diagnostics

MDPI

Article

Methods of Visualizing the Results of an Artificial-Intelligence-Based Computer-Aided Detection System for Chest Radiographs: Effect on the Diagnostic Performance of Radiologists

Sungho Hong [1], Eui Jin Hwang [1,2,*], Soojin Kim [1], Jiyoung Song [1], Taehee Lee [1], Gyeong Deok Jo [1], Yelim Choi [1], Chang Min Park [1,2,3] and Jin Mo Goo [1,2,3]

[1] Department of Radiology, Seoul National University Hospital, Seoul 03082, Republic of Korea
[2] Department of Radiology, Seoul National University College of Medicine, Seoul 03082, Republic of Korea
[3] Institute of Radiation Medicine, Seoul National University Medical Research Center, Seoul 03082, Republic of Korea
* Correspondence: ken921004@hotmail.com; Tel.: +82-2-2072-2057

Abstract: It is unclear whether the visualization methods for artificial-intelligence-based computer-aided detection (AI-CAD) of chest radiographs influence the accuracy of readers' interpretation. We aimed to evaluate the accuracy of radiologists' interpretations of chest radiographs using different visualization methods for the same AI-CAD. Initial chest radiographs of patients with acute respiratory symptoms were retrospectively collected. A commercialized AI-CAD using three different methods of visualizing was applied: (a) closed-line method, (b) heat map method, and (c) combined method. A reader test was conducted with five trainee radiologists over three interpretation sessions. In each session, the chest radiographs were interpreted using AI-CAD with one of the three visualization methods in random order. Examination-level sensitivity and accuracy, and lesion-level detection rates for clinically significant abnormalities were evaluated for the three visualization methods. The sensitivity ($p = 0.007$) and accuracy ($p = 0.037$) of the combined method are significantly higher than that of the closed-line method. Detection rates using the heat map method ($p = 0.043$) and the combined method ($p = 0.004$) are significantly higher than those using the closed-line method. The methods for visualizing AI-CAD results for chest radiographs influenced the performance of radiologists' interpretations. Combining the closed-line and heat map methods for visualizing AI-CAD results led to the highest sensitivity and accuracy of radiologists.

Keywords: chest radiography; artificial intelligence; deep learning; computer-aided detection; diagnostic accuracy

Citation: Hong, S.; Hwang, E.J.; Kim, S.; Song, J.; Lee, T.; Jo, G.D.; Choi, Y.; Park, C.M.; Goo, J.M. Methods of Visualizing the Results of an Artificial-Intelligence-Based Computer-Aided Detection System for Chest Radiographs: Effect on the Diagnostic Performance of Radiologists. *Diagnostics* **2023**, *13*, 1089. https://doi.org/10.3390/diagnostics13061089

Academic Editors: Chiara Romei and Emanuele Neri

Received: 19 January 2023
Revised: 2 March 2023
Accepted: 12 March 2023
Published: 13 March 2023

1. Introduction

Chest radiography is at the forefront of the recent trend of applying artificial intelligence (AI) technology in daily clinical practice. Indeed, AI-based software that can identify various types of abnormalities has been developed and utilized in clinical practice [1–3]. Among the various clinical applications, the use of artificial-intelligence-based software as a computer-aided detection (CAD) tool to help radiologists or physicians identify subtle abnormalities has been most widely accepted [1,2,4–6].

For artificial-intelligence-based computer-aided detection (AI-CAD) tools, the primary aim is to enhance the detection performance of interpreting radiologists or physicians [2,4–7]. Therefore, in addition to intrinsic performance, the method of delivering the results of the analysis to physicians is the key component of an AI-CAD tool to demonstrate its efficacy and value in clinical practice. Typically, an AI-CAD tool provides its results with annotations highlighting the location of the detected abnormality overlaid on the input image, as well as a confidence score for the detection [8,9]. Color-coded heat maps and closed lines along

the boundary of the abnormality are two representative methods for visualizing AI-CAD results [3,10,11].

Choosing one of the two methods of visualization, or using them in combination, may have an influence on the interaction between the AI-CAD and interpreting physician and the performance of the interpreting physician. However, to date, most studies on the development and validation of AI-CAD focused on the performance of the AI itself rather than on the method of visualization of the result. Therefore, it remains unclear the method of visualization that is optimal for improving the performance of radiologists' interpretations.

Therefore, the purpose of our study was to investigate the accuracy of radiologists' interpretation in identifying clinically relevant abnormalities on chest radiographs using AI-CAD with different visualization methods, and to explore the optimal visualization method for AI-CAD for chest radiographs.

2. Materials and Methods

2.1. Patients

We retrospectively included patients who met the following inclusion criteria: (a) visited the emergency department of a tertiary referral hospital in South Korea between 1 January and 30 June 2017; (b) underwent chest radiography for evaluation of acute respiratory symptoms; and (c) underwent chest CT during their stay in the emergency department.

A total of 249 chest radiographs were obtained from 249 patients (male-to-female ratio, 148:101; mean age ± standard deviation, 62 ± 17 years). Table 1 shows the patients' demographic and clinical information. A total of 189 (75.9%) chest radiographs were obtained by using a fixed radiography scanner. The most common chief complaint of visiting the emergency department was dyspnea (18.9%), followed by chest pain (15.3%) and fever (11.6%).

Table 1. Demographic and clinical characteristics of the patients included in the study.

Variables	Number of Patients (%)
Male patients	148 (59.4%)
Chest radiographs from fixed scanner	189 (75.9%)
Chief complaint for visiting emergency department	
Dyspnea	47 (18.9%)
Chest pain	38 (15.3%)
Fever	29 (11.6%)
Hemoptysis	26 (10.4%)
Cough	19 (7.6%)
Generalized weakness	13 (5.2%)
Others	77 (30.9%)

In the present study, 49.8% (124/249) of patients were reported in previous studies [12,13]. However, the purpose of previous studies was to evaluate the performance of an AI-CAD to identify clinically significant abnormalities among chest radiographs from patients in the emergency department [12] and to evaluate the calibration of the AI-CAD [13], which was entirely different from that of the present study.

