Journal of Biomedical Informatics 156 (2024) 104673

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Original Research

# Clinical domain knowledge-derived template improves post hoc AI explanations in pneumothorax classification

Han Yuan [a,1], Chuan Hong [b,1], Peng-Tao Jiang [c], Gangming Zhao [d], Nguyen Tuan Anh Tran [e], Xinxing Xu [f], Yet Yen Yan [g], Nan Liu [a,h,i,*]

[a] *Centre for Quantitative Medicine, Duke-NUS Medical School, Singapore*
[b] *Department of Biostatistics and Bioinformatics, Duke University, USA*
[c] *College of Computer Science, Nankai University, China*
[d] *Faculty of Engineering, The University of Hong Kong, China*
[e] *Department of Diagnostic Radiology, Singapore General Hospital, Singapore*
[f] *Institute of High Performance Computing, Agency for Science, Technology and Research, Singapore*
[g] *Department of Radiology, Changi General Hospital, Singapore*
[h] *Programme in Health Services and Systems Research, Duke-NUS Medical School, Singapore*
[i] *Institute of Data Science, National University of Singapore, Singapore*

ARTICLE INFO

*Keywords:*
Pneumothorax Diagnosis
Convolutional Neural Networks
Explainable Artificial Intelligence
Saliency Map
Grad-CAM
Integrated Gradients

ABSTRACT

*Objective:* Pneumothorax is an acute thoracic disease caused by abnormal air collection between the lungs and chest wall. Recently, artificial intelligence (AI), especially deep learning (DL), has been increasingly employed for automating the diagnostic process of pneumothorax. To address the opaqueness often associated with DL models, explainable artificial intelligence (XAI) methods have been introduced to outline regions related to pneumothorax. However, these explanations sometimes diverge from actual lesion areas, highlighting the need for further improvement.

*Method:* We propose a template-guided approach to incorporate the clinical knowledge of pneumothorax into model explanations generated by XAI methods, thereby enhancing the quality of the explanations. Utilizing one lesion delineation created by radiologists, our approach first generates a template that represents potential areas of pneumothorax occurrence. This template is then superimposed on model explanations to filter out extraneous explanations that fall outside the template's boundaries. To validate its efficacy, we carried out a comparative analysis of three XAI methods (Saliency Map, Grad-CAM, and Integrated Gradients) with and without our template guidance when explaining two DL models (VGG-19 and ResNet-50) in two real-world datasets (SIIM-ACR and ChestX-Det).

*Results:* The proposed approach consistently improved baseline XAI methods across twelve benchmark scenarios built on three XAI methods, two DL models, and two datasets. The average incremental percentages, calculated by the performance improvements over the baseline performance, were 97.8% in Intersection over Union (IoU) and 94.1% in Dice Similarity Coefficient (DSC) when comparing model explanations and ground-truth lesion areas. We further visualized baseline and template-guided model explanations on radiographs to showcase the performance of our approach.

*Conclusions:* In the context of pneumothorax diagnoses, we proposed a template-guided approach for improving model explanations. Our approach not only aligns model explanations more closely with clinical insights but also exhibits extensibility to other thoracic diseases. We anticipate that our template guidance will forge a novel approach to elucidating AI models by integrating clinical domain expertise.

[*] Corresponding author at: Centre for Quantitative Medicine, Duke-NUS Medical School, 8 College Road, Singapore 169857, Singapore.
*E-mail address:* liu.nan@dukenus.edu.sg (N. Liu).

[1] These authors contributed equally.

https://doi.org/10.1016/j.jbi.2024.104673
Received 27 March 2024; Received in revised form 1 June 2024; Accepted 7 June 2024
Available online 9 June 2024

## 1. Introduction

Pneumothorax is an acute thoracic disease caused by abnormal air collection in the pleural space between the lungs and chest wall [1]. Timely intervention is crucial to prevent pneumothorax from evolving into a life-threatening emergency [2]. In clinical practice, pneumothorax is usually diagnosed by radiologists on a chest radiograph – a process that demands considerable expertise and expert efforts. Recent advancements suggest that this process can be automated using artificial intelligence (AI), especially deep learning (DL) models such as convolutional neural networks (CNNs). For instance, EfficientNet B3 [3] has demonstrated high accuracy in classifying pneumothorax of various sizes, with the area under the receiver operating characteristic curve (AUROC) ranging from 88% to 96% [2]. Xception [4] further advanced the classification capability, achieving an AUROC of 99% on an open-access dataset [5]. While these DL-based classifiers have exhibited high-fidelity classification ability, their complexity poses a challenge: Comprising numerous interconnected neurons with intricate relationships, their decision-making processes are often opaque and challenging to interpret [6]. This complexity can hinder radiologists' acceptance and trust in these AI tools, thereby affecting their practical application in real-world settings [7,8].

To solve this problem, researchers have introduced various explainable artificial intelligence (XAI) methods to chest radiograph analysis. For instance, Mosquera et al. [9] applied class activation maps (CAM) [10] to identify regions in chest radiographs that significantly influence the disease diagnosis. Feng et al. [11] and Wang et al. [4] used Grad-CAM [12], a variant of CAM, to pinpoint the specific pixels on chest radiographs that contributed most to model predictions. These heatmaps partially alleviate radiologists' concerns regarding the trustworthiness of DL models [6]. However, a recent benchmarking study pointed out a notable result: Even with a high-accuracy DenseNet-121 [13] achieving an AUROC of 99.3% in the pneumothorax classification, the areas highlighted by the model only coincided with 7.0% of the actual lesion areas as delineated by radiologists [14]. Similarly, Rocha et al. developed a ResNet-50 [15] with an AUROC of 85.4% in classifying pneumothorax, yet its explanations attained an Intersection over Union (IoU) of 17.6% when assessed using lesion areas delineated by coarse bounding boxes [16]. Giachanou et al. reported IoUs ranging from 3.1% to 15.1% across a variety of model explanations for pneumothorax diagnoses [17]. These identified discrepancies between classification and explanation capabilities underline the urgent need to improve existing model explanations [14].

Leveraging prior clinical knowledge is one promising direction to enhance model explanations. Specifically, pneumothorax occurs in the pleural space between the lungs and chest walls [1]. This clinically validated information could serve as invaluable prior knowledge to improve model explanations. Previous studies have successfully utilized disease location information in pneumothorax classification and localization. Crosby et al. [18] capitalized on the observation that pneumothorax typically occurs in apex areas of chest radiographs. Therefore, they segmented the upper third of chest radiographs for pneumothorax classification, achieving enhanced accuracy. However, it remains unclear whether model explanations can also take advantage of the location information. To address this, Jung et al. [7] identified common thoracic disease patterns on chest radiographs, directing models to focus on typical disease locales, which in turn enhanced both classification and explanation quality. However, their method requires an exhaustive labeling of eight common thoracic diseases and is inappropriate for resource-limited settings where only diagnostic labels of a single disease are available.

To overcome aforementioned limitations, we propose a template-guided approach that crafts a template covering potential occurrence areas of pneumothorax to guide model explanations generated by XAI methods. We illustrate the performance of our approach through comparative experiments of three XAI methods with and without our template guidance. Our template-guided approach may provide a novel perspective for incorporating clinical knowledge into the explanation of other thoracic conditions.

**Statement of Significance**

**Problem** Current XAI explanations sometimes diverge from actual lesion areas in the context of pneumothorax diagnoses.

**What is Already Known** Pneumothorax occurs in the pleural space between the lungs and chest walls, which could serve as invaluable prior knowledge to improve model explanations.

**What this Paper Adds** Our study leverages a template derived from clinical knowledge of pneumothorax occurrence information to improve model explanations. This approach not only aligns model explanations more closely with clinical insights but also exhibits extensibility to other thoracic diseases.

## 2. Methods

AI models, especially CNNs, have become the mainstream backbones for chest radiograph classification, with various XAI methods accompanied to interpret their diagnostic processes [12,19,20]. Despite these advancements, a recent study [14] indicates that model explanations provided by the pneumothorax classifier fail to match ground truth lesion areas, suggesting a need for further improvement. To bridge this gap, we propose a template-guided approach for existing XAI methods in the context of pneumothorax diagnoses. This section outlines our methodology, starting with an introduction of notations followed by a detailed description of CNNs' training strategy. We then illustrate three well-established explanation methods for CNNs. The section concludes with our proposed approach that guides model explanations with a clinical knowledge-derived template.

