Medicine®

OPEN

Evaluation of a deep learning-based computer-aided detection algorithm on chest radiographs
Case–control study

Soo Yun Choi, MS[a] [iD], Sunggyun Park, PhD[b] [iD], Minchul Kim, MS[b] [iD], Jongchan Park, ME[b] [iD],
Ye Ra Choi, MD[c] [iD], Kwang Nam Jin, MD[a,c,*] [iD]

Abstract

Along with recent developments in deep learning techniques, computer-aided diagnosis (CAD) has been growing rapidly in the medical imaging field. In this work, we evaluate the deep learning-based CAD algorithm (DCAD) for detecting and localizing 3 major thoracic abnormalities visible on chest radiographs (CR) and to compare the performance of physicians with and without the assistance of the algorithm. A subset of 244 subjects (60% abnormal CRs) was evaluated. Abnormal findings included mass/nodules (55%), consolidation (21%), and pneumothorax (24%). Observer performance tests were conducted to assess whether the performance of physicians could be enhanced with the algorithm. The area under the receiver operating characteristic (ROC) curve (AUC) and the area under the jackknife alternative free-response ROC (JAFROC) were measured to evaluate the performance of the algorithm and physicians in image classification and lesion detection, respectively. The AUCs for nodule/mass, consolidation, and pneumothorax were 0.9883, 1.000, and 0.9997, respectively. For the image classification, the overall AUC of the pooled physicians was 0.8679 without DCAD and 0.9112 with DCAD. Regarding lesion detection, the pooled observers exhibited a weighted JAFROC figure of merit (FOM) of 0.8426 without DCAD and 0.9112 with DCAD. DCAD for CRs could enhance physicians' performance in the detection of 3 major thoracic abnormalities.

Abbreviations: AUC = area under the ROC curve, CAD = computer-aided diagnosis, CI = confidence interval, CR = chest radiograph, DCAD = deep learning-based CAD algorithm, FOM = figure of merit, ROC = receiver operating characteristic, wJAFROC = weighted jackknife alternative free-response ROC.

Keywords: computer-aided, deep learning, diagnosis, radiography, thorax

1. Introduction

Chest radiography has been the most ubiquitous diagnostic examination for screening thoracic diseases because of its relatively low cost and wide availability. The detection of abnormal pathological lesions on chest radiograph (CR) often leads to the primary diagnosis of fatal pulmonary diseases such as lung cancer.[1] However, the interpretation of CR is still a demanding task in clinical settings. A retrospective study discovered that 1 out of every 5 errors in diagnostic radiology occurred during the interpretation of CRs.[2] Considering that variable factors such as low proficiency or heavy workload can degrade the performance of physicians, it might be unsurprising that the interpretation of CRs by humans could not guarantee consistent and intelligible readings to patients.

Along with recent developments in deep learning techniques, computer-aided diagnosis (CAD) has been growing rapidly in the medical imaging field. Artificial intelligence has been introduced to enhance the detection or measurement of breast cancer,[3,4] brain tumors,[5] liver lesions,[6] vessel border,[7] and blood flow dynamics.[8] For thoracic diseases, automated detection algorithms have been designed to detect major diseases, such as tuberculosis, and to classify normal and abnormal CRs.[9] Localizing exact lesions has been another major concern of researchers. Singh et al[10] reported a relatively high accuracy (area under the receiver operating characteristic curve [AUC] 0.837–0.929) of automated detection in classifying pulmonary

opacities, hilar prominence, cardiomegaly, and pleural effusion. However, this paper could not prove the efficacy of the algorithm in detecting specific CR findings, such as pulmonary nodules, masses, and fibrosis. A recently developed DR-based automated detection algorithm has shown impressive results in classifying 4 major thoracic diseases.[11] However, regarding the dataset used for validation, which was experimentally designed, further investigation is necessary to show the utility of the deep learning algorithm with variable datasets. To our knowledge, there were few studies on deep learning-based automated diagnosis algorithms for thoracic lesions indicating case–control data collection from consecutive image datasets.[12,13] It is necessary for accurate interpretation of AI-supported diagnosis of thoracic lesions. The aims of this study were to evaluate the deep learning-based computer-aided detection algorithm for detecting and localizing thoracic abnormalities visible on CRs and to compare the performance of physicians with and without the assistance of the algorithm.

2. Methods

The present retrospective case–control study was approved by the Seoul National University Boramae Medical Center Institutional Review Board (Seoul, Korea, IRB no. 10-2019-48), which waived the informed consent requirement. All methods were performed in accordance with the relevant guidelines and regulations. Lunit (Seoul, Korea) provided technical support for analyzing chest radiographs with the DL algorithm and obtaining outputs.

2.1. Deep learning algorithm and sample size calculation

A DL algorithm (Lunit INSIGHT for Chest Radiography; Lunit [accessible at https://insight.lunit.io]) used in this study was designed to classify chest radiographs of patients with 3 major abnormal findings, including nodule/mass, consolidation, and pneumothorax, and enhance the performance of human readers.[11]

We expect that the AUC between physician alone test and deep learning-based CAD algorithm (DCAD)-aided test will be different. Physician alone and DCAD-aided tests were performed by panels without and with the aid of the algorithm. Based on the pilot study (AUC – physician alone test = 0.9289; AUC – DCAD-aided test = 0.9445), we expect that the effect size will be 0.0156. Under the assumption that the power is 0.8 and the alpha value is 0.05, the minimum sample size is 49 for normal and 73 for abnormal findings if the number of panels is 6. The final sample size was 244 with a calculated power of 0.9902 (normal: abnormal = 0.4:0.6).