2.2. Chest Radiographs

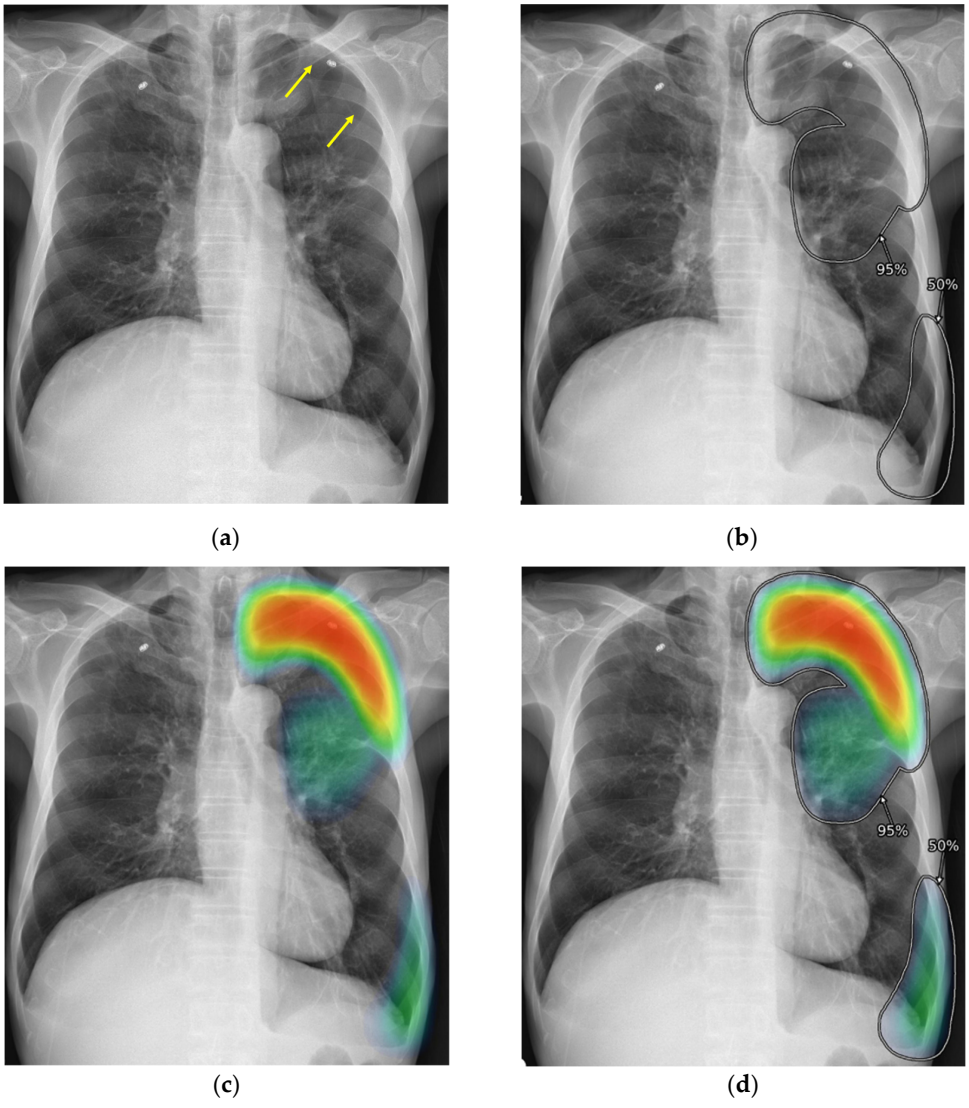
The study included only initial chest radiographs obtained in the emergency department (one radiograph per patient). In cases of multiple visits to the emergency department during the study period, chest radiographs obtained at the initial visit were included.

Posteroanterior and anteroposterior radiographs were included in this study. Posteroanterior radiographs were obtained in an erect position using a single fixed radiog-

raphy unit (Multix FD; Siemens Helthineers, Erlangen, Germany), while anteroposterior radiographs were obtained in the supine position using a portable radiography scanner (DRX-Revolution; Carestream Health, Rochester, NY, USA).

