Received: 27 September 2024 | Accepted: 4 November 2024

DOI: 10.1002/ird3.113

ORIGINAL ARTICLE

iRADIOLOGY

Anatomic boundary-aware explanation for convolutional neural networks in diagnostic radiology

Han Yuan

Duke-NUS Medical School, National University of Singapore, Singapore

Correspondence
Han Yuan.
Email: yuan.han@u.duke.nus.edu

Abstract

Background: Convolutional neural networks (CNN) have achieved remarkable success in medical image analysis. However, unlike some general-domain tasks where model accuracy is paramount, medical applications demand both accuracy and explainability due to the high stakes affecting patients' lives. Based on model explanations, clinicians can evaluate the diagnostic decisions suggested by CNN. Nevertheless, prior explainable artificial intelligence methods treat medical image tasks akin to general vision tasks, following end-to-end paradigms to generate explanations and frequently overlooking crucial clinical domain knowledge.

Methods: We propose a plug-and-play module that explicitly integrates anatomic boundary information into the explanation process for CNN-based thoracopathy classifiers. To generate the anatomic boundary of the lung parenchyma, we utilize a lung segmentation model developed on external public datasets and deploy it on the unseen target dataset to constrain model explanations within the lung parenchyma for the clinical task of thoracopathy classification.

Results: Assessed by the intersection over union and dice similarity coefficient between model-extracted explanations and expert-annotated lesion areas, our method consistently outperformed the baseline devoid of clinical domain knowledge in 71 out of 72 scenarios, encompassing 3 CNN architectures (VGG-11, ResNet-18, and AlexNet), 2 classification settings (binary and multi-label), 3 explanation methods (Saliency Map, Grad-CAM, and Integrated Gradients), and 4 co-occurred thoracic diseases (Atelectasis, Fracture, Mass, and Pneumothorax).

Conclusions: We underscore the effectiveness of leveraging radiology knowledge in improving model explanations for CNN and envisage that it could inspire future efforts to integrate clinical domain knowledge into medical image analysis.

Abbreviations: AUPRC, area under the precision recall curve; CNN, convolutional neural networks; DL, deep learning; DSC, dice similarity coefficient; IoU, intersection over union; SGD, stochastic gradient descent; XAI, explainable artificial intelligence.

KEYWORDS
atelectasis, convolutional neural networks, diagnostic radiology, explainable artificial intelligence, fracture, grad-cam, integrated gradients, mass, pneumothorax, saliency map

1 | INTRODUCTION

In the last decade, convolutional neural networks (CNN) have reshaped the diagnostic process of thoracopathy due to their revolutionary accuracy [1]. However, clinicians still cannot fully trust their predictive decisions on prospective medical practice because of their black-box characteristics and the high stakes affecting patients' lives [2]. Based on model explanations of their inference logic, clinicians can evaluate and correct the predictions made by CNN [3]. Therefore, various explainable artificial intelligence (XAI) methods, such as Saliency Map [4], Grad-CAM [5], and Integrated Gradients [6], have been proposed to highlight the important pixels (regions) toward CNN decisions [7]. These heatmaps are provided to clinicians to assess whether the model accurately focuses on clinically relevant regions, such as lesion areas, and to determine the suitability of adopting the model's suggestions [8].

Thoracopathy comprises conditions of the heart, lungs, mediastinum, esophagus, chest wall, diaphragm, and great vessels [9]. In this study, we focus on four co-occurring thoracic diseases: Atelectasis, Fracture, Mass, and Pneumothorax [10–12]. Conventional clinical diagnosis is based on the clinician's manual evaluation of chest radiographs while multiple CNN-based models have been proposed to automate this process in the deep learning (DL) era. For example, Chen et al. [13] used two asymmetric CNNs of DenseNet [14] and ResNet [15] to learn complementary features and implemented thoracic disease classification in chest X-rays. However, prior research efforts have primarily followed the end-to-end paradigm in both stages of classification and explanation [16], neglecting the clinical domain knowledge that thoracic diseases mainly occur in the lung parenchyma [17–20]. Prior studies have noticed this gap and presented the enhancement of accuracy through the incorporation of clinical domain knowledge [21]. For example, Jung et al. [22] introduced a spatial attention mechanism to highlight potential disease areas, generating disease masks with precise probability distributions based on 112,120 chest radiographs across 14 disease types. However, this approach may be impractical for resource-limited settings with smaller datasets and scarce disease labels. A comparable attention mechanism was incorporated into Thorax-Net [23], which consisted of both a classification branch and an attention branch that were assembled to produce the final diagnosis. As with the prior method, training the attention module required large-scale annotated datasets. Furthermore, the study presented only qualitative visualizations without providing quantitative evaluations of the heatmap-based pathological abnormal regions.

To explore the effectiveness of incorporating clinical domain knowledge to enhance the model explanations of CNN-based thoracopathy classifiers, we propose a plug-and-play module to constrain the model explanations within the lung parenchyma, tailored for resource-limited environments. To obtain the lung parenchyma, we adopt transfer learning to develop an external lung segmenter from external public datasets. This transfer learning strategy effectively reduces annotation costs associated with unseen target datasets and facilitates deployment on datasets with small sample sizes, where comprehensive annotation of all lung parenchyma samples may not guarantee the convergence of the lung segmentation model. Quantitatively assessed on 3 CNN architectures, 3 XAI methods, 4 thoracic diseases, and 2 classification settings, the proposed approach consistently outperformed the baseline model explanations devoid of domain knowledge. This study underscored the effectiveness of radiology knowledge in improving model explanations for CNN and we envisage that it could inspire future efforts to integrate clinical domain knowledge into medical image analysis [24–26].

2 | MATERIALS AND METHODS

2.1 | Datasets

We conducted a comprehensive evaluation of XAI-based model explanations and demonstrated the effectiveness of the proposed method by utilizing the public dataset of ChestX-Det [27]. We extracted 611 healthy samples and 880 samples were diagnosed with at least one of the four correlated thoracic diseases, including Atelectasis, Fracture, Mass, and Pneumothorax. All chest radiographs were resized to the resolution of 224 × 224 pixels to meet the requirements of most pre-trained DL backbones. In addition to binary diagnostic labels indicating the presence or absence of thoracic diseases, clinical experts enriched this dataset with pixel-level lesion annotations for each disease. Pixel-level lesion annotations were utilized solely for evaluating model explanations, while the training of CNN classifiers relied exclusively on image-level diagnostic labels. We randomly split the extracted samples into training, validation, and test datasets at 60: 20: 20, and Table 1 shows the precise number of samples for each category. Training and validation datasets were used to develop thoracopathy classifiers while the test dataset was used to evaluate classification and explanation performances of the developed classifiers and XAI methods. Two classification settings were explored for each disease: The first targeted training a multi-label classifier for all 4 diseases based on the entire training and validation datasets; The second scenario entailed training individual binary classifiers for each disease, using solely healthy samples and diseased samples diagnosed with the respective disease.