2.2. Data collection

For the evaluation of DLAD, 98 CRs with normal results and 146 CRs with abnormal results were retrospectively collected (Supplementary Fig. 1, http://links.lww.com/MD2/A91). The number of each abnormal finding was collected following the prevalence and distribution rate of the X-ray 14 dataset from the National Institute of Health (NIH) Clinical Center. The X-ray 14 dataset is an open-source dataset released by the NIH, containing CRs from 32,177 patients with 14 labels based on the presence and absence of pathologies.[10,14] Based on the distribution of abnormal findings in the X-ray 14 dataset, the number of abnormal findings was calculated as 80 (55%) for mass/nodules,

31 (21%) for consolidation, and 35 (24%) for pneumothorax. All CRs were collected from patients over 19 years old. CRs with abnormal findings were classified into 1 of 3 categories of nodule/mass, localized consolidation, and pneumothorax and selected by following inclusion and exclusion criteria. First, all CRs had normal or only 1 category of abnormal findings. Second, the nodules/masses that were pathologically or clinically diagnosed were visible on CR. Third, regardless of diagnosis, consolidations were localized on CR or CT. To ensure the reliability of the nodule/mass or consolidation, patients who underwent CR and chest CT within 1 month were included. If underlying diffuse lung lesions or concurrent pathologic lesions were identified on CR, they were excluded from the abnormal findings cases. If lesions were smaller than 5 mm or could not be identified on CR, those were excluded from abnormal findings. If the CRs showed more than 4 lesions of a single type or different types of lesions, they were also excluded from the dataset. CRs with pneumothorax and chest draining catheter were excluded. Supplement 1, http://links.lww.com/MD2/A91 describes the inclusion and exclusion criteria in data collection. To select cases with normal and abnormal results, 12-years experienced thoracic radiologist retrospectively reviewed images and radiologic reports on PACS. Normal cases were included from subjected who visited health screening center or oncology outpatient clinic. Cases with any abnormal findings on CR or CT were excluded. Abnormal cases with pulmonary nodules were selected from the pathologically confirmed cases in subjects who underwent CT-guided biopsy. Abnormal cases with localized consolidation were selected from subjects who visited the emergency department or respiratory medicine. Abnormal cases with pneumothorax were selected from subjects who visited the department of thoracic surgery.

2.3. Establishing the standard of reference

To establish the standard of reference, all CRs were labeled and annotated by board-certified radiologists. They confirmed whether the CRs were classified correctly and marked locations of any abnormal findings on the image. Normal CRs and abnormal CRs with nodule/mass and consolidation were reviewed by 1 board-certified radiologist, while CRs with pneumothorax were reviewed by 2 board-certified radiologists. All annotated lesions in CRs with nodule/mass or consolidation were considered true lesions, but only the lesions annotated by 2 radiologists in consensus were selected as the reference standards of CRs with pneumothorax.

2.4. The observer performance test

The observer performance test was conducted to compare the performance of the algorithm and physicians and to assess whether the performance of physicians can be enhanced with the aid of the algorithm. The reader panel comprised 6 physicians with various backgrounds: 2 board-certified radiologists, 2 non-radiology physicians, and 2 general practitioners. The physicians who participated in the establishment of the standard of reference were excluded from the observer performance test. The test proceeded in 2 sessions. In session 1, any diagnostic information related to the CRs was concealed, and the observers independently assessed each CR without the help of the algorithm. First, they were required to determine whether significant abnormal findings that need further examination or treatment existed on the CRs and provide confidence levels from 1 to 5. Then, they

marked the location of abnormalities and scored the confidence level based on their certainty. In session 2, with the assistance of the algorithm, the observers re-evaluated each CR. They could change or confirm their first decision on classification, localization, or lesion-based abnormality score.

Each assessment result of 6 different physicians for 244 CRs was treated independently. For both image classification and lesion localization, the cutoff value of the confidence level was established at 1, so if any confidence level marked on an image or lesion was above 1, it was considered abnormal. Finally, the results of each session were compared with the standard of reference to measure the agreement on the existence and localization of lesions.

2.5. Statistical analysis

Receiver operating characteristic (ROC) analysis was applied to evaluate the performance of the algorithm and physicians in image classification. The maximum value of the lesion-based abnormality score was considered the patient-based abnormality score, and the cutoff value was chosen between 1 and 5 points. If the patient-based abnormality score was higher than the cutoff value, the CR was classified as positive, which means there was a significant lesion, but if the score was lower, the CR was classified as negative. Then, compared with the standard of reference, the ROC curve was plotted with the true positive rate (TPR) and false-positive rate (FPR) values to measure the AUC. For both sessions 1 and 2 in the observer performance test, the AUCs were calculated at a 95% confidence interval.

Likewise, the quality of lesion detection was evaluated based on the following standards: the area under the JAFROC curve, sensitivity, specificity, positive/negative predictive value, accuracy, and false-positive rate. For the JAFROC analysis, while the lesion marked by the observer was categorized as lesion localization (LL) or non-lesion localization (NL), the AFROC curve was plotted with the lesion-localization fraction (LLF) against the probability of at least 1 FP per normal CR. Both the readers and CRs were treated as random effects. For the analysis of sensitivity and specificity, the threshold for the output of the algorithm was defined as 15%. Thus, values below 15% were considered negative, meaning normal, and values above 15% were considered positive, meaning abnormal.

These statistical analyses for image classification and lesion detection were performed for both the standalone test and reader test. While the standalone test measured the efficacy of the algorithm, the reader test determined whether the performance of physicians was enhanced with the assistance of the algorithm. To compare the AUCs between physician alone test and DCAD-aided test, the DeLong test was used for computing P-values. The Dorfman-Berbaum-Metz test was used for the comparison of the weighted jackknife alternative free-response ROC (wJAFROC)

figure of merit (FOM) between physician alone test and DCAD-aided test and generalized estimating equations for the comparison of sensitivity, specificity, positive or negative predictive value, accuracy, and false-positive rates. A P-value ≤ 0.05 was considered significant. All statistical analyses were conducted using R software, version 3.5.3.

