nature communications

Article

https://doi.org/10.1038/s41467-024-54043-1

PACT-3D, a deep learning algorithm for pneumoperitoneum detection in abdominal CT scans

Received: 28 November 2023

Accepted: 29 October 2024

Published online: 07 November 2024

Check for updates

I-Min Chiu[1,2] ✉, Teng-Yi Huang[3], David Ouyang[1], Wei-Che Lin[4,5,6], Yi-Ju Pan[7,8], Chia-Yin Lu[4] & Kuei-Hong Kuo[9,10] ✉

Delays or misdiagnoses in detecting pneumoperitoneum can significantly increase mortality and morbidity. We developed and validated a deep learning model designed to identify pneumoperitoneum in computed tomography images. The model is trained on abdominal scans from Far Eastern Memorial Hospital (January 2012–December 2021) and evaluated using a simulated test set (14,039 scans) and a prospective test set (6351 scans) collected from the same center between December 2022 and May 2023. External validation included 480 scans from Cedars-Sinai Medical Center. Overall, the model achieves a sensitivity of 0.81–0.83 and a specificity of 0.97–0.99 across retrospective, prospective, and external validation; sensitivity improves to 0.92–0.98 when cases with a small amount of free air (total volume <10 ml) are excluded. These findings suggest that the model can deliver accurate and consistent predictions for pneumoperitoneum in computed tomography scans with segmented masks, potentially accelerating diagnostic and treatment workflows in emergency care.

Pneumoperitoneum, which refers to the presence of extraluminal free air in the peritoneal space, is a potentially life-threatening condition that represents a differential diagnosis when managing acute abdominal pain in the Emergency Department (ED). In adults, perforated viscus is the leading cause of pneumoperitoneum, representing 85–95% of cases, and among these, surgical pneumoperitoneum comprises 85–90%[1,2]. Diagnostic tools for identifying pneumoperitoneum include plain radiographs, ultrasound, and Computed Tomography (CT) scan, with the latter remaining the gold standard, exhibiting reported sensitivity levels of approximately 96–100%[3]. Timely diagnosis of pneumoperitoneum is crucial, as delayed recognition can lead to sepsis and result in increased mortality and morbidity[4,5]. However, prolonged CT interpretation times are frequently observed in crowded EDs, with previous reports indicating an average delay of approximately 2 h[6]. Moreover, the use of CT scans during ED visits has dramatically increased in the past decade, with a 330% rise reported in the US from 3.2% of encounters (95% confidence interval [CI] 2.9% to 3.6%) in 1996 to 13.9%[7].

Diagnosing pneumoperitoneum from a CT scan is highly dependent on the reader's expertise and the amount of free air present. According to previous research, only 62.8% of postgraduate year resident feel confident about diagnosing acute pathological findings from CT scans, such as pneumoperitoneum or bowel obstruction[8]. Moreover, studies have shown that discrepancy rates in the

[1]Department of Cardiology, Smidt Heart Institute, Cedars-Sinai Medical Center, Los Angeles, CA, USA. [2]Department of Emergency Medicine, Kaohsiung Chang Gung Memorial Hospital, Kaohsiung, Taiwan. [3]Department of Electrical Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan. [4]Department of Diagnostic Radiology, Kaohsiung Chang Gung Memorial Hospital and Chang Gung University College of Medicine, Kaohsiung, Taiwan. [5]Thyroid Head and Neck Ablation Center, Kaohsiung Chang Gung Memorial Hospital, Kaohsiung, Taiwan. [6]School of Medicine, College of Medicine, National Sun Yat-Sen University, Kaohsiung, Taiwan. [7]Department of Psychiatry, Far Eastern Memorial Hospital, Banciao, Taiwan. [8]Department of Chemical Engineering and Materials Science, Yuan Ze University, Taoyuan, Taiwan. [9]Division of Medical Image, Far Eastern Memorial Hospital, Banciao, Taiwan. [10]National Yang Ming Chiao Tung University School of Medicine, Taipei, Taiwan. ✉e-mail: outofray@hotmail.com; goman178@gmail.com

interpretation of emergency CT scans between residents and attending radiologists vary significantly based on the level of training, ranging from 13.5% to 30.0%[9,10]. Misinterpretations can have a direct negative impact on patient management, with adverse effects noted in 7.2% of patients[11]. These factors may contribute to considerable delays in the recognition of critical pathologies like pneumoperitoneum, potentially leading to poorer patient outcomes.

Artificial Intelligence (AI) has greatly advanced healthcare in recent years, particularly in medical imaging technologies such as computed tomography (CT) scans, X-rays, and ultrasonography[12–18]. AI has also contributed to increased speed and efficiency in medical image analysis, reducing the workload of healthcare professionals and improving patient outcomes. Recent studies have investigated the potential of deep learning algorithms in assisting the detection of pneumoperitoneum on CT scans[19,20]. However, the performance of these AI models varies and is dependent on the selection of datasets. For assessment of the AI model, it is critical to use a dataset that mirrors the actual incidence rate of pneumoperitoneum. Moreover, a prospective evaluation is necessary, along with ongoing enhancements to improve model performance.