2.2 | CNN-based thoracopathy classification

Given the dataset containing small-scale samples, we developed thoracopathy classifiers using three lightweight CNN backbones: VGG-11 [28], ResNet-18 [15], and AlexNet [29]. The rationale for selecting these shallow-layer backbones was to mitigate overparameterization, given the limited size of our training set [16, 30], and to address potential spatial information loss in XAI when interpreting CNN architectures with deeper layers [16, 31]. For model training, we employed stochastic gradient descent (SGD) [32] with a learning rate of 0.001, a momentum of 0.9, and a decay of 0.9 with a patience parameter of 10. Inverse probability weights were introduced in the training of both binary and multi-label classifiers to eliminate the impact of dominating classes [33, 34]. Each CNN model underwent training for 100 epochs and was evaluated on the test dataset. Due to data imbalance, the area under the precision recall curve (AUPRC) was utilized as the primary evaluation metric for model classification [35]. Additional assessments included accuracy, precision, and recall. For all metrics, the standard deviation was calculated using bootstrapped samples from the test dataset to ensure comprehensive reporting [36].

2.3 | XAI for CNN-based thoracopathy classifiers

To explain the model decision logic behind the trained classifiers, various XAI methods were applied to derive each pixel's importance towards the model's predictive classification and the focus areas were further outlined by aggregating the most significant pixels with the top 5% of importance. In this study, we utilized three pixel-level XAI techniques of Saliency Map [4], Grad-CAM [5], and Integrated Gradients [6] because of their representativeness

TABLE 1 An overview of the data split in thoracopathy classification tasks.

Thoracopathy						
Atelectasis	**Fracture**	**Mass**	**Pneumothorax**	**Training set**	**Validation set**	**Test set**
✓	✓	✓	✓	0	0	0
✓	✓	✓		0	1	0
✓	✓		✓	1	0	1
	✓	✓	✓	0	0	0
✓	✓			23	8	7
✓		✓		6	2	2
✓			✓	12	4	4
	✓	✓		10	3	3
	✓		✓	26	9	8
		✓	✓	6	2	2
✓				136	45	45
	✓			173	58	58
		✓		67	22	22
			✓	68	23	23
No thoracopathy				367	122	122

and wide application as the baseline XAI algorithms [37] and compared the focus areas with the ground truth lesion areas based on the intersection over union (IoU) and dice similarity coefficient (DSC), accompanied by their bootstrapped standard deviations.

2.4 | Anatomic boundary-aware model explanation

According to clinical domain knowledge [22], thoracic diseases occur in the lung parenchyma on a 2D projection of a chest radiograph, and therefore, the model explanations should be constrained within the anatomic boundary of the lung parenchyma. Figure 1 outlines the proposed method which develops an auxiliary lung segmenter based on the external lung segmentation dataset of the Japanese Society of Radiological Technology dataset [38], the Shenzhen dataset [39], and the Montgomery County dataset [39]. The training configuration for the lung segmentation model was consistent with that of thoracopathy classifiers elaborated above, but the segmentation model employed a U-Net architecture [40] with a VGG-11 backbone [28] specifically designed for image segmentation tasks. Upon the completion of segmenter training, each chest radiograph from the unseen target dataset was supplemented with a boundary constraint. This boundary constraint enforced the model focus area within the predicted lung region to enhance the model explanations. The post hoc nature of this method enables seamless integration with any XAI technique, allowing for a straightforward plug-and-play application without complex modifications. For reproduction, the code has been publicly released on GitHub (https://github.com/Han-Yuan-Med/constrained-explanation).

3 | RESULTS

First, we quantitatively showed the performance of various CNN backbones in different classification settings in Table 2. VGG-11 consistently outperformed ResNet-18 and AlexNet across scenarios concerning AUPRC, accuracy, and in most scenarios, precision. The highest average recall of 0.676 was achieved by ResNet-18, surpassing VGG-11's 0.619 and AlexNet's 0.546. We also investigated binary and multi-label classification scenarios to assess whether additional information improves the discriminative capability of CNN classifiers. Binary classifiers showcased superior performance in terms of AUPRC, precision, and in most scenarios, accuracy compared to multi-label classifiers. Nonetheless, multi-label classifiers demonstrated an average recall of 0.628, marginally better than the 0.596 achieved by binary classifiers.

Tables 3–6 show the explanation performance of VGG-11, ResNet-18, and AlexNet utilizing different XAI methods for various thoracic diseases under binary or multi-label settings.

The binary VGG-11 model with boundary-aware Saliency Map achieved optimal explanation performance for Atelectasis, while the binary AlexNet model utilizing boundary-aware Grad-CAM excelled in Fracture classification explanation, the binary ResNet-18 with boundary-aware Grad-CAM demonstrated superior explanation

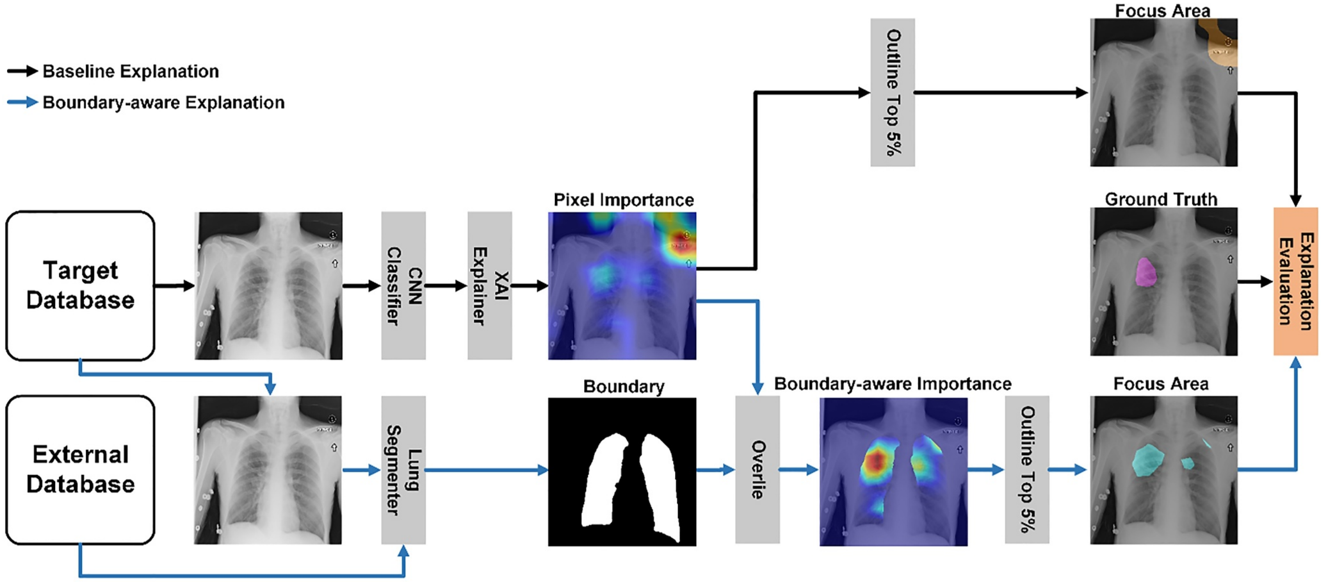

FIGURE 1 Schematic diagram of the proposed boundary-aware explanation. These figures are open-access from Deepwise AI Lab under the Apache-2.0 license, permitting use, modification, and distribution. CNN, convolutional neural networks; XAI, explainable artificial intelligence.

