Diagn Interv Radiol 2024; DOI: 10.4274/dir.2025.253567

ARTIFICIAL INTELLIGENCE AND INFORMATICS

ORIGINAL ARTICLE

Automated evaluation of pulmonary lesion changes on chest radiograph during follow-up using semantic segmentation

Youngjae Kim[1,2*]
Yura Ahn[3*]
Sang Min Lee[3]
Han Na Noh[4]
Jongjun Won[2]
Chaewon Kim[2]
Hyunna Lee[5]

[1]University of Ulsan Faculty of Medicine, Department of Biomedical Engineering, AMIST, Asan Medical Center, Seoul, Republic of Korea

[2]University of Ulsan Faculty of Medicine, Department of Convergence Medicine, Asan Medical Institute of Convergence Science and Technology, Asan Medical Center, Seoul, Republic of Korea

[3]University of Ulsan Faculty of Medicine, Department of Radiology and Research Institute of Radiology, Asan Medical Center, Seoul, Republic of Korea

[4]University of Ulsan Faculty of Medicine, Health Screening and Promotion Center, Asan Medical Center, Seoul, Republic of Korea

[5]Bigdata Research Center, Asan Institute for Life Science, Asan Medical Center, Seoul, Republic of Korea

PURPOSE

To develop and validate a deep learning-based model utilizing lesion-specific segmentation to determine the changed/unchanged status of consolidation and pleural effusion in paired chest radiographs (CRs).

METHODS

The model was trained using 5.178 CRs from a single institution for lesion segmentation. Paired CRs from the emergency department (ED) and intensive care unit (ICU) were used to determine the thresholds for change and temporal validation. Model performance was evaluated through the area under the receiver operating characteristic curve (AUC), and its accuracy was compared with that of a thoracic radiologist.

RESULTS

In the ED, the model achieved AUCs of 0.988 and 0.883 for consolidation and pleural effusion, respectively, with accuracies of 0.900 (36/40) and 0.825 (33/40). The radiologist showed accuracies of 0.975 (39/40) and 0.950 (38/40), respectively. In the ICU, model AUCs were 0.970 (consolidation) and 0.955 (pleural effusion), with accuracies of 0.875 (35/40) and 0.800 (32/40), respectively. Radiologist performance was 0.975 (39/40) for consolidation and 1.000 (40/40) for pleural effusion. No significant accuracy differences were observed between the model and radiologist for consolidation in the ICU or both targets in the ED (all $P > 0.05$), except for pleural effusion in the ICU ($P = 0.01$).

CONCLUSION

The lesion-specific deep learning model was feasible for identifying interval changes in consolidation and pleural effusion on follow-up CRs.

CLINICAL SIGNIFICANCE

It could potentially be utilized for prioritizing interpretation, generating alerts, and extracting time-series data from multiple follow-up CRs.

KEYWORDS

Radiography, thoracic, follow-up studies, diagnosis, computer-assisted, artificial intelligence, segmentation

Chest radiography is a widely used medical imaging modality due to its cost-effectiveness and low radiation exposure. Chest radiographs (CRs) detect thoracic abnormalities and track changes during follow-ups. Monitoring abnormalities such as pleural effusion or consolidation is crucial for evaluating disease progression and treatment response.[1-4] However, frequent follow-up CRs increase workload. For example, in intensive care units (ICUs), CR is often performed daily for patients who are critically ill or after device adjustments, generating millions of ICU CRs annually in the United States.[5,6] Consequently, the timely and accurate interpretation of follow-up CRs is becoming more challenging.

***Joint first authors**

Corresponding author: Sang Min Lee

E-mail: sangmin.lee.md@gmail.com

Received 22 July 2025; revision requested 24 August 2025; accepted 23 September 2025.

Epub: 01.12.2025

DOI: 10.4274/dir.2025.253567

You may cite this article as: Kim Y, Ahn Y, Lee SM, et al. Automated evaluation of pulmonary lesion changes on chest radiograph during follow-up using semantic segmentation. *Diagn Interv Radiol.* XXX November 2025 DOI: 10.4274/dir.2025.253567 [Epub Ahead of Print].

Since follow-up CRs primarily detect changes between exams, analyzing CR pairs rather than relying solely on single-image abnormality detection is necessary. One line of currently developed deep learning methods detects overall changes using image registration to identify all CR findings.[7,8] It operates independently of detectable abnormality types and lesion-specific segmentation performance. However, it lacks information on which lesions have changed and the nature of these changes, which are essential in clinical practice. Furthermore, in settings such as the ICU, where various medical devices are attached, even simple repositioning, addition, or removal of a device may be recorded as a change, making it difficult to accurately determine whether a true change has occurred in the finding of interest.

Some methods have targeted specific abnormalities. For example, Li et al.[9] compared lung infiltration on serial CRs of patients with Coronavirus Disase-19, Huang et al.[10] quantified pleural effusion severity on individual CRs, and Lim et al.[11] estimated lung nodule volume from serial CRs. Although these studies demonstrated the feasibility or potential applicability of abnormality-specific monitoring, their scope was restricted to a single lesion type. An alternative approach that enables the simultaneous tracking of different abnormalities is lesion segmentation. Singh et al.[12] developed a deep learning algorithm that segments specific abnormalities and determines their changed/unchanged status based on the persistence of segmentation masks for lesions. The study reported an area under the receiver operating characteristic curve (AUC) of 0.758 for evaluating changes in pulmonary opacities over follow-up CRs. However, the algorithm was unable to determine the changed/unchanged status when their extent varied despite persistence. Despite its limitations, an algorithm that autonomously detects, segments, and assesses the changed/unchanged status of various abnormalities based on the degree of observed changes would be valuable.

Therefore, this study aims to develop a deep learning-based classifier for determining changed/unchanged status in paired CRs, using automatic lesion segmentation and extent comparison for consolidation and pleural effusion, and to validate its feasibility.

Main points

- Using lesion-specific segmentation, a deep-learning model determines consolidation and pleural effusion changes in chest radiographs (CRs) by assessing changes in their extent.
- A deep-learning model achieved an area under the curve of 0.970–0.988 for determining the changed/unchanged status of consolidation and 0.883–0.955 for pleural effusion in follow-up CRs from emergency department and intensive care unit datasets.
- With a predefined threshold, the model demonstrated an accuracy of 0.875–0.900 for changed/unchanged determination in consolidation and 0.800–0.825 for pleural effusion.

Methods

This retrospective study was approved by the institutional review board of Asan Medical Center, which waived the requirement for written informed consent (approval number: 2023-0810, date: 2023-07-01). Of the 5.178 CRs used for training, 4.593 were utilized in a previous study to develop a model for detecting five abnormalities.[13] However, our model is not related to the model from that study.

Training and validation datasets

In the classifier pipeline, the training set for abnormality segmentation was derived from CRs of adult patients (≥18 years) obtained at a tertiary referral hospital between January 2015 and December 2018 (Figure 1). The training set consisted of three types: normal CR, abnormal CR (with consolidation or pleural effusion), and CR with medical devices (Appendix S1). Radiologist-labeled lesion masks that had been developed and validated in the previous work were used.[13] However, the lesion segmentation algorithm, the paired radiograph comparison, and the change-detection framework were newly developed in this study. During the training process for the segmentation component of the model, the training set was further divided into a 9:1 ratio for model development and tuning.

After developing a lesion segmentation algorithm, CRs obtained from the emergency department (ED) and ICU between January 2019 and December 2019 were collected to determine the changed/unchanged classifier threshold. For each patient, one pair of CRs was randomly selected while maintaining the chronological order. The pairing principle was applied regardless of the CR projection type (posteroanterior or anteroposterior). However, due to the nature of the ED and ICU settings with patients who are critically ill, most radiographs were anteroposterior. Two thoracic radiologists (BLINDED and BLINDED, with 7 and 17 years of experience in thoracic imaging, respectively), blinded to the radiologic report, interpreted the changed/unchanged status, as well as the presence of target abnormality (i.e., consolidation and pleural effusion), in queried CR pairs in a random order until the target number of each dataset was reached. Both the changed/unchanged status and type of abnormality were determined in consensus by the two radiologists.