In this study, we introduced PACT-3D, a 3-dimensional U-Net algorithm specifically tailored for 3D medical image segmentation. This convolutional neural network excels at capturing spatial hierarchy and information across both the transverse and vertical axes of biomedical images. The PACT-3D model is designed to automatically segment areas of pneumoperitoneum from CT scans, providing predictions at the patient level and visualizations at the pixel level. It is engineered to detect pneumoperitoneum with high accuracy, and its performance has been thoroughly evaluated using both a simulated test dataset and in a prospective observational setting.

Result

Demographic characteristics

In this study, we retrospectively analyzed 140,339 abdominal CT scans from 2012 to 2021. After exclusions, 139,781 were eligible for analysis. Pneumoperitoneum was identified in 973 of these and the studies were randomly allocated to training, validation, and test datasets in a 5:1:1 ratio (Fig. 1). The training set comprised 695 scans with pneumoperitoneum, alongside a randomly selected equivalent number of negative scans. The validation set included 139 scans with pneumoperitoneum, matched with an equal number of negative cases. To evaluate the performance of the PACT-3D model, the test set was designed to mirror a real-world prevalence ratio of approximately 1:100, consisting of 139 scans with pneumoperitoneum and a larger pool of 13,900 negative scans. Additionally, we conducted a prospective clinical evaluation using abdominal CT scans from December 2022 to May 2023 at the same hospital, resulting in a prospective test set of 6351 CT scans. This approach aims to thoroughly evaluate the model's performance under conditions that closely resemble those of clinical settings.

The mean age of patients in the simulated test set was 54 years with a standard deviation (SD) of 13.1, while the prospective test set had a slightly higher mean age of 59 years (SD = 16.9). Females represented 48.2% (n = 6767) of the simulated test set and 47.2% (n = 3000) of the prospective test set. The incidence of pneumoperitoneum detected was set to 1.0% in the simulated test set. Analyzing all CT scans in ER, the incidence of pneumoperitoneum was 1.3% (n = 82) in the prospective test set (Table 1).

Distribution of CT vendors

Regarding the distribution of CT vendors, there were noticeable differences between the simulated and prospective test sets. In the simulated test set, Philips Brilliance 64 scanners were used in 8.0% of cases, while Siemens Somatom Definition and Definition Flash scanners were used in 10.7% and 5.5% of cases, respectively. GE LightSpeed VCT scanners accounted for 15.7% of the scans. A significant portion, 60.1%, involved Siemens Somatom Definition AS scanners (Table 1).

In contrast, the prospective test set exhibited a varied distribution. Siemens Somatom Definition AS scanners were used less frequently, constituting 43.6% of the scans. The GE Revolution Frontier became more prevalent, representing 24.8% of scans in this set. This shift in vendor distribution indicates a temporal change in scanner preference or availability between the two test sets. The image acquisition setting of different CT vendors was shown in Supplementary Table 1.

Model performance

The trained 3D U-Net model demonstrated satisfactory performance in detecting pneumoperitoneum on the validation set. The Dice score for

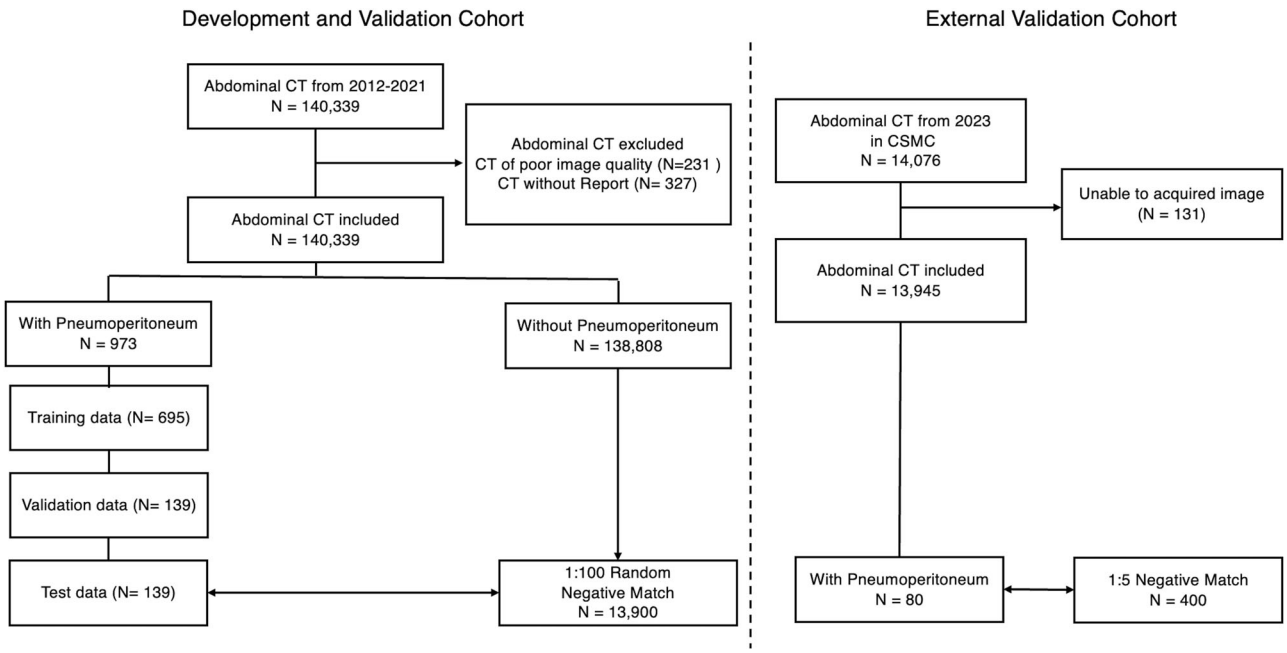

Fig. 1 | The inclusion flowchart of this study. 'N' represents the number of CT studies in each step.

