

PACT-3D, a Deep Learning Algorithm for Pneumoperitoneum Detection in Abdominal CT Scans

Corresponding Author: Dr Kuei-Hong Kuo

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

This paper describes the application of a 3D unit for detection of pneumoperitoneum on CT scans acquired in the emergency setting. The developed network achieved good specificity but rather poor sensitivity which is a concern if used for triage purposes. It would seem better to have higher sensitivity and lower specificity if this is used as a triage mechanism. The description of the machine learning methods is fine but the description of the data set is confusing. They state: "139,781 CT scans for further analysis. Notably, 973 of these scans were 59 radiologically confirmed to depict pneumoperitoneum. We randomly allocated the CT scans 60 that presented pneumoperitoneum. The training set consisted of 1,390 scans, with 695 scans 61 showing pneumoperitoneum. The validation set included 278 scans, of which 139 displayed 62 pneumoperitoneum." Figure 1 helps to clarify this, so I think it was done correctly, but they should update their words to match the figure.

The distribution of scanners is also described in a confusing fashion period. It appears that the training set was on one set of scanners and the validation or testing was done on another set of scanners period.

This is a rare but important problem, and I see potential in this manuscript, but the data sourcing methods needs to be clarified. The low sensitivity reduces the interest level in this manuscript.

Reviewer #2

(Remarks to the Author)

General Comments:

This paper aims to investigate the accuracy detection of pneumoperitoneum of a deep learning model (PACT 3D), based on 3D U-NET. In this prospective single centre study, the authors developed a 3D U-Net based deep learning model with the goal to improve the pneumoperitoneum detection in CT images using CT scans from January 2012 to December 2021; the model was trained with post contrast CT scans, in comparison with radiologist reports.

The authors have evaluated PACT 3D using a simulated test set and a real-world prospective validation in the same centre. PACT 3D showed a sensitivity in pneumoperitoneum detection of 0.81 and a specificity of 0.99. Also, the model demonstrated a higher sensitivity for gastroduodenal and small bowel perforations, offering a potential significant diagnostic tool.

The object of the study is interesting and adequately presented; however, the manuscript has some limitations that need to be addressed.

Please see specific comments.

Specific comments:

Title: I suggest to shorten the title

Abstract:

- P1L13: Please add statistics.
- Briefly define inclusion and exclusion criteria of your study
- I suggest modifying the conclusions section by providing a more concise and consistent take-home message.

Introduction:

- P2L50: Please define "real world clinical settings"

- I suggest to briefly define in this chapter what a 3D U-Net model is.
- Introduction is more focused on time efficiency than detection. The latter being heavily dependant on reader's expertise and magnitude of free air. Please add some words on that.

Results:

- Overall, the section is well written and the section's division in subheadings helps the reader in understanding the main results of the study
- I suggest dividing the Paragraph 3.1 in two different paragraphs, one for the Demographic characteristics and for the Distributions of Ct scans.
- Exclusion criteria are mentioned in this section but not described in materials and methods.
- Diagnostic accuracy: confidence intervals must be reported.
- There is no reference in the text regarding Table 1. Please add

Discussion:

- P5L118: please add references
- P5L124: add references
- P7L150: please define how you've evaluated all these aspects; in particular have you performed a satisfactory analysis of these model among the radiologists that have used it?
- Also, how have you evaluated the eventual change of patients' outcomes? Please clarify
- I suggest enhancing more the clinical value that your paper might have in clinical setting.

Materials and Methods:

- Please add details of the acquisition (i.e. what post-contrast phase of acquisition have you chosen for your analysis? How you've evaluated the amount of contrast media administered to the patient?)
- "dataset was enriched with CT scans indicating the presence or absence of pneumoperitoneum, a condition diagnosed using formal radiologist reports." Do the authors mean that diagnosis relied only on CT reports? Please clarify.
- "poor image quality" ought to be defined.
- How have you evaluated the image quality of the Ct scans? Please clarify
- I suggest enhancing the statistical analysis paragraph with more details.
- Are the authors sure W/L settings were 380/40? That sounds like a non-optimal choice.

References: Please see comments above.

Tables: ok

Figures: ok

Linguistic and typewriting: English writing and punctuation needs some significant improvements.