3. Results

Among 80 cases of nodules/masses, there were 75 CRs with a single nodule/mass and 5 CRs with 2 nodules/masses. For CRs with consolidations, lesions appeared on a single lobe in 22 CRs, multiple lobes of a single lung in 1 CR, and both lungs in 7 CRs. The demographic features of the CR dataset are described in Table 1.

3.1. Performance of the algorithm

In the standalone test using the standard reference of 244 CRs, the algorithm achieved an AUC of 0.9935 (95% confidence interval [CI], 0.9868–1.000), an wJAFROC FOM of 0.9735 (0.9539–0.9916), sensitivity of 97.26% (95% CI, 95.21–99.31), specificity of 92.86% (95% CI, 89.63–96.09), positive rate of 95.30% (95% CI, 92.65–97.96), negative rate of 92.86% (95% CI, 89.63–96.09), and accuracy of 95.49% (95% CI, 92.89–98.10). The performance of the model in both image classification and lesion localization was measured to be higher than that of 3 observer groups, including board-certified radiologists, non-radiology physicians, and general practitioners (Figs. 1 and 2). The AUC for detecting nodule/mass, consolidation, and pneumothorax was 0.9883 (95% CI, 0.99762–1.000), 1.000 (95% CI, 1.000–1.000), and 0.9997 (95% CI, 0.9989–1.000), respectively.

3.2. The observer performance test

For the primary evaluation of image classification, the overall AUC of pooled physicians was 0.8679 (95% CI, 0.8507–0.8850) in physician alone test and 0.9112 (95% CI, 0.8769–0.9454) in DCAD-aided test. This increase from physician alone test to DCAD-aided test was statistically significant (95% CI, 0.0487–0.0725). Regarding the speciality of observers, the AUCs of board-certified radiologists, non-radiology physicians, and general practitioners in physician alone test were 0.9313 (95% CI, 0.9097–0.9529), 0.9152 (95% CI, 0.8917–0.9389), and 0.7686 (95% CI, 0.7311–0.8062), respectively; in DCAD-aided test, with the assistance of the algorithm, the AUCs of the 3 groups increased to 0.9586 (95% CI, 0.9399–0.9730), 0.9440 (95% CI, 0.9425–0.9747), and 0.8940 (95% CI, 0.8663–0.9218) (Fig. 3). The increments of AUC (95% CI) between physician alone test and DCAD-aided test were 0.0237 (0.0139–0.0407) in board-certified radiologists, 0.0287 (0.0158–0.0416) in non-radiology physi-

Table 1

Demographic description of the dataset.