Table 1 | Demographics and CT vendor distributions in simulated and prospective test sets

	Simulated Test Set Mean (SD) / *N* (%)	Prospective Test Set Mean (SD) / *N* (%)	External Test Set Mean (SD) / *N* (%)
Total CT scans	14,039	6351	480
Age	54 (13.1)	59 (16.9)	57 (19.0)
Female	6767 (48.2%)	3000 (47.2%)	204 (42.5%)
CT Vendors			
Philips Brilliance 64	1123 (8.0%)		
Siemens Somatom definition	1502 (10.7%)		
Siemens Somatom definition Flash	772 (5.5%)	524 (8.3%)	8 (1.7%)
Siemens Somatom definition AS	624 (60.1%)	2772 (43.6%)	
GE LightSpeed VCT	2204 (15.7%)	1479 (23.3%)	33 (6.9%)
GE Revolution Frontier		1576 (24.8%)	193 (40.2%)
GE Discovery			99 (20.6%)
Toshiba Aquilion ONE			136 (28.3%)
Pneumoperitoneum	139 (1.0%)	82 (1.3%)	80 (16.7%)

CT Computed Tomography, *PPV* Positive Predictive Value.

Table 2 | Performance of PACT-3D in test set

Performance Metrics	Simulated Test Set value (95% CI)	Prospective Test Set value (95% CI)	External Test Set value (95% CI)
Sensitivity	0.81 (0.75–0.86)	0.83 (0.77–0.90)	0.81 (0.71–0.88)
Specificity	0.99 (0.98–1.0)	0.99 (0.98–0.99)	0.97 (0.94–0.98)
PPV	0.41 (0.34–0.48)	0.44 (0.37–0.52)	0.79 (0.69–0.87)
F1-score	0.54 (0.47–0.61)	0.58 (0.51–0.65)	0.80 (0.74–0.86)
Sensitivity in etiology			
Gastro-duodenal	0.93 (0.82–0.98)	0.87 (0.73–0.94)	
Small Bowel	1.0 (0.87–1.0)	0.88 (0.63–0.98)	
Large Intestine	0.64 (0.41–0.77)	0.73 (0.50–0.89)	
Trauma	1.0 (0.57–1.0)	0.83 (0.45–0.97)	
Post-operative	0.59 (0.33–0.84)	0.8 (0.55–0.93)	
Sensitivity in total volume of free air			
Total volume > 1 ml	0.89 (0.84–0.93)	0.91 (0.86–0.95)	0.86 (0.75–0.93)
Total volume > 10 ml	0.95 (0.90–0.98)	0.98 (0.93–1.0)	0.92 (0.80–0.97)

pneumoperitoneum segmentation was 0.81, indicating a high degree of overlap between the predicted and ground truth regions. Throughout the training process, we meticulously balanced the number of negative CT scans against positive ones at varying ratios to refine the model's sensitivity and positive predictive value (PPV). Our objective was to optimize the F1-score, which harmonizes sensitivity and PPV, as reflected in Supplementary Table 2. The data revealed that a balanced ratio of positive to negative cases (1:1) yielded the highest F1-score.

In the simulated test set, our model achieved a F1-score of 0.54 (95% CI: 0.47–0.61), with a sensitivity of 0.81 (95% CI: 0.75–0.86), a specificity of 0.99 (95% CI: 0.98–1.0), and a PPV of 0.41 (95% CI: 0.34–0.38). Of the 139 CT scans positive for pneumoperitoneum, the model identified 112 and missed 27. Among the 13,900 negative scans, 167 were incorrectly classified as pneumoperitoneum. In the prospective test set at ER, the model's performance yielded F1-score of 0.58 (95% CI: 0.51–0.65), with a sensitivity of 0.83 (95% CI: 0.77–0.90), specificity of 0.99 (95% CI: 0.98–0.99), and a PPV of 0.44 (95% CI: 0.37–0.52). Out of the 69 CT scans with confirmed pneumoperitoneum, the model detected 54 and misclassified 88 out of 8,451 negative scans (Table 2).

External validation

At CSMC, a total of 14,076 abdominal CT scans were identified in 2023. Among these, 80 scans were documented as positive for pneumoperitoneum in the reports. We included 400 negative control scans, matched for age and sex. In this external validation cohort, the mean age was 57 years (SD = 19.0), and 204 (42.5%) of the participants were female. There were notable differences in the distribution of CT vendors within the CSMC cohort, with most scans performed using GE Revolution (40.2%), GE Discovery (20.6%), and Toshiba Aquilion (28.3%).