2.3. AI-CAD

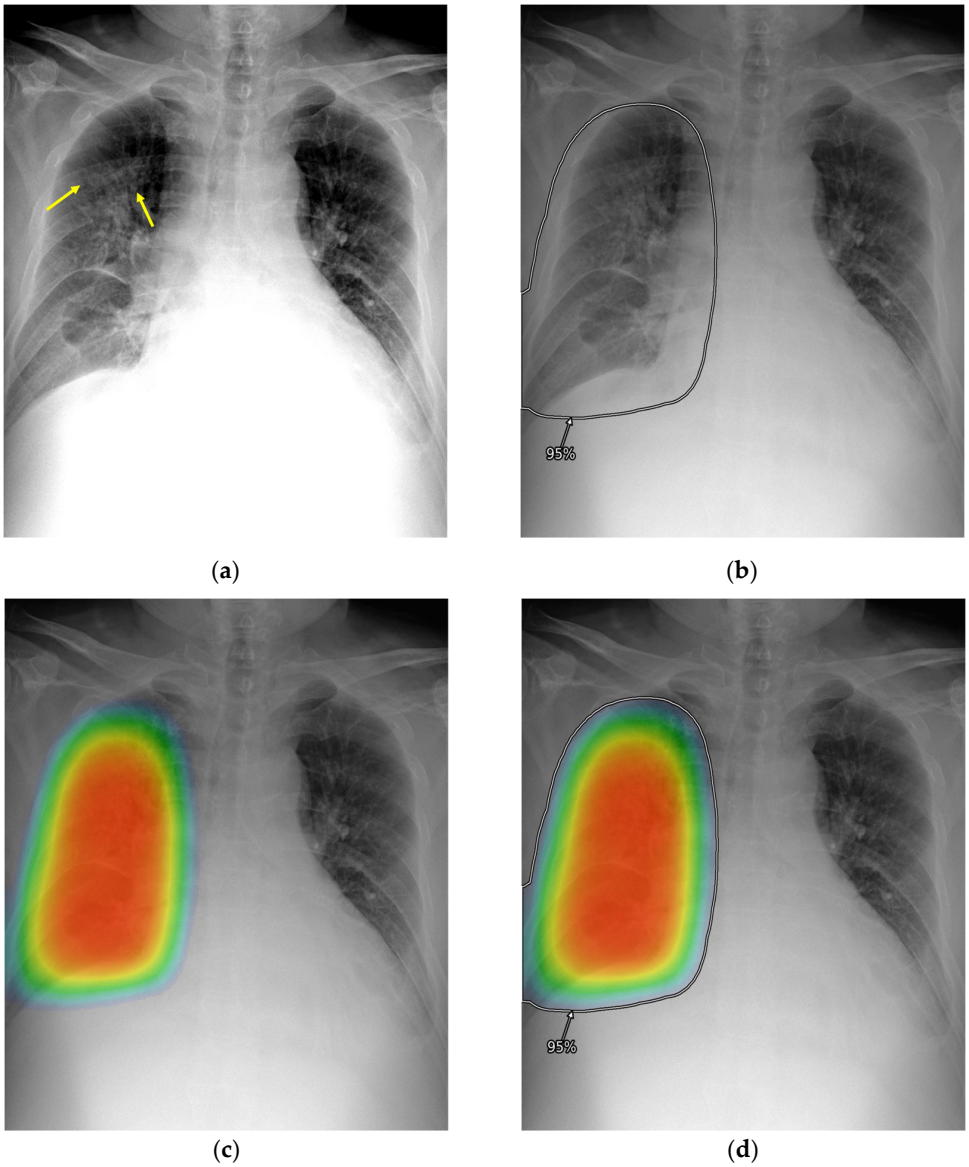
Commercially available AI-CAD (Lunit INSIGHT for CXR, version 2.0.3.0; Lunit, Seoul, Korea) was retrospectively applied to chest radiographs. AI-CAD was designed to identify pulmonary nodules, pulmonary infiltration, and pneumothorax on a single frontal chest radiograph, with a confidence score (0–100%) for the presence of identified abnormality [14]. Three different methods were utilized for the visualization of the AI-CAD analysis results (Figures 1 and 2).



(**a**) (**b**)

(**c**) (**d**)

Figure 1. Representative examples showing that the advantages of the heat map method can complement the disadvantage of the closed-line method. (**a**) Chest radiography of a 75-year-old man who visited an emergency department with chest pain showed right pneumothorax (arrows). The artificial-intelligence-based computer-aided detection (AI-CAD) detected pneumothorax with a probability score of 95%. With visualization of AI-CAD result by closed-line method (**b**), one of five radiologists missed the pneumothorax. Meanwhile, with heat map method (**c**) and combined method (**d**), all five trainee radiologists identified the abnormality.



(**a**) (**b**)

(**c**) (**d**)

Figure 2. Representative examples showing that the advantages of the closed-line method can complement the disadvantage of the heat map method. (**a**) Chest radiography of a 61-year-old man who visited an emergency department with a fever showed increased opacity at the right upper lung field (arrows) suggesting pneumonia, and bilateral pleural effusion. The artificial-intelligence-based computer-aided detection (AI-CAD) detected the opacity with a probability score of 95%. With visualization of AI-CAD result by closed-line method (**b**), heat map method (**c**), and combined method (**d**), the increased parenchymal opacity was identified by three, four, and five trainee radiologists, respectively, whereas bilateral pleural effusion was identified by two, two, and three radiologists, respectively.

Heat map method: A heat map was overlaid on the identified abnormalities. The color of the heat map represents the confidence score. Higher confidence scores are coded in red, while lower confidence scores are coded in blue.

Closed-line method: A closed line without any color information is displayed along the boundary of the identified abnormality. Confidence scores were directly visualized with numbers next to the closed curve.