TABLE 2 Thoracopathy classification performance of VGG-11 and ResNet-18 on the test set.

Disease	Model	Setting	AUPRC	Accuracy	Precision	Recall
Atelectasis	VGG-11	Binary	0.845 (0.041)	0.851 (0.023)	0.758 (0.053)	0.797 (0.043)
		Multi-label	0.494 (0.046)	0.729 (0.020)	0.373 (0.046)	0.517 (0.068)
	ResNet-18	Binary	0.708 (0.053)	0.762 (0.031)	0.608 (0.055)	0.763 (0.047)
		Multi-label	0.382 (0.055)	0.569 (0.024)	0.296 (0.028)	0.833 (0.043)
	AlexNet	Binary	0.630 (0.064)	0.720 (0.038)	0.560 (0.055)	0.700 (0.055)
		Multi-label	0.294 (0.053)	0.635 (0.031)	0.257 (0.044)	0.433 (0.064)
Fracture	VGG-11	Binary	0.873 (0.031)	0.819 (0.028)	0.860 (0.043)	0.636 (0.064)
		Multi-label	0.570 (0.033)	0.599 (0.024)	0.364 (0.031)	0.696 (0.058)
	ResNet-18	Binary	0.549 (0.055)	0.598 (0.034)	0.485 (0.047)	0.623 (0.053)
		Multi-label	0.276 (0.057)	0.599 (0.028)	0.215 (0.037)	0.816 (0.056)
	AlexNet	Binary	0.648 (0.056)	0.650 (0.031)	0.538 (0.042)	0.731 (0.045)
		Multi-label	0.326 (0.033)	0.619 (0.026)	0.345 (0.039)	0.494 (0.053)
Mass	VGG-11	Binary	0.528 (0.096)	0.768 (0.033)	0.385 (0.101)	0.345 (0.098)
		Multi-label	0.404 (0.043)	0.408 (0.027)	0.107 (0.021)	0.667 (0.082)
	ResNet-18	Binary	0.236 (0.058)	0.609 (0.037)	0.266 (0.057)	0.586 (0.098)
		Multi-label	0.150 (0.037)	0.786 (0.021)	0.207 (0.048)	0.400 (0.084)
	AlexNet	Binary	0.187 (0.038)	0.563 (0.043)	0.206 (0.057)	0.448 (0.098)
		Multi-label	0.086 (0.018)	0.144 (0.018)	0.102 (0.016)	0.967 (0.027)
Pneumothorax	VGG-11	Binary	0.843 (0.048)	0.894 (0.021)	0.769 (0.072)	0.789 (0.060)
		Multi-label	0.398 (0.054)	0.726 (0.022)	0.232 (0.051)	0.500 (0.094)
	ResNet-18	Binary	0.411 (0.078)	0.619 (0.035)	0.328 (0.056)	0.579 (0.086)
		Multi-label	0.276 (0.057)	0.599 (0.028)	0.215 (0.037)	0.816 (0.056)
	AlexNet	Binary	0.254 (0.044)	0.688 (0.032)	0.250 (0.087)	0.158 (0.053)
		Multi-label	0.129 (0.030)	0.632 (0.029)	0.147 (0.027)	0.395 (0.067)

Note: The evaluation metrics are presented, accompanied by their respective standard errors enclosed within parentheses.

Abbreviation: AUPRC, area under the precision recall curve.

performance for Mass, and the multi-label AlexNet model employed boundary-aware Saliency Map for Pneumothorax. Also, anatomic boundary-aware explanations consistently exhibited improvements in IoU and DSC compared to baseline explanations, except for the Grad-CAM explanation of VGG-11 in multi-label Fracture classification. Additionally, binary classifiers outperformed multi-label classifiers in 26 out of 36 scenarios for baseline explanations and in 25 out of 36 scenarios for anatomic boundary-aware explanations. Lastly, although boundary-aware XAI methods produced the best explanation results, the anatomic boundary without any classification training showcased even better performance than classifier-based explanations in certain scenarios. Particularly in Fracture, only 4 classifier-based explanations achieved better IoU and DSC compared to the training-free boundary.

To provide a comparative visualization of model explanations with and without anatomic boundary, Figures 2–4 depict visual comparisons of Saliency Map, Grad-CAM, and Integrated Gradients based on binary VGG-11 classifiers. The images from the first to the fifth column are original images, ground-truth lesion area, anatomic boundary, baseline explanation, and boundary-aware explanation. Compared with baseline explanations, the anatomic boundary constrained model explanations within the lung parenchyma and therefore enhanced the consistency between the ground truth lesion regions and model explanations of focus areas.

Additionally, we performed an extensive analysis to examine how the quality of anatomic boundaries influence boundary-aware XAI performance. Specifically, we saved intermediate checkpoints of the segmentation

TABLE 3 Atelectasis explanation performance of CNN models by various XAI methods on the test dataset.

Disease	Model	Setting	XAI	IoU	DSC
Atelectasis	VGG-11	Binary	Saliency map	5.69 (0.51)	10.44 (0.92)
			Saliency map + boundary	5.89 (0.54)	10.69 (0.94)
		Multi-label	Saliency map	1.92 (0.38)	3.61 (0.66)
			Saliency map + boundary	3.84 (0.51)	7.09 (0.89)
		Binary	Grad-CAM	1.63 (0.48)	2.93 (0.82)
			Grad-CAM + boundary	2.74 (0.59)	4.95 (1.05)
		Multi-label	Grad-CAM	1.49 (0.38)	2.74 (0.69)
			Grad-CAM + boundary	1.76 (0.64)	3.17 (1.07)
		Binary	Integrated gradients	3.81 (0.36)	7.18 (0.64)
			Integrated gradients + boundary	4.92 (0.51)	9.05 (0.84)
		Multi-label	Integrated gradients	0.99 (0.26)	1.91 (0.48)
			Integrated gradients + boundary	3.29 (0.51)	6.09 (0.89)
	ResNet-18	Binary	Saliency map	2.28 (0.33)	4.36 (0.61)
			Saliency map + boundary	3.56 (0.46)	6.59 (0.77)
		Multi-label	Saliency map	3.80 (0.46)	7.07 (0.82)
			Saliency map + boundary	4.56 (0.56)	8.29 (0.94)
		Binary	Grad-CAM	4.12 (0.77)	7.41 (1.35)
			Grad-CAM + boundary	4.42 (0.71)	7.91 (1.22)
		Multi-label	Grad-CAM	3.77 (1.07)	6.31 (1.63)
			Grad-CAM + boundary	5.55 (0.99)	9.73 (1.66)
		Binary	Integrated gradients	2.82 (0.36)	5.32 (0.64)
			Integrated gradients + boundary	3.56 (0.46)	6.61 (0.79)
		Multi-label	Integrated gradients	4.08 (0.54)	7.51 (0.89)
			Integrated gradients + boundary	4.67 (0.56)	8.46 (0.97)
	AlexNet	Binary	Saliency map	2.94 (0.41)	5.49 (0.77)
			Saliency map + boundary	4.43 (0.69)	7.96 (1.15)
		Multi-label	Saliency map	0.01 (0.00)	0.03 (0.03)
			Saliency map + boundary	0.92 (0.31)	1.74 (0.59)
		Binary	Grad-CAM	1.21 (0.33)	2.21 (0.61)
			Grad-CAM + boundary	4.42 (1.10)	7.29 (1.61)
		Multi-label	Grad-CAM	0.08 (0.05)	0.16 (0.08)
			Grad-CAM + boundary	0.86 (0.26)	1.62 (0.46)
		Binary	Integrated gradients	1.77 (0.31)	3.36 (0.61)
			Integrated gradients + boundary	3.95 (0.59)	7.21 (1.02)
		Multi-label	Integrated gradients	0.01 (0.00)	0.02 (0.03)
			Integrated gradients + boundary	1.14 (0.36)	2.16 (0.61)
	—		Boundary	3.80 (0.66)	6.89 (1.10)