Version 1:

Reviewer comments:

Reviewer #3

(Remarks to the Author)

Reviewer 1 comments have been sufficiently addressed. However, I do have additional comments. The major concern I have is that the specificity and sensitivity have been manipulated (with reason) by removing the small pneumoperitoneum and specifying the prevalence of pneumoperitoneum to improve the performance. This change while I understand is in response to the reviewer comment I think is misleading. By making this change, I am now increasingly concerned about the real world performance (in a different institution) of this model. I feel that the paper contributes new knowledge specifically to these areas and most important on how this task appears to be a challenging task.

Novelty of the work includes.

- Testing on multiple machines across different time points
- Prospective validation of the model
- Making the code available

- Cohort building: There is a gap in the next evaluation to show how many of these cases were missed. This will likely be deployed in a triage environment so real world performance would be good (Can look at the report concordance and discordance in the simulated and prospective cohorts.

Overall in making this change , the writing of the paper is now challenging for a reader to understand – should we use AI for this task ? and when should we use it? Focussing on the strengths and strategies that can improve the model performance can allow us to really understand the value of AI for this task. The content is there , but I worry that its not clear in the writing and I had to infer it . I would expound on the following section

1. 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. – add these examples or have the radiologist systematically evaluate why these cases were difficult

Some minor comments

1. Clarify this new reference as to radiologists or surgical residents or ER doctors - According to previous research, only 62.8% of junior physicians feel confident about diagnosing acute pathological findings from CT scans, such as pneumoperitoneum or bowel obstruction

Reviewer #4

(Remarks to the Author)

The authors did a good job responding to the raised comments. The document is significantly improved.

A minor issue remains, there is a mismatch between line 270 "The data was divided into training, validation, and test sets in a 5:1:1 ratio" and line 83 "training, validation, and test datasets in an 8:1:1 ratio"

Congratulations on this interesting study!

Version 2:

Reviewer comments:

Reviewer #3

(Remarks to the Author)

Thank you again for the opportunity to review the revised paper. My concerns have been addressed, and I believe that the paper is much more robust with good external validation, outside of the original geographic region with the inclusion of Mount Sinai. There is now more clarification even showing the failure modes of the algorithm and a lot of consideration in terms of the sensitivity and the impact of specificity on how real-world deployment would matter for this.

Major comment

1. My main concern is that the references used do not consistently match the content. For example, on line 352, reference 27 is cited for the statement that "this approach aided in addressing class imbalance and enhanced accuracy for hard-to-classify examples by using this combination of dice loss and focal loss." However, the paper referenced is inaccurate and not relevant. I did not have the opportunity to cross-check every reference, but I recommend that the team thoroughly reviews and verifies all references to ensure their accuracy and relevance.

Minor comments:

1. On line 85, I think that sentence is not clear and needs to be rewritten. This is the original sentence: "In this study, we introduced PACT- 3D, a three-dimensional U-Net algorithm is a convolutional neural network." It probably should be: "In this study, we introduced PACT- 3D, a three-dimensional U-net algorithm based on a convolutional neural network...." Feel free to edit with your suggestion.

2. On line 99, we see the split of 5.1.1. This is usually a little bit odd because typically, everything should add up to a total of 100%. I will defer to the editors as this is an unusual way of representing the data splits.

3. On line 124, there is an error: "AS scanners were used less frequently, constituting 43.6.9% of the scans." This needs to be corrected

4. Clarify if the numbers in Figure 1 represent studies or patient numbers

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Reviewer #1 (Remarks to the Author):

This paper describes the application of a 3D unit for detection of pneumoperitoneum on CT scans acquired in the emergency setting. The developed network achieved good specificity but rather poor sensitivity which is a concern if used for triage purposes. It would seem better to have higher sensitivity and lower specificity if this is used as a triage mechanism.

Answer:

We appreciate the thoughtful feedback regarding the sensitivity and specificity balance in our study. In response to your concerns about its application as a triage tool, we have undertaken a thorough evaluation of the model's performance. Our approach involved optimizing the F1-score by varying the the ratio of positive to negative cases during training since simply adjusting threshold for model output did not significant impact the outcome. As shown in supplemental table 2, when we adjusted the training ratio to increase sensitivity, this was accompanied by a significant reduction in specificity, positive predictive value, and also F1-score. Given that pneumoperitoneum is a relatively rare condition, a PPV below 0.1 would likely result in an unacceptably high rate of false positives, which could lead to the clinical dismissal of the tool due to frequent false alarms. Therefore, in the manuscript, we have chosen to report both the original model performance and the adjusted results that highlight the improved sensitivity. We believe this dual reporting method provides a comprehensive view of the model's capabilities and allows for an informed assessment of its utility in a triage scenario.