### 2.1. Notations

We first introduce key notations for subsequent illustrations of classifier training and explanation. For the pneumothorax classification task, we denote the nonoverlapping training, validation, and test datasets as $D^{train}$, $D^{val}$, and $D^{test}$, respectively. Each dataset consists of pairs of images and corresponding image-level binary labels, structured identically. As an illustrative example, we consider the training dataset $D^{train}$, which includes $N_{train}$ samples:

$$D_{train} = \left\{ \left( I_i^{train}, Y_i^{train} \right), i = 1, 2, \cdots, N_{train} \right\}.$$

$I_i^{train}$ designates a two-dimensional image with a width of $W_0$ and a height of $H_0$. $Y_i^{train} \in \{0, 1\}$ is the ground truth label by radiologists and $Y_i^{train} = 1$ states that $I_i^{train}$ is diagnosed with pneumothorax. $I_i^{train}$ consists of $W_0 \times H_0$ pixels and $p_{w,h}(I_i^{train})$ denotes a pixel in $I_i^{train}$ whose coordinate of width and height is $(w, h)$:

$$I_i^{train} = \{p_{w,h}(I_i^{train}), w = 1, 2, \cdots, W_0, h = 1, 2, \cdots, H_0\}.$$

Each $p_{w,h}(I_i^{train})$ comprises three elements $e_{w,h,c}(I_i^{train})$ standing for pixel values in channel $c$ of red, green, or blue:

$$p_{w,h}\left(I_i^{train}\right) = \left\{e_{w,h,c}\left(I_i^{train}\right), c = red, green, blue\right\}.$$

Based on the input of $e_{w,h,c}(I_i^{train})$ and the output target of $Y_i^{train}$, the pneumothorax classifier is trained and subsequently explained. Model explanations are typically generated by initially calculating the importance of pixels and then shortlisting pixels with importance values exceeding a pre-determined threshold to constitute the important subregion [14,21]. Our template-guided approach relies on a radiologists-annotated lesion delineation $A_T^{train}$ of a single image $I_T^{train}$ from $D_{train}$. Additionally, for assessing the alignment between model explanations and real lesion areas, pneumothorax samples $I_i^{test}$ in test dataset $D^{test}$ are also annotated with pixel-level lesion areas $A_i^{test}$. $A_T^{train}$ holds the same structure as $A_i^{test}$ and we use $A_i^{test}$ as an illustrative instance. $A_i^{test}$ is a two-dimensional image and consists of $W_0 \times H_0$ elements $a_{w,h}(A_i^{test}) \in \{0, 1\}$.

Decided by radiologists, $a_{w,h}\left(A_i^{test}\right) = 1$ denotes that the pixel with the coordinate of $(w,h)$ in $I_i^{test}$ belongs to the lesion areas:

$$A_i^{test} = \left\{a_{w,h}\left(A_i^{test}\right), w = 1,2,\cdots,W_0, h = 1,2,\cdots,H_0\right\}.$$

It is important to note that lesion annotations $A_T^{train}$ and $A_i^{test}$ are exclusively employed for model explanations. The model training of pneumothorax classifiers follows the standard paradigm that uses binary diagnostic labels $Y_i^{train}$ and $Y_i^{val}$ [22].

### 2.2. Image classifier training

CNNs have achieved outstanding performance in various thoracic disease classification tasks [23]. In general, the image classifier training is to find a set of parameters that minimizes the difference between CNNs' predictions and ground truth labels in the training set. Formally, with the training dataset $D^{train}$, we aim to optimize a model $f_\theta$ parameterized by $\theta$. The model takes input $e_{w,h,c}(I_i^{train})$, $f_\theta$ produces an output $f_\theta\left(I_i^{train}\right)$. The optimization objective is to minimize the difference $d$ between $f_\theta\left(I_i^{train}\right)$ and sample labels $Y_i^{train}$ for all samples in $D^{train}$. The cumulative difference over all training samples known as loss function $l\left(\theta; D^{train}\right)$ is expressed as:

$$l\left(\theta; D^{train}\right) = \frac{1}{N_{train}}\sum\nolimits_i d\left(f_\theta\left(I_i^{train}\right), Y_i^{train}\right).$$

To avoid overfitting of $f_\theta$, the validation dataset $D^{val}$ is applied to early stop the optimization procedure. If the loss $l\left(\theta; D^{val}\right)$ has not decreased for a pre-defined epoch number $N_{epoch}$, the iteration of $\theta$ will be terminated. The last $\theta$ that led to a decrease in $l\left(\theta; D^{val}\right)$ is retained as the optimal parameter $\theta^*$:

$$\theta^* = argmin_\theta\left(l\left(\theta; D^{train}\right) | l\left(\theta; D^{val}\right)\right).$$

After the determination of $\theta^*$, we measure the classification performance $M$ on the unseen test dataset $D^{test}$. An evaluation metric $m$ is used to assess the model performance by comparing the model prediction $f_{\theta^*}\left(I_i^{test}\right)$ and the true label $Y_i^{test}$:

$$M\left(\theta^*; D^{test}\right) = \frac{1}{N_{test}}\sum\nolimits_i m\left(f_{\theta^*}\left(I_i^{test}\right), Y_i^{test}\right).$$

### 2.3. Image classifier explanation

The developed model $f_{\theta^*}$ classifies an unseen image $I_i^{test}$ from the test dataset $D^{test}$ as $f_{\theta^*}\left(I_i^{test}\right)$. We aim to further explain $f_{\theta^*}\left(I_i^{test}\right)$ to both uncover the model decision logic and evaluate its trustworthiness [24]. A commonly used explanation paradigm calculates each pixel's importance $E\left(p_{w,h}\left(I_i^{test}\right)\right)$ to the prediction $f_{\theta^*}\left(I_i^{test}\right)$, and further identifies the focus area $R\left(I_i^{test}\right)$ consisting of the most discriminative pixels towards model outputs [25]. Explanations are considered reliable if focus areas precisely match disease lesion areas annotated by human experts [14]. Within the explanation paradigm using focus areas [26], we select three mainstream XAI techniques, Saliency Map [19], Grad-CAM [12], and Integrated Gradients [20], given their frequent use as XAI baselines in the existing literature [27,28]. Technical details of these techniques can be found in their original publications [12,19,20]. Here we provide a concise overview to facilitate the downstream illustration of our template-guided approach.

As a pioneering method in image classifier explanation, Saliency Map [19] calculates the importance of $p_{w,h}\left(I_i^{test}\right)$ through its forthright gradient towards $f_{\theta^*}\left(I_i^{test}\right)$. Specifically, it computes $f_{\theta^*}\left(I_i^{test}\right)$'s gradients with respect to every element $e_{w,h,c}\left(I_i^{test}\right)$ in pixel $p_{w,h}\left(I_i^{test}\right)$ and derives the pixel importance $E\left(p_{w,h}\left(I_i^{test}\right)\right)$ as the largest absolute gradient among all channels.

Saliency Map depicts the impact of each pixel towards final model outputs while possibly outlines all recognizable objects in $I_i^{test}$ and fails to spotlight $R\left(I_i^{test}\right)$ towards $f_{\theta^*}\left(I_i^{test}\right)$ [10]. Grad-CAM [12] conjectures that the problem can be resolved by initially computing the pixel importance $E\left(p_{w,h}\left(I_{i,conv}^{test}\right)\right)$ on the last convolutional layer $I_{i,conv}^{test}$, and subsequently reconstructing $E\left(p_{w,h}\left(I_i^{test}\right)\right)$ through the bilinear interpolation of $E\left(p_{w,h}\left(I_{i,conv}^{test}\right)\right)$.

Both Saliency Map and Grad-CAM depict the local changes in $f_{\theta^*}\left(I_i^{test}\right)$ with respect to a small range of pixel values. However, if a pixel's possible values within a narrow range are always important towards $f_{\theta^*}\left(I_i^{test}\right)$, the gradient saturates to zero, indicating the opposite conclusion that the pixel is trivial [20]. Integrated Gradients [20] solve this problem via computing the gradients sum of $m$ pseudo images interpolated between $I_i^{test}$ and a reference image $I_{ref}$ obtained by fusing all training images. Same as the previous two methods, Integrated Gradients output the pixel importance $E\left(p_{w,h}\left(I_i^{test}\right)\right)$.