For temporal validation of the changed/unchanged classifier, CRs obtained from the ED and ICU between January 2020 and December 2020 were collected, each containing a single abnormality (consolidation or pleural effusion). To compare the performance between the model and radiologist, another thoracic radiologist (BLINDED, with 27 years of experience in thoracic imaging) independently reviewed the temporal validation set and determined the changed/unchanged status. This review was conducted blinded to the reference standard result but with knowledge of the target abnormality type (consolidation vs. pleural effusion).

Architecture of the lesion-specific classifier

Our model included two pipelines: 1) abnormality segmentation and 2) lesion area quantification and decision-making within pairs (Figure 2). First, the nnU-Net, a U-Net-based medical segmentation model known for its robust and high performance, served as the base model. Its structure and training options were modified for enhanced pulmonary lesion segmentation performance.[14] To improve the model's generalization ability, a multi-task learning (MTL) approach that jointly performs segmentation and classification was adopted, thereby improving the model's capability to differentiate between lesions in similar anatomical locations and medical devices and reducing potential segmentation errors. Two auxiliary classifiers were incorporated at the nnU-Net bottleneck for MTL: one for lesion presence classification and the other for lesion type classification (Appendix S2). The modified nnU-Net was trained for 1,000 epochs using 5-fold cross-validation, and the final lesion segmentation masks were generated by ensembling the inferred masks from each fold.

In the changed/unchanged classifier, lesion areas in each generated mask were quantified by multiplying the number of pixels in each lesion class by the pixel spacing of the corresponding CR. Changes in lesion quantities were calculated as the absolute

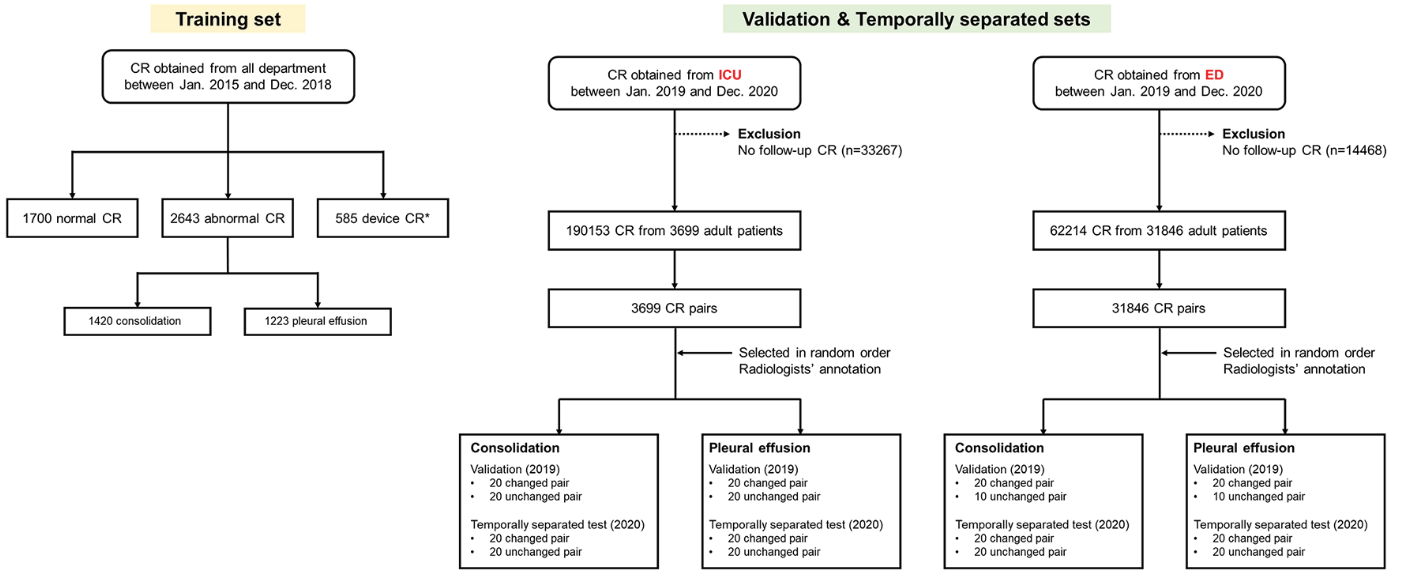


Figure 1. Flowchart illustrating dataset inclusion. *CRs containing medical devices include endotracheal tubes, drainage catheters, central lines, peripherally inserted central catheters, nasogastric tubes, chemoports, and electrocardiogram leads. CR, chest radiograph; ED, emergency department; ICU, intensive care unit.