		Abnormal			
	Normal	Nodule/mass	Consolidation	Pneumothorax	Total
Number of patients	98	80	31	35	244
Female	59	39	18	33	149
Male	39	41	13	2	95
Age (yr)	48.84 ± 12.54	66.17 ± 11.16	58.55 ± 18.12	25.43 ± 9.83	52.40 ± 18.29
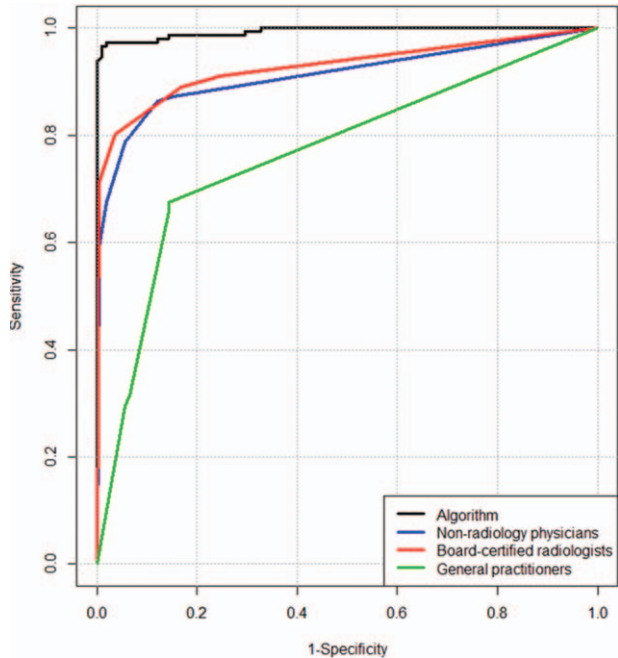
Figure 1. Performance of observers in image classification. The areas under the receiver operating characteristic receiver operating characteristic curves (AUCs) of the deep learning algorithm and observer groups are shown for the detection of 3 major thoracic abnormalities, including pulmonary nodules or masses, consolidation, and pneumothorax. AUC (95% confidence interval) was 0.9935 (0.9868–1.000) in deep learning-based CAD algorithm (DCAD), 0.9313 (0.9097–0.9529) in board-certified radiologists, 0.9153 (0.8917–0.9389) in non-radiology physicians, and 0.7686 (0.7311–0.8062) in general practitioners, respectively.
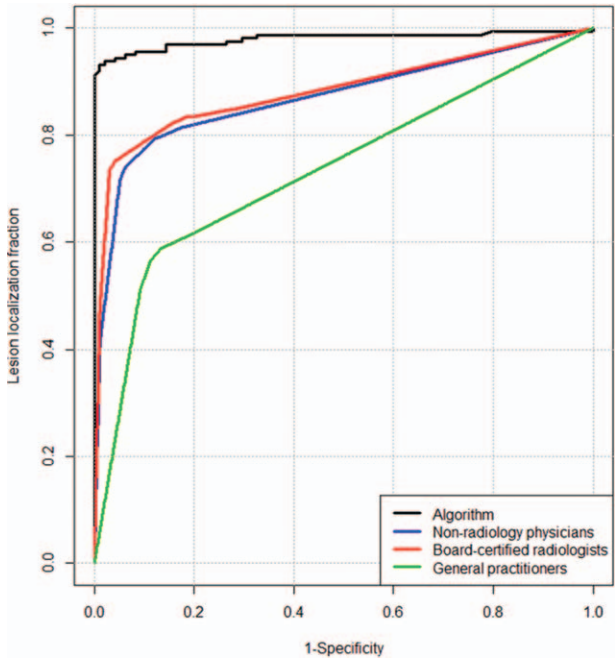
Figure 2. Performance of observers in lesion localization. A weighted jackknife alternative free-response ROC curves (wJAFROC) figure of merit (FOM) of the algorithm and observer groups are shown for the localization of 3 major thoracic abnormalities, including pulmonary nodules or masses, consolidation, and pneumothorax. A wJAFROC FOM (95% confidence interval) was 0.9735 (0.9539–0.9916) in deep learning-based CAD algorithm (DCAD), 0.8951 (0.8608–0.9293) in board-certified radiologists, 0.8859 (0.8518–0.9200) in non-radiology physicians, and 0.7686 (0.7311–0.8062) in general practitioners, respectively.
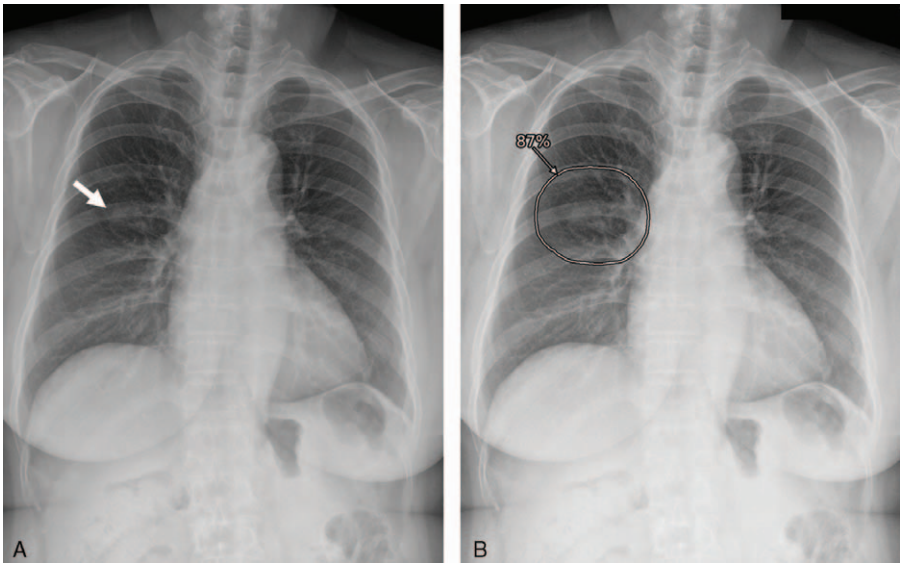
Figure 3. Chest radiograph with pulmonary nodule and result of the deep learning-based computer-aided diagnosis (CAD) algorithm. (A) A small ill-defined nodular opacity in the right middle lung zone is noted on the chest radiograph (arrow); (B) the deep learning-based CAD algorithm correctly detected and localized the nodule with a corresponding annotated circle and abnormality score of 87%. Among 6 observers, 3 classified the image as normal. After reviewing the results of the deep learning-based CAD algorithm, all 3 observers changed the result to abnormal with localization of the pulmonary nodule.

Table 2

Performance of observers in image classification.

Observer group	AUC (95% CI)		
	Physician alone test	DCAD-aided test	DCAD-aided test – physician alone test[*]
Board-certified radiologists	0.9313 (0.9097–0.9529)	0.9586 (0.9399–0.9730)	0.0237 (0.0139–0.0407)
Non-radiology physicians	0.9153 (0.8917–0.9389)	0.9440 (0.9425–0.9747)	0.0287 (0.0158–0.0416)
General practitioners	0.7686 (0.7311–0.8062)	0.8940 (0.8663–0.9218)	0.1248 (0.0960–0.1548)
Total	0.8679 (0.8507–0.8850)	0.9285 (0.9157–0.9413)	0.0606 (0.0487–0.0725)

Tests 1 and 2 were performed by observers without and with the aid of the algorithm. AUC = area under the receiver operating characteristic curve, CI = confidence interval, DCAD = deep learning-based computer-aided diagnosis algorithm.

[*] DeLong test for comparison of the AUC.

Table 3

Performance of observers in lesion localization.

	Board-certified radiologists			Non-radiology physicians			General practitioners			Total		
	Physician alone test	DCAD-aided test	P	Physician alone test	DCAD-aided test	P	Physician alone test	DCAD-aided test	P	Physician alone test	DCAD-aided test	P[*]
wJAFROC FOM	0.8951 (0.8608–0.9293)	0.9307 (0.9032–0.9582)	.0001	0.8859 (0.8518–0.9200)	0.9222 (0.8950–0.9494)	<.0001	0.7686 (0.7311–0.8062)	0.8940 (0.8663–0.9218)	<.0001	0.8426 (0.7635–0.9217)	0.9112 (0.8769–0.9454)	.0287
Sensitivity %	91.72 (89.27–94.17)	93.75 (91.60–95.90)	.0004	87.93 (85.03–90.83)	92.41 (90.06–94.77)	<.0001	67.93 (63.78–72.08)	86.21 (83.14–89.27)	<.0001	82.53 (80.58–84.48)	91.72 (90.31–93.14)	<.0001
Specificity %	75.51 (71.69–79.33)	85.89 (82.79–88.98)	.316	84.69 (81.49–87.89)	91.33 (88.82–93.83)	.0002	85.71 (82.60–88.83)	91.33 (88.82–93.83)	.0023	81.97 (80.00–83.95)	86.39 (84.63–88.15)	<.0001
Positive rate %	84.71 (81.51–87.91)	76.53 (72.76–80.30)	.0005	89.47 (86.75–92.20)	94.04 (91.93–96.14)	<.0001	87.56 (84.62–90.49)	93.63 (91.46–95.80)	<.0001	87.14 (85.42–88.85)	90.89 (89.41–92.37)	<.0001
Negative rate %	86.05 (82.97–89.13)	96.55 (94.93–98.17)	.0003	82.59 (79.22–85.96)	89.05 (86.28–91.83)	<.0001	64.37 (60.11–68.63)	81.74 (78.30–85.17)	<.0001	76.03 (73.83–78.22)	87.59 (85.59–89.28)	<.0001
Accuracy %	85.19 (82.03–88.34)	88.48 (85.64–91.32)	.0004	86.63 (83.60–89.65)	91.98 (89.56–94.39)	<.0001	75.10 (71.26–78.95)	88.27 (85.41–91.13)	<.0001	82.30 (80.35–84.26)	89.57 (88.01–91.14)	<.0001