After obtaining $E\left(p_{w,h}\left(I_i^{test}\right)\right)$ by different methods, a binarization cutoff value $v^*$ is used to outline the most important pixels and constitute the model focus region $R_{v^*}\left(I_i^{test}\right)$. Explanations are considered reliable when $R_{v^*}\left(I_i^{test}\right)$ highly overlaps with lesion areas $A_i^{test}$ [14]. Different metrics, e.g. $IoU$, are applied to quantify the performance $Q$ of model explanations on the test dataset $D^{test}$:

$$Q\left(\theta^*; D^{test}\right) = \frac{1}{N_{test}}\sum\nolimits_i IoU\left(R_{v^*}\left(I_i^{test}\right), A_i^{test}\right).$$

### 2.4. Proposed template-guided explanation

As illustrated above, baseline XAI methods outline the important region $R_{v^*}\left(I_i^{test}\right)$ from the whole area of $I_i^{test}$. However, domain knowledge elucidates that pneumothorax typically occurs in the pleural space between the lungs and chest walls [7,18]. Particularly, on an upright frontal radiograph, pneumothorax is recognized by non-dependent lucency that parallels the chest wall and displaces the visceral pleural line medially. It usually localizes to the lung apices and lateral aspect of the lungs. Based on this prior clinical knowledge, we propose a template-guided approach that integrates the disease occurrence areas with baseline model explanations. The proposed approach requires minimal human involvement and yields explanations that align better with the clinical understanding of pneumothorax. Fig. 1**(a)** shows the overview of our template guidance as a plug-and-play module for existing XAI methods. To depict the pleural space from the clinical experts' view, a canonical lesion annotation by radiologists is extracted as the basis for template generation. Then several morphological operations are implemented to further refine the pleural space – potential occurrence areas of pneumothorax. After that, we shepherd the original explanations using the generated template region: Only the pixel within the template boundaries will be included in model focus areas. Finally, focus areas with or without template guidance are compared with the ground truth lesion annotations.

The first step in the proposed template guidance is to generate the optimal template carrying the location information of disease occurrence. Fig. 1**(b)** summarizes the details of template generation: Using one canonical lesion delineation $A_T^{train}$ as the starting point, the candidate templates are generated by flipping, overlap, and dilation. Selected by radiologists, $A_T^{train}$ contours at least the pleural space on one side. Then the step of flipping turns over the original lesion delineation horizontally to generate $A_{T,flip}^{train}$ on the other side. After that, considering the domain knowledge that pneumothorax potentially occurs in both the left and the right pleural space, the step of overlapping is implemented to generate $A_{T,overlap}^{train}$ spotlighting both left and right pleural spaces [29]. A pixel $p_{w,h}\left(A_{T,overlap}^{train}\right)$ is included in the template area if it is within either $p_{w,h}\left(A_T^{train}\right)$ or $p_{w,h}\left(A_{T,flip}^{train}\right)$. Another factor affecting the depiction of

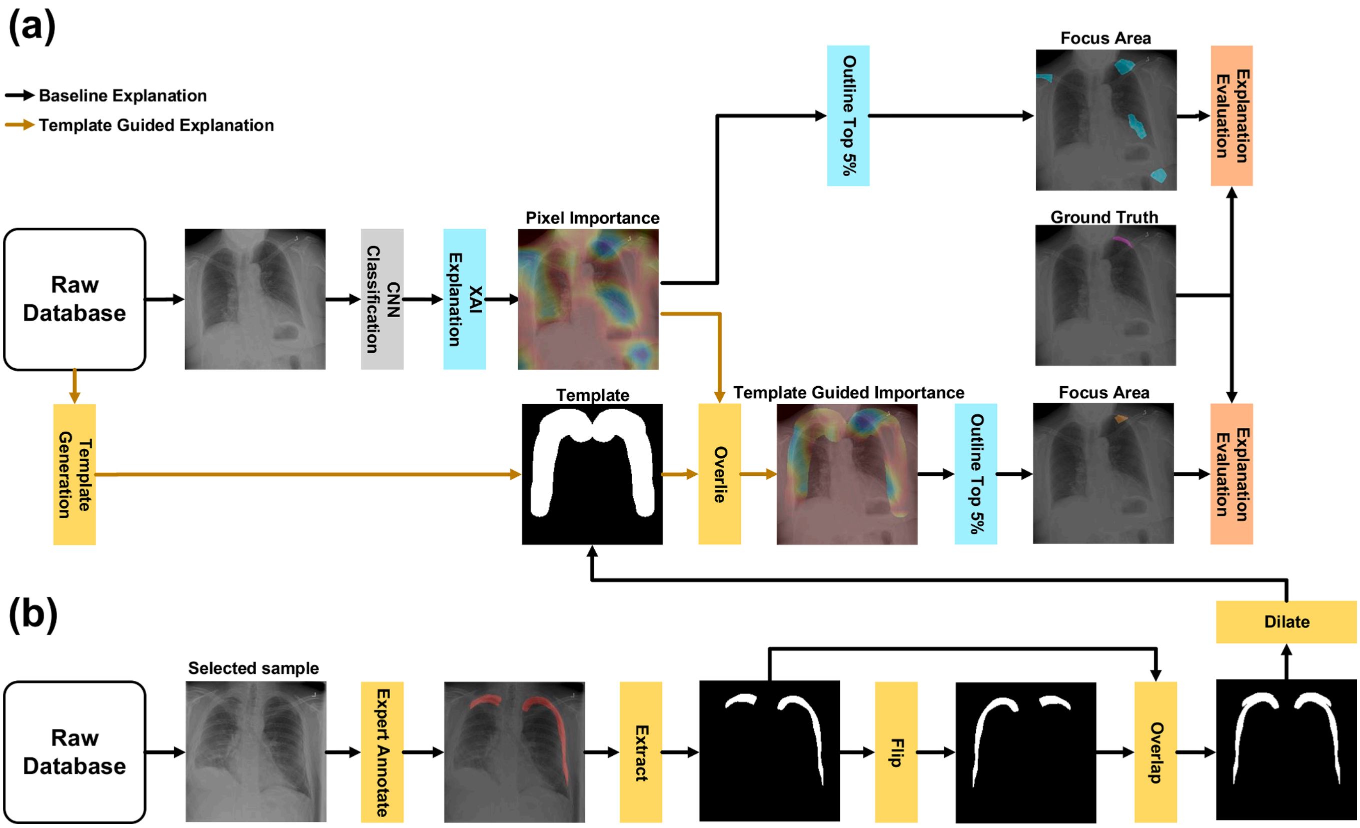

**Fig. 1.** Overview of the proposed template-guided explanation pipeline. (a) The application of template guidance to improve XAI explanation. (b) Detailed process of template generation, which includes expert annotation, extraction, flipping, overlap, and dilation.

pleural space is that the chest radiographs are captured at different distances and angles. Thus, the position and scale of pleural space vary in different radiographs [30]. To address this issue, we introduce the step of dilation to eliminate the problem of deformation through enlarging the template area to cover a broader space. Following a previous work [31], a 15 × 15 ellipse kernel sweeps each pixel on the original image, and a pixel will be included in the dilated template area if one of its neighbor pixels within the kernel belongs to $A^{train}_{T,overlap}$. After that, we obtain the final template $T^*$ wherein $a_{w,h}(T^*) = 1$ denotes the coordinate of $(w,h)$ in $I_i^{test}$ belongs to the pleural space. Through the element-wise product, the template-guided pixel importance $E^*\left(p_{w,h}\left(I_i^{test}\right)\right)$ is calculated:

$$E^*\left(p_{w,h}\left(I_i^{test}\right)\right) = T^* \odot E\left(p_{w,h}\left(I_i^{test}\right)\right).$$

Finally, the identical approach as the baseline explanation is employed to extract the model focus region from $E^*\left(p_{w,h}\left(I_i^{test}\right)\right)$.

## 3. Experiments

In this section, we first introduced the datasets. Then we provided details on the training and explanation of pneumothorax classifiers, and clarified the relevant evaluation metrics. After that, we presented the experimental results of pneumothorax classification and explanation. We demonstrated that the proposed template-guided approach consistently improved the baseline XAI methods. To provide a comprehensive assessment, we visualized both successful and failed cases of model explanations. All experiments were conducted using Python, and the code has been made publicly available on GitHub for reproducibility [32].

### 3.1. Datasets

The performance of pneumothorax classification and explanation was demonstrated using two real-world datasets, SIIM-ACR Pneumothorax Segmentation Challenge [33] and ChestX-Det [34]. These two datasets were chosen as they have been widely used as the benchmark datasets for pneumothorax classification and lesion recognition [35–40]. The SIIM-ACR dataset comprises a total of 12,047 chest radiographs, among which 2,669 instances are diagnosed as positive, indicating the presence of pneumothorax. Unlike the SIIM-ACR dataset, the ChestX-Det dataset is notably smaller, consisting of 611 healthy images and 189 pneumothorax-positive images. Besides the binary pneumothorax diagnosis at the image level, both datasets provide pixel-level lesion delineations in positive cases, which are not available in other recent chest radiograph datasets [41,42].

