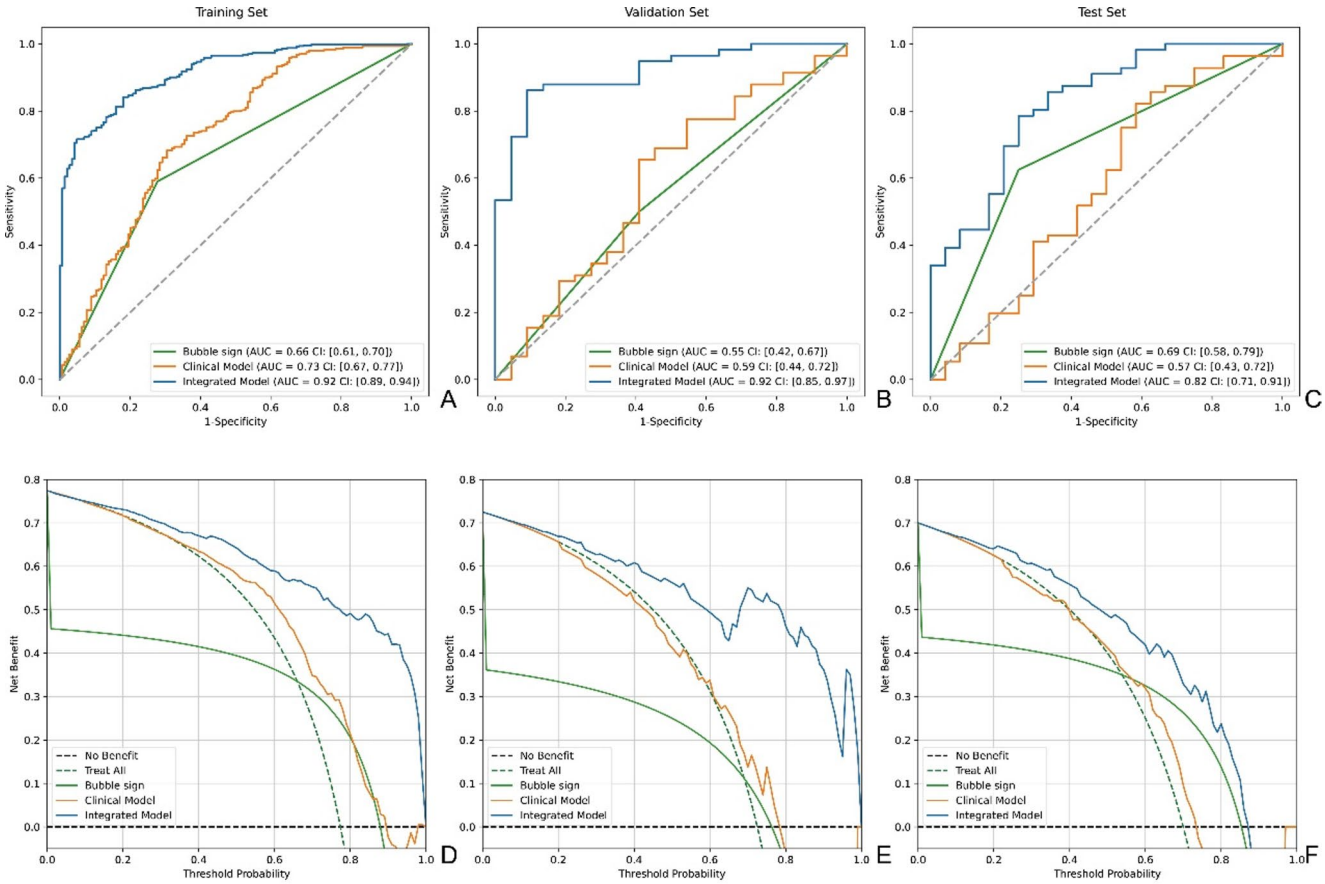


Fig. 4 Diagnostic Performance and Clinical Utility Analysis. **A**–**C** Receiver operating characteristic (ROC) curves comparing the diagnostic performance of the integrated AI model, clinical model, and the "gas bubble sign" in the training (**A**), validation (**B**), and external test (**C**) sets. The Bubble Bed Based Learning Engine for Abdominal Infection (BUBBLE-AI) achieved superior performance across all datasets with an area under the curve (AUC) of 0.92 in training, 0.92 in validation, and 0.82 in testing. **D**–**F** Decision curve analysis (DCA) for the three diagnostic approaches in the training (**D**), validation (**E**), and external test (**F**) sets. The BUBBLE-AI (blue line) demonstrates the highest net clinical benefit across the widest range of threshold probabilities, substantially outperforming the clinical model (orange line), the "gas bubble sign" (green line), and both the treat-all (gray dashed line) and treat-none (horizontal black line) strategies