Furthermore, we have conducted a subanalysis on the total volume of free air present in scans. Our findings indicate that by excluding scans with less than 1ml of free air—volumes typically associated with conditions like diverticulitis or appendicitis—the model's sensitivity can be further increased to approximately 0.9. Interestingly, this refined sensitivity also correlates with a higher rate of urgent surgeries among cases identified by the model, underscoring the potential clinical significance of the detected pneumoperitoneum. We trust that these revisions address your concerns and clarify the potential clinical application of our deep learning model in emergency settings.

The description of the machine learning methods is fine but the description of the data set is confusing. They state: "139,781 CT scans for further analysis. Notably, 973 of these scans were 59 radiologically confirmed to depict pneumoperitoneum. We randomly allocated the CT scans 60 that presented pneumoperitoneum. The training set consisted of 1,390 scans, with 695 scans 61 showing pneumoperitoneum. The validation set included 278 scans, of which 139 displayed 62 pneumoperitoneum. " Figure 1 helps to clarify this, so I think it was done correctly, but they should update their words to match the figure.

Answer:

Thank you for the feedback. We understand the description of our dataset division may not be clear as it could be. We have revised the paragraph to better and to align with Figure 1 as follow:

"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 an 8:1:1 ratio (Figure 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.”

The distribution of scanners is also described in a confusing fashion period. It appears that the training set was on one set of scanners and the validation or testing was done on another set of scanners period.

Answer:

Thank you for your feedback regarding the distribution of scanners. To clarify, the simulated test dataset, along with training and validation set was compiled from scans conducted between January 2012 and December 2021, utilizing a variety of CT scanners available during that period. The prospective test dataset, however, includes scans from December 2022 to May 2023. During this more recent timeframe, certain CT scanner models that were previously in use had been phased out. This temporal distinction between the datasets ensures that the model's performance is assessed across a range of CT technologies, including both older and current models, thereby enhancing the evaluation of its generalizability. We acknowledge the importance of this point and will provide a more detailed discussion in the manuscript to address any potential confusion.

This is a rare but important problem, and I see potential in this manuscript, but the data sourcing methods needs to be clarified. The low sensitivity reduces the interest level in this manuscript.

Answer:

Thank you for your valuable efforts in reviewing this article and providing helpful suggestions. In the revised manuscript, we have addressed most of the issues and highlighted them in red color. We have also provided a point-by-point response to your comments. Please note that during the revision process, we discovered some duplicated scan counts in the prospective dataset. We have re-conducted the analysis and presented the correct results in the revised manuscript.

Reviewer #2 (Remarks to the Author):

General Comments:

This paper aims to investigate the accuracy detection of pneumoperitoneum of a deep learning model (PACT 3D), based on 3D U-NET. In this prospective single centre study, the authors developed a 3D U-Net based deep learning model with the goal to improve the pneumoperitoneum detection in CT images using CT scans from January 2012 to December 2021; the model was trained with post contrast CT scans, in comparison with radiologist reports.

The authors have evaluated PACT 3D using a simulated test set and a real-world prospective validation in the same centre. PACT 3D showed a sensitivity in pneumoperitoneum detection of 0.81 and a specificity of 0.99. Also, the model demonstrated a higher sensitivity for gastroduodenal and small bowel perforations, offering a potential significant diagnostic tool.

The object of the study is interesting and adequately presented; however, the manuscript has some limitations that need to be addressed.

Thank you for your valuable efforts in reviewing this article and providing helpful suggestions. In the revised manuscript, we have addressed most of the issues and highlighted them in red color. Please note that during the revision process, we discovered some duplicated scan counts in the prospective dataset. We have re-conducted the analysis and presented the correct results in the revised manuscript. Below, we provide a point-by-point response to your comments.

Please see specific comments.

Specific comments:

Title: I suggest to shorten the title

Answer:

Thank you for the suggestion. We've shorten the title in the revised manuscript. Now the title is "PACT-3D, a Deep Learning Algorithm for Pneumoperitoneum Detection in Abdominal CT Scans"

Abstract:

- P1L13: Please add statistics.
- Briefly define inclusion and exclusion criteria of your study
- I suggest modifying the conclusions section by providing a more concise and consistent take-home message.

Answer: Thank you for the recommendation. We've revised the abstract under these suggestions.

Introduction:

- P2L50: Please define "real world clinical settings"

Answer: We've rephrased this term in the revised manuscript and give more detailed describing how the model should be validated.