Physician alone test and DCAD-aided test were performed by observers without and with the aid of the algorithm. DCAD = deep learning-based computer-aided diagnosis algorithm, FOM = figure of merit, ROC = receiver operator characteristic, wJAFROC = weighted jackknife alternative free-response ROC.

[*] Dorfman-Berbaum-Metz test for comparison of the wJAFROC FOM and generalized estimating equations for the comparison of sensitivity, specificity, positive or negative predictive value, and accuracy.

cians, and 0.1248 (0.0960–0.1548) in general practitioners, which shows that the algorithm was effective in improving the performance of readers (Table 2).

The results for the secondary evaluation of lesion detection are described in Table 3. The pooled observers exhibited a wJAFROC FOM of 0.8426 (95% CI, 0.7653–0.9217) in physician alone test and 0.9112 (95% CI, 0.8769–0.9454) in DCAD-aided test, which also indicates significant improvement between tests ($P < .05$). Specifically, in physician alone test, the wJAFROC FOM of board-certified radiologists, non-radiology physicians, and general practitioners were 0.8951 (95% CI, 0.8608–0.9293), 0.8859 (95% CI, 0.8518–0.9200), and 0.7469 (95% CI, 0.7049–0.7889), respectively; in DCAD-aided test, these statistics significantly increased to 0.9307 (95% CI, 9032–0.9582), 0.9222 (95% CI, 0.8950–0.9494), and 0.8806 (95% CI, 0.8479–0.9132), meaning that the assistance of the algorithm had a meaningful impact on the performance of readers ($P < .05$).

In the case of board-certified radiologists, the sensitivity, negative rate, and accuracy showed significant improvement between physician alone test and DCAD-aided test ($P < .05$). The positive rate decreased significantly, while the increment in specificity was not meaningful. For non-radiology physicians and general practitioners, significant improvements in DCAD-aided tests were observed for sensitivity, specificity, positive rate, negative rate, and accuracy ($P < .05$). In terms of the false-positive rate, both non-radiology physicians and general practitioners achieved significantly lower scores in DCAD-aided test than in the physician alone test ($P < .01$) (Table 4).

Regarding types of pathology, the sensitivities of observers in localizing nodule/mass, consolidation, and pneumothorax were 78.13 (95% CI, 74.43–81.82), 86.02 (95% CI, 81.04–91.00), and 89.71 (95% CI, 85.54–93.88), respectively; these values increased to 88.75 (95% CI, 85.92–91.58), 91.94 (88.02–95.85), and 98.53 (96.88–100) in DCAD-aided test (Figs. 4 and 5).

Table 4

False-positive rates of observers.

	False-positive rate (95% CI)		
	Physician alone test	DCAD-aided test	P[*]
Board-certified radiologists	24.49 (18.47–30.51)	23.47 (17.54–29.40)	.316
Non-radiology physicians	15.31 (10.27–20.35)	8.67 (4.73–12.61)	.0002
General practitioners	14.29 (9.39–19.18)	8.67 (4.73–12.61)	.0023
Total	18.03 (14.92–21.13)	13.61 (10.83–16.38)	<.0001

The physician alone test and the DCAD-aided test were performed by observers without and with the aid of the algorithm. CI = confidence interval, DCAD = deep learning-based computer-aided diagnosis algorithm.