0.69–1.00.69.00) in postoperative cases. In stark contrast, its performance was unreliable for the small, heterogeneous "other etiologies" subgroup (AUC 0.53, $n = 11$). This finding has crucial clinical implications: while the model is a robust tool for the vast majority of patients (who comprised 86% of the study population), its use is not recommended for rare or uncommon etiologies. For these specific cases, traditional clinical judgment must supersede the model's output pending further validation in larger, dedicated cohorts.

Decision curve analysis and clinical utility

Decision curve analysis (DCA) of the external test set quantified the superior clinical utility of the BUBBLE-AI model. It was the only approach to provide a positive net benefit across nearly the entire range of clinically relevant threshold probabilities (0.1 to 0.93), consistently outperforming the limited benefit ranges of the clinical model (0.37–0.90, 0.20–0.75 and 0.25–0.73) and the "gas bubble sign" (0.54–0.85). To illustrate this in a practical scenario, at a 30% treatment threshold, the BUBBLE-AI would lead to an appropriate treatment decision for 89% of patients—a net improvement of 13% and 18% over the "gas bubble sign" and clinical assessment, respectively (Fig. 4).

The model's broad range of net benefit indicates its flexibility and robustness, suggesting it can be a valuable decision-support tool across diverse clinical scenarios and for physicians with varying risk tolerances. The substantial net benefit observed at commonly used clinical thresholds (20–40%) underscores its potential for meaningful, real-world clinical impact, facilitating more confident and individualized treatment decisions.