- I suggest to briefly define in this chapter what a 3D U-Net model is.

Answer: Thank you for the suggestion. We've added a paragraph on brief definition of our model and what it's designed for.

- Introduction is more focused on time efficiency than detection. The latter being heavily dependant on reader's expertise and magnitude of free air.

Answer:

Thank you for your insightful comments regarding the focus of the introduction in our manuscript. As suggested, I have revised the section to place greater emphasis on the challenges associated with the detection of pneumoperitoneum, particularly highlighting the dependency on the radiologist's expertise and the volume of free air detectable by CT scan. I have also included findings from a study that directly correlates misinterpretations with adverse patient management outcomes, thus addressing the importance of detection over mere time efficiency. I believe that these changes align the introduction more closely with the realities of clinical practice and the intricacies of diagnostic processes in emergency settings.

Results:

- Overall, the section is well written and the section's division in subheadings helps the reader in understanding the main results of the study
- I suggest dividing the Paragraph 3.1 in two different paragraphs, one for the Demographic characteristics and for the Distributions of Ct scans.

Answer: We've divided the 2 paragraphs into 3.1 and 3.1 subheadings.

- Exclusion criteria are mentioned in this section but not described in materials and methods.

Answer: The exclusion criteria were described in Method 2.2 Image Data Acquisition with figured 2 cited. We've revised the term poor quality to detailed as "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."

- Diagnostic accuracy: confidence intervals must be reported.

Answer: We've added confidence intervals in reporting the diagnostic metrics.

- There is no reference in the text regarding Table 1. Please add

Answer: Thank you for noticing it. We've added the reference of Table 1 in the paragraph in 3.1 and 3.2.

Discussion:

- P5L118: please add references

Answer: The reference was added.

- P5L124: add references

Answer: The reference was added.

- P7L150: please define how you've evaluated all these aspects; in particular have you performed a satisfactory analysis of these model among the radiologists that have used it?
- Also, how have you evaluated the eventual change of patients' outcomes? Please clarify
- I suggest enhancing more the clinical value that your paper might have in clinical setting.

Answer:

Our study was indeed focused on the retrospective and prospective analysis of the PACT-3D model's performance. While the interaction between radiologists and the model was not directly examined during this phase of our research, we acknowledge the significance of such engagement and we are committed to conducting future studies that will assess the impact of PACT-3D on radiologists' and emergency clinicians' diagnostic processes, specifically their confidence and satisfaction levels. This will yield a more detailed understanding of the model's practical utility in clinical contexts.

Regarding the integration of our model's output into clinical workflows, it is true that this was not within the scope of the current study. However, our additional analyses have revealed a meaningful clinical correlation with the model's predictions. Notably, PACT-3D demonstrated expert-level accuracy (0.95-0.98) in identifying free air volumes greater than 10ml, commonly associated with conditions necessitating urgent surgical intervention. This finding aligns with our sub-analysis, which indicated that cases of pneumoperitoneum predicted by the model were more frequently associated with the need for urgent surgery. These promising results bolster our confidence in considering PACT-3D's integration into clinical workflows. The manuscript will be updated to reflect this clarification and to detail our prospective research initiatives.

Materials and Methods:

- Please add details of the acquisition (i.e. what post-contrast phase of acquisition have you chosen for your analysis? How you've evaluated the amount of contrast media administered to the patient?)

Answer: Thank you for the question. We've added a supplement table describing image acquisition regarding 6 CT scanner utilized in the study hospital.

- "dataset was enriched with CT scans indicating the presence or absence of pneumoperitoneum, a condition diagnosed using formal radiologist reports." Do the authors mean that diagnosis relied only on CT reports? Please clarify.

Answer:

For scans indicative of pneumoperitoneum, dual radiologist verification was implemented to confirm the presence of free air. For scans without pneumoperitoneum, we relied on the accuracy of the original radiologist reports.

- “poor image quality” ought to be defined.
- How have you evaluated the image quality of the Ct scans? Please clarify

Answer:

We acknowledge that "poor image quality" was imprecisely defined. Specifically, we excluded CT scans that encountered errors during retrieval from our database, such as those that could not be read by our image reader due to file corruption or format errors. We have clarified this criterion and refined the relevant descriptions in the Methods section of our manuscript.

- I suggest enhancing the statistical analysis paragraph with more details.

Answer:

Thanks for the suggestion, we've elaborated more details in the paragraph about how data were presented, about subgroup analysis, and what software and framework we used.