We randomly split the SIIM-ACR dataset into three parts at 60:20:20. Specifically, the training set consisted of 7,226 images (60%, containing 1,600 positive samples), the validation set comprised 2,410 images (20%, containing 534 positive samples), and the test set included 2,410 images (20%, containing 534 positive samples). To validate the generalizability of the proposed method, we evenly partitioned the ChestX-Det dataset into validation (50%, containing 95 positive samples) and test sets (50%, containing 94 positive samples) for external validation. Table 1 gives an overview of the used data sets, annotations, and functions in our study. Detailed information is elaborated in the subsequent two sections.

### 3.2. Classifier training and evaluation

We implemented the pneumothorax classifier with two lightweight architectures: VGG-19 [43] and ResNet-50 [15] to avoid over-parameterization in small-sample datasets [44,45] and spatial

**Table 1**
An overview of the used datasets, annotation, and function.

| Dataset | | Annotation | Function |
|---|---|---|---|
| SIIM-ACR | Training set | Binary diagnosis | Classifier training |
| | | Lesion delineation | Template generation |
| | Validation set | Binary diagnosis | Classifier training<br>Binarization cutoff calculation |
| | Test set | Binary diagnosis | Internal evaluation of classifier's classification capability |
| | | Lesion delineation | Internal evaluation of XAI's explanation capability |
| ChestX-Det | Validation set | Binary diagnosis | Binarization cutoff calculation |
| | Test set | Binary diagnosis | External evaluation of classifier's classification capability |
| | | Lesion delineation | External evaluation of XAI's explanation capability |

information loss of XAI when explaining architectures with deeper layers [46,47]. Model output layers were modified for binary classification. A Stochastic Gradient Descent (SGD) optimizer was employed with a learning rate of 1e-3 and a momentum of 0.9. Model training was conducted in batches of 16 images, using weighted cross-entropy as the loss function to counterbalance the predominance of negative samples [48]. The training was set as 100 epochs on the training set of SIIM-ACR with an early-stop initiated if no improvement was observed on the validation set of SIIM-ACR over 10 consecutive epochs. After the training, the model classification performance was evaluated on both the internal test set of SIIM-ACR and the external test set of ChestX-Det. Evaluation metrics included AUROC, the area under the precision recall curve (AUPRC), accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). Binarization cutoffs were chosen as the points closest to the upper-left corner in the ROC curves on the respective validation sets [49]. For each metric, standard errors were calculated using the nonparametric bootstrap method [50].

### 3.3. XAI explanation and evaluation

After the pneumothorax classification training and evaluation, model explanations play a pivotal role in clinical application [21]. Our study utilized three model explanation methods: Saliency Map [19], Grad-CAM [12], and Integrated Gradients [20]. Our template-guided approach worked as a plug-and-play module on the basis of the three XAI methods, necessitating only one lesion delineation from the training set of SIIM-ACR. Therefore, we had a total of six explanation methods. The direct production of the six explanation methods was the pixel importance, from which focus areas were further outlined as the final explanation using the threshold value $v^*$ of 0.95 [51]. We leveraged IoU and Dice Score Coefficient (DSC) to quantify the alignment between the generated focus areas and ground truth lesion delineations on both the internal test set of SIIM-ACR and the external test set of ChestX-Det. The standard errors of IoU and DSC were computed through the nonparametric bootstrap method [50].

### 3.4. Experimental results

In this section, we showed the evaluation results followed by their respective standard errors enclosed within parentheses. Table 2 quantifies the model classification performance on the internal test set of SIIM-ACR. The VGG-19 classifier achieved results of an AUROC of 0.864 (0.008), an AUPRC of 0.660 (0.023), an accuracy of 80.5% (0.8%), a sensitivity of 78.3% (1.8%), a specificity of 81.1% (0.9%), a PPV of 54.1% (1.9%), and an NPV of 92.9% (0.7%). The ResNet-50 discriminator attained an AUROC of 0.842 (0.007), an AUPRC of 0.630 (0.023), an accuracy of 77.8% (0.8%), a sensitivity of 75.7% (1.5%), a specificity of 78.4% (0.9%), a PPV of 49.9% (2.0%), and an NPV of 91.9% (0.6%). Following the evaluation of model classification, Table 3 illustrates the model explanation performance of the baseline XAI methods and their template-guided versions. Under the framework of VGG-19, the original Saliency Map achieved an IoU of 2.2% (0.2%) and a DSC of 4.1% (0.3%). The original Grad-CAM obtained an IoU of 1.4% (0.1%) and a DSC of 2.6% (0.2%). The original Integrated Gradients achieved an IoU of 3.1% (0.2%) and a DSC of 5.9% (0.3%). Adding template guidance consistently resulted in performance improvements in terms of IoU and DSC: 1.0% and 1.9% for Saliency Map, 0.9% and 1.7% for Grad-CAM, and 1.4% and 2.3% for Integrated Gradients. Based on ResNet-50, the performance enhancements were 1.7% and 3.1% for Saliency Map, 3.0% and 5.1% for Grad-CAM, and 2.6% and 4.5% for Integrated Gradients. In the internal test scenarios, the incremental percentages of IoU and DSC, calculated by the performance improvements over the baseline performance, ranged from 41.7% to 168.4% and 30.7% to 114.9%, respectively.

Table 4 presents the classification performance of developed VGG-19 and ResNet-50 on the external test set of ChestX-Det. Specifically, VGG-19 without fine-tuning presented an AUROC of 0.942 (0.016), an AUPRC of 0.896 (0.025), an accuracy of 89.7% (1.5%), a sensitivity of 86.2% (3.3%), a specificity of 90.8% (1.6%), a PPV of 74.3% (4.4%), and an NPV of 95.5% (1.1%). The directly-deployed ResNet-50 also showed satisfactory performance. Regarding the explanation performance, akin to the internal validation on SIIM-ACR, our template-guided approach consistently improved all three baseline XAI methods as showcased in Table 5. In terms of IoU and DSC, the template-guided explanation of VGG-19 achieved improvements of 1.6% and 3.0% for Saliency Map,

**Table 3**
Internal explanation evaluation of various deep learning models by XAI methods. The evaluation metrics on the test set are presented, accompanied by their respective standard errors enclosed within parentheses.

| Model | XAI | Template-Guidance | IoU (%) | DSC (%) |
|---|---|---|---|---|
| VGG-19 | Saliency Map | ✘ | 2.2 (0.2) | 4.1 (0.3) |
| | | ✔ | 3.2 (0.2) | 6.0 (0.3) |
| | Grad-CAM | ✘ | 1.4 (0.1) | 2.6 (0.2) |
| | | ✔ | 2.3 (0.2) | 4.3 (0.3) |
| | Integrated Gradients | ✘ | 3.1 (0.2) | 5.9 (0.3) |
| | | ✔ | 4.5 (0.2) | 8.2 (0.3) |
| ResNet-50 | Saliency Map | ✘ | 2.3 (0.1) | 4.3 (0.2) |
| | | ✔ | 4.0 (0.2) | 7.4 (0.3) |
| | Grad-CAM | ✘ | 1.7 (0.2) | 3.1 (0.3) |
| | | ✔ | 4.7 (0.3) | 8.2 (0.4) |
| | Integrated Gradients | ✘ | 2.1 (0.1) | 4.0 (0.2) |
| | | ✔ | 4.7 (0.2) | 8.5 (0.4) |

**Table 2**
Internal classification evaluation of various deep learning models. The evaluation metrics on the test set are presented, accompanied by their respective standard errors enclosed within parentheses.

| Model | AUROC | AUPRC | Accuracy (%) | Sensitivity (%) | Specificity (%) | PPV (%) | NPV (%) |
|---|---|---|---|---|---|---|---|
| VGG-19 | 0.864 (0.008) | 0.660 (0.023) | 80.5 (0.8) | 78.3 (1.8) | 81.1 (0.9) | 54.1 (1.9) | 92.9 (0.7) |
| ResNet-50 | 0.842 (0.007) | 0.630 (0.023) | 77.8 (0.8) | 75.7 (1.5) | 78.4 (0.9) | 49.9 (2.0) | 91.9 (0.6) |