[*] Generalized estimating equations for the comparison of the false-positive rate.
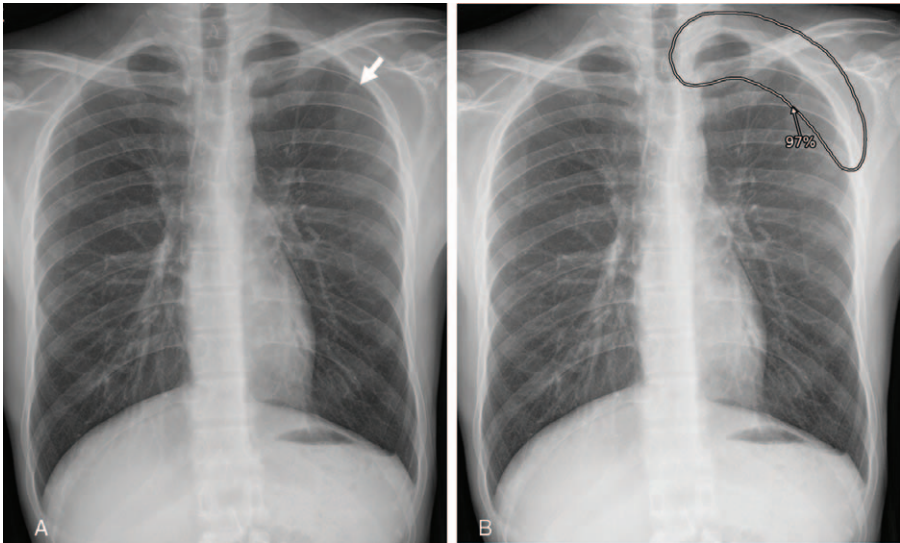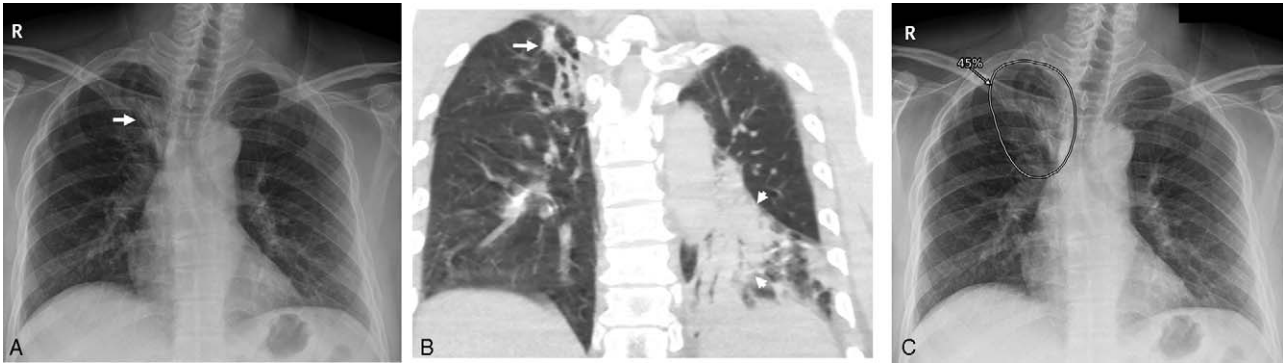
Figure 4. Chest radiograph with pneumothorax and result of the deep learning-based computer-aided diagnosis (CAD) algorithm. (A) A small left pneumothorax and pleural line in the right apex is noted on the chest radiograph (arrow); (B) the deep learning-based CAD algorithm correctly detected and localized the nodule with a corresponding annotated circle and abnormality score of 97%. Among 6 observers, 4 classified the image as normal. After reviewing the results of the deep learning-based CAD algorithm, 3 observers changed the result to abnormal with localization of the left pneumothorax.

4. Discussion

The purpose of this study was to validate the clinical impact of the algorithm in enhancing physicians' performance in detecting major thoracic lesions. The results proved that with the assistance of the algorithm, 3 reader groups made significant improvements in their performance, while the algorithm itself also showed notable performance in the stand-alone test. The improvement of performance in both image classification (AUC, 0.8679–0.9285; $P=.0606$) and lesion localization (AUC, 0.8426–0.9112; $P=.0287$) was shown in this study. Our results were very similar to those of a recent study showing the outperforming value of a deep learning-based algorithm on chest radiographs with multicenter datasets.[11] It showed improvements in diagnostic performance in both image-wise classifications (AUC, 0.814–0.932 to 0.904–0.958; all $P<.005$) and lesion-wise localization (AUC, 0.781–

0.907 to 0.873–0.938; all $P<.001$) with the assistance of the algorithm. Although our results were obtained with a single center dataset, we calculated the sample size and referred to the known distribution of major thoracic abnormalities on CRs in the open-source image database.[14]

Regarding detecting thoracic lesions, we found that the sensitivities of physicians were improved from 78.13% to 88.75% for nodule/mass, from 86.02 to 88.75% for consolidation, and from 89.71 to 98.53% for pneumothorax. Another study also revealed that the sensitivity of radiologists improved (from 65.1% to 70.3%, $P<.001$) and the number of false-positive findings per radiograph declined (from 0.2 to 0.18, $P<.001$) when the radiologists re-reviewed radiographs with the deep learning algorithm for pulmonary nodules.[15] However, pulmonary nodules are one of the abnormal findings frequently

Figure 5. Chest radiograph with pneumonia and result of the deep learning-based computer-aided diagnosis (CAD) algorithm. (A) Small right supra-hilar patchy opacity is noted on the chest radiograph (arrow); (B) coronal image of chest CT image obtained 3 d after the chest radiograph shows the bronchiectasis and atelectasis in the right upper lobe (arrow) and patchy consolidation in the left lower lobe (short arrows). (C) The deep learning-based CAD algorithm did not detect the pneumonic consolidation in the left lower lobe, while the algorithm detected and localized the bronchiectasis and atelectasis with a corresponding annotated circle and abnormality score of 45%. Among 6 observers, 4 classified the image as normal. After reviewing the results of the deep learning-based CAD algorithm, 2 observers changed the result to abnormal with localization of bronchiectasis and atelectasis in the right upper lobe, while none of observers detected the consolidation in the left lower lobe in both the physician alone test and DCAD-aided test.

noted in daily clinical practice. The strength of our study lies in the distinct algorithm, which evaluates 3 major thoracic abnormalities, including nodules or masses, consolidation, and pneumothorax. Another notable achievement of this algorithm is its high accuracy in localizing major pulmonary lesions, which enables physicians to confirm more accurate and reasonable diagnoses. The decision-making process of the algorithm is often compared to the black box due to its weak and sometimes defying explanation.[16,17] For instance, previous algorithms for classifying thoracic diseases from radiographs calculated probabilities or simply identified the existence of diseases without any reasoning.[18,19] Unlike human doctors who can list legitimate reasons for their diagnosis, how these neural networks arrive at successful results is untraceable. Even researchers who develop algorithms cannot explain whether algorithms learn the relationship between lesions and diseases or extracted patterns from irrelevant features, which makes physicians hesitant about adopting computer-aided detection.