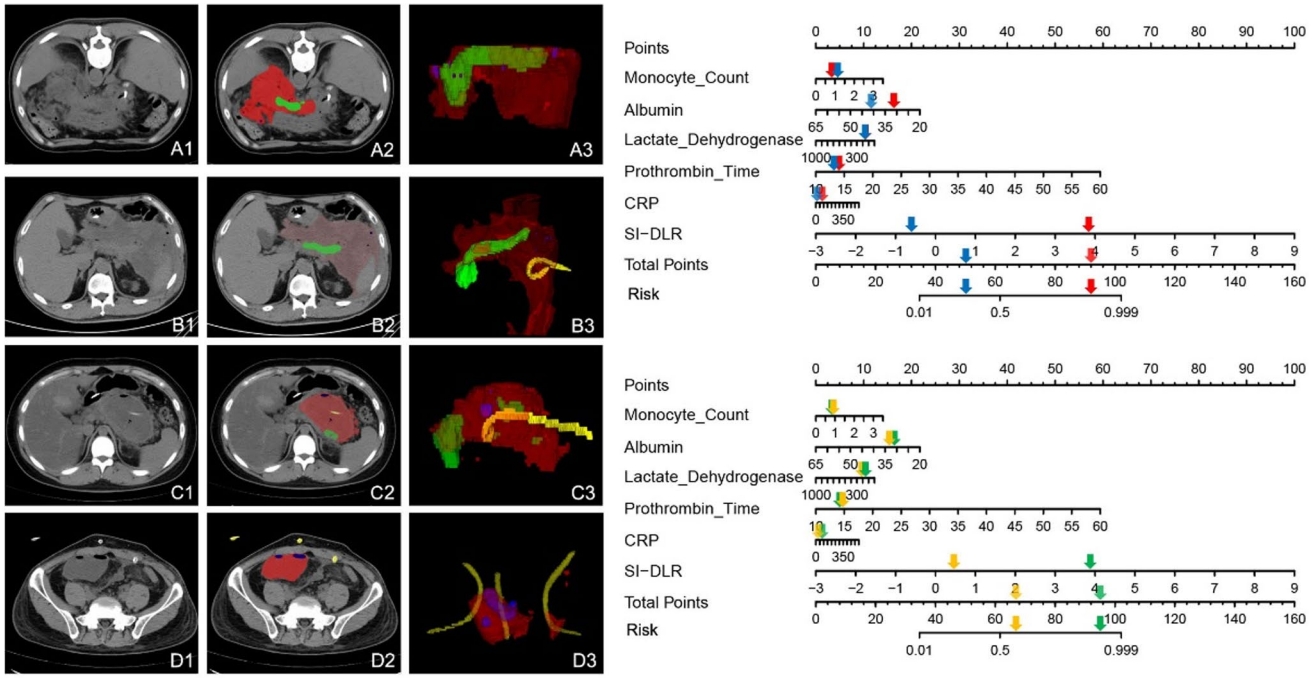


Fig. 5 Representative Clinical Cases Illustrating Model Performance. **A** A true-positive case of a 58-year-old male with infected pancreatic necrosis, where the model predicted a high infection risk (DLR-Score: 0.78) confirmed by culture (E. coli, K. pneumoniae). **B** A true-negative case of a 52-year-old female with a sterile postoperative collection, where the model predicted a low infection risk (DLR-Score: 0.15) confirmed by sterile culture. **C** A false-negative case of a 61-year-old male with an early-stage infection, where the model predicted a borderline low risk (DLR-Score: 0.31) despite a positive culture (*E. faecalis*), illustrating a key limitation in detecting infections without clear morphological changes. DLR-Score = Deep Learning Radiomics Score

Misdiagnosis analysis and clinical implications

A comprehensive misdiagnosis analysis across the 797-patient cohort quantified the substantial real-world advantage of the BUBBLE-AI model. It achieved the lowest total misdiagnosis rate at 24.1% (192/797 cases), a stark improvement over both the traditional bubble sign (38.5%) and the clinical-only model (34.3%). The model's key strength was its ability to minimize false-positive diagnoses, drastically reducing them to just 13 cases compared to 64 for the clinical model and 55 for the bubble sign. This represents an 80% reduction in unnecessary positive diagnoses versus the clinical model, an improvement with profound implications for antimicrobial stewardship.

Analysis of the error patterns revealed distinct and clinically informative failure modes for each approach. The bubble sign's errors were predictable (Supplementary Fig. 8): its high false-negative rate ($n = 252$) stemmed from non-gas-producing infections, while its false positives ($n = 55$) were due to iatrogenic or residual gas. In contrast, the BUBBLE-AI's few errors occurred at the boundaries of diagnostic certainty; false positives were typically associated with drainage tube artifacts and had borderline prediction scores (mean DLR-Score 0.48 ± 0.12), while false negatives represented the most challenging early-stage collections lacking discernible morphological changes (mean DLR-Score 0.31 ± 0.18).

Beyond overall accuracy, the model's robust risk stratification capabilities enable a clear clinical implementation framework. The model demonstrated exceptional certainty at the extremes of its predictions, with misdiagnosis rates of only 2.1% for very low-risk cases (probability < 0.2) and 7.2% for very high-risk cases (probability > 0.8). These findings, validated by a detailed decision curve analysis confirming substantial clinical utility (see Fig. 4), support a risk-stratified management protocol: watchful waiting for low-risk patients, consideration of empirical treatment for high-risk patients, and recommending further diagnostic workup for the intermediate-risk group where clinical correlation is most critical (Fig. 6 and Supplementary Fig. 9).

Discussion

In this study, we introduce and validate a comprehensive AI-based diagnostic framework that significantly improves infection prediction in intra-abdominal fluid collections (IAFCs) using non-contrast CT, addressing a critical unmet clinical need. Our primary finding is that the BUBBLE-AI model, which combines automated image analysis with