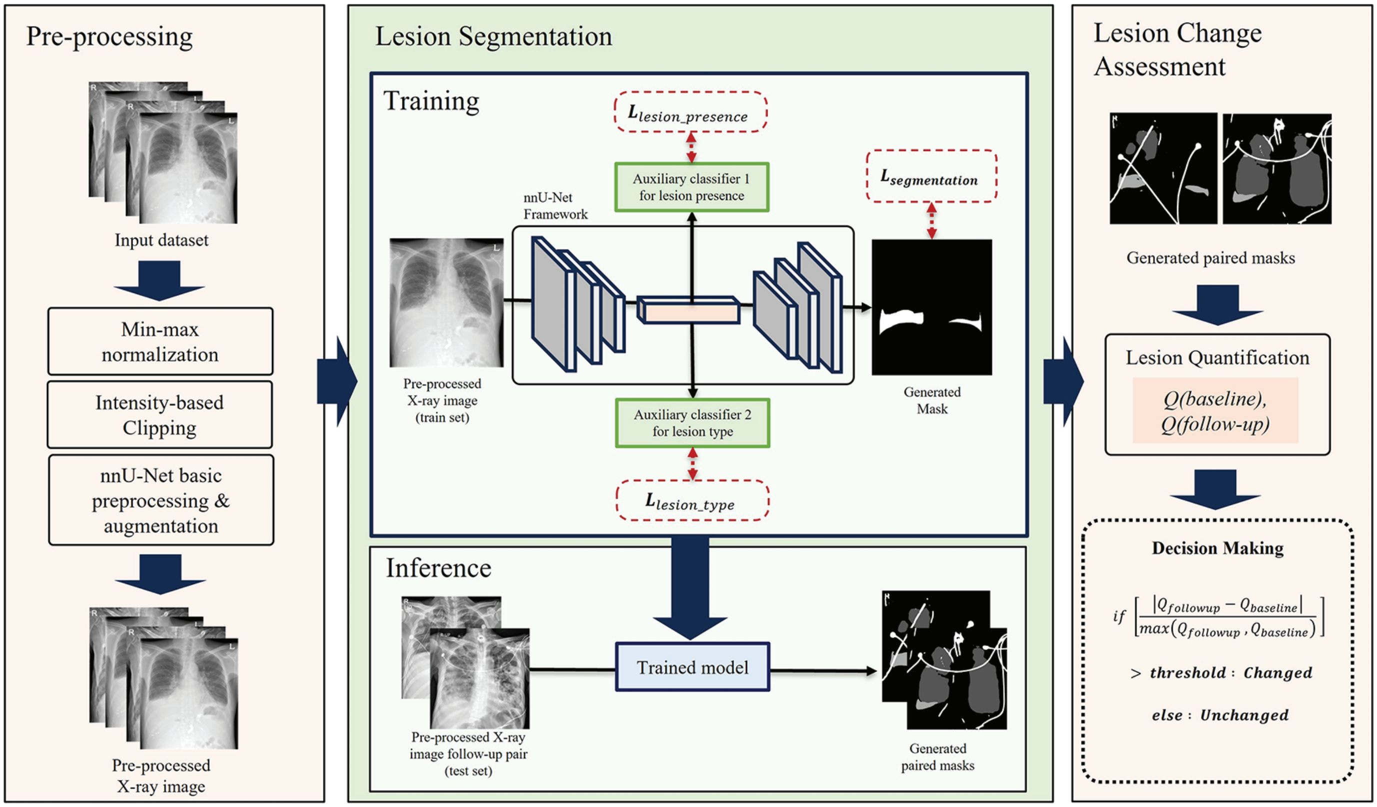

Pre-processing
Input dataset
Min-max normalization
Intensity-based Clipping
nnU-Net basic preprocessing & augmentation
Pre-processed X-ray image
Lesion Segmentation
Training
L_lesion_presence
Auxiliary classifier 1 for lesion presence
nnU-Net Framework
L_segmentation
Pre-processed X-ray image (train set)
Generated Mask
Auxiliary classifier 2 for lesion type
L_lesion_type
Inference
Trained model
Pre-processed X-ray image follow-up pair (test set)
Generated paired masks
Lesion Change Assessment
Generated paired masks
Lesion Quantification
Q(baseline), Q(follow-up)
Decision Making
if [|Q_followup − Q_baseline| / max(Q_followup, Q_baseline)] > threshold : Changed
else : Unchanged

Figure 2. Schematic illustrating the model workflow. The process includes three stages: preprocessing (image normalization and augmentation), lesion segmentation (using a modified nnU-Net with auxiliary classifiers to generate lesion masks), and lesion change assessment (quantifying lesion changes to classify them as changed or unchanged).

difference in quantified lesion areas divided by the larger of the two values to determine the relative change. The tuning set was used to optimize the threshold for changed/unchanged decision-making (Appendix S3). The calculated ratio was then used to classify each paired image as changed/unchanged based on a predefined threshold (Supplementary Figure 1).

$$\textit{if}\left[\frac{|Q_{followup}-Q_{baseline}|}{\max(Q_{followup},Q_{baseline})}\right]>\textit{threshold}:\text{"changed"}$$
$$\textit{else}:\text{"unchanged"}$$

Model training was done using the Pytorch framework and NVIDIA NVIDIA TITAN RTX 24GB GPU (NVIDIA Corporation, Santa Clara, CA, USA). The code for the model architecture is available on GitHub (https://github.com/ provbs/CR_DL_FU/).

Statistical analysis

The segmentation performance of the model for consolidation and pleural effusion was evaluated using Dice scores. T-tests were conducted to compare models, and *P* values were calculated for their differences. The performance of the model in classifying changed/unchanged was evaluated using the radiologists' results as the reference standard. The AUC was calculated, and the optimal threshold was determined using the Youden index based on the tuning set results. The accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F1 score of the temporal validation set were then calculated using the predetermined threshold. The model and radiologist accuracy were compared using the McNemar test. All statistical analyses were conducted using R version 4.3.1 (R Foundation for Statistical Computing).

Results

Dataset characteristics

The training dataset consisted of 1.700 normal CRs, 1.223 with pleural effusion, 1.420 with consolidation, and 585 with medical devices. For the changed/unchanged classifier tuning and temporal validation, 3.699 CR pairs from the ICU and 31.846 from the ED were generated after excluding single CRs without follow-up.

In the ED dataset, 30 pleural effusion pairs (20 changed, 10 unchanged) and 30 consolidation pairs (20 changed, 10 unchanged) from 2019 were included for the changed/unchanged classifier tuning. For temporal validation, 40 pleural effusion pairs (20 changed, 20 unchanged) and 40 consolidation pairs (20 changed, 20 unchanged) from 2019 were selected. In the ICU dataset, 40 pleural effusion pairs (20 changed, 20 unchanged) and 40 consolidation pairs (20 changed, 20 unchanged) from 2019 were used for the changed/unchanged classifier tuning, whereas the same numbers of pleural effusion and consolidation pairs from 2020 were used for temporal validation.

The median interval between CR pairs was 10 days (interquartile range: 1–100 days) in the tuning set and 15 days (interquartile range: 1–72 days) in the temporal validation set. Table 1 shows the demographics in detail.

Performance of lesion segmentation

The nnU-Net with MTL (using two auxiliary classifiers) and medical equipment masks in the training dataset was the best-performing segmentation model, with Dice scores of 0.848 for pleural effusion and 0.841 for consolidation (Table 2 and Supplementary Figure 2).

Incorporating medical equipment labels during training enhanced the Dice score for consolidation by approximately 0.044, although it decreased that for pleural effusion by 0.043, resulting in no major change in the average Dice score. Nevertheless, the qualitative results showed that the model trained with medical equipment labelling considerably reduced misclassification of medical equipment as lesions, a critical distinction in ICU and ED settings. Furthermore, integrating MTL and medical equipment labels improved the average Dice score by 0.015, reducing the difference between lesion types and achieving a more balanced performance (Figure 3).

Performance of lesion-specific change detection

In the tuning set, the AUCs of the model were 0.747 for consolidation and 0.850 for pleural effusion in the ED, and 0.980 for consolidation and 0.800 for pleural effusion in the ICU (Supplementary Figure 3). To account for different clinical settings, thresholds were determined separately for the ED and ICU. The optimal thresholds derived from the tuning set were 0.26 for consolidation and 0.29 for pleural effusion in the ED and 0.40 for consolidation and 0.55 for pleural effusion in the ICU.

In the temporal validation set, the AUCs of the model were 0.988 for consolidation and 0.883 for pleural effusion in the ED and 0.970 for consolidation and 0.955 for pleural effusion in the ICU (Figure 4). The AUC for consol-

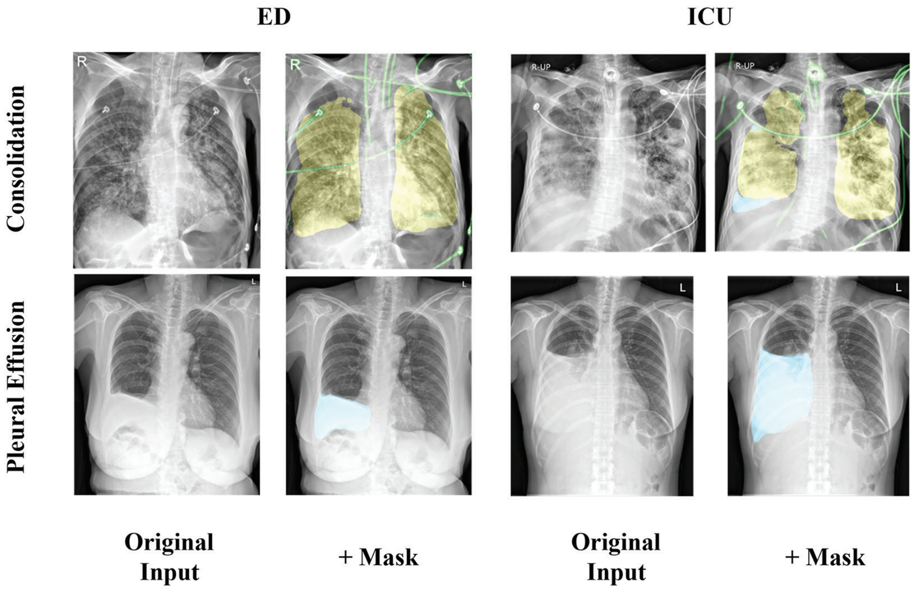


Figure 3. Lesion segmentation results using the best-performing model. The first column shows the original images, and the second column shows the model output in the emergency department (ED) and intensive care unit (ICU). Yellow indicates consolidation, and sky blue indicates pleural effusion.