In this context, rather than displaying only numbers, drawing heatmaps that visualize confidence levels for each lesion may earn more trust from physicians. By examining what the algorithm labeled abnormal on a CR, physicians can strengthen their reasoning behind the diagnosis and feel more confident with this assistance. Our algorithm is designed to detect 3 major pulmonary lesions, which may require further examination and may lead to the discovery of thoracic diseases such as lung cancer and tuberculosis. Indeed, the results proved that it leads to a higher level of detection than radiologists. Therefore, this lesion-localization function of our algorithm may unbox the black-box system and give more confidence to its users.

As several researchers have already suggested,[9,20,21] the most established application of the deep learning algorithm in chest radiography is the second reader. Compared to unaided reading, the assistance of computer-aided detection as the second reader increases the accuracy and sensitivity of screening thoracic diseases. According to Donald and Barnard,[2] perceptual errors occupied 81% of diagnostic errors that occurred on chest X-ray evaluation. If the second read by the algorithm follows after the initial interpretation by physicians, overlooking lung lesions can be prevented in advance. Indeed, White et al[20] revealed that 47% of lung cancers missed in the initial interpretation of radiographs were identified by computer-aided detection software. Hence, our algorithm, which exhibited outstanding performance in the observer performance test, may facilitate decision-making procedures in the clinical setting as the second reader.

Additionally, the results of the standalone test indicate that utilizing our algorithm as the concurrent reader can be clinically useful. Prior to interpretation by humans, the algorithm can automatically detect abnormalities and mark their locations with shorter reading times than the second-reader mode.[22] It can serve as a substitute in cases of urgent situations such as when skilled radiologists are not available, and by prioritizing between CRs with and without risk markers, physicians can examine patients suspicious of serious illnesses first. Additionally, the efficacy of our algorithm as the primary reader in the clinical setting has been validated by the recent research of Hwang et al,[23] which tested the same deep learning algorithm used in our study. The algorithm was used to retrospectively review CRs from consecutive patients who visited the emergency department. Although the dataset was obtained from consecutive patients at a tertiary hospital with a different population from our study, this algorithm showed outstanding performance in classifying abnormal CRs, achieving an AUC of 0.95.

Considering that the influence of the algorithm differed according to the proficiency of physicians in the observer performance test, 2 different scenarios can be proposed for the application of our algorithm. First, the feedback of the algorithm can prevent radiologists from making errors or biases. Regardless of experience, physicians cannot avoid the fact that they are humans who are prone to a variety of mistakes, such as interpretation errors, omission errors, and cognitive biases.[2,24–26] Though the failure to correctly interpret CRs often leads to fatal consequences, problems such as a shortage of radiologists and heavy workloads still augment visual and mental fatigue, increasing radiologic errors.[27] On the other hand, our detection algorithm only needs appropriate hardware and electricity to produce stable outputs. Thus, by providing a reliable reference, our algorithm may lighten the burden of physicians and improve both the working environment and healthcare service for patients. Second, in comparison with the other 2 groups, the general practitioner group showed a significantly higher increase in their performance after receiving the assistance of the algorithm. This finding indicates that the algorithm will be more helpful to medical care centers suffering shortages of skilled radiologists. Rural areas or undeveloped regions are often isolated from medical services, lacking both trained radiologists and advanced medical equipment. As written in the WHO 2016 report, CXR is recommended as a useful triaging and diagnostic tool due to its low operating costs, easy operation, and low radiation dose.[28] If computer-aided detection is adopted in those rural areas, people will be able to access more qualified and reasonable medical treatment without expensing large costs to equipment or relying on distant hospitals.[29] Similarly, in local clinics, the algorithm trained with various cases of large hospitals can supplement the diagnosis of doctors as a subsidiary tool.

There are several limitations in our study. First, the dataset used for training and validating the algorithm was formed experimentally. Though the collected dataset followed the prevalence rate of pathologies, each CR only included 1 type of pathology. Additionally, whether the algorithm can concurrently distinguish rare and clinically irrelevant types of lesions in a single CR is unknown. Second, the algorithm covered only 3 major thoracic lesions, which occupy a large proportion but not the whole spectrum of all lesions existing in real-world situations. Third, the results of our study are limited to a single center, so the generalizability to other institutions is uncertain. Nevertheless, as mentioned earlier, Hwang et al,[23] who used the same algorithm, determined that this algorithm exhibited a successful performance in classifying lung abnormalities in CRs of consecutive patients while identifying even nontarget diseases. Nevertheless, prospectively designed datasets will greatly enhance the methodological quality of the research.

In this study, we demonstrated better performance of physicians in the diagnosis of major pulmonary lesions assisted by the deep learning-based CAD algorithm in terms of image classification and lesion localization. The improvement of diagnostic performance was significant in both radiologists and non-radiology physicians or general practitioners. In conclusion, our study presented an algorithm that can detect 3 major pulmonary lesions at high accuracy and contribute to the enhancement of physicians' performance. Further studies are expected to validate the efficacy of this algorithm in a prospective setting.

Author contributions

Conceptualization: Sunggyun Park, Kwang Nam Jin.
Data curation: Sunggyun Park, Minchul Kim, Jongchan Park, Kwang Nam Jin.
Formal analysis: Soo Yun Choi, Sunggyun Park, Ye Ra Choi, Kwang Nam Jin.
Funding acquisition: Kwang Nam Jin.
Investigation: Sunggyun Park, Ye Ra Choi, Kwang Nam Jin.
Methodology: Sunggyun Park, Minchul Kim, Jongchan Park, Kwang Nam Jin.
Project administration: Kwang Nam Jin.
Resources: Soo Yun Choi, Minchul Kim, Jongchan Park, Kwang Nam Jin.
Software: Sunggyun Park, Minchul Kim, Jongchan Park, Kwang Nam Jin.
Supervision: Kwang Nam Jin.
Validation: Sunggyun Park, Ye Ra Choi, Kwang Nam Jin.
Visualization: Soo Yun Choi, Kwang Nam Jin.
Writing – original draft: Soo Yun Choi, Kwang Nam Jin.
Writing – review & editing: Kwang Nam Jin.