Note: The evaluation metrics are presented, accompanied by their respective standard errors enclosed within parentheses.

Abbreviations: CNN, convolutional neural networks; DSC, dice similarity coefficient; IoU, intersection over union; XAI, explainable artificial intelligence.

TABLE 4 Fracture explanation performance of CNN models by various XAI methods on the test dataset.

Disease	Model	Setting	XAI	IoU	DSC
Fracture	VGG-11	Binary	Saliency map	0.40 (0.05)	0.79 (0.10)
			Saliency map + boundary	0.94 (0.15)	1.85 (0.26)
		Multi-label	Saliency map	0.29 (0.08)	0.58 (0.20)
			Saliency map + boundary	0.86 (0.18)	1.67 (0.33)
		Binary	Grad-CAM	0.38 (0.15)	0.73 (0.31)
			Grad-CAM + boundary	1.40 (0.48)	2.43 (0.79)
		Multi-label	Grad-CAM	0.51 (0.15)	0.98 (0.28)
			Grad-CAM + boundary	0.46 (0.13)	0.89 (0.23)
		Binary	Integrated gradients	0.28 (0.05)	0.56 (0.10)
			Integrated gradients + boundary	0.85 (0.15)	1.66 (0.28)
		Multi-label	Integrated gradients	0.15 (0.05)	0.30 (0.10)
			Integrated gradients + boundary	0.93 (0.20)	1.80 (0.41)
	ResNet-18	Binary	Saliency map	0.77 (0.10)	1.51 (0.23)
			Saliency map + boundary	1.62 (0.23)	3.10 (0.41)
		Multi-label	Saliency map	0.58 (0.10)	1.15 (0.15)
			Saliency map + boundary	1.29 (0.18)	2.51 (0.36)
		Binary	Grad-CAM	1.52 (0.33)	2.82 (0.59)
			Grad-CAM + boundary	2.27 (0.46)	4.19 (0.84)
		Multi-label	Grad-CAM	0.35 (0.15)	0.66 (0.28)
			Grad-CAM + boundary	1.18 (0.31)	2.20 (0.56)
		Binary	Integrated gradients	0.54 (0.08)	1.06 (0.15)
			Integrated gradients + boundary	1.24 (0.20)	2.38 (0.38)
		Multi-label	Integrated gradients	0.37 (0.05)	0.73 (0.13)
			Integrated gradients + boundary	0.96 (0.18)	1.87 (0.36)
	AlexNet	Binary	Saliency map	0.41 (0.10)	0.80 (0.20)
			Saliency map + boundary	1.38 (0.31)	2.60 (0.56)
		Multi-label	Saliency map	0.22 (0.08)	0.41 (0.18)
			Saliency map + boundary	1.21 (0.23)	2.32 (0.41)
		Binary	Grad-CAM	0.10 (0.08)	0.20 (0.13)
			Grad-CAM + boundary	1.94 (0.54)	3.51 (0.94)
		Multi-label	Grad-CAM	0.14 (0.08)	0.26 (0.15)
			Grad-CAM + boundary	0.81 (0.28)	1.49 (0.51)
		Binary	Integrated gradients	0.30 (0.08)	0.60 (0.15)
			Integrated gradients + boundary	1.22 (0.23)	2.35 (0.43)
		Multi-label	Integrated gradients	0.28 (0.10)	0.54 (0.18)
			Integrated gradients + boundary	1.27 (0.20)	2.42 (0.38)
	—		Boundary	1.42 (0.18)	2.75 (0.33)

Note: The evaluation metrics are presented, accompanied by their respective standard errors enclosed within parentheses.

Abbreviations: CNN, convolutional neural networks; DSC, dice similarity coefficient; IoU, intersection over union; XAI, explainable artificial intelligence.

TABLE 5 Mass explanation performance of CNN models by various XAI methods on the test dataset.

Disease	Model	Setting	XAI	IoU	DSC
Mass	VGG-11	Binary	Saliency map	1.46 (0.48)	2.75 (0.87)
			Saliency map + boundary	3.64 (0.79)	6.76 (1.38)
		Multi-label	Saliency map	1.65 (0.59)	3.05 (1.10)
			Saliency map + boundary	4.23 (1.07)	7.60 (1.79)
		Binary	Grad-CAM	0.62 (0.41)	1.12 (0.71)
			Grad-CAM + boundary	5.75 (1.94)	9.20 (2.83)
		Multi-label	Grad-CAM	1.10 (0.74)	1.85 (1.20)
			Grad-CAM + boundary	4.28 (1.58)	6.89 (2.30)
		Binary	Integrated gradients	1.79 (0.43)	3.42 (0.84)
			Integrated gradients + boundary	5.26 (0.99)	9.51 (1.66)
		Multi-label	Integrated gradients	0.45 (0.15)	0.88 (0.31)
			Integrated gradients + boundary	4.26 (0.82)	7.86 (1.45)
	ResNet-18	Binary	Saliency map	3.99 (0.56)	7.47 (0.99)
			Saliency map + boundary	6.30 (0.94)	11.35 (1.68)
		Multi-label	Saliency map	3.68 (0.48)	6.97 (0.89)
			Saliency map + boundary	6.32 (1.07)	11.37 (1.81)
		Binary	Grad-CAM	6.06 (1.86)	10.11 (2.98)
			Grad-CAM + boundary	9.36 (2.40)	15.08 (3.78)
		Multi-label	Grad-CAM	3.27 (1.05)	5.76 (1.79)
			Grad-CAM + boundary	5.71 (1.66)	9.73 (2.70)
		Binary	Integrated gradients	5.21 (0.77)	9.55 (1.30)
			Integrated gradients + boundary	7.48 (1.15)	13.29 (1.89)
		Multi-label	Integrated gradients	3.99 (0.64)	7.44 (1.15)
			Integrated gradients + boundary	7.21 (1.07)	12.81 (1.86)
	AlexNet	Binary	Saliency map	0.45 (0.18)	0.89 (0.31)
			Saliency map + boundary	2.61 (0.51)	4.95 (0.97)
		Multi-label	Saliency map	0.04 (0.03)	0.08 (0.05)
			Saliency map + boundary	1.88 (0.74)	3.40 (1.30)
		Binary	Grad-CAM	0.58 (0.20)	1.11 (0.43)
			Grad-CAM + boundary	1.34 (0.43)	2.55 (0.82)
		Multi-label	Grad-CAM	0.00 (0.00)	0.00 (0.00)
			Grad-CAM + boundary	0.43 (0.23)	0.82 (0.41)
		Binary	Integrated gradients	1.62 (0.36)	3.10 (0.69)
			Integrated gradients + boundary	5.24 (0.97)	9.47 (1.61)
		Multi-label	Integrated gradients	0.02 (0.03)	0.04 (0.03)
			Integrated gradients + boundary	2.07 (0.64)	3.78 (1.12)
	—		Boundary	5.51 (0.74)	10.07 (1.30)