- Are the authors sure W/L settings were **380/40**? That sounds like a non-optimal choice.

Answer:

Thank you for the noticing. The W/L settings were 600/40 during the annotation stage, and we applied maximal-minimal normalization rather than W/L during the training process. We'll correct it in the revision.

References: Please see comments above.

Tables: ok

Figures: ok

Linguistic and typewriting: English writing and punctuation needs some significant improvements.

Answer:

Thank you for your valuable. We have sought the assistance of an English editing service to thoroughly revise the language used throughout our document. We trust that these revisions have addressed your concerns, and we appreciate the opportunity to enhance the quality of our work.

Reviewer #3

Reviewer 1 comments have been sufficiently addressed. However, I do have additional comments. The major concern I have is that the specificity and sensitivity have been manipulated (with reason) by removing the small pneumoperitoneum and specifying the prevalence of pneumoperitoneum to improve the performance. This change while I understand is in response to the reviewer comment I think is misleading. By making this change, I am now increasingly concerned about the real world performance (in a different institution) of this model. I feel that the paper contributes new knowledge specifically to these areas and most important on how this task appears to be a challenging task.

Novelty of the work includes.

- Testing on multiple machines across different time points
- Prospective validation of the model
- Making the code available
- Cohort building: There is a gap in the next evaluation to show how many of these cases were missed. This will likely be deployed in a triage environment so real world performance would be good (Can look at the report concordance and discordance in the simulated and prospective cohorts.

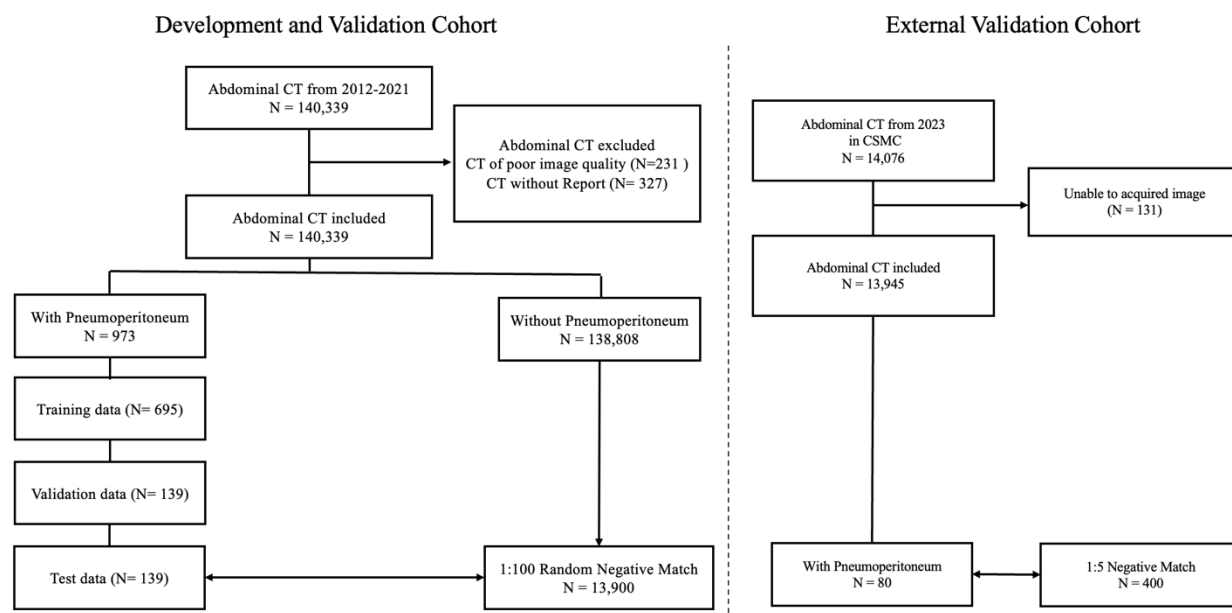
Overall in making this change , the writing of the paper is now challenging for a reader to understand – should we use AI for this task ? and when should we use it? Focussing on the strengths and strategies that can improve the model performance can allow us to really understand the value of AI for this task. The content is there , but I worry that its not clear in the writing and I had to infer it .