In the CSMC test set, PACT-3D achieved an F1-score of 0.80 (95% CI: 0.74–0.86), with a sensitivity of 0.81 (95% CI: 0.71–0.88), specificity of 0.97 (95% CI: 0.94–0.98), and a positive predictive value (PPV) of 0.79 (95% CI: 0.69–0.87). Of the 80 CT scans positive for pneumoperitoneum, the model correctly identified 65.

Subgroup analysis

When analyzing performance by etiological subgroup, the PACT-3D model displayed high accuracy for gastroduodenal and small bowel perforations, as well as trauma-related cases, achieving sensitivities of 0.93 (0.82–0.98), 1.0 (0.87–1.0), and 1.0 (0.57–1.0), respectively. In contrast, the model demonstrated relatively lower sensitivities for large intestine perforation and post-operative changes, recording values of 0.64 (0.41–0.77), 0.59 (0.33–0.84). During the prospective observational period, a consistent pattern was observed. The sensitivities for gastroduodenal, small bowel, and trauma-related perforations were 0.87 (0.73–0.94), 0.88 (0.63–0.98), and 0.83 (0.45–0.97), while those for large intestine perforation and post-operative changes were 0.73 (0.50–0.89) and 0.8 (0.55–0.93), respectively (Table 2).

In subgroup analyses evaluating performance across various total volumes of free air, we observed improvement in the sensitivity of PACT-3D. Specifically, sensitivity increased to 0.89 (95% CI: 0.84–0.93), 0.91 (95% CI: 0.86–0.95), and 0.86 (95% CI: 0.75–0.93) on the simulated test set, prospective test sets, and external test set respectively, when scans with a total free air volume of less than 1 ml were excluded. This sensitivity further escalated to 0.95 (95% CI: 0.90–0.98), 0.98 (95% CI: 0.93–1.0), and 0.92 (95% CI: 0.80–0.97) among three test sets upon excluding scans with less than 10 ml of total free air volume, indicating a correlation between detection capability and the quantity of free air present (Table 2).

From another point of view, an association was found between scans predicted as positive by the model and a heightened rate of urgent surgeries, defined as surgeries conducted within 24 h following the CT scan. After excluding post operation scans, in the simulated test set, urgent surgeries were performed on 84 (85.8%) of the patients out of 99 whose pneumoperitoneum was identified by the model. In contrast, among the patients with missed pneumoperitoneum diagnoses by the model, 10 (55.6%) out of 18 underwent urgent surgeries ($p < 0.001$). Within the prospective test set, 40 (75.5%) of the 53 patients diagnosed with pneumoperitoneum by the model received urgent surgeries, as opposed to 8 (57.1%) of the 14 patients with pneumoperitoneum that the model failed to detect ($p < 0.001$).

Discussion

In this study, we introduced PACT-3D, a 3D U-Net-based deep learning model, designed for detecting pneumoperitoneum on abdominal CT scans. The robustness of PACT-3D is demonstrated by its training on scans from a wide array of CT scanner models, its prospective and external testing, ensuring consistent performance despite geographic differences and the evolving landscape of medical imaging technology. PACT-3D demonstrated robust performance, characterized by

high sensitivity and specificity. The model's high specificity and satisfactory PPV are particularly noteworthy given the rarity of pneumoperitoneum in routine settings, which is crucial for minimizing false positives and thus reducing the risk of alarm fatigue. The consistent performance of PACT-3D, observed in a prospective test set that included newer CT scanner models, and its external validation across an international dataset, further supports its generalizability. By providing a prediction mask in addition to binary classification for pneumoperitoneum, the model enhances its trustworthiness and reliability, offering significant potential to accelerate clinical decision-making across various scenarios and timeframes.

Historically, AI algorithms have encountered challenges when attempting to detect free air in CT scans. They often exhibit reduced sensitivity, even if their specificity is commendable[19,20]. Previous studies, focusing on the utilization of 2D segmentation models for pneumoperitoneum detection, have highlighted challenges in differentiating free air from the common place bowel gas[20]. While 2D models have been a cornerstone in healthcare deep learning applications, this is largely because many medical imaging modalities, such as X-rays, ultrasound, and specific MRI or CT slices, intrinsically generate 2D images, making these models a natural choice[21–24]. Additionally, 2D models tend to be computationally less demanding than 3D models, suiting institutions with restricted computational capabilities. The extensive availability of pretrained 2D models, which have been trained on diverse and vast datasets, further contributes to their dominance[25]. By fine-tuning these models for specific medical tasks, performance can often be enhanced, benefiting from features learned across various domains. However, despite the prevalence of 2D architectures in healthcare, the detection of pneumoperitoneum, with its inherent risk of confusion with bowel air, greatly benefits from the depth of understanding offered by 3D morphology. The use of a 3D segmentation model allows better recognition of free air morphological patterns, distinguishing them from bowel gas with enhanced accuracy. This adaptation, coupled with the model's rapid inference capability, heightens its potential to augment diagnostic precision and efficiency.