Note: The evaluation metrics are presented, accompanied by their respective standard errors enclosed within parentheses.

Abbreviations: CNN, convolutional neural networks; DSC, dice similarity coefficient; IoU, intersection over union; XAI, explainable artificial intelligence.

TABLE 6 Pneumothorax explanation performance of CNN models by various XAI methods on the test dataset.

Disease	Model	Setting	XAI	IoU	DSC
Pneumothorax	VGG-11	Binary	Saliency map	1.71 (0.28)	3.30 (0.51)
			Saliency map + boundary	2.65 (0.31)	5.09 (0.56)
		Multi-label	Saliency map	0.46 (0.15)	0.89 (0.28)
			Saliency map + boundary	1.89 (0.38)	3.59 (0.71)
		Binary	Grad-CAM	1.17 (0.43)	2.14 (0.77)
			Grad-CAM + boundary	1.59 (0.46)	2.97 (0.82)
		Multi-label	Grad-CAM	0.37 (0.18)	0.72 (0.36)
			Grad-CAM + boundary	1.82 (0.82)	3.18 (1.33)
		Binary	Integrated gradients	1.34 (0.41)	2.56 (0.74)
			Integrated gradients + boundary	2.64 (0.54)	4.98 (0.97)
		Multi-label	Integrated gradients	0.22 (0.13)	0.44 (0.23)
			Integrated gradients + boundary	2.02 (0.41)	3.86 (0.77)
	ResNet-18	Binary	Saliency map	0.45 (0.10)	0.89 (0.20)
			Saliency map + boundary	1.61 (0.20)	3.12 (0.41)
		Multi-label	Saliency map	0.51 (0.13)	1.02 (0.23)
			Saliency map + boundary	1.32 (0.20)	2.58 (0.43)
		Binary	Grad-CAM	1.45 (0.71)	2.63 (1.25)
			Grad-CAM + boundary	1.81 (0.74)	3.32 (1.35)
		Multi-label	Grad-CAM	0.69 (0.36)	1.28 (0.66)
			Grad-CAM + boundary	1.25 (0.56)	2.25 (0.99)
		Binary	Integrated gradients	0.68 (0.20)	1.32 (0.41)
			Integrated gradients + boundary	1.59 (0.31)	3.04 (0.56)
		Multi-label	Integrated gradients	0.72 (0.28)	1.35 (0.54)
			Integrated gradients + boundary	1.48 (0.38)	2.82 (0.69)
	AlexNet	Binary	Saliency map	0.14 (0.08)	0.29 (0.15)
			Saliency map + boundary	4.16 (0.64)	7.62 (1.07)
		Multi-label	Saliency map	0.49 (0.13)	0.98 (0.26)
			Saliency map + boundary	5.53 (0.74)	10.01 (1.28)
		Binary	Grad-CAM	0.52 (0.31)	0.94 (0.56)
			Grad-CAM + boundary	0.84 (0.26)	1.59 (0.48)
		Multi-label	Grad-CAM	0.28 (0.20)	0.50 (0.38)
			Grad-CAM + boundary	3.75 (1.30)	5.95 (1.86)
		Binary	Integrated gradients	0.09 (0.05)	0.18 (0.10)
			Integrated gradients + boundary	4.67 (0.69)	8.60 (1.25)
		Multi-label	Integrated gradients	0.19 (0.05)	0.39 (0.13)
			Integrated gradients + boundary	4.74 (0.74)	8.66 (1.20)
	—		Boundary	2.39 (0.31)	4.61 (0.56)

Note: The evaluation metrics are presented, accompanied by their respective standard errors enclosed within parentheses.

Abbreviations: CNN, convolutional neural networks; DSC, dice similarity coefficient; IoU, intersection over union; XAI, explainable artificial intelligence.