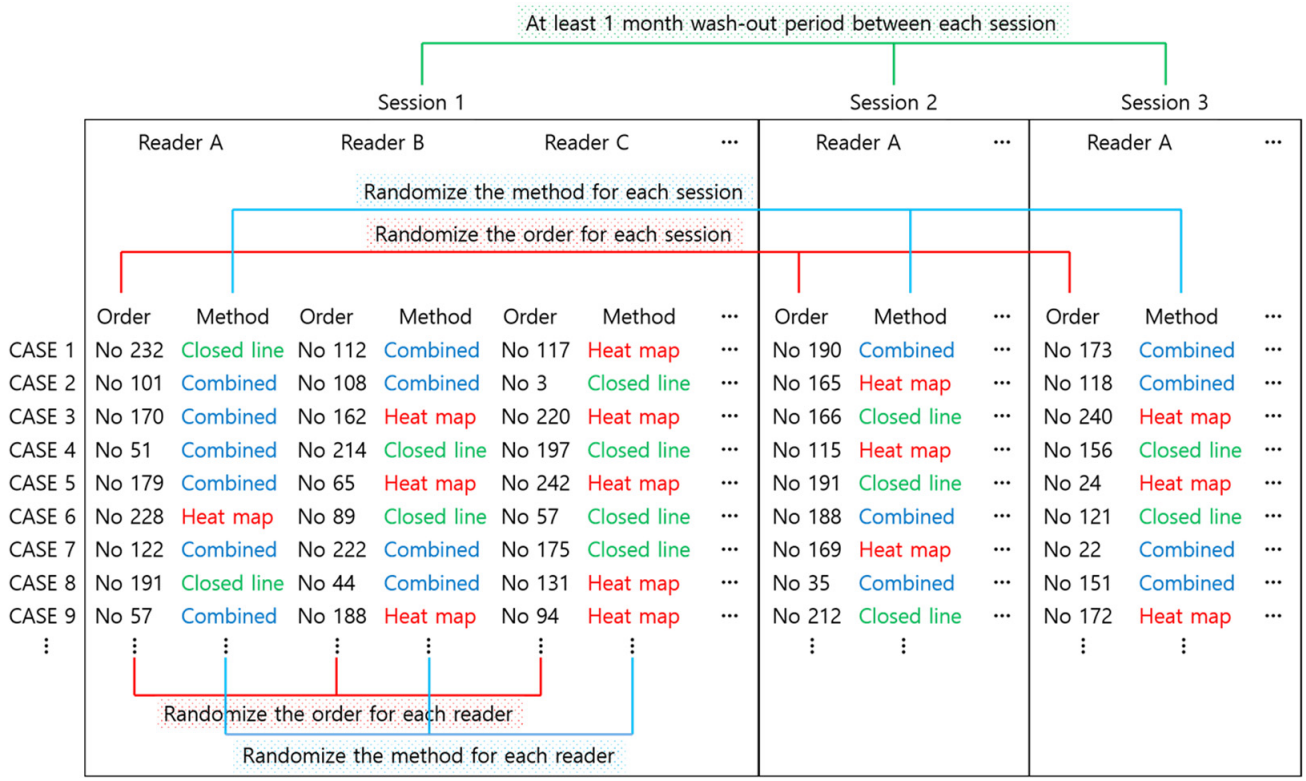
Combined method: Information visualized using both the heat map and closed-line methods was visualized on a single image.

The threshold confidence score for visualization was 15%, the default value provided by the manufacturer.

2.4. Reader Test

Five trainee radiologists (S.K., J.S., T.L., G.D.J., and Y.C.; 1st to 3rd year of residency training) participated in the reader test. The reader test consisted of three interpretations. In each interpretation session, readers interpreted the chest radiographs with AI-CAD results using one of the three visualization methods. Readers were informed that all chest radiographs were obtained from patients who visited an emergency department with acute respiratory symptoms, and the chief complaint of visiting the emergency department was also provided. First, the readers were asked to answer whether there were any clinically significant abnormalities requiring further evaluation or treatment. In case of an abnormality, readers were asked to describe up to three abnormal findings.

In order to minimize the bias caused by repeated interpretation of a single chest radiograph, the sequence of the utilization of each visualization method was randomized for each reader and the chest radiograph. Further, the sequences of chest radiographs in each interpretation session were randomly reshuffled in each interpretation session for all readers (Scheme 1). Finally, a wash-out period of at least one month was set up between each interpretation session.



Scheme 1. A scheme showing the randomization of the method and order for each reader and each session.

2.5. Reference Standard and Performance Metrics

To define the reference standard for the presence of any clinically significant abnormality, a single thoracic radiologist (E.J.H.; 11 years of experience in the interpretation of chest radiographs and chest CTs) reviewed the chest radiographs and the corresponding chest CTs. With reference to chest CT, up to two key abnormal findings that may be associated with patients' respiratory symptoms and require further evaluation or treatment were defined as the reference standards. Key abnormal findings were classified into six categories: (a) pulmonary nodule or mass, (b) pulmonary air-space opacity, (c) pulmonary interstitial

opacity, (d) pleural effusion, (e) pneumothorax, and (f) others. Subtle abnormalities that could not be identified on chest radiographs, even in the retrospective review of chest CT, were excluded from the reference standard.

Sensitivity, specificity, and accuracy were used to evaluate examination-level classification performance. For evaluation of sensitivity and accuracy, the interpretation of readers or AI-CAD results was regarded as true-positive only when at least one key abnormal finding was correctly identified. Sensitivity was defined as the proportion of true-positive interpretations among the radiographs with positive reference standards, and accuracy was defined as the proportion of true-positive and true-negative interpretations among all radiographs.

For evaluation of the lesion-level detection performance, the detection rate (the proportion of correctly identified abnormalities among all clinically significant abnormalities by the reference standard) was used. The detection rate of each abnormality type was also investigated.

2.6. Preference Survey

After completing all three sessions of the reader test, readers were asked to complete a questionnaire to survey their subjective preferences for each visualization method. The questionnaire comprised seven items: (a) conspicuity of the result, (b) interpretability of the result, (c) convenience for the correlation between the original image and AI-CAD result, (d) degree of visual fatigue, (e) subjective impression to improve interpretation speed, (f) subjective impression to improve interpretation accuracy, and (g) overall preference. Readers answered each question with five-point-scale scores.

2.7. Statistical Analysis

Statistical analyses were conducted using IBM SPSS Statistics (version 25; IBM, Armonk, NY, USA) and R (version 3.6.3; R Foundation for Statistical Computing, Vienna, Austria). To consider the clustering effect caused by multiple evaluations of single chest radiographs by multiple readers and multiple visualization methods, we used binary logistic regression with generalized estimating equations to estimate the average sensitivity, specificity, accuracy, and detection rate of the readers [15]. The detection rate of each abnormality type was also investigated. The variability of each performance metric among the five readers was evaluated using the coefficient of variation and compared using Levene's F-test. p-values < 0.05 were considered statistically significant.

3. Results

3.1. Patient Demographics and Clinical Characteristics

Among 249 chest radiographs included in the study, 162 (65.1%) show clinically significant abnormalities according to the reference standard. Pulmonary air-space opacity (54.9%) is the most common abnormality, followed by pleural effusion (18.8%) and pulmonary nodule or mass (16.0%) (Table 2).