Table 1. Baseline characteristics for the tuning and temporal validation sets

	Tuning set		Temporal validation set	
Characteristics	ED	ICU	ED	ICU
Number of patients	80	60	80	80
Age, years[a]	66.8 ± 13.7	64.4 ± 13.6	72.0 ± 14.2	61.3 ± 12.6
Sex				
Male	49 (61.2%)	29 (48.3%)	50 (62.5%)	51 (63.8%)
Female	31 (38.8%)	31 (51.7%)	30 (37.5%)	29 (36.2%)
Interval between baseline and follow-up CR[b]	4.0 (1.0, 84.5)	23.0 (5.5, 156.5)	13.5 (1.0, 124.0)	15.5 (3.5, 44.0)

[a]Data are mean ± standard deviation.
[b]Data are median with interquartile range in parentheses.
CR, chest radiograph; ED, emergency department; ICU, intensive care unit.

idation was similar between the ED and ICU, whereas the AUC for pleural effusion in the ED was slightly lower than that in the ICU.

Comparisons between the model and the thoracic radiologist

In the ED, the model achieved an accuracy of 0.900 (36/40) for consolidation, with a sensitivity of 1.000 for "changed" and a specificity of 0.800 for "unchanged." For pleural effusion, the accuracy was 0.825 (33/40), with a sensitivity of 0.850 and specificity of 0.800 (Figure 5). The accuracy of the thoracic radiologist was 0.975 (39/40) for consolidation and 0.950 (38/40) for pleural effusion (Table 3).

In the ICU, the model achieved an accuracy of 0.875 (35/40) for consolidation, with a sensitivity of 0.900 for "changed" and a specificity of 0.850 for "unchanged." For pleural effusion, the accuracy of the model was 0.800 (32/40), with a sensitivity of 0.600 and specificity of 1.000 (Supplementary Figures 4 and 5). The accuracy of the thoracic radiologist was 0.975 (39/40) for consolidation and 1.000 (40/40) for pleural effusion (Table 3 and Supplementary Figure 5).

When comparing the accuracy of the model and the thoracic radiologist, no significant difference was found for consolidation in the ED [0.900 (36/40) vs. 0.975 (39/40), *P* = 0.371], pleural effusion in the ED [0.825 (33/40) vs. 0.950 (38/40), *P* = 0.182], and consolidation in the ICU [0.875 (35/40) vs. 0.975 (39/40), *P* = 0.221]. However, for pleural effusion in the ICU, the radiologist outperformed the model [1.000 (40/40) vs. 0.800 (32/40), *P* = 0.013] (Supplementary Figure 6).

Discussion

Multiple CRs for follow-up are common in clinical practice. However, current interpretation techniques are largely limited to single images, and automated methods for follow-up CR analysis remain underdeveloped. In this study, we developed and validated a deep-learning model for assessing the changed/unchanged status through lesion-specific segmentation. In validation within the ED and ICU settings, the model classified the changed/unchanged status with an accuracy of 0.875–0.900 for consolidation and 0.800–0.825 for pleural effusion, comparable to that of the radiologist, except for pleural effusion in the ICU.

Interpreting follow-up CRs poses challenges for both radiologists and deep-learning algorithms due to changes in the thoracic cage caused by variations in posture or inspiration status, as well as background changes such as alterations in medical devices. Consequently, few studies focus on the automated interpretation of CR pairs.[7,8,12] The approach of determining changes or no changes in the overall image landscape can help prioritize worklists and improve workflow efficiency,[7,8] although it lacks details on the specific objects involved or the extent of the changes. Unlike previous approaches, we aimed to develop a model that identifies specific abnormal changes. Lesion-specific interpretation is straightforward and enables the detection of clinically relevant changes, such as consolidation increases in patients with pneumonia. With further refinement, it could become a component of an autonomous reporting system. To achieve this, we focused on two major abnormalities—consolidation and pleural effusion—that are commonly monitored for treatment response. These abnormalities were tested in the ED and ICU settings, where they are more prevalent and dynamic than in outpatient clinics or general wards.

Our model achieved an AUC of 0.883–0.988, outperforming a previous study (AUC: 0.687 for pulmonary opacity changes and 0.782 for pleural effusion changes) that determined changed/unchanged status sole-

Table 2. Comparison of lesion segmentation performance across different training settings

	Dice score		
Learning method	Consolidation	Pleural effusion	Average
nnU-Net (without device labels)	0.821*	0.838	0.830
nnU-Net (with device labels)	**0.866****	0.794***	0.830
nnU-Net + MTL (with device labels)	0.841	**0.848**	**0.845**

*$P < 0.10$, **$P < 0.05$, ***$P < 0.001$. MTL, multi-task learning.

Table 3. Evaluation metrics assessed on the temporal validation set

Evaluator	Lesion type	AUC	Accuracy	Sensitivity	Specificity	PPV	NPV	F1 score
ED								
Model	Consolidation	0.988 [0.968, 1.000]	0.900 (36/40) [0.763, 0.972]	1.000 (20/20) [0.832, 1.000]	0.800 (16/20) [0.563, 0.943]	0.833 (20/24) [0.626, 0.953]	1.000 (16/16) [0.794, 1.000]	0.909
	Pleural effusion	0.883 [0.800, 0.993]	0.825 (33/40) [0.672, 0.927]	0.850 (17/20) [0.621, 0.968]	0.800 (16/20) [0.563, 0.943]	0.810 (17/21) [0.581, 0.946]	0.842 (16/19) [0.604, 0.966]	0.829
Radiologist	Consolidation	-	0.975 (39/40) [0.868, 0.999]	1.000 (20/20) [0.832, 1.000]	0.950 (19/20) [0.751, 0.999]	0.952 (20/21) [0.762, 0.999]	1.000 (19/19) [0.824, 1.000]	0.976
	Pleural effusion	-	0.950 (38/40) [0.831, 0.994]	0.950 (19/20) [0.751, 0.999]	0.950 (19/20) [0.751, 0.999]	0.950 (19/20) [0.751, 0.999]	0.950 (19/20) [0.751, 0.999]	0.950
ICU								
Model	Consolidation	0.970 [0.931, 1.000]	0.875 (35/40) [0.732, 0.958]	0.900 (18/20) [0.683, 0.988]	0.850 (17/20) [0.621, 0.968]	0.857 (18/21) [0.637, 0.970]	0.895 (17/19) [0.669, 0.987]	0.878
	Pleural effusion	0.955 [0.908, 1.000]	0.800 (32/40) [0.644, 0.909]	0.600 (12/20) [0.361, 0.809]	1.000 (20/20) [0.832, 1.000]	1.000 (12/12) [0.735, 1.000]	0.714 (20/28) [0.513, 0.868]	0.750
Radiologist	Consolidation	-	0.975 (39/40) [0.868, 0.999]	1.000 (20/20) [0.832, 1.000]	0.950 (19/20) [0.751, 0.999]	0.952 (20/21) [0.762, 0.999]	1.000 (19/19) [0.824, 1.000]	0.976
	Pleural effusion	-	1.000 (40/40) [0.912, 1.000]	1.000 (20/20) [0.832, 1.000]	1.000 (20/20) [0.832, 1.000]	1.000 (20/20) [0.832, 1.000]	1.000 (20/20) [0.832, 1.000]	1.000

Data in the parentheses are the number of CR pairs. AUC, area under curve; ED, emergency department; ICU, intensive care unit; NPV, negative predictive value; PPV, positive predictive value.

ly based on lesion persistence.[12] It was also similar to prior non-lesion-specific models, which have AUCs of 0.800–0.858.[7,8] This performance may be due to the accurate lesion segmentation of our model, achieving a Dice score of up to 0.845 in the training set, and its reduced misclassification of medical devices as lesions. Singh et al.[12] reported that the mis-segmentation of medical devices as pulmonary opacities is a challenge. To address this, we specifically trained our model on CRs with medical devices, ensuring robust performance in the ICU and ED settings where they are almost always present.