In the subgroup analysis, PACT-3D particularly excelled in detecting gastroduodenal and small bowel origin pneumoperitoneum, with sensitivities exceeding 0.9 in both test sets. For large intestine origin cases, sensitivity ranged between 0.64 and 0.73. We surmise this disparity arises from the inherently larger air bubble sizes in the upper gastrointestinal tract, facilitating differentiation from standard bowel gas. In contrast, large intestine perforations, frequently linked with inflammatory processes, present greater interpretative challenges, even for seasoned radiologists[26,27]. Consistent with this, the model demonstrated improved sensitivity when CT scans with minimal free air volume were excluded, showing an increase to 0.89–0.91 for total free air volumes greater than 1 ml, and further to 0.95–0.98 for volumes greater than 10 ml (Table 2).

The missed cases in both the simulated and prospective test sets highlight an important aspect of the model's performance in real-world settings. Upon reviewing the cases that PACT-3D failed to predict, we found that most missed instances involved free air that was scattered and appeared in retroperitoneal areas, which can easily be mistaken for other bowel gas at first glance. Specifically, the model may miss cases with smaller air bubbles, but it reliably identifies cases with larger, cumulated volumes of free air, which typically require urgent intervention. On the other hand, the model's high specificity demonstrates that it won't easily trigger false alarms, reducing the risk of clinician fatigue. In cases where PACT-3D incorrectly identified pneumoperitoneum, a review of the prediction masks revealed that most errors were due to the model mistakenly identifying air-containing abscesses, subcutaneous emphysema, air within fluid collections, distended bowel gas, or air density artifacts related to artificial implants (Supplementary Table 3). Although these cases were not correctly diagnosed, many still required medical intervention. This selective performance could make PACT-3D a valuable triage tool in emergency and critical care, where the primary goal is to quickly identify and prioritize cases that necessitate immediate surgical intervention.

When assessing the model's predictions in relation to clinical outcomes, specifically the necessity for urgent surgery, we observed a significant correlation. Patients with pneumoperitoneum detected by the PACT-3D model underwent urgent surgery at a higher rate (75.5–85.8%) compared to those where pneumoperitoneum was not detected (55.6–57.1%). This suggests that the model is more adept at identifying larger volumes of free air, particularly those originating from the upper gastrointestinal tract, where emergency surgical intervention is often imperative. Conversely, smaller volumes of free air, typically resulting from inflammatory conditions like acute diverticulitis, are usually managed with conservative treatment or elective surgery in patients who are hemodynamically stable[28]. These findings indicate that the PACT-3D model can serve as a valuable tool for risk stratification by illustrating the perforated area alongside the volume of free air. This makes the model particularly useful in emergency settings, where timely diagnosis is critical. However, to fully harness the potential of AI in this domain, ongoing efforts should focus on improving the model's sensitivity to smaller pneumoperitoneum cases.

Comparatively, our model's performance exceeds prior deep learning endeavors in detecting pneumoperitoneum or related abdominal pathologies on CT scans. This corroborates the robustness and superiority of our 3D U-Net-based approach. The sensitivity and specificity position PACT-3D as a valuable tool for radiologists, especially in urgent situations where prompt and accurate detection is critical. Notably, the model maintained consistent performance in both sensitivity and specificity in the external test cohort, reinforcing its reliability across geographically diverse datasets. Several factors contribute to PACT-3D's performance, including the implementation of the 3D U-Net architecture, renowned for its efficacy in diverse medical image segmentation tasks, and the amalgamation of Dice loss and focal loss to counteract training set imbalances.

Several limitations are inherent to our study. Firstly, while our model demonstrated robust performance in detecting pneumoperitoneum overall, its efficacy in discerning smaller or more subtle instances was found to be lower compared to larger pneumoperitoneum cases. Given that these nuanced cases often present a significant diagnostic challenge, this limitation underscores the need for further refinement of the model. Future research should focus on enhancing the model's capability to accurately detect smaller instances, thereby maximizing its potential as a reliable diagnostic aid across all presentations. Secondly, although PACT-3D has undergone validation across multiple institutions, its potential impact on clinical practice—such as optimizing diagnostic workflows or improving patient outcomes—has yet to be thoroughly evaluated. Future studies should explore how integrating PACT-3D into clinical settings might influence decision-making processes, workflow efficiency, and overall patient care, ensuring that its benefits are fully realized in real-world applications.

In conclusion, this study highlighted the feasibility of developing a deep learning model that accurately identify pneumoperitoneum in abdominal CT scans. As a 3-dimensional model in medical image segmentation, PACT-3D maintained consistent performance across different testing periods. Its high specificity helps to avoid clinician fatigue due to false alarms, while its high sensitivity is particularly noteworthy in cases with larger volumes of free air. The model holds significant potential to aid rapid decision-making in emergency care, which could lead to improved patient outcomes.

Methods

The study follows the STARD protocol and has been approved by the Institutional Review Boards of Far Eastern Memorial Hospital (IRB

number: 111086-F) and Cedars-Sinai Medical Center (IRB number: STUDY00003494). All participant records were de-identified and anonymized before analysis. Informed consent was waived by the IRBs due to the retrospective nature of the study and that the dataset was de-identified before access. During the prospective phase, the model was deployed without any clinical interventions or changes to standard care. After the prospective period, data were retrieved from the research database, along with our model's predictions, for downstream analysis.