Table 2. Radiographic findings of patients.

Variables	Number of Patients (%)
Chest radiographs with significant abnormality	162 (65.1%)
Type of abnormality	
Pulmonary air-space opacity	89 (54.9%)
Pulmonary nodule or mass	27 (16.0%)
Pulmonary interstitial opacity	13 (7.7%)
Pleural effusion	32 (18.9%)
Pneumothorax	7 (4.3%)
Others *	7 (4.3%)

* Included two aortic dilatations, two endotracheal tube malposition, one mediastinal mass, one pneumomediastinum, and one rib fracture.

3.2. Examination-Level Classification Performances

Sensitivities, specificities, and accuracies of the interpretation by readers using AI-CAD with different visualization methods are described in Table 3. The highest sensitivity is observed in the combined method (71.5%; 95% CI, 65.4–76.8%), which is significantly higher than that in the closed-line method (68.2%; 95% CI, 62.2–73.6%; $p = 0.007$). Sensitivity in the heat map method does not significantly differ from that in the other two methods (70.3%; 95% CI, 64.3–75.7%; $p = 0.383$ [vs. combined method], $p = 0.129$ [vs. closed-line method]). The specificities of the interpretations do not significantly differ across the visualization methods. The accuracy of interpretation is highest in the combined method (77.0%; 95% CI, 72.6–80.9%), which is significantly higher than that in the closed-line method (75.2%; 95% CI, 70.7–79.2%; $p = 0.037$). Accuracy in the heat map method does not significantly differ from that in the other two methods (76.5%; 95% CI, 72.0–80.5%) (Table 3).

Table 3. Sensitivity, specificity, and accuracy of interpretation by readers and stand-alone AI-CAD.

Interpretation Method	Sensitivity	Specificity	Accuracy
Readers with closed-line method	68.2% (547/810; 62.2–73.6%)	89.4% (385/435; 83.6–93.4%)	75.2% (932/1295; 70.7–79.2%)
Readers with heat map method	70.3% (564/810; 64.3–75.7%)	89.2% (384/435; 83.0–93.4%)	76.5% (948/1295; 72.0–80.5%)
Readers with combined method	71.5% (573/810; 65.4–76.8%)	89.0% (383/435; 83.1–93.0%)	77.0% (956/1295; 72.6–80.9%)
p-value (closed-line vs. heat map method)	0.129	0.884	0.224
p-value (closed-line vs. combined method)	0.007	0.745	0.037
p-value (heat map vs. combined method)	0.383	0.881	0.516
Standalone AI-CAD	84.6% (137/162; 78.1–89.8%)	70.1% (61/87; 59.4–79.5%)	77.8% (198/249; 70.8–83.4%)
p-value (vs. readers with closed-line method)	<0.001	0.004	0.521
p-value (vs. readers with heat map method)	<0.001	0.002	0.485
p-value (vs. readers with combined method)	<0.001	0.003	0.492

Abbreviation: AI-CAD, artificial-intelligence-based computer-aided detection. Numbers in parentheses indicate numerators/denominators; 95% confidence intervals.

The performance of the stand-alone AI-CAD is also described in Table 3. The sensitivity (84.6%; 95% CI, 78.1–89.8%) of the stand-alone AI-CAD is significantly higher than the interpretation by readers for all visualization methods (all $p < 0.001$). Meanwhile, specificity (70.1%; 95% CI, 59.4–79.5%) is significantly lower than the interpretation by readers for all visualization methods (all $p < 0.05$). The accuracy of the stand-alone AI-CAD (77.8%; 95% CI, 70.8–83.4%) does not significantly differ from that of readers, regardless of the visualization methods.

Figure 3 and Table S1 in the Supplementary Materials show the accuracy of the interpretations of individual readers.

Figure 3. Accuracy of interpretation of individual readers, and stand-alone AI-CAD.

3.3. Lesion-Level Detection Performances

For identification of all types of abnormalities, detection rates of readers using the heat map method (66.8%; 95% CI, 61.0–72.1%) and combined method (67.5%; 95% CI, 61.7–72.8%) are significantly higher than those using the closed-line method (63.9%; 58.1–69.3%; $p = 0.043$ [vs. heat map method], $p = 0.004$ [vs. combined method]) (Table 4). The detection rates for the different types of abnormalities are described in Table 4.

Table 4. Detection rates of readers and stand-alone AI-CAD for different types of abnormalities.

Interpretation Method	All Abnormalities	Pulmonary Air-Space Opacity	Pulmonary Nodule or Mass	Pulmonary Interstitial Opacity	Pleural Effusion	Pneumothorax	Others
Readers with closed-line method	63.9% (562/885; 58.1–69.3%)	66.0% (290/445; 57.8–73.3%)	61.8% (89/145; 46.3–75.3%)	53.0% (34/65; 33.0–72.0%)	74.6% (117/160; 60.5–84.9%)	82.9% (29/35; 66.7–92.1%)	8.6% (3/35; 2.8–23.4%)
Readers with heat map method	66.8% (587/885; 61.0–72.1%)	69.5% (305/445; 61.4–76.5%)	57.5% (83/145; 42.4–71.3%)	60.0% (38/65; 38.0–78.0%)	79.0% (124/160; 66.0–87.9%)	97.1% (34/35; 82.3–99.6%)	8.6% (3/35; 2.8–23.4%)
Readers with combined method	67.5% (593/885; 61.7–72.8%)	71.3% (313/445; 63.0–78.5%)	61.8% (89/145; 46.7–75.0%)	58.0% (37/65; 38.0–76.0%)	75.2% (118/160; 62.5–84.7%)	97.1% (34/35; 82.3–99.6%)	5.7% (2/35; 1.4–20.2%)
p-value (closed-line vs. heat map method)	0.043	0.067	0.168	0.454	0.152	0.040	>0.999
p-value (closed-line vs. combined method)	0.004	0.003	>0.999	0.395	0.808	0.040	0.642
p-value (heat map vs. combined method)	0.580	0.343	0.147	0.736	0.122	>0.999	0.642
Standalone AI-CAD	81.4% (144/177; 74.8–86.8%)	89.9% (80/89; 81.7–95.3%)	72.4% (21/29; 52.8–87.3%)	100% (13/13; 75.3–100%)	68.8% (22/32; 50–83.9%)	100% (7/7; 59.0–100%)	14.3% (1/7; 0.4–57.9%)
p-value (vs. readers with closed-line method	<0.001	<0.001	0.110	<0.001	0.058	0.006	0.907
p-value (vs. readers with heat map method)	<0.001	<0.001	0.021	<0.001	0.032	0.280	0.907
p-value (vs. readers with combined method)	<0.001	<0.001	0.125	<0.001	0.060	0.280	0.673