**Table 4**
External classification evaluation of various deep learning models. The evaluation metrics on the test set are presented, accompanied by their respective standard errors enclosed within parentheses.

| Model | AUROC | AUPRC | Accuracy (%) | Sensitivity (%) | Specificity (%) | PPV (%) | NPV (%) |
|---|---|---|---|---|---|---|---|
| VGG-19 | 0.942 (0.016) | 0.896 (0.025) | 89.7 (1.5) | 86.2 (3.3) | 90.8 (1.6) | 74.3 (4.4) | 95.5 (1.1) |
| ResNet-50 | 0.943 (0.013) | 0.870 (0.029) | 89.7 (1.6) | 84.0 (3.8) | 91.5 (1.7) | 75.2 (3.9) | 94.9 (1.2) |

**Table 5**
External explanation evaluation of various deep learning models by XAI methods. The evaluation metrics on the test set are presented, accompanied by their respective standard errors enclosed within parentheses.

| Model | XAI | Template-Guidance | IoU (%) | DSC (%) |
|---|---|---|---|---|
| VGG-19 | Saliency Map | ✘ | 1.3 (0.2) | 2.5 (0.4) |
| | | ✔ | 2.9 (0.3) | 5.5 (0.6) |
| | Grad-CAM | ✘ | 1.1 (0.4) | 1.9 (0.6) |
| | | ✔ | 1.9 (0.4) | 3.6 (0.7) |
| | Integrated Gradients | ✘ | 2.3 (0.3) | 4.4 (0.5) |
| | | ✔ | 3.9 (0.4) | 7.3 (0.7) |
| ResNet-50 | Saliency Map | ✘ | 1.7 (0.2) | 3.2 (0.4) |
| | | ✔ | 3.7 (0.4) | 6.8 (0.7) |
| | Grad-CAM | ✘ | 1.5 (0.4) | 2.7 (0.8) |
| | | ✔ | 3.5 (0.6) | 6.3 (0.9) |
| | Integrated Gradients | ✘ | 1.8 (0.2) | 3.4 (0.4) |
| | | ✔ | 4.2 (0.5) | 7.6 (0.8) |

0.8% and 1.7% for Grad-CAM, and 1.6% and 2.9% for Integrated Gradients. Based on ResNet-50, the performance enhancements were 2.0% and 3.6% for Saliency Map, 2.0% and 3.6% for Grad-CAM, and 2.4% and 4.2% for Integrated Gradients. Notably, the incremental percentages of IoU and DSC, when compared with baseline methods, ranged from 71.3% to 130.9% and 66.5% to 134.1%, respectively.

These quantitative metrics elucidated the explanation improvements attributable to the proposed template-guided approach. To further compare XAI methods with and without template guidance, Figs. 2 and 3 visualize their explanations on the internal test set of SIIM-ACR and the external test set of ChestX-Det, respectively. From the left to the right, each figure displays the lesion areas delineated by radiologists (Ground truth), important regions outlined by the original explanations (Baseline), and the enhanced explanations (Our method). Figs. 2**(b)** and 3**(b)** show the samples on which the proposed approach can upgrade the original explanation quality. However, the proposed method fails to upgrade the baseline in Figs. 2**(c)** and 3**(c)**. Such a performance contrast demonstrated that our template-guided approach fails when the pneumothorax exists outside the template region. Figs. 2**(a)** and 3**(a)** illustrate scenarios where both XAI methods with and without template guidance perform well, whereas Figs. 2**(d)** and 3**(d)** depict situations where both XAI methods with and without template guidance exhibit poor performance. Also, we identified that either method presented a lower performance for small pneumothorax compared with the large one, which has been reported by other studies [2].

## 4. Discussion

This study proposed a template-guided approach to improve AI model explanation in the context of pneumothorax diagnoses. Based on clinical knowledge that pneumothorax occurs in the pleural space, we generated a template covering the pleural space based on a canonical lesion annotation by radiologists. Then the template was superimposed on the baseline explanations to filter out extraneous model explanations that fall outside the template's boundaries. This straightforward approach effectively constrained model explanations within the potential areas of pneumothorax occurrence, avoiding clinically irrelevant correlations in extraneous areas outside lesion regions, thereby consistently improving baseline XAI methods across twelve benchmark scenarios [52].

Beyond the investigated pneumothorax, our template guidance holds applicability to other thoracic diseases characterized by clinically validated disease locations. Cardiomegaly, the heart enlargement evident at the cardiac region [53], serves as another use case for the proposed template-guided approach. According to the radiological knowledge, the cardiac region encompasses the central area of a frontal chest radiograph [54]. To derive a comprehensive occurrence template of cardiomegaly, radiologists are invited to meticulously analyze radiograph samples and collaborate closely with AI engineers to ascertain the details of morphological operations. With the derived template, model explanation aligns better with clinical knowledge, underscoring the value of embedding domain knowledge in DL for healthcare.

Our method focuses on incorporating domain knowledge into model explanations. Researchers have also proposed various methods to enhance model explanations for chest radiograph analysis. For instance, Zou et al. [55] integrated SHapley Additive exPlanations (SHAP) [56] with Grad-CAM++ [57] to produce augmented explanations for chest radiograph-based pneumonia diagnosis [58]. Similar ensembling of Saliency Map and Grad-CAM has also been proven effective in improving the robustness and accuracy of model explanations pertaining to prostate lesion diagnosis. Nevertheless, it is worth noting that some professionals hold the perspective that a segmentation model assists clinicians better than a classification model supplemented with enhanced explanations [59]. Unlike a classification model that outputs a single diagnostic probability, a segmentation model explicitly delineates disease lesion areas. Yet an accurate segmentation model is largely dependent on the availability of large-scale pixel-level annotations, which are time-consuming and hard to acquire [60]. Potential solutions to this dilemma are semi-supervised learning and weakly-supervised learning [61].

Validated through comprehensive studies, both semi-supervised learning and weakly-supervised learning stand out as effective methods for alleviating the annotation burden during the development of an accurate segmentation model. Madani et al. [59] proposed a semi-supervised approach for cardiac disease prediction that achieved high accuracy using only a small amount of lesion annotations. Based on only 4% labeled data, they achieved 85% of the accuracy by the fully-supervised model on 100% labeled data. Semi-supervised learning still requires few pixel-level annotations while weakly-supervised learning aims to build a segmentation model using only image-level labels. Ouyang et al. [62] derived the pixel-level segmentations through focus areas extracted from a classification model and corrected the noisy segmentation label by a spatial annotation smoothing technique. They showed that the weakly-supervised approach upgraded the segmentation model training significantly without any pixel-level annotations. While these methods are promising in reducing the labeling cost, several studies have reported that semi- or weakly-supervised learning failed to reach the baseline by a fully-supervised model [63,64]. In medical AI, how to achieve a balance between the annotation cost and AI accuracy remains an unsolved conundrum in resource-limited settings [24]. With the recent release of versatile foundation models, a potential solution could be the Segment Anything Model (SAM), known for its capability to cut out any object in any image with a single click [65]. Hence, under the same budget, SAM facilitates the annotation of a larger number of samples and the development of a more accurate segmentation model [65].