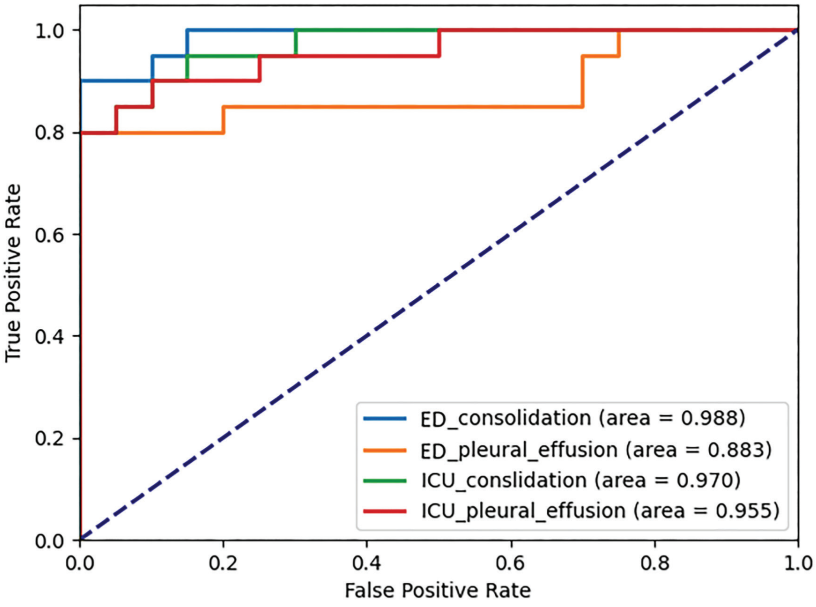


Figure 4. ROC curves illustrating algorithm performance for each department and lesion type in the temporal validation set. The shaded areas represent the area under the ROC curves. ED, emergency department; ICU, intensive care unit; ROC, receiver operating characteristic.

The accuracy of our model was similar to that of the radiologist for consolidation in the ICU and ED and for pleural effusion in the ED, though slightly lower. The decision of the radiologist on whether a condition had changed or remained unchanged closely aligns with the reference standard. Although consolidation and pleural effusion are typically assessed qualitatively in routine practice, the threshold of readers may be interchangeable. Our model showed considerably lower performance than that of the radiologist for pleural effusion in the ICU. This may be related to the position of the patient in the ICU. In patients in the supine position, both consolidation and pleural effusion can appear as diffusely increased opacity, making differentiation difficult. Pleural fluid tends to spread under gravity, making the margins of effusion indistinct. Radiologists also assess changes in pleural effusion while considering positional changes, which may be challenging for our model. Notably, all incorrect ICU pleural effusion classifications occurred in "changed" cases, whereas the model correctly identified all stable cases. We therefore consider that the model can

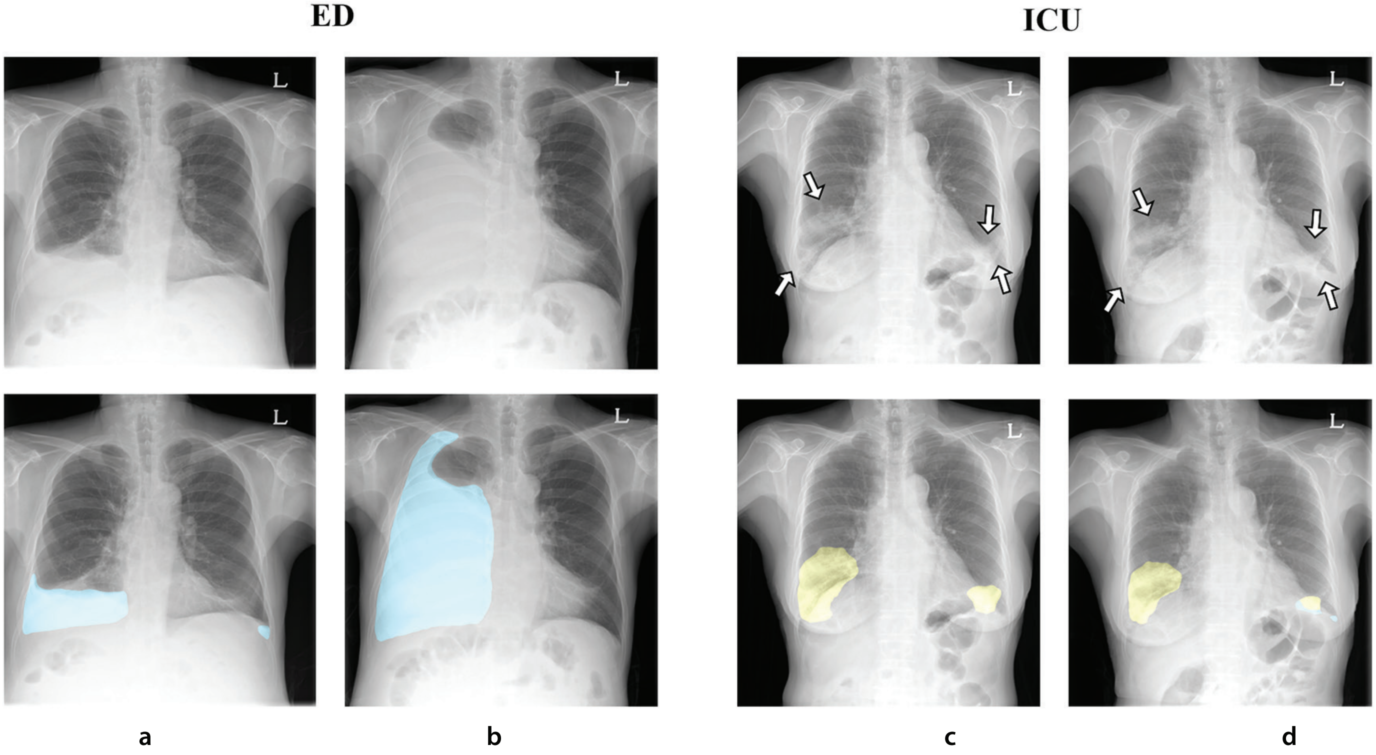


Figure 5. Examples of determining changed/unchanged status in the emergency department (ED) and intensive care unit (ICU). **(a, b)** In the ED, a right pleural effusion increased between baseline **(a)** and follow-up chest radiograph (CR) **(b)**. The model detected and segmented the effusion (sky-blue areas) and classified it as changed. The pixel difference was 19.299 (75% ratio). **(c, d)** In the ICU, patchy consolidation in both lower lobes (arrows) remained unchanged between baseline **(c)** and follow-up CR **(d)**, matching the reference standard. The model segmented the consolidation (yellow areas) and classified it as unchanged. The pixel difference was –2.253 (36.9% ratio).

adequately triage stable pleural effusion, but reduced sensitivity in supine patients remains an important limitation that warrants further refinement. In addition, the limited size of the tuning set (40 pairs) may have led to overfitting of the threshold for pleural effusion, contributing to skewed results. Although this should be addressed in future studies, our findings provide proof of concept for the feasibility of lesion-specific segmentation in change status detection.

Unlike preexisting non-lesion-specific models, which primarily filter grossly stable CR pairs, our lesion-specific model offers dual functionality. It can inform and prioritize changes for physician review while simultaneously filtering stable cases. Previous approaches based on registration and subtraction within pairs are limited compared with a segmentation-based method, which has the potential to be applied to multiple CRs in longitudinal follow-up, enabling the extraction of lesion extent as time-series data with quantification. Recent advances in language models have enabled training on large-scale, weakly labeled data for multi-label, multiclass change detection and even automated report generation.[15,16] In contrast, our model leverages radiologist-provided hard labeling, which ensures disease-specific accuracy and offers interpretable, intuitive visual explanations of the degree of change. These strengths may provide potential applicability, working synergistically with text generation models as part of automated reporting systems. However, our model is currently limited to two abnormalities: consolidation and pleural effusion. Expanding its capabilities to include other major abnormalities, such as nodules or interstitial opacities, may be a valuable next step. Furthermore, improving lesion segmentation and consideration of position change are warranted.