Study setting

In our research, we employed a dataset of post contrast abdominal CT scans from a single medical center, collected over a period spanning from January 2012 to December 2021. This dataset was enriched with CT scans indicating the presence or absence of pneumoperitoneum, a condition diagnosed using formal radiologist reports. For scans identified as positive for pneumoperitoneum, verification was performed by two radiologists who confirmed the presence of free air during the annotation process. To assess its applicability in a clinical setting, the model was prospectively validated from December 2022 to May 2023 in the same hospital for its performance in real-world data.

Image data acquisition

Abdominal CT scans during the study period were collected, and so does the corresponding reports. We included only the CT scans with contrast injection, axial plane scan, and reformatting slice thickness of 5 mm, with the field of view including the abdomen. CT scans with image acquisition and processing error, and CT scan without reports were excluded from this study. Figure 1 illustrates the recruitment and analysis flowchart.

Dataset collection and splitting

We employed natural language processing (NLP) methods to retrieve reports with and without a positive description of pneumoperitoneum from the image database (Supplementary 1). Initially, we utilized the NLP results as CT labeling and subsequently made minor revisions based on a random check of 1/5 of the CTs. We enrolled all CT scans that displayed pneumoperitoneum. The data was divided into training, validation, and test sets in a 5:1:1 ratio. To ensure no data leakage, CT scans from the same patient were exclusively allocated to the training set. For CT scans without pneumoperitoneum, we randomly selected non-duplicated patient scans, ensuring a 1:1 match with the pneumoperitoneum scans for both the training and validation sets. To mimic real-world conditions, our test set was formulated with a clinical ratio of 1:100 for positive to negative cases, reflecting an annual prevalence.

Image annotation

Two senior radiologists with both 13 years of experience radiologist manually segmented the free gas bubble on the axial section with a window width and center of 600 HU and 40 HU, respectively. Contouring of bowel gas was strictly prohibited. Later, the labeled pixels with CT number of corresponding image higher than −150HU were removed. Finally, another radiologist checked and revised all pneumoperitoneum annotations. Prior to using the data for training, we standardized all CT images by removing the window width settings and applying pixel normalization based on the maximal and minimal values.

Deep learning model and training

For pneumoperitoneum segmentation, we developed a 3D U-Net based neural network to predict the segmented mask of bowel gas (Fig. 2)[29]. Its design incorporates a contracting path to capture context,

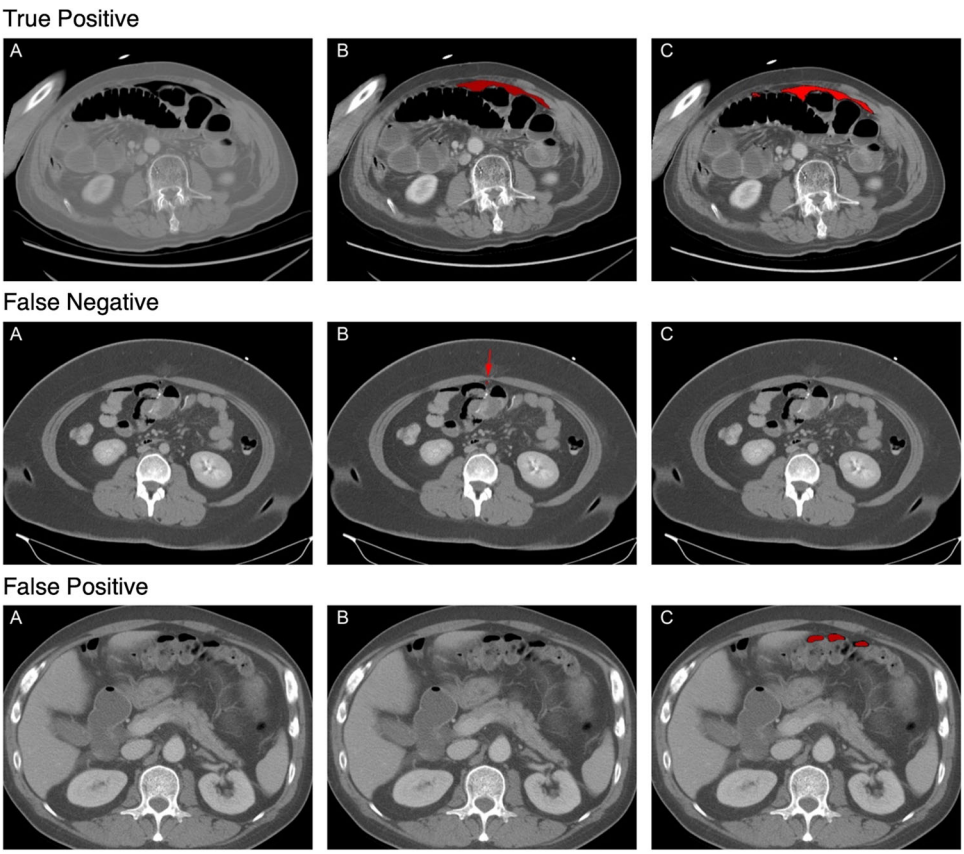

Fig. 2 | This figure illustrates three distinct outcomes of the model inference in the simulated test set, namely, "True Positive", "False Negative", and "False Positive". For each scenario: (**A**) represents the original CT scan image, (**B**) denotes the ground truth labeling, and (**C**) illustrates the mask generated by the trained segmentation model.

juxtaposed with a symmetric expanding path, which facilitates precise localization. In enhancing the network, successive layers replace traditional pooling operations with up-sampling operators, thereby refining the output resolution.