Our study has limitations that warrant future investigation. First, we

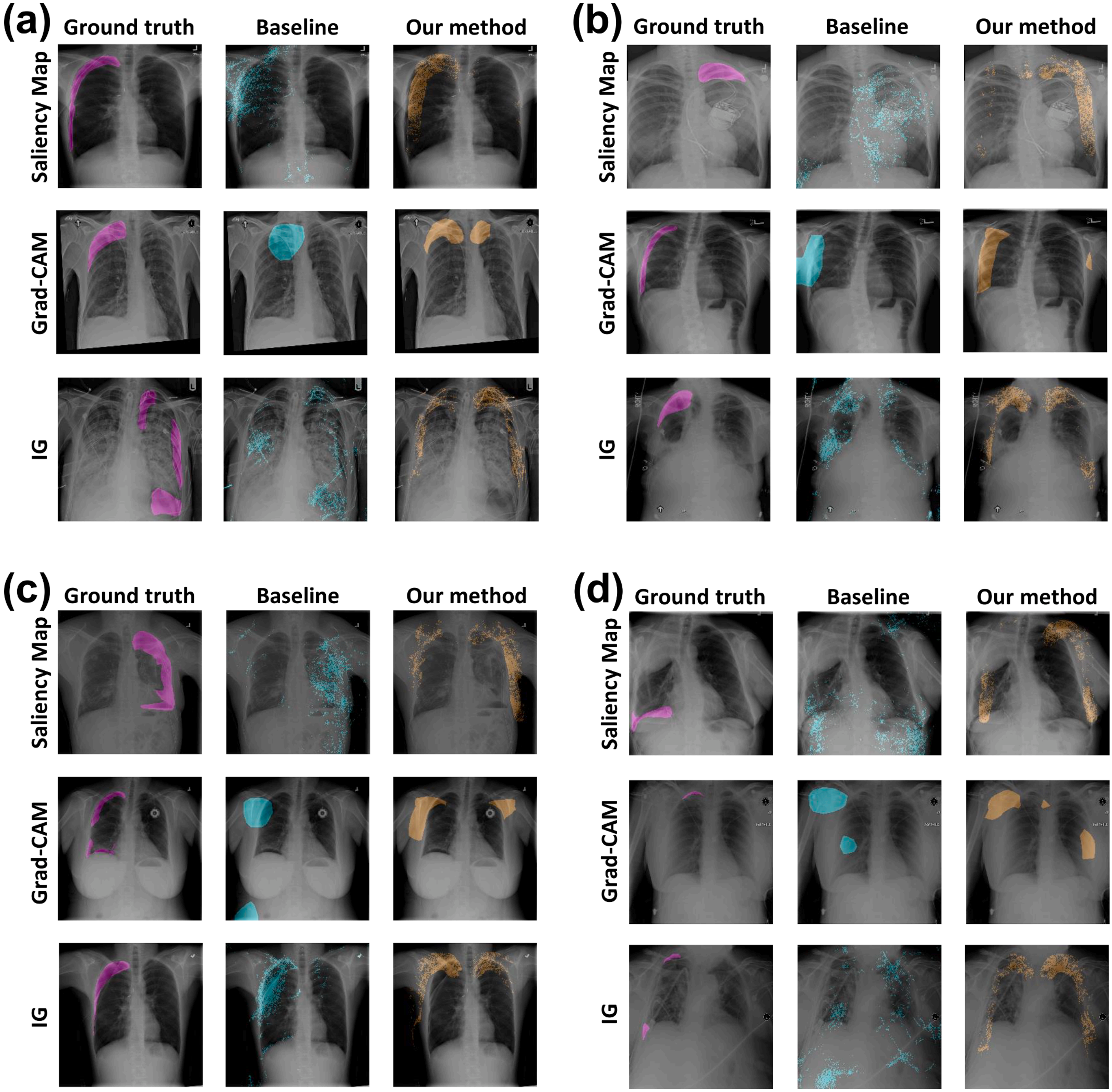

**Fig. 2.** Visualization comparison of pneumothorax radiograph explained by original and the template-guided Saliency Map, Grad-CAM, and Integrated Gradients (IG) on the internal SIIM-ACR test set.

employed a static template as a prior in guiding model explanations. Although the proposed method improved baseline explanations, the current performance is still unable to meet the deployment standards required by some regulatory agencies [66]. For example, the Korea Ministry of Food and Drug Safety mandates a minimum Dice coefficient of 20% for clinically relevant explanations [67]. Future research may explore the integration of the current approach with image transformation, which has been proven valuable in modifying the scale, angle, and displacement of the fixed template, thereby improving the explanation performance [30,68]. Second, we evaluated the performance of the template-guided approach within a limited set of experimental configurations, comprising three XAI baselines, two DL models, one thoracic disease, and one annotated lesion area. Future endeavors will encompass alternative XAI methods like LayerCAM [69], extra DL models including vision transformer [70], additional thoracic diseases such as cardiomegaly [40,71,72], and templates generated by combining multiple annotated lesion images.

## 5. Conclusion

Clinical domain knowledge has been under-investigated in the DL community when designing XAI methods and applications. In this study, we showcase the value of clinical knowledge, especially potential areas of disease occurrence, in consistently improving model explanations across twelve benchmark scenarios. It is imperative to highlight that our template-guided approach necessitates only a single lesion delineation crafted by radiologists, obviating the need for extensive annotation efforts. We anticipate that template guidance will forge a novel approach

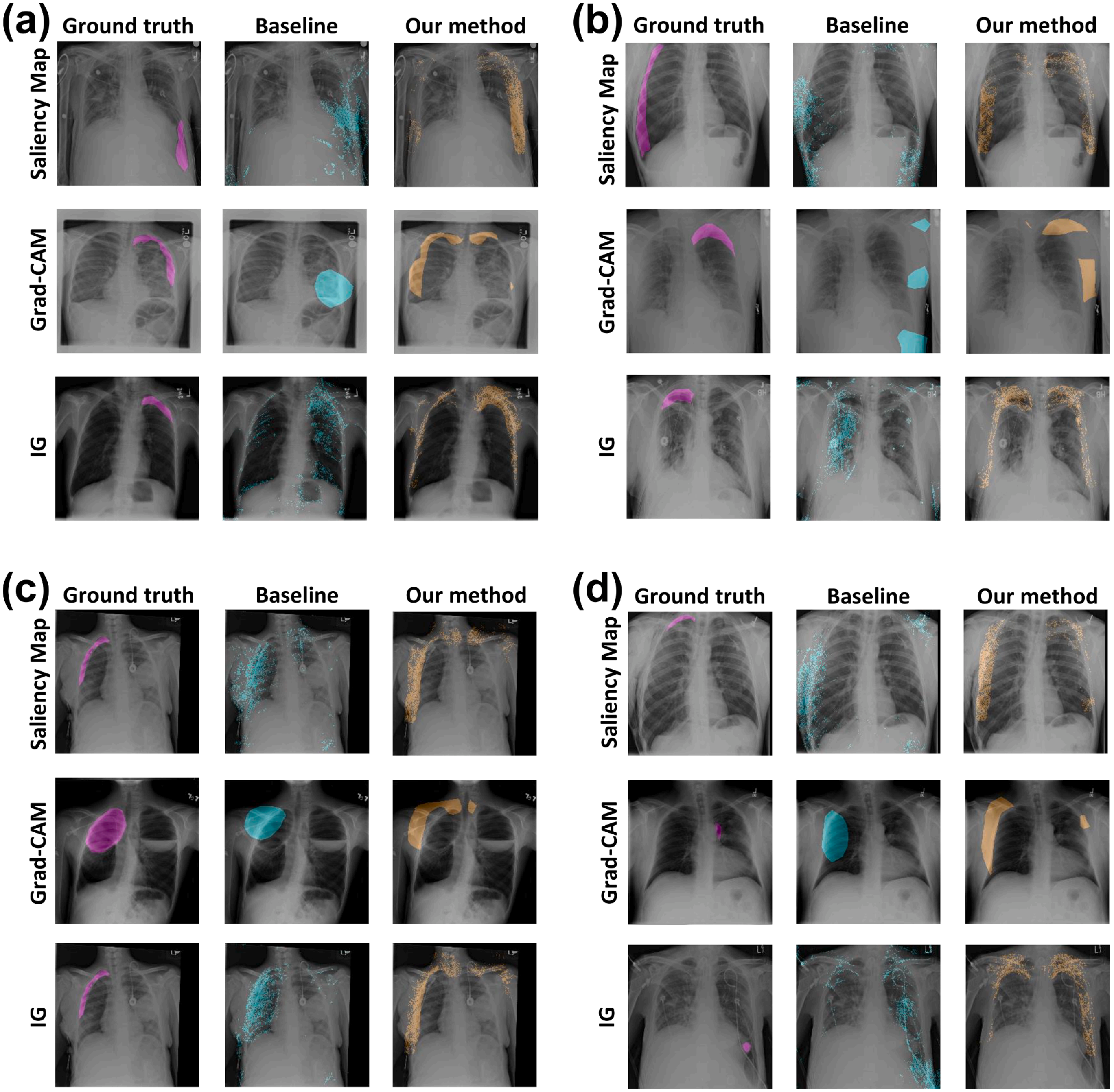

**Fig. 3.** Visualization comparison of pneumothorax radiograph explained by original and the template-guided Saliency Map, Grad-CAM, and Integrated Gradients (IG) on the external ChestX-Det test set.

to elucidate AI models with the integration of clinical domain knowledge.

## Declaration of generative AI in scientific writing

During the revision of the initial draft, Han Yuan used GPT-3.5 to check grammar. After using this tool, Han Yuan and other authors reviewed and edited the content as needed. Han Yuan takes full responsibility for the content of the publication.