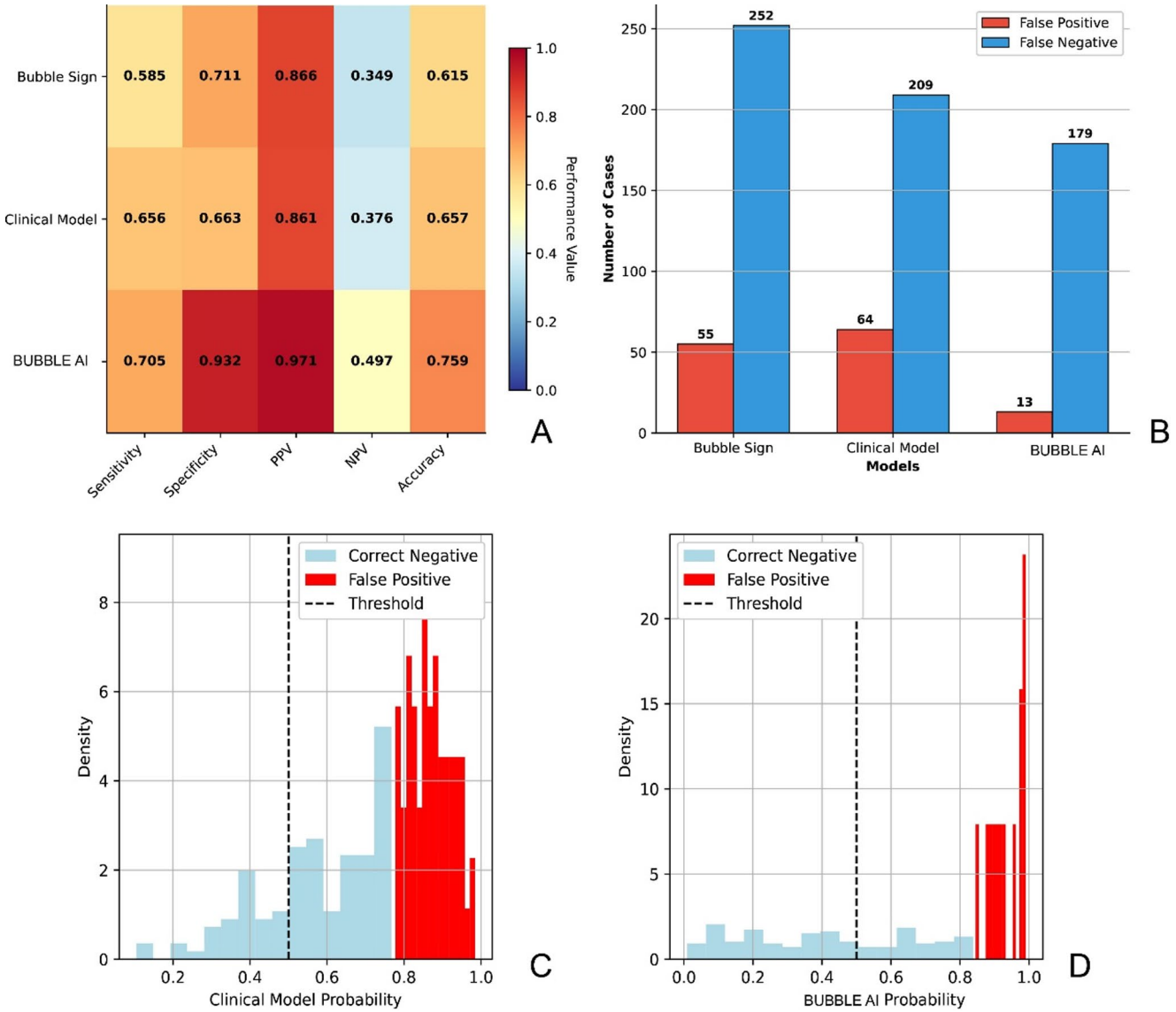


Fig. 6 Comprehensive Misdiagnosis and Risk Stratification Analysis. **A** A heatmap comparing key performance metrics (sensitivity, specificity, PPV, NPV, and accuracy) across the Bubble Bed Based Learning Engine for Abdominal Infection (BUBBLE-AI), clinical model, and "gas bubble sign". **B** A bar chart detailing the absolute counts of misdiagnosed cases, with breakdowns for false-positive and false-negative errors for each method. **C**, **D** Misdiagnosis rates stratified by the BUBBLE-AI's predicted probability ranges, showing the false positive rates (**C**) and false negative rates (**D**) across different risk strata. These panels demonstrate the model's robust risk stratification capability, with the lowest error rates observed in the extreme low-risk (<0.2) and high-risk (>0.8) probability ranges. PPV = positive predictive value; NPV = negative predictive value

clinical data, substantially outperforms all traditional diagnostic methods. With an AUC of 0.82 in a rigorous external validation, our model demonstrates a statistically significant improvement over both the conventional "gas bubble sign" (AUC 0.69, $P<0.001$) and clinical assessment alone (AUC 0.57, $P<0.001$). More importantly, this represents the potential for a paradigm shift from subjective, experience-dependent assessment to objective, data-driven diagnosis standardizable across institutions.