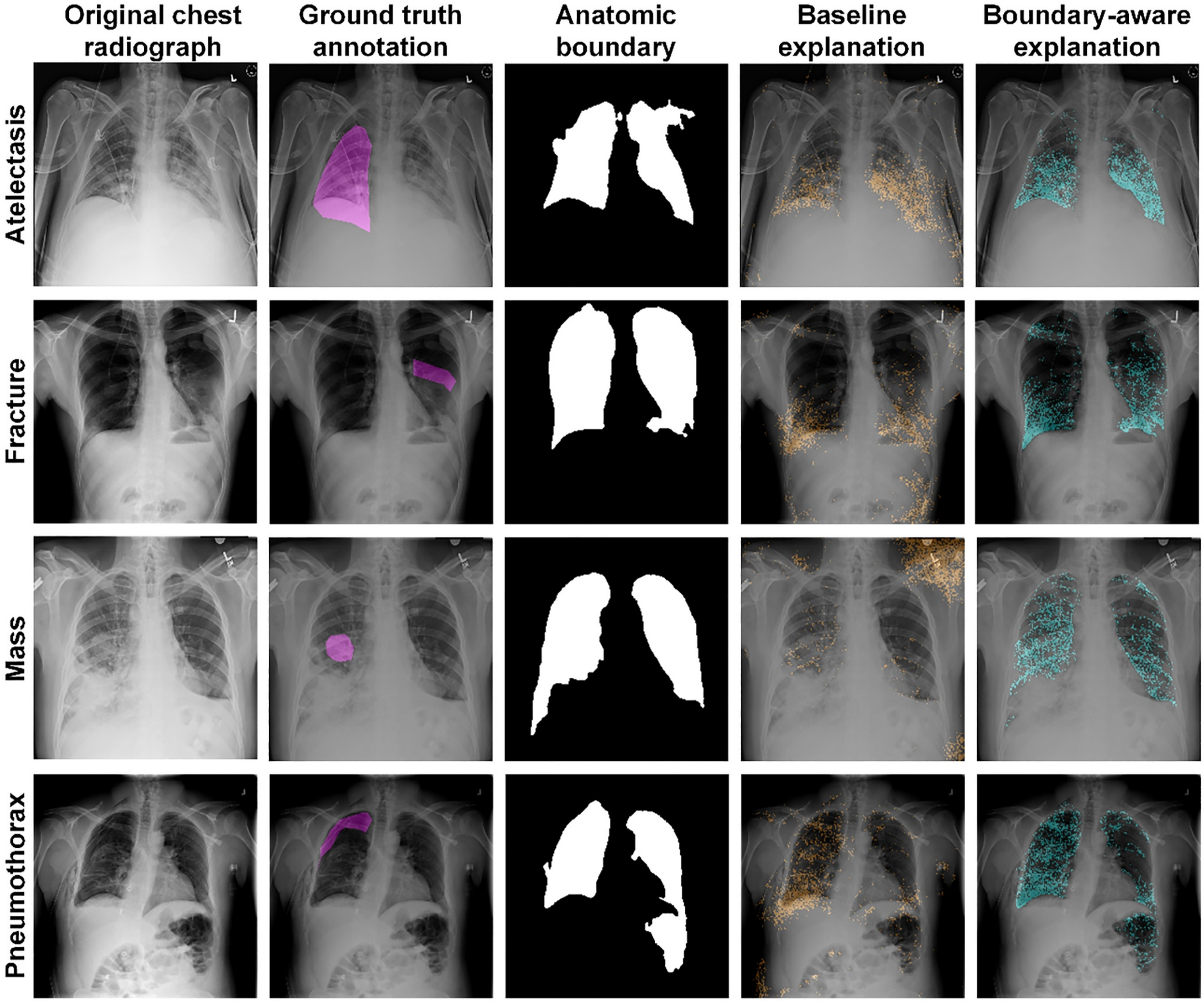

FIGURE 2 Visualization comparison of thoracopathy radiograph explained by baseline and boundary-aware Saliency Map. These figures are open-access from Deepwise AI Lab under the Apache-2.0 license, permitting use, modification, and distribution.

model before it converged on external lung segmentation datasets. These checkpoints, which varied in segmentation accuracy, enabled us to assess their effect on downstream model explanations. Table 7 reports the IoU and DSC of different checkpoints applied to VGG-11 classifiers for binary diagnosis of Atelectasis, demonstrating a positive correlation between the quality of boundary segmentation and anatomic boundary-aware XAI performance.

Lastly, we empirically compared the computational latency of our proposed method with baseline XAI approaches applied to VGG-11 classifiers for binary diagnosis of Atelectasis on an affordable NVIDIA GeForce RTX 2080 Super GPU. Our method, encompassing segmenter inference and boundary overlay, required an additional 0.058, 0.057, and 0.057 s per image, compared to 0.095, 0.128, and 0.250 s for the baseline Saliency Map, Grad-CAM, and Integrated Gradients. Given that experienced radiologists typically take approximately 34 s to interpret one chest radiograph [41], this increase was clinically acceptable and can be optimized through more advanced devices such as the NVIDIA H100 Tensor Core GPU.

4 | DISCUSSION

In this study, we evaluated the performance of three popular XAI methods and proposed a plug-and-play method using the anatomic boundary of the lung parenchyma to enhance model explanations. Under diverse combinations of CNN architectures, XAI methods, and classification settings, the proposed method consistently improved baseline model explanations while maintaining acceptable computational latency.

In our experiments, VGG-11 outperformed ResNet-18, a sophisticated architecture with more layers, in terms of

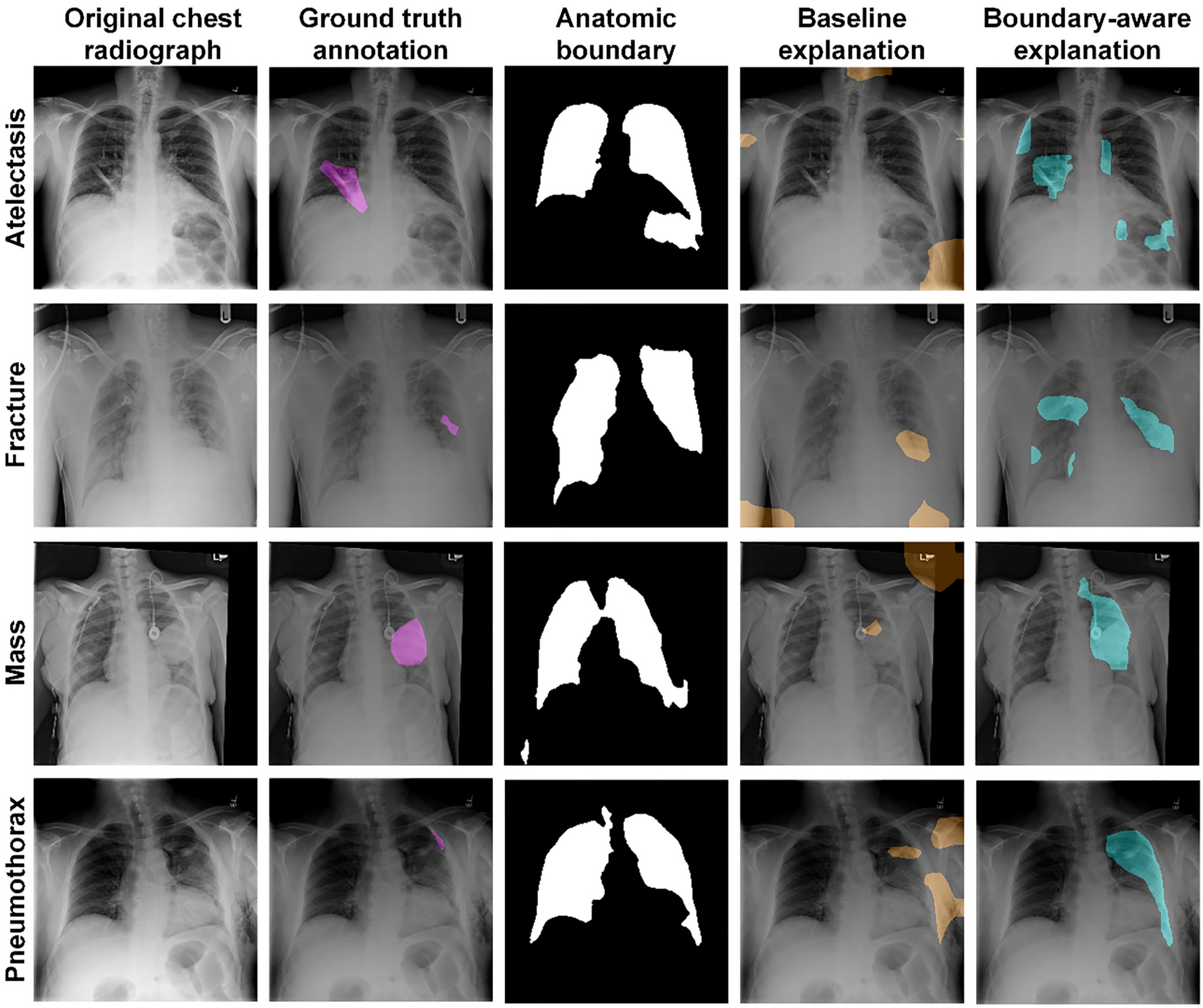

FIGURE 3 Visualization comparison of thoracopathy radiograph explained by baseline and boundary-aware Grad-CAM. These figures are open-access from Deepwise AI Lab under the Apache-2.0 license, permitting use, modification, and distribution.