Response: Thank you for your valuable comments. We have conducted additional analysis by externally validating the model on a geographically diverse international dataset and revised the manuscript accordingly. We found that the PACT-3D model demonstrated consistent performance across institutions, with sensitivity ranging from 0.81 to 0.83 and specificity ranging from 0.97 to 0.99 across the hold-out test set, prospective test set, and external test set. The model's performance was reliable across different institutions, consistently detecting pneumoperitoneum that required urgent intervention, which will help to accelerate the diagnostic and treatment workflow in emergency care setting.

The inclusion flowchart, demographics and performance of the external validation cohort were shown in the method and result,

“

Figure 1.



3.4 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.

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	6,351	480
Age	54 (13.1)	59 (16.9)	57 (19.0)
Female	6,767 (48.2%)	3,000 (47.2%)	204 (42.5%)
CT Vendors			

Philips Brilliance 64	1,123 (8.0%)		
Siemens Somatom definition	1,502 (10.7%)		
Siemens Somatom definition Flash	772 (5.5%)	524 (8.3%)	8 (1.7%)
Siemens Somatom definition AS	624 (60.1%)	2,772 (43.6%)	
GE LightSpeed VCT	2,204 (15.7%)	1,479 (23.3%)	33 (6.9%)
GE Revolution Frontier		1,576 (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%)

Table 2. Performance of PACT-3D in Test Set

Performance Metrics	Simulated Test Set	Prospective Test Set	External Test Set
	value (95% CI)	value (95% CI)	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 > 1ml	0.89 (0.84-0.93)	0.91 (0.86-0.95)	0.86 (0.75-0.93)
Total volume > 10ml	0.95 (0.90-0.98)	0.98 (0.93-1.0)	0.92 (0.80-0.97)

Additionally, we've clarified the use case and included the results in both the abstract and discussion sections.

In abstract,

"...Additionally, external validation was conducted on an international cohort using 480 CT scans from Cedars-Sinai Medical Center. PACT-3D achieved a sensitivity of 0.81 and a specificity of 0.99 in retrospective testing. In prospective validation, the model yielded similar performance with a sensitivity of 0.83 and a specificity of 0.99. External validation further demonstrated a sensitivity of 0.81 and a specificity of 0.97. Sensitivity improved to 0.95, 0.98, and 0.92 in the simulated, prospective, and external test sets, respectively, when cases with a small amount of free air (total volume < 10 ml) were excluded. By delivering accurate and consistent predictions, along with providing segmented masks, PACT-3D holds the potential to accelerate the diagnostic and treatment workflow in emergency care setting."

In 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.....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....

... 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 the model incorrectly identified

pneumoperitoneum, a review of the prediction masks revealed that most errors were due to the model mistakenly identifying air-containing abscesses, air bubbles in the lung, around distended bowel gas, or air density artifacts related to artificial implants. Although these cases were not correctly diagnosed as pneumoperitoneum, 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.

”

1. 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. – add these examples or have the radiologist systematically evaluate why these cases were difficult

Response: Thank you for the excellent suggestion! We have manually reviewed both the false positive and false negative predictions and provided some of the common reasons for these errors in the revised manuscript. Additionally, we have attached the original CT slices along with the prediction masks as a supplemental figure to help readers understand more easily.