References

[1] Koo HJ, Choi C-M, Park S, et al. Chest radiography surveillance for lung cancer: results from a National Health Insurance database in South Korea. Lung Cancer 2019;128:120–6.
[2] Donald JJ, Barnard SA. Common patterns in 558 diagnostic radiology errors. J Med Imaging Radiat Oncol 2012;56:173–8.
[3] Rodriguez-Ruiz A, Krupinski E, Mordang JJ, et al. Detection of breast cancer with mammography: effect of an artificial intelligence support system. Radiology 2019;290:305–14.
[4] Sadoughi F, Kazemy Z, Hamedan F, et al. Artificial intelligence methods for the diagnosis of breast cancer by image processing: a review. Breast Cancer (Dove Med Press) 2018;10:219–30.
[5] Rudie JD, Rauschecker AM, Bryan RN, et al. Emerging applications of artificial intelligence in neuro-oncology. Radiology 2019;290:607–18.
[6] Zhou LQ, Wang JY, Yu SY, et al. Artificial intelligence in medical imaging of the liver. World J Gastroenterol 2019;25:672–82.
[7] Gao Z, Chung J, Abdelrazek M, et al. Privileged modality distillation for vessel border detection in intracoronary imaging. IEEE Trans Med Imaging 2020;39:1524–34.
[8] Gao Z, Wang X, Sun S, et al. Learning physical properties in complex visual scenes: an intelligent machine for perceiving blood flow dynamics from static CT angiography imaging. Neural Networks 2020;123:82–93.
[9] Hwang EJ, Park S, Jin KN, et al. Development and validation of a deep learning-based automatic detection algorithm for active pulmonary tuberculosis on chest radiographs. Clin Infect Dis 2019;69:739–47.
[10] Singh R, Kalra MK, Nitiwarangkul C, et al. Deep learning in chest radiography: detection of findings and presence of change. PLoS One 2018;13:e0204155.
[11] Hwang EJ, Park S, Jin KN, et al. Development and validation of a deep learning-based automated detection algorithm for major thoracic diseases on chest radiographs. JAMA Netw Open 2019;2:e191095.
[12] Park S, Lee SM, Kim N, et al. Application of deep learning-based computer-aided detection system: detecting pneumothorax on chest radiograph after biopsy. Eur Radiol 2019;29:5341–8.
[13] Cho Y, Kim YG, Lee SM, et al. Reproducibility of abnormality detection on chest radiographs using convolutional neural network in paired radiographs obtained within a short-term interval. Sci Rep 2020;10:17417.
[14] Wang X, Peng Y, Lu L, et al. Chest X-ray 8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. Paper presented at: 2017 IEEE Conference on Computer Vision and Pattern Recognition 2017.
[15] Sim Y, Chung MJ, Kotter E, et al. Deep convolutional neural network-based software improves radiologist detection of malignant lung nodules on chest radiographs. Radiology 2020;294:199–209.
[16] Razzak MI, Naz S, Zaib A. Dey N, Ashour AS, Borra S. Deep learning for medical image processing: overview, challenges and the future. Classification in Bioapps: Automation of Decision Making Cham: Springer International Publishing; 2018;323–50.
[17] Wang X, Li L, Liu W, et al. An interactive system for computer-aided diagnosis of breast masses. J Digit Imaging 2012;25:570–9.
[18] Smidth M, Sokolowski I, Kærsvang L, et al. Developing an algorithm to identify people with chronic obstructive pulmonary disease (COPD) using administrative data. BMC Med Inform Decis Mak 2012;12:38.
[19] Rohilla A, Hooda R, Mittal A. TB detection in chest radiograph using deep learning architecture. Paper presented at: Proceeding of 5th International Conference on Emerging Trends in Engineering, Technology, Science and Management (ICETETSM-17) 2017.
[20] White CS, Flukinger T, Jeudy J, et al. Use of a computer-aided detection system to detect missed lung cancer at chest radiography. Radiology 2009;252:273–81.
[21] Peldschus K, Herzog P, Wood SA, et al. Computer-aided diagnosis as a second reader: spectrum of findings in CT studies of the chest interpreted as normal. Chest 2005;128:1517–23.
[22] Matsumoto S, Ohno Y, Aoki T, et al. Computer-aided detection of lung nodules on multidetector CT in concurrent-reader and second-reader modes: a comparative study. Eur J Radiol 2013;82:1332–7.
[23] Hwang EJ, Nam JG, Lim WH, et al. Deep learning for chest radiograph diagnosis in the emergency department. Radiology 2019;293:573–80.
[24] Waite S, Scott J, Gale B, et al. Interpretive error in radiology. Am J Roentgenol 2017;208:739–49.
[25] McLain PL, Kirkwood CR. The quality of emergency room radiograph interpretations. J Fam Pract 1985;20:443–8.
[26] Olivetti L, Fileni A, De Stefano F, et al. The legal implications of error in radiology. Radiol Med 2008;113:599–608.
[27] Brady AP. Error and discrepancy in radiology: inevitable or avoidable? Insights Imaging 2017;8:171–82.
[28] World Health O. Chest Radiography in Tuberculosis Detection: Summary of Current WHO Recommendations and Guidance on Programmatic Approaches. Geneva: World Health Organization; 2016.
[29] Cecchetti AA. Why introduce machine learning to rural health care? Marshall J Med 2018;4:2.