To augment the data, we normalized all CTs to 512 × 512 × z-axis and randomly cubed them to 384 × 384 × z-axis using the 'albumentations' library for each image in the training set[30]. The loss function we employed for the model combined Dice loss and Focal loss, each weighted at 50%. This approach aided in addressing class imbalance and enhanced accuracy for hard-to-classify examples[31]. We used an adaptive moment estimation (Adam) optimizer with parameter settings of $\beta1 = 0.9$ and $\beta2 = 0.999$, and a CosineAnnealingLR scheduler with parameter settings of T_max=8 and eta_min=3×10^{-6}. The model was trained with the Nvidia RTX A6000 GPU, with mini-batches of size 1 and an initial learning rate of 3×10^{-4}.

External validation

To ensure the generalizability and robustness of the PACT-3D model, we conducted an external validation using an international dataset of CT scans from Cedars-Sinai Medical Center (CSMC). This dataset included abdominal CTs with intravenous contrast injections performed between January and December 2023. We first identified positive CT scans for pneumoperitoneum using the same NLP method employed in the development dataset, which involved searching for positive descriptions of pneumoperitoneum in the CT reports. Negative control scans were then randomly selected in a 1:5 ratio and matched for age and gender. For each study, we analyzed post-contrast, axial plane scans. All CT scans were standardized by removing window width settings and applying pixel normalization based on the maximum and minimum values before model inference to generate prediction masks.

Performance evaluation and statistical analysis

The study aimed to evaluate the performance of the PACT-3D model in diagnosing pneumoperitoneum from abdominal CT scans, with continuous variables reported as means and SD, and categorical variables as counts and percentages. The model was trained to minimize loss within the validation dataset, and the optimized model weights were preserved for subsequent inference.

To assess the model's efficacy, we evaluated its predictive performance on both a simulated test set and a prospective test set. Our primary metrics for evaluation included F1-score, sensitivity, specificity, and PPV, were calculated alongside their 95% confidence intervals. Additionally, we conducted a subgroup analysis to explore how the model's performance varied across different etiologies such as gastroduodenal, small bowel, large intestine perforations, trauma, and post-operative cases. The modeling pipeline was implemented using Python (3.9) with PyTorch (2.0) and MONAI (1.3.0) as the deep learning framework. Image processing and data analysis were facilitated by Python libraries such as SimpleITK (2.2.1), scikit-image (0.20.0), pandas (2.0.2), and matplotlib (3.7.1), while SPSS was utilized for all subsequent statistical analyses.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Given the sensitive nature of patient data and privacy regulations, we maintain strict control over data access to ensure compliance with institutional policies and legal requirements. To request access to the data, researchers must provide a detailed research proposal outlining the intended use of the data. Access requests can be directed to goman178@gmail.com, and we aim to respond within 30 days. Approval will be contingent upon ethical review and the establishment of a data use agreement to safeguard patient privacy. Data usage will be restricted to non-commercial research purposes, and no attempts to re-identify individuals will be permitted.

Code availability

The code developed to support the findings of this study is available online at https://github.com/IMinChiu/pact-3d.