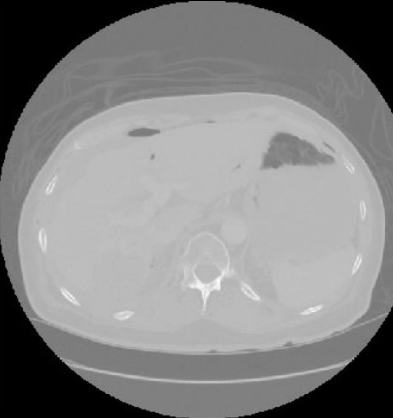
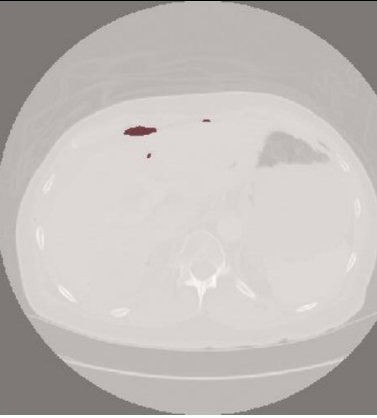
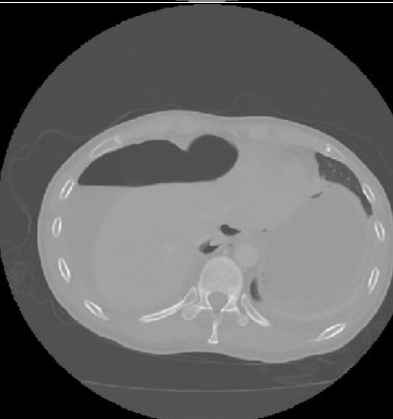
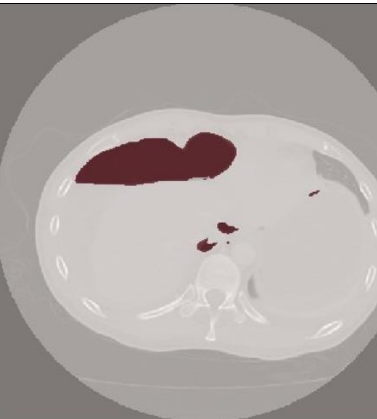
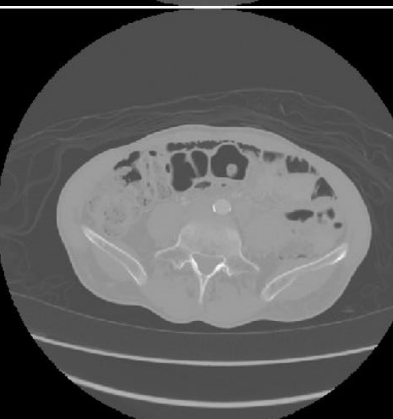
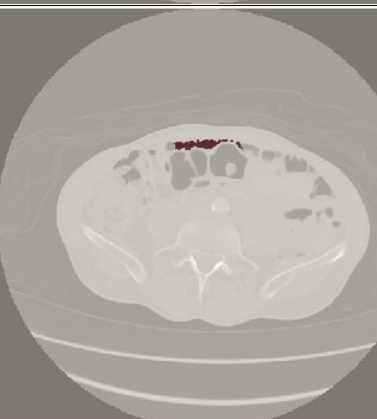
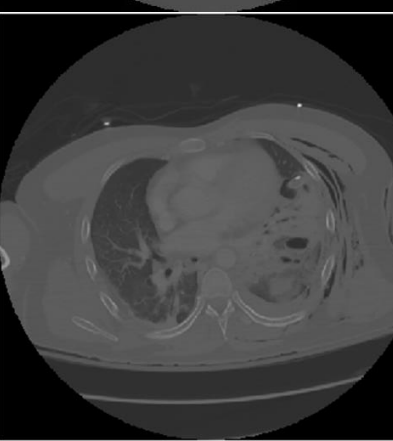
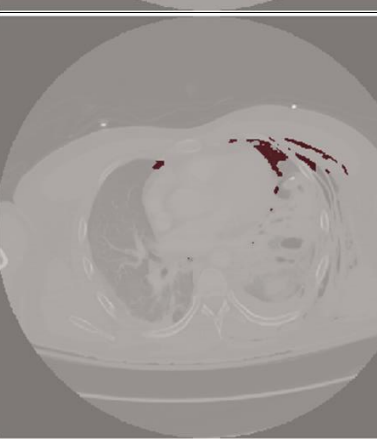
In the discussion,

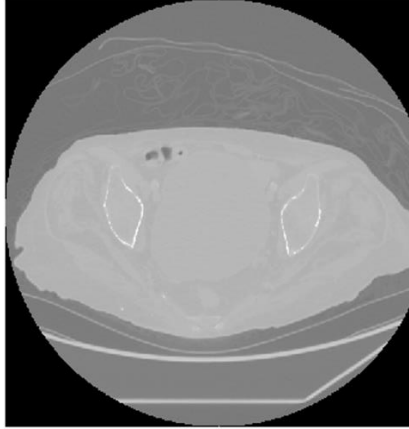
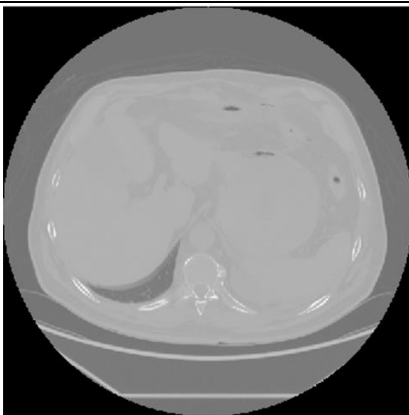
“The missed cases in both the simulated and prospective test sets highlight an important aspect of the model's performance in real-world settings. Upon reviewing the cases that PACT-3D failed to predict, we found that most missed instances involved free air that was scattered and appeared in retroperitoneal 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 (Supplemental 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.

in Supplemental figure 1,

Supplemental Table 3. Manual review of CT scans in true positive, false positive, and false negative prediction.