Our study has some limitations. First, as a single-center retrospective study, it may have selection bias and limited generalizability. Second, the experiment was conducted using datasets from the same institution. Although the datasets do not overlap, true external validation was not performed. Third, the tuning and temporal validation sets were relatively small. Since our model was designed specifically for consolidation and pleural effusion, only patients with at least one of these abnormalities were eligible. This may contribute to the performance differences between the tuning and temporal validation sets. Further validation in a larger population is necessary.

In conclusion, lesion-specific segmentation enables the deep-learning-based model to determine the changed/unchanged status of consolidation and pleural effusion based on changes in their extent.

Footnotes

Conflict of interest disclosure

The authors declared no conflicts of interest.

Funding

This work was supported by the Korea Medical Device Development Fund grant funded by the Korea government (the Ministry of Science and ICT, the Ministry of Trade, Industry and Energy, the Ministry of Health & Welfare, the Ministry of Food and Drug Safety) (Project Number: 2710002589, RS-2023-00254209).

References

1. Expert Panel on Thoracic Imaging; Morris MF, Henry TS, Raptis CA, et al. ACR appropriateness criteria® workup of pleural effusion or pleural disease. *J Am Coll Radiol*. 2024;21(6S):S343-S352. [Crossref]
2. Krishna R, Antoine MH, Alahmadi MH, Rudrappa M. Pleural Effusion. 2024 Aug 31. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2025 Jan–. [Crossref]
3. Karkhanis VS, Joshi JM. Pleural effusion: diagnosis, treatment, and management. *Open Access Emerg Med*. 2012;4:31-52. [Crossref]
4. Little BP, Gilman MD, Humphrey KL, Alkasab TK, Gibbons FK, Shepard JA, et al. Outcome of recommendations for radiographic follow-up of pneumonia on outpatient chest radiography. *AJR Am J Roentgenol*. 2014;202(1):54-9. [Crossref]
5. Gershengorn HB, Wunsch H, Scales DC, Rubenfeld GD. Trends in use of daily chest radiographs among us adults receiving mechanical ventilation. *JAMA Netw Open*. 2018;1(4):e181119. [Crossref]
6. Oba Y, Zaza T. Abandoning daily routine chest radiography in the intensive care unit: meta-analysis. *Radiology*. 2010;255(2):386-95. [Crossref]
7. Cho K, Kim J, Kim KD, et al. Music-ViT: a multitask Siamese convolutional vision transformer for differentiating change from no-change in follow-up chest radiographs. *Med Image Anal*. 2023;89:102894. [Crossref]
8. Yun J, Ahn Y, Cho K, et al. Deep learning for automated triaging of stable chest radiographs in a follow-up setting. *Radiology*. 2023;309(1):e230606. [Crossref]
9. Li MD, Arun NT, Gidwani M, et al. Automated assessment and tracking of COVID-19 pulmonary disease severity on chest radiographs using convolutional siamese neural networks. *Radiol Artif Intell*. 2020;2(4):e200079. [Crossref]
10. Huang T, Yang R, Shen L, et al. Deep transfer learning to quantify pleural effusion severity in chest X-rays. *BMC Med Imaging*. 2022;22(1):100. [Crossref]
11. Lim CY, Cha YK, Chung MJ, et al. Estimating the volume of nodules and masses on serial chest radiography using a deep-learning-based automatic detection algorithm: a preliminary study. *Diagnostics (Basel)*. 2023;13(12):2060. [Crossref]
12. Singh R, Kalra MK, Nitiwarangkul C, et al. Deep learning in chest radiography: detection of findings and presence of change. *PLoS One*. 2018;13(10):e0204155. [Crossref]
13. Park B, Cho Y, Lee G, et al. A curriculum learning strategy to enhance the accuracy of classification of various lesions in chest-PA X-ray screening for pulmonary abnormalities. *Sci Rep*. 2019;9(1):15352. [Crossref]
14. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods*. 2021;18(2):203-211. [Crossref]
15. Yu K, Ghosh S, Liu Z, Deible C, Poynton CB, Batmanghelich K. Anatomy-specific progression classification in chest radiographs via weakly supervised learning. *Radiol Artif Intell*. 2024;6(5):e230277. [Crossref]
16. Wang Z, Deng Q, So TY, Chiu WH, Lee K, Hui ES. Disease probability-enhanced follow-up chest X-ray radiology report summary generation. *Sci Rep*. 2025;15(1):26930. [Crossref]

Appendix S1. Medical equipment data in final segmentation training

Using the 4.593 chest radiographs (CRs) of general patients, we trained a nnU-Net model with default settings for lesion segmentation. The training was conducted for 1.000 epochs using the Pytorch framework on an NVIDIA TITAN RTX 24GB GPU. The 250 images from the additional CR dataset, containing only labeled medical devices from patients in the intensive care unit (ICU), were then processed using the trained segmentation model. This inference generated 250 CR images with both pulmonary lesions and medical devices altogether.

We did not evaluate or optimize the lesion segmentation performance specifically for this 250 CR dataset, as it represents a relatively small proportion of the entire training dataset in terms of lesion data. Its primary purpose is to provide the final model with explicit information on medical devices rather than lesion-related information.

By including this dataset in our final training dataset, we aimed to enable the model to semantically learn to better differentiate medical equipment from lesions in ICU and emergency department (ED) datasets, thereby reducing the misclassification of medical devices as lesions, resulting in better segmentation performance in general. As shown in Table 1, this additional dataset contributed to improving the performances of our final segmentation model by addressing this challenge effectively.

Appendix S2. Details of the segmentation model

Input preprocessing

The CRs in the training/tuning/temporal validation datasets were all preprocessed with the following steps: Firstly, intensity-based clipping was implemented, where the top and bottom 0.5% of pixel intensities were clipped. This was done to mitigate the influence of high-intensity outliers, such as L&R markers or other unexpected artifacts on the CR. Secondly, min-max normalization was conducted to scale pixel values of the images within the range of 0–1. Both steps were done to ensure consistency and facilitated convergence during the segmentation model training explained after. Other preprocessing and augmentations were conducted in accordance with the nnU-Net methodology, accounting for median shape, distribution of spacings, intensity distribution, and image modality within the training dataset.

Model structure and training details

The model employed in this study is a modified nnU-Net with two auxiliary classifiers incorporated between the encoder and segmentation decoder. The auxiliary classifiers used here are quite simple: two fully connected layers with a ReLu activation in between. These auxiliary classifiers are intended to enable the shared encoder to progressively focus on features related to lesion type and presence during its training, thereby passing more pertinent information to the segmentation decoder. The total loss was calculated as the sum of the original nnU-Net segmentation loss and the losses from the auxiliary classifiers, lesion presence and lesion type classifier, as follows:

$$\mathbf{L}_{\text{total}} = \mathbf{L}_{\text{seg}} + (\mathbf{L}_{lesion\_presence} + \mathbf{L}_{lesion\_type}) \qquad (1)$$

Severe augmentations, including Gaussian noise and Gaussian blur transformations, were adopted to enhance segmentation performance on the noisier ICU/ED CRs.

During inference, we excluded regions with fewer than 50 pixels in the predicted mask to eliminate insignificant noise, as most mask sizes exceed 2000 × 2000 pixels, with some surpassing 3000 pixels on one side. This threshold was chosen based on the observation that smaller regions often represent false positives or artifacts rather than meaningful predictions.