Abbreviation: AI-CAD, artificial-intelligence-based computer-aided detection. Numbers in parentheses indicate numerators/denominators; 95% confidence intervals.

The stand-alone AI-CAD exhibits a detection rate of 81.4% (95% CI, 74.8–86.8%), which is significantly higher than that of readers for all visualization methods (all $p < 0.05$) (Table 4). The detection rates of stand-alone AI-CAD for different types of abnormalities are described in Table 4.

Table S2 in the Supplementary Materials shows the detection rates of individual readers.

3.4. Variation of Performances across Readers

The coefficients of variation for sensitivity, specificity, and detection rates across the five readers are described in Table 5. The sensitivity, specificity, and detection rate show the highest degree of variation in the closed-line method (sensitivity, 0.162; specificity, 0.070; accuracy, 0.087; detection rate, 0.171) and the lowest degree of variation in the combined method (sensitivity, 0.116; specificity, 0.060; accuracy, 0.055; detection rate, 0.133). However, statistical evidence of these differences is not observed.

Table 5. Coefficients of variation for performances of readers.

	Sensitivity	Specificity	Accuracy	Detection Rate
Closed-line method	0.162	0.070	0.087	0.171
Heat map method	0.142	0.062	0.067	0.154
Combined method	0.116	0.060	0.055	0.133
p-value	0.930	0.957	0.893	0.978

p-values are the results of Leven's F test for differences in the coefficient of variation.

3.5. Preference Survey

Table 6 shows the results of the preference survey for three visualization methods. The rating for the overall preference is highest in the combined method. Three of five readers most preferred the combined method, while one reader preferred the closed-line method and the other preferred the heat map method. Regarding each survey question, the combined method receives the highest rating for conspicuity, interpretability, and subjective impression to improve interpretation accuracy. Meanwhile, the score for visual fatigue is also highest in the combined method.

Table 6. Result of preference survey of the three visualizing methods.

Survey Item	Closed-Line Method	Heat Map Method	Combined Method
Conspicuity of result	4 (3–5)	4 (3–5)	5 (5–5)
Interpretability of result	4 (3–4)	4 (2–5)	5 (5–5)
Convenience of the correlation between original image and AI-CAD result	4 (3–4)	4 (4–5)	4 (4–5)
Degree of visual fatigue	2 (1–3)	2 (1–3)	3 (1–5)
Subjective impression to improve interpretation speed	4 (3–5)	4 (1–5)	4 (2–5)
Subjective impression to improve interpretation accuracy	4 (3–5)	4 (4–5)	5 (4–5)
Overall preference	4 (3–5)	4 (2–4)	5 (3–5)

Abbreviation: AI-CAD, artificial-intelligence-based computer-aided detection. All numbers indicate median (range).

Table S3 in the Supplementary Materials shows the correlation between the subjective overall preference and impression of improved accuracy versus the accuracy of interpretations.

4. Discussion

To enhance the performance of interpreting radiologists, the appropriate delivery and visualization of results are key components of AI-CAD. However, the optimal method for visualizing the results of AI CAD analyses has rarely been investigated. In the present study, we evaluated the performance of trainee radiologists for the identification of abnormalities in chest radiographs using AI-CAD with three different methods of visualization: (a) closed-lines along the boundary of the abnormality, (b) color-coded heat maps overlaid on the abnormality, and (c) a combination of closed-lines and heat maps. The average examination-level sensitivities are 68.2%, 70.3%, and 71.5% for the closed-line, heat map, and combined methods, respectively. A statistically significant difference is observed between the closed-line and combined methods. Meanwhile, the average specificities are similar among the three methods (89.0–89.4%).

Limited explainability is an important drawback of deep-learning-based AI algorithms, as the difficulty in understanding the logical background and factors associated with the output from the algorithm may hinder its reliability, especially for AI in healthcare [16–18]. AI algorithms for the detection of specific objects or findings in medical images address this explainability problem relatively simply by highlighting the location of the detected object [19–21]. In this regard, most currently used AIs in the field of medical imaging are designed to identify specific findings in medical images to assist physicians in practice [1,2,22]. However, for those AI-CAD applications, an appropriate explanation of results by AI-CAD and its delivery to physicians is still important because it may influence the interaction between the AI-CAD and physicians, and the performance of the physicians using the AI-CAD. In our study, the stand-alone AI-CAD exhibits higher examination-level sensitivity and lesion-level detection rate than readers using AI-CAD, indicating that a substantial proportion of true-positive detections by AI-CAD were rejected by the readers. The results suggest that improving the reliability of readers might be as important as improving the performance of AI-CAD to enhance the accuracy of interpretation by readers, which is the primary goal of AI-CAD [16,23].