## Funding

This study was supported by Duke-NUS Medical School, Singapore. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## Data availability

We used de-identified chest radiographs from SIIM-ACR Pneumothorax Segmentation Challenge and ChestX-Det, which are available at https://kaggle.com/competitions/siim-acr-pneumothorax-segmentation and https://github.com/Deepwise-AILab/ChestX-Det10-Dataset, respectively.

## CRediT authorship contribution statement

**Han Yuan:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Chuan Hong:** Writing – review & editing, Writing – original draft, Methodology, Investigation. **Peng-Tao Jiang:** Writing – review & editing, Methodology, Investigation. **Gangming Zhao:** Writing – review & editing, Methodology, Investigation. **Nguyen Tuan Anh Tran:** Writing – review & editing, Methodology, Investigation. **Xinxing Xu:** Writing – review & editing, Methodology, Investigation. **Yet Yen Yan:** Writing – review & editing, Methodology, Investigation. **Nan Liu:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References

[1] J.B. Imran, A.L. Eastman, Pneumothorax, J. Am. Med. Assoc. 318 (10) (2017), 974–974.

[2] Y.L. Thian, D. Ng, J.T.P.D. Hallinan, P. Jagmohan, S.Y. Sia, C.H. Tan, Y.H. Ting, P. L. Kei, G.G. Pulickal, Tiong VTY: deep learning systems for pneumothorax detection on chest radiographs: a multicenter external validation study, Radiol. Artif. Intell. 3 (4) (2021) e200190.

[3] M. Tan, Q. Le, Efficientnet: rethinking model scaling for convolutional neural networks, Proc. Int. Conf. Mach. Learn. (2019) 6105–6114.

[4] F. Chollet, Xception: deep learning with depthwise separable convolutions, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2017) 1251–1258.

[5] Y. Wang, L. Sun, Q. Jin, Enhanced diagnosis of pneumothorax with an improved real-time augmentation for imbalanced chest X-rays data based on DCNN, IEEE/ACM Trans. Comput. Biol. Bioinf. 18 (3) (2019) 951–962.

[6] T. Dhar, N. Dey, S. Borra, R.S. Sherratt, Challenges of deep learning in medical image analysis—improving explainability and trust, IEEE Trans. Technol. Soc. 4 (1) (2023) 68–75.

[7] H.-G. Jung, W.-J. Nam, H.-W. Kim, S.-W. Lee, Weakly supervised thoracic disease localization via disease masks, Neurocomputing 517 (2023) 34–43.

[8] M. Liu, S. Li, H. Yuan, M.E.H. Ong, Y. Ning, F. Xie, S.E. Saffari, Y. Shang, V. Volovici, B. Chakraborty, et al., Handling missing values in healthcare data: a systematic review of deep learning-based imputation techniques, Artif. Intell. Med. 102587 (2023).

[9] C. Mosquera, F.N. Diaz, F. Binder, J.M. Rabellino, S.E. Benitez, A.D. Beresñak, A. Seehaus, G. Ducrey, J.A. Ocantos, D.R. Luna, Chest x-ray automated triage: a semiologic approach designed for clinical implementation, exploiting different types of labels through a combination of four Deep Learning architectures, Comput. Methods Programs Biomed. 206 (2021) 106130.

[10] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2016) 2921–2929.

[11] S. Feng, Q. Liu, A. Patel, S.U. Bazai, C.K. Jin, J.S. Kim, M. Sarrafzadeh, D. Azzollini, J. Yeoh, E. Kim, Automated pneumothorax triaging in chest X-rays in the New Zealand population using deep-learning algorithms, J. Med. Imaging Radiat. Oncol. 66 (8) (2022) 1035–1043.

[12] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: visual explanations from deep networks via gradient-based localization, Proc. IEEE Int. Conf. Comput. Vis. (2017) 618–626.

[13] G. Huang, Z. Liu, L. Van Der aaten, K.Q. Weinberger, Densely connected convolutional networks, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2017) 4700–4708.

[14] A. Saporta, X. Gui, A. Agrawal, A. Pareek, S.Q. Truong, C.D. Nguyen, V.-D. Ngo, J. Seekins, F.G. Blankenberg, A.Y. Ng, Benchmarking saliency methods for chest X-ray interpretation, Nat. Mach. Intell. 4 (10) (2022) 867–878.

[15] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2016) 770–778.

[16] J. Rocha, A.M. Mendonça, S.C. Pereira, A. Campilho, Confident-CAM: improving heat map interpretation in chest X-ray image classification, Proc. Int. Conf. Bioinformatics Biomed. (2023) 4116–4123.

[17] G. Bel Bordes, Fairness and Explainability in Chest X-ray Image Classifiers, Universiteit Utrecht CENTAI, 2023.

[18] J. Crosby, T. Rhines, F. Li, H. MacMahon, M. Giger, Deep learning for pneumothorax detection and localization using networks fine-tuned with multiple institutional datasets, Proc. SPIE Med. Imag. (2020).

[19] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: visualising image classification models and saliency maps, Proc. Int. Conf. Learn. Represent. (2013).

[20] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, Proc. Int. Conf. Mach. Learn. (2017) 3319–3328.

[21] E. Mosqueira-Rey, E. Hernández-Pereira, D. Alonso-Ríos, J. Bobes-Bascarán, Á. Fernández-Leal, Human-in-the-loop machine learning: a state of the art, Artif. Intell. Rev. 56 (4) (2023) 3005–3054.

[22] W. Rawat, Z. Wang, Deep convolutional neural networks for image classification: a comprehensive review, Neural Comput. 29 (9) (2017) 2352–2449.

[23] B. Norgeot, B.S. Glicksberg, A.J. Butte, A call for deep-learning healthcare, Nat. Med. 25 (1) (2019) 14–15.

[24] F. Xie, H. Yuan, Y. Ning, M.E.H. Ong, M. Feng, W. Hsu, B. Chakraborty, N. Liu, Deep learning for temporal data representation in electronic health records: a systematic review of challenges and methodologies, J. Biomed. Inform. 103980 (2021).

[25] S. Nazir, D.M. Dickson, M.U. Akram, Survey of explainable artificial intelligence techniques for biomedical imaging with deep neural networks, Comput. Biol. Med. (2023:) 106668.

[26] B.H. Van der Velden, H.J. Kuijf, K.G. Gilhuijs, M.A. Viergever, Explainable artificial intelligence (XAI) in deep learning-based medical image analysis, Med. Image Anal. 79 (2022) 102470.

[27] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, B. Kim, Sanity checks for saliency maps, Proc. Adv. Neural Inf. Process. Syst.: 2018 (2018).

[28] J.P. Amorim, P.H. Abreu, J. Santos, M. Cortes, V. Vila, Evaluating the faithfulness of saliency maps in explaining deep learning models using realistic perturbations, Inf. Process. Manag. 60 (2) (2023) 103225.

[29] A. Quincho-Lopez, D.L. Quincho-Lopez, F.D. Hurtado-Medina, Case report: pneumothorax and pneumomediastinum as uncommon complications of COVID-19 pneumonia—literature review, Am. J. Trop. Med. Hyg. 103 (3) (2020) 1170.

[30] J. Liu, G. Zhao, Y. Fei, M. Zhang, Y. Wang, Y. Yu, Align, attend and locate: chest x-ray diagnosis via contrast induced attention network with limited supervision, Proc. IEEE Int. Conf. Comput. Vis. (2019) 10632–10641.

[31] J.C. Souza, J.O.B. Diniz, J.L. Ferreira, G.L.F. Da Silva, A.C. Silva, A.C. de Paiva, An automatic method for lung segmentation and reconstruction in chest X-ray using deep neural networks, Comput. Methods Programs Biomed. 177 (2019) 285–296.

[32] Template guidance code [https://github.com/Han-Yuan-Med/template-explanation].

[33] SIIM-ACR Pneumothorax Segmentation Challenge [https://kaggle.com/competitions/siim-acr-pneumothorax-segmentation].

[34] J. Liu, J. Lian, Y. Yu, Chestx-det10: chest x-ray dataset on detection of thoracic abnormalities, arXiv 2020.

[35] X. Wang, S. Yang, J. Lan, Y. Fang, J. He, M. Wang, J. Zhang, X. Han, Automatic segmentation of pneumothorax in chest radiographs based on a two-stage deep learning method, IEEE Trans. Cogn. Develop. Syst. 14 (1) (2020) 205–218.

[36] Y. Wang, K. Wang, X. Peng, L. Shi, J. Sun, S. Zheng, F. Shan, W. Shi, L. Liu, DeepSDM: boundary-aware pneumothorax segmentation in chest X-ray images, Neurocomputing 454 (2021) 201–211.