Appendix S3. Optimal threshold determination for assessment decision-making

To select the optimal threshold, we employed Youden's J statistic, which measures the effectiveness of a threshold in terms of maximizing the true positive rate while minimizing the false positive rate. For each threshold t_i, where t_i increases by 0.01 from 0.00 to 1.00, the true positive ratio (TPR) and false positive ratio (FPR) can be computed, and the value of J at t_i is calculated as:

$$J(t_i) = TPR(t_i) - FPR(t_i) \qquad (2)$$

The optimal threshold can be determined by identifying the highest value of J, representing the most balanced threshold between sensitivity and specificity:

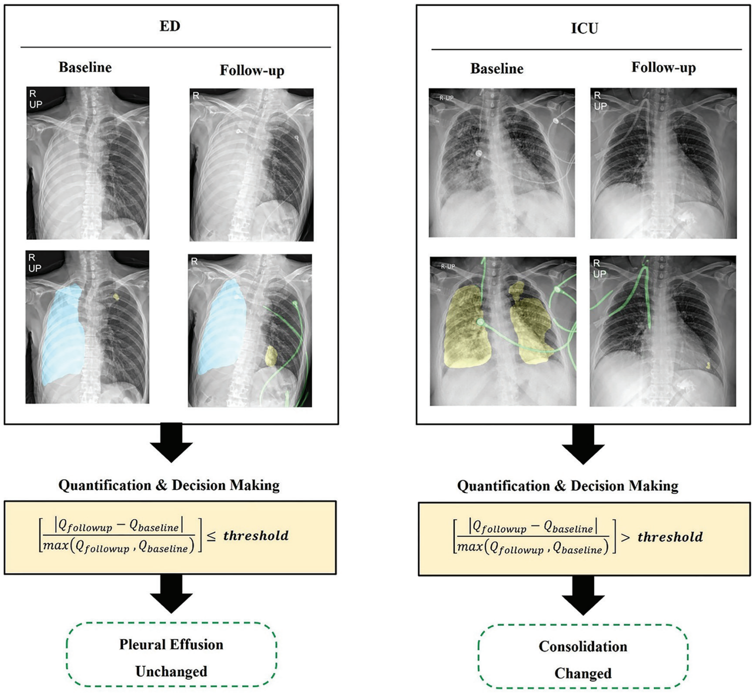
$$t_{optimal} = argmax_{ti}\, J\,(t_i) \qquad (3)$$

If multiple values of $t_{optimal}$ exist, we selected the lowest threshold value to increase sensitivity and reduce the chance of false negatives:

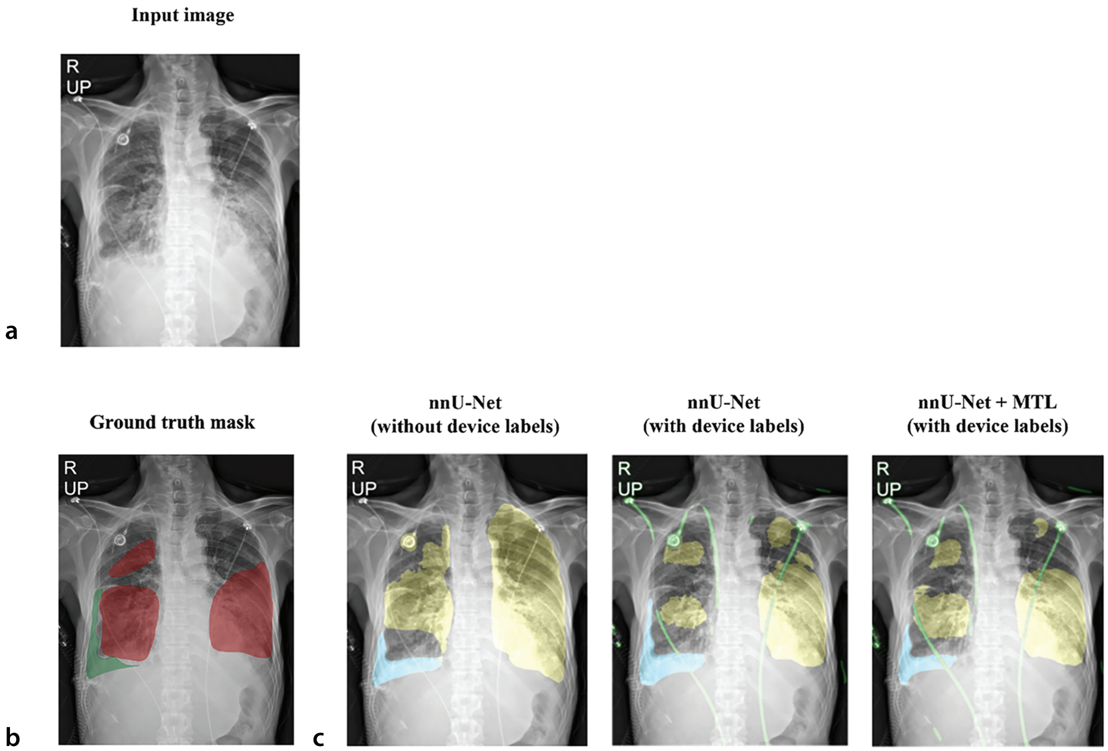
$$t_{selected} = \min(t_{optimal,i}), \quad \text{for } i = 1,2...,n \qquad (4)$$

This approach minimizes the misclassification of true positives as negatives, which is crucial in settings such as the ICU/ED, where timely intervention is essential. We have inferenced the tuning set using the trained segmentation model and then calculated the optimal threshold accordingly.

Supplementary Figures

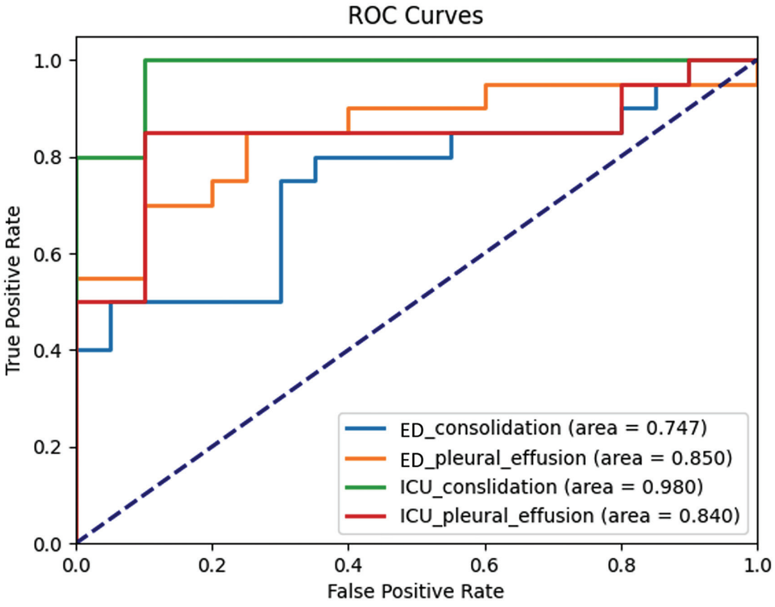

ED
Baseline
Follow-up
R UP
ICU
Baseline
Follow-up
R UP
Quantification & Decision Making
$\left[\frac{|Q_{followup} - Q_{baseline}|}{max(Q_{followup}, Q_{baseline})}\right] \le threshold$
Pleural Effusion
Unchanged
Quantification & Decision Making
$\left[\frac{|Q_{followup} - Q_{baseline}|}{max(Q_{followup}, Q_{baseline})}\right] > threshold$
Consolidation
Changed

Supplementary Figure 1. Determination of unchanged/changed in the ED and ICU. The yellow area represents consolidation, while the sky-blue area indicates pleural effusion. ED, emergency department; ICU, intensive care unit.