The clinical implications of this enhanced accuracy are multi-faceted, particularly for antimicrobial stewardship and clinical workflow optimization. The model's high specificity of 79% is its most salient clinical asset, as our analysis indicates it can help clinicians avoid inappropriate antibiotic therapy by reducing false-positive diagnoses. This corresponds to a 16.7% relative reduction in false positives within our external test set compared to traditional assessment—a crucial step in combating antimicrobial resistance. However, this advantage must be weighed against its 66% sensitivity, a figure that defines the model's intended role as a powerful decision-support tool, not a substitute for clinical reasoning. For high-risk patients with clear signs of infection, a

negative AI prediction should never delay empirical treatment. Instead, we envision a risk-stratified workflow where AI-predicted low-risk patients undergo active surveillance, maximizing the model's strength in identifying individuals who can be safely managed with de-escalation strategies.

Methodologically, our work offers both a technical advance for clinical translation and novel insights into the pathophysiology of infection. The implementation of the nnMamba architecture enabled fully automated segmentation, a critical prerequisite for moving an AI tool from research to routine practice. More fundamentally, we proposed and validated the novel "bubble bed" concept, based on the hypothesis that the true imaging signature of infection lies not in the gas bubbles themselves, but in the surrounding inflammatory microenvironment they induce. This microenvironment, characterized by complex textural heterogeneity on CT, was validated by our results: features derived from the "bubble bed" were the most prominent predictors in the final model (35.8%). This finding confirms that advanced deep learning can decode subtle, infection-related changes in perilesional tissue that are imperceptible to the human eye, thereby linking the macroscopic imaging phenotype to underlying pathophysiology.

Placing this work in the context of the current literature highlights its contribution. Previous AI studies in this domain were often hampered by single-center designs, user-dependent manual segmentation, and a lack of external validation—all significant barriers to clinical translation. Our study, with its multicenter, fully automated, and externally validated design, provides a higher level of evidence and establishes a more solid foundation for future clinical adoption. Furthermore, our deliberate focus on non-contrast CT, the safest and most accessible modality for critically ill patients with potential renal dysfunction, significantly broadens the model's real-world applicability.

A candid acknowledgement of this study's limitations is essential. First, as a retrospective study, we could not standardize antibiotic administration prior to drainage procedures. Many patients likely received empirical antibiotic therapy, which could potentially suppress bacterial growth and lead to false-negative cultures, misclassifying some truly infected cases as sterile. However, this challenge actually strengthens our conclusions: the robust performance of our model (AUC 0.82 in external validation) despite this potential label noise demonstrates that the "bubble bed" imaging features capture infection-related tissue changes that persist even when bacterial loads are suppressed below culture detection thresholds. This reflects real-world clinical scenarios where imaging-based infection assessment is most needed—when patients have already received antibiotics and culture reliability is compromised. Second, the model's generalizability is not infinite; its poor performance in the small, heterogeneous "other etiologies" subgroup (AUC 0.53) underscores that rare conditions may require dedicated training datasets. Third, the single-timepoint analysis cannot capture the dynamic nature of infection, as a static CT image is merely a snapshot. Finally, despite our interpretability analysis, the inherent "black box" nature of deep learning components remains a challenge for building complete clinical trust. Fourth, "Our peri-lesional feature extraction approach encompasses heterogeneous adjacent tissues, which could be viewed as a limitation requiring tissue-specific analysis. However, we argue this reflects the true biological nature of infectious inflammation, which affects all tissue types within its microenvironmental sphere. Future studies could explore tissue-specific subgroup analyses to further refine this approach. Finally, we position this study as a proof-of-concept investigation acknowledging that abdominal fluid collections, unlike neoplasms, represent complex entities influenced by multiple interacting factors beyond infection alone—including recurrence status, surrounding fibrosis, anatomical location, previous treatments, and patient comorbidities. Our primary goal was to validate whether, amidst this complexity, an AI-detectable imaging signature specifically related to infection exists. The robust performance of our "bubble bed" concept confirms this hypothesis and establishes a foundation for future comprehensive models that will systematically integrate these multidimensional clinical variables to provide individualized patient risk assessment.

These limitations directly inform a clear agenda for future research. The most crucial next step is a prospective, multicenter randomized controlled trial (RCT) to definitively assess the model's real-world impact on clinical outcomes, such as antibiotic-free days, length of hospital stay, and mortality. Furthermore, exploring longitudinal imaging analysis, where models learn from the temporal evolution of IAFCs, could enable earlier infection prediction. Integrating our imaging score with novel biomarkers from plasma or drained fluid (e.g., proteomics, cell-free DNA) also holds promise for pushing diagnostic accuracy to new heights.