Prediction	Reason	Original image	Image with prediction mask
------------	--------	----------------	----------------------------

<p>True positive</p>			
<p>True positive</p>			
<p>False positive</p>	<p>Incorrectly identifies air in a distended, overlapping bowel lumen as pneumoperitoneum.</p>		
<p>False positive</p>	<p>Incorrectly identifies air in the lung and subcutaneous emphysema as pneumoperitoneum.</p>		

False negative	Report: a few punctate foci of pneumoperitoneum of unclear etiology		
False negative	Report: punctate pneumoperitoneum in the upper abdomen		

Some minor comments

1. Clarify this new reference as to radiologists or surgical residents or ER doctors - According to previous research, only 62.8% of junior physicians feel confident about diagnosing acute pathological findings from CT scans, such as pneumoperitoneum or bowel obstruction⁸

Response: Thank you for the comment. The previous research conducted a survey targeting postgraduate year residents to assess their confidence in interpreting radiology images, including X-rays and CTs, with a specific focus on abdominal CTs. We will rephrase the sentence to clarify that it refers to postgraduate year residents.

In the introduction,

“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 obstruction⁸”

Reviewer #4 (Remarks to the Author):

The authors did a good job responding to the raised comments. The document is significantly improved.

A minor issue remains, there is a mismatch between line 270 “The data was divided into training, validation, and test sets in a **5:1:1 ratio**” and line 83 “training, validation, and test datasets in an **8:1:1 ratio**”

Reply: thank you for pointing this out. The data was divided into training, validation, and test in 5:1:1 ratio. We have edited the typo in the revised manuscript.

Congratulations on this interesting study!

Reviewer #3

Major comment

1. My main concern is that the references used do not consistently match the content. For example, on line 352, reference 27 is cited for the statement that "this approach aided in addressing class imbalance and enhanced accuracy for hard-to-classify examples by using this combination of dice loss and focal loss." However, the paper referenced is inaccurate and not relevant. I did not have the opportunity to cross-check every reference, but I recommend that the team thoroughly reviews and verifies all references to ensure their accuracy and relevance.

Response: Thank you for the notice. We found that the citations in the methods section were not converted correctly along with the other paragraphs. We have revised this in the manuscript.

In the method,

“To augment the data, we normalized all CTs to $512 \times 512 \times z$ -axis and randomly cubed them to $384 \times 384 \times z$ -axis using the ‘alumentations’ library for each image in the training set³⁰. 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 examples³¹.”

30. Buslaev A, Iglovikov VI, Khvedchenya E, Parinov A, Druzhinin M, Kalinin AA. Alumentations: Fast and Flexible Image Augmentations. *Information*. 2020;11(2):125.

31. Lin T. Focal Loss for Dense Object Detection. *arXiv preprint arXiv:170802002*. 2017;”

Minor comments:

1. On line 85, I think that sentence is not clear and needs to be rewritten. This is the original sentence: "In this study, we introduced PACT- 3D, a three-dimensional U-Net algorithm is a convolutional neural network." It probably should be: "In this study, we introduced PACT- 3D, a three-dimensional U-net algorithm based on a convolutional neural network...." Feel free to edit with your suggestion.

Response: Thank you for pointing this out. The sentence was indeed unclear. We have revised it as follows in the updated version.

Introduction,

“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.”

2. On line 99, we see the split of 5:1:1. This is usually a little bit odd because typically, everything should add up to a total of 100%. I will defer to the editors as this is an unusual way of representing the data splits.

Response: We understand your points. We used a 5:1:1 split because the number of positive scans in the dataset is 973, which is divisible by 7. This allowed us to have a balanced and sufficient number of cases for both validation and testing.

3. On line 124, there is an error: "AS scanners were used less frequently, constituting 43.6.9% of the scans." This needs to be corrected

Response: Thank you for pointing this out, we've corrected the typo in the revised manuscript.

In result,

“Siemens Somatom Definition AS scanners were used less frequently, constituting 43.6% of the scans.”

4. Clarify if the numbers in Figure 1 represent studies or patient numbers

Response: The numbers represent studies in Figure 1. We've added this clarification in the figure legend.

Figure Legend of Figure 1,

“The inclusion flowchart of this study. 'N' represents the number of CT studies at each step.”