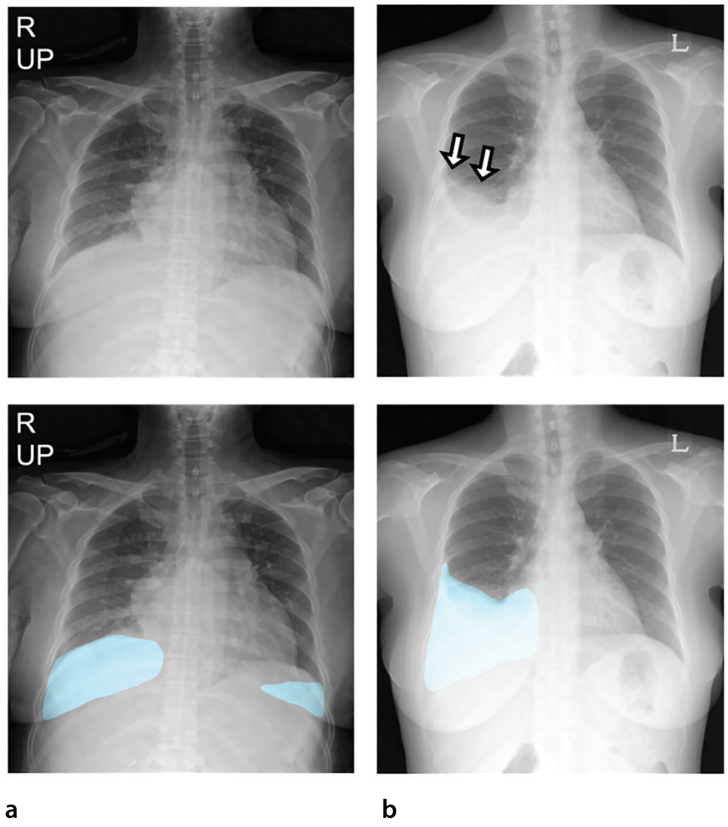

Input image
R UP
a
Ground truth mask
nnU-Net (without device labels)
nnU-Net (with device labels)
nnU-Net + MTL (with device labels)
b
c

	nnU-Net (without device labels)	nnU-Net (with device labels)	nnU-Net + MTL (with device labels)
Dice for Consolidation :	0.6397	0.6745	**0.7710**
Dice for Pleural effusion :	0.6866	0.6777	**0.7361**

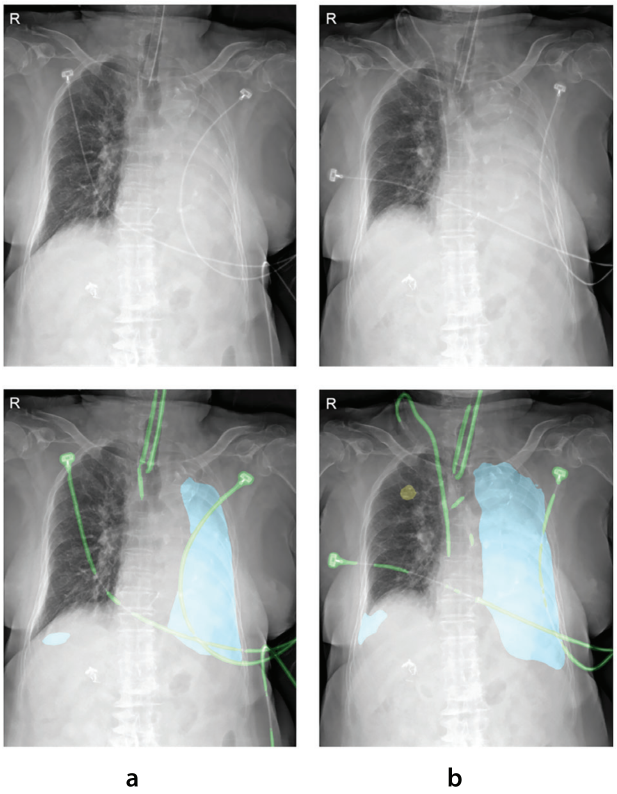
Supplementary Figure 2. Segmentation results across models with various training settings. **(a)** Original input image. **(b)** The ground-truth mask by a thoracic radiologist. The red area represents a consolidation, while the green area indicates pleural effusion. **(c)** Segmentation results from three different settings. The yellow area represents consolidation, while the sky-blue area indicates pleural effusion. Green areas represent the segmentation of medical devices. The Dice score was calculated for each abnormality. MTL, multi-task learning.



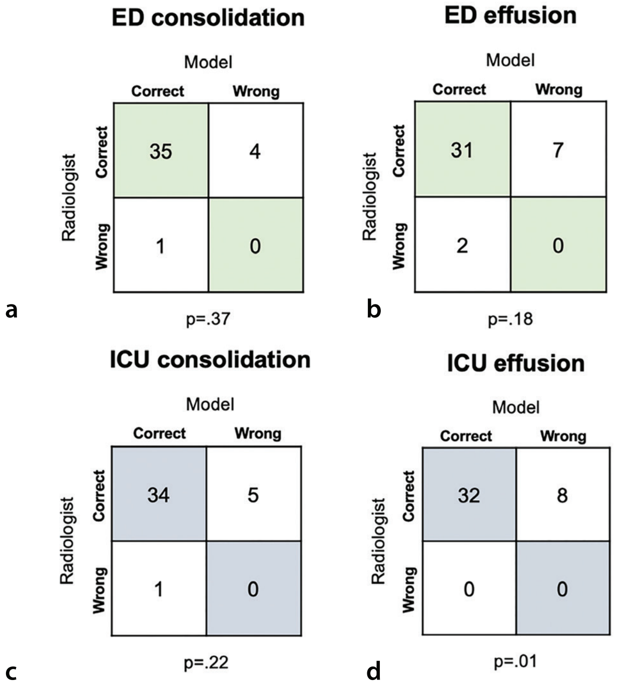
Supplementary Figure 3. Receiver operating characteristic (ROC) curves show the performance of the algorithm for each department and the abnormality lesion in the tuning set. Area means the area under the ROC curves. ED, emergency department; ICU, intensive care unit.



Supplementary Figure 4. Example of failure in detecting "changed" in pleural effusion in the intensive care unit. **(a)** Baseline chest radiograph (CR) obtained in a semi-supine position showed no clear fluid level, but the model segmented increased opacity along both hemidiaphragms as pleural effusion (sky-blue areas). The presence of true effusion could not be confirmed on CR alone. **(b)** In the follow-up CR acquired in an upright position, overt right pleural effusion (arrows) was evident and correctly segmented by the model (sky-blue areas). The radiologist interpreted this case as "changed," whereas the model classified it as "unchanged."



Supplementary Figure 5. Example of failure in detecting "unchanged" pleural effusion in the intensive care unit. **(a)** Baseline chest radiograph (CR) and **(b)** follow-up CR obtained in the supine position both showed diffuse left pleural effusion opacifying the left hemithorax. Due to slight rightward rotation on the baseline CR, the segmented area for pleural effusion (sky-blue areas) appeared smaller, leading the model to classify the case as "changed," whereas the radiologist interpreted it as "unchanged."



a ED consolidation

Radiologist \ Model	Correct	Wrong
Correct	35	4
Wrong	1	0

p=.37

b ED effusion

Radiologist \ Model	Correct	Wrong
Correct	31	7
Wrong	2	0

p=.18

c ICU consolidation

Radiologist \ Model	Correct	Wrong
Correct	34	5
Wrong	1	0

p=.22

d ICU effusion

Radiologist \ Model	Correct	Wrong
Correct	32	8
Wrong	0	0

p=.01

Supplementary Figure 6. Comparisons between the model and radiologist in assessing change/unchanged status of consolidation or pleural effusion. **(a)** for consolidation in the emergency department (ED), **(b)** for effusion in the ED, **(c)** for consolidation in the intensive care unit (ICU), and **(d)** for effusion in the ICU. *P* values were calculated using the McNemar test to compare accuracy.