In conclusion, this AI-based framework represents a significant and tangible advance in the diagnosis of IAFC infections. By providing an objective and accurate tool, it offers a clear pathway toward more precise clinical management, improved antimicrobial stewardship, and a meaningful step forward for data-driven precision medicine in the challenging field of infectious diseases.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00261-025-05260-9.

Author contributions Author Contributions:Conceptualization: Y.G.; Methodology: Y.G.; Formal Analysis: Y.G., Y.Z.; Investigation: J.Y., J.L.; Data Curation: Y.S., D.Z.; Writing – Original Draft: Y.G., Y.B.;

Writing – Review & Editing: C.S., Y.B.; Visualization: L.W.; Software: C.C.; Validation: M.Y., F.L.; Resources: J.L., Y.B.; Funding Acquisition: C.S., Y.B.; Project Administration: C.S., Y.B.; Supervision: T.W., X.F.All authors reviewed and approved the final manuscript.

Funding This work was supported in part by the National Science Foundation for Scientists of China (81871352, 82171915, 82171930, 82271972, 82371955, 82202125, 82572215 and 82202145), Natural Science Foundation of Shanghai Science and Technology Innovation Action Plan (21ZR1478500, 21Y11910300), Clinical Research Plan of SHDC (SHDC2022CRD028), Shanghai Municipal Health Commission (2024ZZ1015), Plan for Promoting Scientific Research Paradigm Reform through Artificial Intelligence (2024RGZD001) and 2025 Special Project for Clinical Research from the Shanghai Municipal Health Commission (202540148).

Data availability To ensure reproducibility, the model code is publicly available at [https://github.com/CHANGHAI-AILab/AP_Task2]. While the clinical dataset cannot be publicly shared due to patient privacy regulations, it may be available to qualified researchers upon reasonable request and appropriate data use agreements.The nnMamba segmentation model and prediction model code are available at [GitHub repository: https://github.com/institution/iafc-infection-prediction]. Data are available through corresponding authors. Computer codes are available as online inference codes on GitHub (https://github.com/CHANGHAI-AILab/AP_Task2).

Declarations

Competing interests The authors declare no competing interests.

Ethical approval Institutional Review Board approval was obtained by Changhai hospital.

Informed consent Written informed consent was waived by the Institutional Review Board.