References

1. Mularski, R. A., Sippel, J. M. & Osborne, M. L. Pneumoperitoneum: a review of nonsurgical causes. *Crit. Care Med.* **28**, 2638–2644 (2000).
2. Makki, A. M. The pattern of causes of pneumoperitoneum-induced peritonitis: results of an empirical study. *J. Microsc Ultrastruct.* **5**, 28–31 (2017).
3. Larsen, N. E., Mikkelsen, E., Knudsen, A. R. & Larsen, L. P. Low-dose CT for diagnosing intestinal obstruction and pneumoperitoneum; need for retakes and diagnostic accuracy. *Acta Radio. Open* **10**, 2058460121989313 (2021).
4. Ordoñez, C. A. & Puyana, J. C. Management of peritonitis in the critically ill patient. *Surg. Clin. North Am.* **86**, 1323–1349 (2006).
5. van Ruler, O. et al. Comparison of on-demand vs planned relaparotomy strategy in patients with severe peritonitis: a randomized trial. *Jama* **298**, 865–872 (2007).
6. Mills, A. M. et al. The impact of crowding on time until abdominal CT interpretation in emergency department patients with acute abdominal pain. *Postgrad. Med.* **122**, 75–81 (2010).
7. Kocher, K. E. et al. National trends in use of computed tomography in the emergency department. *Ann. Emerg. Med.* **58**, 452–462.e3 (2011).
8. Saha, A., Roland, R. A., Hartman, M. S. & Daffner, R. H. Radiology medical student education: an outcome-based survey of PGY-1 residents. *Acad. Radio.* **20**, 284–289 (2013).
9. Wechsler, R. J. et al. Effects of training and experience in interpretation of emergency body CT scans. *Radiology* **199**, 717–720 (1996).
10. Tieng, N., Grinberg, D. & Li, S. F. Discrepancies in interpretation of ED body computed tomographic scans by radiology residents. *Am. J. Emerg. Med.* **25**, 45–48 (2007).
11. Ruchman, R. B. et al. Preliminary radiology resident interpretations versus final attending radiologist interpretations and the impact on patient care in a community hospital. *Am. J. Roentgenol.* **189**, 523–526 (2007).
12. Immonen, E. et al. The use of deep learning towards dose optimization in low-dose computed tomography: A scoping review. *Radiography* **28**, 208–214 (2022).
13. Meedeniya, D. et al. Chest X-ray analysis empowered with deep learning: a systematic review. *Appl. Soft Comput.* 109319 (2022).
14. Shen, Y.-T., Chen, L., Yue, W.-W. & Xu, H.-X. Artificial intelligence in ultrasound. *Eur. J. Radiol.* **139**, 109717 (2021).
15. Cheng, C. Y. et al. Deep learning assisted detection of abdominal free fluid in Morison's pouch during focused assessment with sonography in Trauma. *Front. Med. (Lausanne)* **8**, 707437 (2021).
16. Chiu, I.-M. et al. Use of a deep-learning algorithm to guide novices in performing focused assessment with sonography in trauma. *JAMA Netw. Open* **6**, e235102–e235102 (2023).
17. Lu, C.-Y., et al. Artificial intelligence application in skull bone fracture with segmentation approach. *J. Imag. Inform. Med.* 1–16 (2024).
18. Chiu, I. M. et al. Prospective clinical evaluation of deep learning for ultrasonographic screening of abdominal aortic aneurysms. *npj Digit. Med.* **7**, 282 (2024).
19. Taubmann, O. et al. *Automatic detection of free intra-abdominal air in computed tomography* 232–241 (Springer, 2020).
20. Brejnebøl, M. W., Nielsen, Y. W., Taubmann, O., Eibenberger, E. & Müller, F. C. Artificial Intelligence based detection of

pneumoperitoneum on CT scans in patients presenting with acute abdominal pain: a clinical diagnostic test accuracy study. *Eur. J. Radiol.* **150**, 110216 (2022).
21. Wessel, J. et al. Sequential rib labeling and segmentation in chest X-ray using Mask R-CNN. *arXiv preprint arXiv:190808329* (2019).
22. Cheng, C.-Y. et al. Development and validation of a deep learning pipeline to measure pericardial effusion in echocardiography. *Front. Cardiovasc. Med.* **10**, 1195235 (2023).
23. Almotairi, S., Kareem, G., Aouf, M., Almutairi, B. & Salem, M. A.-M. Liver tumor segmentation in CT scans using modified SegNet. *Sensors* **20**, 1516 (2020).
24. Tiwari, A., Srivastava, S. & Pant, M. Brain tumor segmentation and classification from magnetic resonance images: Review of selected methods from 2014 to 2019. *Pattern Recognit. Lett.* **131**, 244–260 (2020).
25. Baheti, B., Pati, S., Menze, B. & Bakas, S. Leveraging 2D deep learning imagenet-trained models for native 3D medical image analysis. *Brainlesion* **13769**, 68–79 (2023).
26. Thorisson, A. et al. Diagnostic accuracy of acute diverticulitis with unenhanced low-dose CT. *BJS Open* **4**, 659–665 (2020).
27. Ali, M., Iqbal, J. & Sayani, R. Accuracy of computed tomography in differentiating perforated from nonperforated appendicitis, taking histopathology as the gold standard. *Cureus* **10**, e3735 (2018).
28. Martín-Román, L. et al. Relevance of pneumoperitoneum in the conservative approach to complicated acute diverticulitis. A retrospective study identifying risk factors associated with treatment failure. *Minerva Surg.* **77**, 327–334 (2022).
29. Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T. & Ronneberger, O. *3D U-Net: learning dense volumetric segmentation from sparse annotation* 424–432 (Springer, 2016).
30. Buslaev, A. et al. Albumentations: fast and flexible image augmentations. *Information* **11**, 125 (2020).
31. Lin, T.-Y., Goyal, P., Girshick, R., He, K. Dollar, P. Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2980–2988 *arXiv preprint arXiv:170802002* (2017).

Acknowledgements

Kuei-Hong Kuo contributed equally with I-Min Chiuo as first author. I-Min Chiu contributed equally with Kuei-Hong Kuo as corresponding author. This study was supported by grants from the Ministry of Health and Welfare (MOHW-113-IM-I-212-000013-15), and the Ministry of Science and Technology (NSTC 111-2314-B-418-002 and NSTC 112-2321-B-075A-002), all awarded to KH.H.

Author contributions

Conceptualization: I.M.C., K.H.K., Methodology: I.M.C., T.Y.H. Investigation: I.M.C., W.C.L., Y.J.P., Visualization: D.O., C.Y.L., Validation: D.O., W.C.L., I.M.C., Supervision: T.Y.H., K.H.K., Writing—original draft: I.M.C., Y.J.P., Writing—review & editing: T.Y.H., D.O., W.C.L., C.Y.L., K.H.K.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-024-54043-1.

Correspondence and requests for materials should be addressed to I-Min Chiu or Kuei-Hong Kuo.

Peer review information *Nature Communications* thanks Anselm Schulz, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.