The two representative methods for the visualization of AI-CAD results, the closed-line method and the heat map method, present certain advantages and disadvantages. The most important advantage of the closed-line method is the feasibility of its application in a gray-scale monitoring system. Displaying detection results without color-coded weights can be both a strength and a weakness. Although it cannot provide intuitive information for the confidence of the prediction by AI-CAD, it may help readers avoid neglecting AI-CAD detection results with low confidence. Reviewing detection with low confidence is important because the confidence of AI-CAD detection does not appropriately reflect the probability of the presence of an abnormality [13]. Meanwhile, the heat map method can help readers focus quickly on the abnormality by AI-CAD and can provide intuitive information regarding the confidence of the detection. However, detection with low confidence using AI-CAD may be neglected by the reader in the heat map method. In our study, the examination-level sensitivity of trainee radiologists was slightly higher in the heat map method than in the closed-line method, although no statistically significant difference was found (70.3% vs. 68.2%; $p = 0.129$). Lesion-level detection rates of individual abnormal findings are significantly higher using the heat map method (66.8% vs. 63.9%; $p = 0.043$). Increased attention to the color-coded heat map may have contributed to the better sensitivity of the readers.

A simple combination of the two visualization methods may have a synergistic effect in terms of the performance of readers, since it can embody the strength of both methods, that is, the increased attraction of readers for AI-CAD detection with both high and low confidence. In our study, both the examination-level sensitivity and lesion-level detection rates are significantly higher in the combined method than in the closed-line method (examination-level sensitivity, 71.5% vs. 68.2% [$p = 0.007$]; 67.5% vs. 63.9% [$p = 0.004$]). Compared to the heat map method, the combined method leads to slightly higher examination-level sensitivity and lesion-level detection rates, but no statistical evidence of a difference is observed.

Reducing inter-reader variability in interpretation accuracy is another important goal of AI-CAD [24]. In our study, although statistical evidence of differences is not observed due to the limited statistical power, the degrees of inter-reader variability of examination-level sensitivity, specificity, accuracy, and lesion-level detection rate are the lowest in the combined method.

The subjective preference of the user might be another important factor for selecting the visualization method, even though the preference or subjective impression does not perfectly correlate with the actual effectiveness (Table S3). In the survey of the readers, the rating for overall preference is the highest in the combined method. However, the rating for visual fatigue is also the highest for the combined method. Repeated exposure to excessive information in a single overlay image (color-coded heat map, close line for the boundary, and confidence scores in numbers) may lead to fatigue in readers, especially in practice, in which a radiologist should interpret many radiographs in a limited interpretation session.

The present study has several limitations. First, our study was conducted using chest radiographs from a single institution, and only five trainee radiologists participated in the study. Therefore, the generalizability of the results is uncertain. Second, because our study was a retrospective experimental reader test, the reproducibility of our results in an actual practice situation cannot be guaranteed. Third, only a limited number of radiologists (five trainee radiologists) participated in the study, which limits the generalization of the result. Future studies with a larger number of participating radiologists might be required. Finally, the statistical power of the study might be limited because the numbers of chest radiographs and readers were relatively small.

5. Conclusions

In conclusion, the method of visualizing the results of AI-CAD influences the performance of radiologists' sensitivity in identifying significant abnormalities on chest radiographs from patients with acute respiratory symptoms. The combination of the closed-line and heat map methods led to the highest examination-level sensitivity and lesion-level detection rate. A prospective study in an actual practice situation might be required to confirm the optimum method for visualizing AI-CAD results.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/diagnostics13061089/s1, Table S1: Sensitivity, specificity, and accuracy of Interpretation of Individual Readers; Table S2: Detection rates of individual readers; Table S3: Correlation between preference, subject impression to improve.

Author Contributions: Conceptualization, E.J.H.; methodology, E.J.H.; software, E.J.H. and C.M.P.; validation, E.J.H.; formal analysis, S.H. and E.J.H.; investigation, S.H., E.J.H., S.K., J.S., T.L., G.D.J. and Y.C.; resources, E.J.H., C.M.P. and J.M.G.; data curation, S.H. and E.J.H.; writing—original draft preparation, S.H. and E.J.H.; writing—review and editing, S.K., J.S., T.L., G.D.J., Y.C., C.M.P. and J.M.G.; visualization, S.H. and E.J.H.; supervision, E.J.H.; project administration, E.J.H. and J.M.G.; funding acquisition, J.M.G. All authors have read and agreed to the published version of the manuscript.

Funding: The present study was supported by the Korea Health Technology R&D Project through the Korea Health Industry Development Institute, funded by the Ministry of Health and Welfare, Republic of Korea (HI19C1129).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board of Seoul National University Hospital (H-2103-015-1201 and 08 March 2021).

Informed Consent Statement: The requirement for informed consent was waived by the Institutional Review Board.

Data Availability Statement: The dataset generated during the current study is available from the corresponding author upon reasonable request.

Acknowledgments: Lunit provided technical support for this study.

Conflicts of Interest: Eui Jin Hwang receives research grants from Lunit, Coreline Soft, and Monitor Corporation outside the present study; Chang Min Park receives a research grant from Lunit outside the present study, and stock of Promedius and stock options of Lunit and Coreline Soft; Jin Mo Goo receives research grants from Infinitt Healthcare, Dongkook Lifescience, and LG Electronics outside the present study.