References

1. Liu, T. H. *et al.* Acute pancreatitis in intensive care unit patients: value of clinical and radiologic prognosticators at predicting clinical course and outcome. *Crit Care Med* 31, 1026–1030 (2003). https://doi.org/10.1097/01.CCM.0000049951.77583.85
2. Werge, M., Novovic, S., Schmidt, P. N. & Gluud, L. L. Infection increases mortality in necrotizing pancreatitis: A systematic review and meta-analysis. *Pancreatology* 16, 698–707 (2016). https://doi.org/10.1016/j.pan.2016.07.004
3. Yasuda, H. *et al.* Etiology and mortality in severe acute pancreatitis: A multicenter study in Japan. *Pancreatology* 20, 307–317 (2020). https://doi.org/10.1016/j.pan.2020.03.001
4. Zhao, Y. *et al.* Early prediction of acute pancreatitis severity based on changes in pancreatic and peripancreatic computed tomography radiomics nomogram. *Quant Imaging Med Surg* 13, 1927–1936 (2023). https://doi.org/10.21037/qims-22-821
5. Banks, P. A. *et al.* Classification of acute pancreatitis-2012: revision of the Atlanta classification and definitions by international consensus. *Gut* 62, 102–111 (2013). https://doi.org/10.1136/gutjnl-2012-302779
6. Mortele, K. J. *et al.* A modified CT severity index for evaluating acute pancreatitis: improved correlation with patient outcome. *AJR Am J Roentgenol* 183, 1261–1265 (2004). https://doi.org/10.2214/ajr.183.5.1831261
7. van Grinsven, J. *et al.* Natural History of Gas Configurations and Encapsulation in Necrotic Collections During Necrotizing Pancreatitis. *J Gastrointest Surg* 22, 1557–1564 (2018). https://doi.org/10.1007/s11605-018-3792-z
8. Arvanitakis, M. *et al.* Endoscopic management of acute necrotizing pancreatitis: European Society of Gastrointestinal Endoscopy (ESGE) evidence-based multidisciplinary guidelines. *Endoscopy* 50, 524–546 (2018). https://doi.org/10.1055/a-0588-5365
9. van Santvoort, H. C. *et al.* A conservative and minimally invasive approach to necrotizing pancreatitis improves outcome. *Gastroenterology* 141, 1254–1263 (2011). https://doi.org/10.1053/j.gastro.2011.06.073
10. Bai, Y., Gao, J., Zou, D. W. & Li, Z. S. Prophylactic antibiotics cannot reduce infected pancreatic necrosis and mortality in acute necrotizing pancreatitis: evidence from a meta-analysis of randomized controlled trials. *Am J Gastroenterol* 103, 104–110 (2008). https://doi.org/10.1111/j.1572-0241.2007.01575.x
11. Dellinger, E. P. *et al.* Early antibiotic treatment for severe acute necrotizing pancreatitis: a randomized, double-blind, placebo-controlled study. *Ann Surg* 245, 674–683 (2007). https://doi.org/10.1097/01.sla.0000250414.09255.84
12. Isenmann, R. *et al.* Prophylactic antibiotic treatment in patients with predicted severe acute pancreatitis: a placebo-controlled, double-blind trial. *Gastroenterology* 126, 997–1004 (2004). https://doi.org/10.1053/j.gastro.2003.12.050
13. Talukdar, R. *et al.* Antibiotic use in acute pancreatitis: an Indian multicenter observational study. *Indian J Gastroenterol* 33, 458–465 (2014). https://doi.org/10.1007/s12664-014-0494-7
14. Baltatzis, M., Jegatheeswaran, S., O'Reilly, D. A. & Siriwardena, A. K. Antibiotic use in acute pancreatitis: Global overview of compliance with international guidelines. *Pancreatology* 16, 189–193 (2016). https://doi.org/10.1016/j.pan.2015.12.179
15. Murata, A. *et al.* A descriptive study evaluating the circumstances of medical treatment for acute pancreatitis before publication of the new JPN guidelines based on the Japanese administrative database associated with the Diagnosis Procedure Combination system. *J Hepatobiliary Pancreat Sci* 18, 678–683 (2011). https://doi.org/10.1007/s00534-011-0375-8
16. Chartrand, G. *et al.* Deep Learning: A Primer for Radiologists. *Radiographics* 37, 2113–2131 (2017). https://doi.org/10.1148/rg.2017170077
17. Park, S. *et al.* Application of deep learning-based computer-aided detection system: detecting pneumothorax on chest radiograph after biopsy. *Eur Radiol* 29, 5341–5348 (2019). https://doi.org/10.1007/s00330-019-06130-x
18. Kumar, V. *et al.* Radiomics: the process and the challenges. *Magn Reson Imaging* 30, 1234–1248 (2012). https://doi.org/10.1016/j.mri.2012.06.010
19. Lambin, P. *et al.* Radiomics: extracting more information from medical images using advanced feature analysis. *Eur J Cancer* 48, 441–446 (2012). https://doi.org/10.1016/j.ejca.2011.11.036
20. 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). https://doi.org/10.1136/bmj.g7594
21. Administration, U. S. F. a. D. Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan. https://www.fda.gov/media/145022/download (2021).

Authors and Affiliations

Yifei Guo[1] · Chengwei Chen[1] · Tiegong Wang[1] · Yixuan Shen[1] · Danqun Zheng[1] · Yilun Zheng[1] · Jieyu Yu[1] · Jing Li[1] · Xu Fang[1] · Fang Liu[1] · Ming Yang[1] · Li Wang[1] · Jianping Lu[1] · Chengwei Shao[1] · Yun Bian[1]

✉ Chengwei Shao
chengweishaoch@163.com

✉ Yun Bian
bianyun2012@foxmail.com

Yifei Guo
1623162930@qq.com

Chengwei Chen
timchen91@aliyun.com

Tiegong Wang
wtiegong@163.com

Yixuan Shen
shenyixuan0423@163.com

Danqun Zheng
770320180@qq.com

Yilun Zheng
827010676@qq.com

Jieyu Yu
295923524@qq.com

Jing Li
lijing_hao2019@163.com

Xu Fang
fangxu20192020@163.com

Fang Liu
liufang217217@163.com

Ming Yang
1065377183@qq.com

Li Wang
wanglichanghai2019@126.com

Jianping Lu
lujianping2019@163.com

[1] Department of Radiology, Changhai Hospital, Shanghai, China