[37] W. Sae-Lim, W. Wettayaprasit, R. Suwannanon, S. Cheewatanakornkul, P. Aiyarak, automated pneumothorax segmentation and quantification algorithm based on deep learning, Intell. Syst. Appl. 22 (2024) 200383.

[38] F. Haghighi, M.R.H. Taher, M.B. Gotway, J. Liang, Self-supervised learning for medical image analysis: discriminative, restorative, or adversarial? Med. Image Anal. 94 (2024) 103086.

[39] C. Kim, G. Kim, S. Yang, H. Kim, S. Lee, H. Cho, Chest X-ray feature pyramid sum model with diseased area data augmentation method, Proc. IEEE/CVF Int. Conf. Comput. Vis.: 2023 (2023) 2757–2766.

[40] J. Lian, J. Liu, S. Zhang, K. Gao, X. Liu, D. Zhang, Y. Yu, A structure-aware relation network for thoracic diseases detection and segmentation, IEEE Trans. Med. Imaging 40 (8) (2021) 2042–2052.

[41] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, Chexpert: a large chest radiograph dataset with uncertainty labels and expert comparison, Proc. AAAI Conf. Artif. Intell. 33 (01) (2019) 590–597.

[42] H.Q. Nguyen, K. Lam, L.T. Le, H.H. Pham, D.Q. Tran, D.B. Nguyen, D.D. Le, C. M. Pham, H.T. Tong, D.H. Dinh, VinDr-CXR: an open dataset of chest X-rays with radiologist's annotations, Sci. Data 9 (1) (2022) 429.

[43] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, Proc. Int. Conf. Learn. Represent. (2015).

[44] E. Agliari, F. Alemanno, A. Barra, G. De Marzo, The emergence of a concept in shallow neural networks, Neural Netw. 148 (2022) 232–253.

[45] S. Oymak, M. Soltanolkotabi, Toward moderate overparameterization: global convergence guarantees for training shallow neural networks, IEEE J. Select. Areas Inf. Theory 1 (1) (2020) 84–105.

[46] W. Liao, B. Zou, R. Zhao, Y. Chen, Z. He, M. Zhou, Clinical interpretable deep learning model for glaucoma diagnosis, IEEE J. Biomed. Health Inform. 24 (5) (2019) 1405–1412.

[47] Z. Qiu, H. Rivaz, Y. Xiao, Is visual explanation with Grad-CAM more reliable for deeper neural networks? a case study with automatic pneumothorax diagnosis, Proc. Int. Workshop Mach. Learn. Med. Imag. (2023) 224–233.

[48] Y. Zhao, H. Yuan, Y. Wu, Prediction of adverse drug reaction using machine learning and deep learning based on an imbalanced electronic medical records dataset, Proc. Int. Conf. Med. Health Inform. (2021) 17–21.

[49] W.J. Youden, Index for rating diagnostic tests, Cancer 3 (1) (1950) 32–35.

[50] B. Efron, R. Tibshirani, Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy, Stat. Sci. 1 (1) (1986) 54–75.

[51] Z. Chen, Z. Tian, J. Zhu, C. Li, S. Du, C-cam: Causal cam for weakly supervised semantic segmentation on medical image, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2022) 11676–11685.

[52] A. Brown, N. Tomasev, J. Freyberg, Y. Liu, A. Karthikesalingam, J. Schrouff, Detecting shortcut learning for fair medical AI using shortcut testing, Nat. Commun. 14 (1) (2023) 4314.

[53] M. Innat, M.F. Hossain, K. Mader, A.Z. Kouzani, A convolutional attention mapping deep neural network for classification and localization of cardiomegaly on chest X-rays, Sci. Rep. 13 (1) (2023) 6247.

[54] K. Truszkiewicz, R. Poręba, P. Gać, Radiological cardiothoracic ratio in evidence-based medicine, J. Clin. Med. 10 (9) (2021) 2016.

[55] M.A. Gulum, C.M. Trombley, M. Kantardzic, Improved deep learning explanations for prostate lesion classification through grad-CAM and saliency map fusion, Proc. IEEE Int. Symp. Comput.-Based Med. Syst. (2021) 498–502.

[56] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Proc. Adv. Neural Inf. Proc. Syst. 30 (2017).

[57] A. Chattopadhay, A. Sarkar, P. Howlader, V.N. Balasubramanian, Grad-cam++: generalized gradient-based visual explanations for deep convolutional networks, Proc. IEEE Winter Conf. Appl. Comput. Vis. (2018) 839–847.

[58] L. Zou, H.L. Goh, C.J.Y. Liew, J.L. Quah, G.T. Gu, J.J. Chew, M.P. Kumar, C.G. L. Ang, A.W.A. Ta, Ensemble image explainable AI (XAI) algorithm for severe community-acquired pneumonia and COVID-19 respiratory infections, IEEE Trans. Artif. Intell. 4 (2) (2022) 242–254.

[59] A. Madani, J.R. Ong, A. Tibrewal, M.R. Mofrad, Deep echocardiography: data-efficient supervised and semi-supervised deep learning towards automated diagnosis of cardiac disease, npj Digital Med. 1 (1) (2018) 1–11.

[60] E. Tiu, E. Talius, P. Patel, C.P. Langlotz, A.Y. Ng, P. Rajpurkar, Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning, Nat. Biomed. Eng. 6 (12) (2022) 1399–1406.

[61] L. Tong, H. Wu, M.D. Wang, CAESNet: convolutional AutoEncoder based Semi-supervised Network for improving multiclass classification of endomicroscopic images, J. Am. Med. Inform. Assoc. 26 (11) (2019) 1286–1296.

[62] X. Ouyang, Z. Xue, Y. Zhan, X.S. Zhou, Q. Wang, Y. Zhou, Q. Wang, J.-Z. Cheng, Weakly supervised segmentation framework with uncertainty: a study on pneumothorax segmentation in chest X-ray, Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv. (2019).

[63] D. Pathak, P. Krahenbuhl, T. Darrell, Constrained convolutional neural networks for weakly supervised segmentation, Proc. IEEE Int. Conf. Comput. Vis. (2015).

[64] H. Kervadec, J. Dolz, M. Tang, E. Granger, Y. Boykov, I.B. Ayed, Constrained-CNN losses for weakly supervised segmentation, Med. Image Anal. 54 (2019) 88–99.

[65] M. Mazurowski, H. Dong, H. Gu, J. Yang, N. Konz, Y. Zhang, Segment anything model for medical image analysis: an experimental study, Med. Image Anal. 89 (2023) 102918.

[66] W. Jin, X. Li, G. Hamarneh, Evaluating explainable AI on a multi-modal medical imaging task: can existing algorithms fulfill clinical requirements? Proc. AAAI Conf. Artif. Intell. 36 (11) (2022) 11945–11953.

[67] S.Y. Lee, S. Ha, M.G. Jeon, H. Li, H. Choi, H.P. Kim, Y.R. Choi, Y.J. Jeong, Y. H. Park, H. Ahn, Localization-adjusted diagnostic performance and assistance effect of a computer-aided detection system for pneumothorax and consolidation, npj Digital Med. 5 (1) (2022) 1–11.

[68] H. Yuan, P. Jiang, G. Zhao, Human-guided design to explain deep learning-based pneumothorax classifier, Proc. Med. Imag. Deep Learn. (2023).

[69] P. Jiang, C. Zhang, Q. Hou, M. Cheng, Y. Wei, Layercam: exploring hierarchical class activation maps for localization, IEEE Trans. Image Process. 30 (2021) 5875–5888.

[70] F. Shamshad, S. Khan, S.W. Zamir, M.H. Khan, M. Hayat, F.S. Khan, H. Fu, Transformers in medical imaging: a survey, Med. Image Anal. 102802 (2023).

[71] R. Zeleznik, J. Weiss, J. Taron, C. Guthier, D.S. Bitterman, C. Hancox, B.H. Kann, D. W. Kim, R.S. Punglia, J. Bredfeldt, Deep-learning system to improve the quality and efficiency of volumetric heart segmentation for breast cancer, npj Digital Med. 4 (1) (2021) 1–7.

[72] J.C. Seah, C.H. Tang, Q.D. Buchlak, X.G. Holt, J.B. Wardman, A. Aimoldin, N. Esmaili, H. Ahmad, H. Pham, J.F. Lambert, Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: a retrospective, multireader multicase study, Lancet Digital Health 3 (8) (2021) e496–e506.