AUPRC, accuracy, and precision. The phenomenon of "The deeper is not the better" has been reported in various medical applications [42, 43] and was double-verified in our experimental results on thoracopathy classification, implying that the conventional VGG-11 was well-suited for handling datasets characterized by small sample sizes. While VGG-11 demonstrated success in classification accuracy, this did not ensure its superiority over ResNet-18 in XAI explanation, which was illustrated by previous studies that deeper architectures tend to possess better interpretability [3, 44]. Furthermore, in our experimentation with VGG-11, ResNet-18, and AlexNet, we trained classifiers for both binary and multi-label classification. Across disease classification and model explanation tasks, binary classifiers showcased superior performance compared to multi-label classifiers. However, these experimental results cannot conclusively demonstrate the superiority of binary classifiers over multi-label ones. The diagnostic process in clinical settings is inherently complex, often necessitating multi-label tasks rather than simplified binary classifications [45]. A promising approach for achieving high accuracy in multi-label applications involves leveraging accurate binary classifiers. Shiraishi et al. [46] proposed a referable method that combines binary classifiers using advanced statistical techniques, including penalized logistic regression, stacking, and a sparsity-inducing penalty, to formulate solutions for multi-class classification. Lastly, the best-achieved explanation performance by anatomic boundary-aware Grad-CAM on Mass still failed to meet the regulatory standards with a minimum DSC of 20% for clinically relevant areas [47], demonstrating the existing gap between our model and real-world deployment standards from the perspective of lesion segmentations [48]. However, our primary focus was the classification of thoracic diseases, with existing XAI

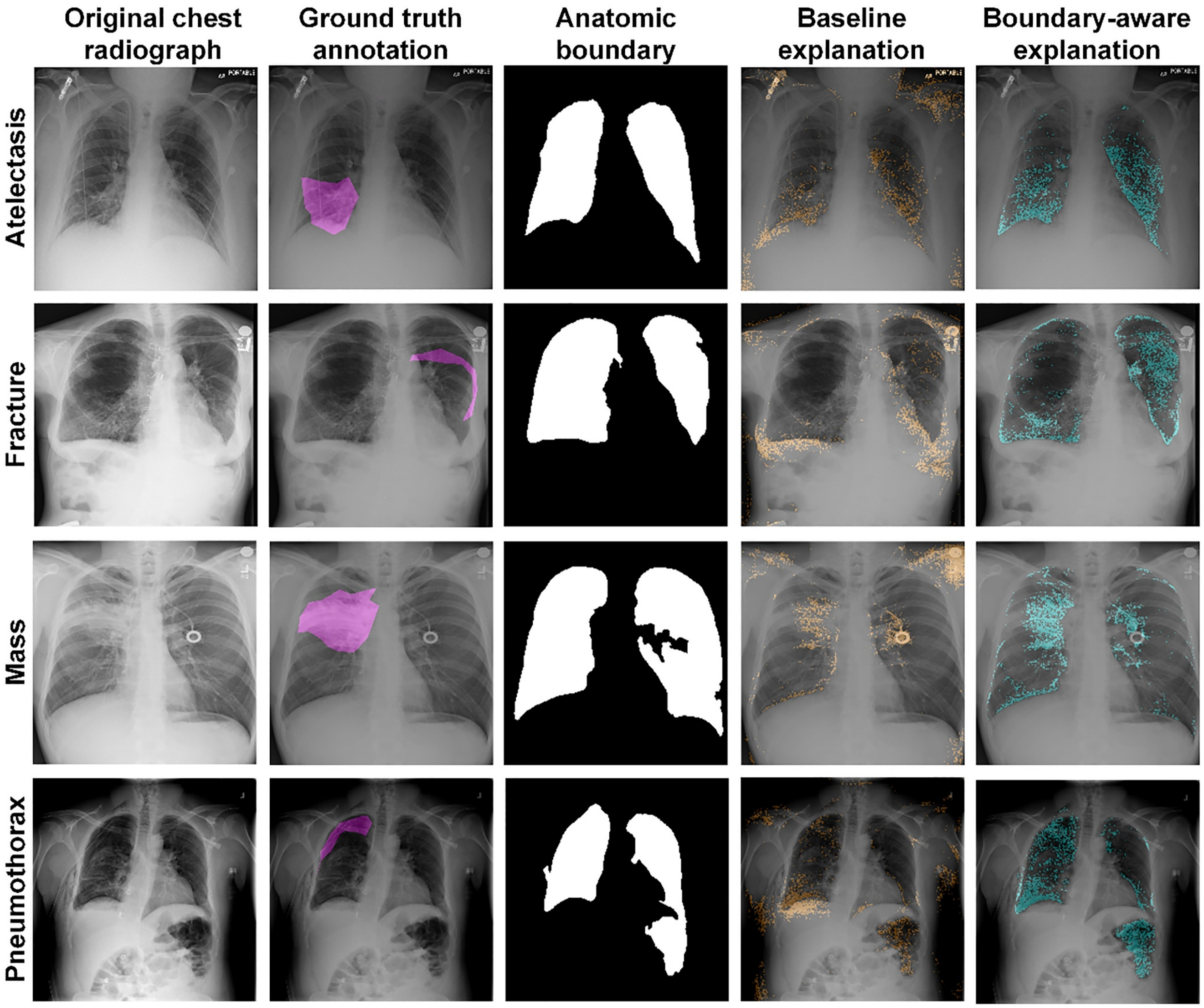

FIGURE 4 Visualization comparison of thoracopathy radiograph explained by baseline and boundary-aware Integrated Gradients. These figures are open-access from Deepwise AI Lab under the Apache-2.0 license, permitting use, modification, and distribution.

TABLE 7 Atelectasis explanation performance by various XAI methods using different segmenters on the test dataset.