References

1. European Society of Radiology. Current practical experience with artificial intelligence in clinical radiology: A survey of the European Society of Radiology. *Insights Imaging* **2022**, *13*, 107. [CrossRef] [PubMed]
2. Hwang, E.J.; Goo, J.M.; Yoon, S.H.; Beck, K.S.; Seo, J.B.; Choi, B.W.; Chung, M.J.; Park, C.M.; Jin, K.N.; Lee, S.M. Use of Artificial Intelligence-Based Software as Medical Devices for Chest Radiography: A Position Paper from the Korean Society of Thoracic Radiology. *Korean J. Radiol.* **2021**, *22*, 1743–1748. [CrossRef]
3. Lee, S.; Shin, H.J.; Kim, S.; Kim, E.K. Successful Implementation of an Artificial Intelligence-Based Computer-Aided Detection System for Chest Radiography in Daily Clinical Practice. *Korean J. Radiol.* **2022**, *23*, 847–852. [CrossRef]
4. Hwang, E.J.; Park, C.M. Clinical Implementation of Deep Learning in Thoracic Radiology: Potential Applications and Challenges. *Korean J. Radiol.* **2020**, *21*, 511–525. [CrossRef]
5. Kapoor, N.; Lacson, R.; Khorasani, R. Workflow Applications of Artificial Intelligence in Radiology and an Overview of Available Tools. *J. Am. Coll. Radiol.* **2020**, *17*, 1363–1370. [CrossRef] [PubMed]
6. Tang, A.; Tam, R.; Cadrin-Chênevert, A.; Guest, W.; Chong, J.; Barfett, J.; Chepelev, L.; Cairns, R.; Mitchell, J.R.; Cicero, M.D.; et al. Canadian Association of Radiologists White Paper on Artificial Intelligence in Radiology. *Can. Assoc. Radiol. J.* **2018**, *69*, 120–135. [CrossRef]
7. Ahn, J.S.; Ebrahimian, S.; McDermott, S.; Lee, S.; Naccarato, L.; Di Capua, J.F.; Wu, M.Y.; Zhang, E.W.; Muse, V.; Miller, B.; et al. Association of Artificial Intelligence-Aided Chest Radiograph Interpretation with Reader Performance and Efficiency. *JAMA Netw. Open* **2022**, *5*, e2229289. [CrossRef] [PubMed]
8. Calli, E.; Sogancioglu, E.; van Ginneken, B.; van Leeuwen, K.G.; Murphy, K. Deep learning for chest X-ray analysis: A survey. *Med. Image Anal.* **2021**, *72*, 102125. [CrossRef]
9. Meedeniya, D.; Kumarasinghe, H.; Kolonne, S.; Fernando, C.; De la Torre Díez, I.; Marques, G. Chest X-ray analysis empowered with deep learning: A systematic review. *Appl. Soft Comput.* **2022**, *126*, 109319. [CrossRef] [PubMed]
10. Hong, W.; Hwang, E.J.; Lee, J.H.; Park, J.; Goo, J.M.; Park, C.M. Deep Learning for Detecting Pneumothorax on Chest Radiographs after Needle Biopsy: Clinical Implementation. *Radiology* **2022**, *303*, 433–441. [CrossRef] [PubMed]
11. Hwang, E.J.; Lee, J.S.; Lee, J.H.; Lim, W.H.; Kim, J.H.; Choi, K.S.; Choi, T.W.; Kim, T.H.; Goo, J.M.; Park, C.M. Deep Learning for Detection of Pulmonary Metastasis on Chest Radiographs. *Radiology* **2021**, *301*, 455–463. [CrossRef]
12. Hwang, E.J.; Nam, J.G.; Lim, W.H.; Park, S.J.; Jeong, Y.S.; Kang, J.H.; Hong, E.K.; Kim, T.M.; Goo, J.M.; Park, S.; et al. Deep Learning for Chest Radiograph Diagnosis in the Emergency Department. *Radiology* **2019**, *293*, 573–580. [CrossRef]
13. Hwang, E.J.; Kim, H.; Lee, J.H.; Goo, J.M.; Park, C.M. Automated identification of chest radiographs with referable abnormality with deep learning: Need for recalibration. *Eur. Radiol.* **2020**, *30*, 6902–6912. [CrossRef]
14. Hwang, E.J.; Park, S.; Jin, K.N.; Im Kim, J.; Choi, S.Y.; Lee, J.H.; Goo, J.M.; Aum, J.; Yim, J.J.; Cohen, J.G.; 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. [CrossRef] [PubMed]
15. Leisenring, W.; Pepe, M.S.; Longton, G. A marginal regression modelling framework for evaluating medical diagnostic tests. *Stat. Med.* **1997**, *16*, 1263–1281. [CrossRef]
16. Ploug, T.; Sundby, A.; Moeslund, T.B.; Holm, S. Population Preferences for Performance and Explainability of Artificial Intelligence in Health Care: Choice-Based Conjoint Survey. *J. Med. Internet Res.* **2021**, *23*, e26611. [CrossRef]
17. Ghassemi, M.; Oakden-Rayner, L.; Beam, A.L. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit. Health* **2021**, *3*, e745–e750. [CrossRef] [PubMed]
18. Reddy, S. Explainability and artificial intelligence in medicine. *Lancet Digit. Health* **2022**, *4*, e214–e215. [CrossRef]
19. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.
20. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 779–788.
21. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
22. van Leeuwen, K.G.; Schalekamp, S.; Rutten, M.J.; van Ginneken, B.; de Rooij, M. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. *Eur. Radiol.* **2021**, *31*, 3797–3804. [CrossRef]

23. Gaube, S.; Suresh, H.; Raue, M.; Merritt, A.; Berkowitz, S.J.; Lermer, E.; Coughlin, J.F.; Guttag, J.V.; Colak, E.; Ghassemi, M. Do as AI say: Susceptibility in deployment of clinical decision-aids. *NPJ Digit. Med.* **2021**, *4*, 1–8. [CrossRef] [PubMed]
24. Hwang, E.J.; Goo, J.M.; Kim, H.Y.; Yi, J.; Yoon, S.H.; Kim, Y. Implementation of the cloud-based computerized interpretation system in a nationwide lung cancer screening with low-dose CT: Comparison with the conventional reading system. *Eur. Radiol.* **2021**, *31*, 475–485. [CrossRef] [PubMed]