	Boundary segmentation		XAI		
Segmenter	**IoU**	**DSC**	**Method**	**IoU**	**DSC**
0	93.23 (0.50)	96.43 (0.30)	Saliency map + boundary	5.89 (0.54)	10.69 (0.94)
			Grad-CAM + boundary	2.74 (0.59)	4.95 (1.05)
			Integrated gradients + boundary	4.92 (0.51)	9.05 (0.84)
1	65.19 (1.35)	78.24 (1.07)	Saliency map + boundary	5.70 (0.48)	10.40 (0.82)
			Grad-CAM + boundary	1.78 (0.48)	3.22 (0.89)
			Integrated gradients + boundary	4.42 (0.41)	8.21 (0.74)
2	42.82 (1.05)	59.37 (1.05)	Saliency map + boundary	5.59 (0.48)	10.26 (0.84)
			Grad-CAM + boundary	1.21 (0.43)	2.15 (0.77)
			Integrated gradients + boundary	3.97 (0.36)	7.46 (0.64)

Note: The evaluation metrics are presented, accompanied by their respective standard errors enclosed within parentheses.

Abbreviations: DSC, dice similarity coefficient; IoU, intersection over union; XAI, explainable artificial intelligence.

methods incorporating model explanations that can only be categorized as weakly supervised lesion segmentation, rather than fully supervised approaches. The standards for evaluating the accuracy of classification models' explanations remain unresolved [49]. Loh et al. [50] highlighted that operator-level performance, along with improved model interpretability, can lead to real-world deployment. Therefore, we suggest that classification models lacking highly accurate XAI capabilities should be considered as auxiliary tools to assist radiologists, rather than as full replacements for their expertise. Alternatively, we suggest the exploration of advanced weakly supervised frameworks to achieve robust performance in both thoracapathy classification and lesion segmentation. For example, Ouyang et al. [51] developed a classification model with advanced abnormality localization capabilities through a hierarchical attention mining framework and showed that their method achieved state-of-the-art results in both tasks.

The low IoU and DSC values of baseline model explanations revealed the issue of spurious correlations between non-pathological regions and radiological diagnoses. For example, the model explanations for Mass in Figures 2–4 and Fracture in Figures 3 and 4 focused on areas outside human bodies, such as medical devices [52] or laterality markers [53], rather than clinically relevant pathology. This phenomenon, known as shortcut learning, is not exclusive to artificial systems of chest radiograph analysis but is also prevalent in biological systems for comparative psychology and behavioral neuroscience [52]. For instance, DeGrave et al. [53] reported similar observations that CNN relied on confounding factors for COVID-19 diagnosis based on chest radiographs, stressing the need for XAI to assess undesired and unintended shortcuts in DL inference logic before real-world deployment [48]. To mitigate the problem of shortcut learning, Ahmed et al. [54] proposed to crop rectangular regions of the lungs as DL inputs to quantitatively improve COVID-19 classification accuracy and qualitatively enhance model interpretability. In contrast, our study offered several advancements: it employed pixel-level segmentation of lung regions, provided a comprehensive assessment of multiple thoracic diseases, and presented a quantitative evaluation of model explanations. While fully eliminating shortcut learning may be unattainable, efforts to mitigate it and better align learned solutions with intended outcomes should be prioritized [52], which is the primary contribution of our method.

Different from the previous applications of domain knowledge in medical image analysis [55, 56] that required additional annotation of clinical knowledge on the target dataset, our method leveraged an external lung segmenter to generate the anatomic boundary and demonstrated its effectiveness through consistent improvements in model explanations. However, we acknowledge the value of additional annotations like the delineation of the lung parenchyma by clinical experts on the target dataset. Given the prevalent domain shift indicated by the anatomic boundary in Figures 2–4, the external segmenter can be fine-tuned using experts' delineation to match the data distribution in the target dataset and offer better constraints for model explanations. Also, the proposed method relied on unified boundaries of the lung parenchyma for different thoracic diseases and future research may tailor fine-grained constraints by considering the characteristics of each thoracic disease. Beyond chest radiograph analysis, boundary-aware XAI methods can be applied across various medical imaging modalities. For example, DL models have successfully segmented organs such as the pancreas, esophagus, stomach, duodenum, liver, spleen, left kidney, and gallbladder from computed tomography images [57]. By focusing on specific organs with suspected lesions segmented by DL models, XAI methods could highlight potential regions of interest, offering radiologists valuable reference points for diagnosis.

There are several other limitations of our work. First, the explanation methods utilized in this study were confined to three explanation methods, including Saliency Map, Grad-CAM, and Integrated Gradients. The DL models were limited to 3 lightweight architectures of VGG-11, ResNet-18, and AlexNet. The segmenter was limited to a default U-Net architecture with a VGG-11 backbone. Considering the instability of XAI methods across different model backbones [58], additional XAI methods such as LayerCAM [59], DL models like Vision Transformer [60], and anatomic boundary segmenters like MedSAM [61] would offer a more comprehensive analysis. Second, our experiments revealed that the anatomic boundary consistently improved model explanations while the potential of anatomic information as a regularization or a reward in supervised learning [3, 62] or reinforcement learning [63, 64], respectively, remains unexplored, presenting a promising avenue for future research. Furthermore, this research exclusively explored 4 thoracic diseases, which could be extended to other diseases such as pleural effusion, edema, and consolidation in future studies [65]. Finally, rigorous statistical tests can be implemented to explore the association between XAI performance and geometric features of pathologies [65] and the association between XAI performance and classification performance for deeper insights into model predictive behaviors and inference logics [66, 67].

5 | CONCLUSION

The black-box nature has long hindered CNN models from gaining the trust of clinicians. In this study, we proposed an anatomic boundary-aware method for improving XAI methods for CNN in diagnostic radiology. We envisage that the consistent improvements in model explanations will inspire future endeavors to integrate clinical domain knowledge into medical image analysis.

AUTHOR CONTRIBUTIONS

Han Yuan: Conceptualization (lead), data curation (lead), formal analysis (lead), investigation (lead), methodology (lead), software (lead), validation (lead), visualization (lead), writing—original draft (lead), writing—review and editing (lead).

ACKNOWLEDGMENTS

None.

CONFLICT OF INTEREST STATEMENT

The author declares that he has no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data that supports the finding of this study is openly available in ChestX-Det at https://doi.org/10.1109/TMI.2021.3070847.

ETHICS STATEMENT

This study is exempt from review by the ethics committee because it does not involve human participants, animal subjects, or sensitive data collection.

INFORMED CONSENT

Not applicable.

ORCID

Han Yuan https://orcid.org/0000-0002-2674-6068

REFERENCES

[1] Yasaka K, Abe O. Deep learning and artificial intelligence in radiology: current applications and future directions. PLoS Med. 2018;15(11):e1002707. https://doi.org/10.1371/journal.pmed.1002707

[2] Xie F, Yuan H, Ning Y, Ong MEH, Feng M, Hsu W, et al. Deep learning for temporal data representation in electronic health records: a systematic review of challenges and methodologies. J Biomed Inf. 2022;126:103980. https://doi.org/10.1016/j.jbi.2021.103980

[3] Zhou B, Bau D, Oliva A, Torralba A. Interpreting deep visual representations *via* network dissection. IEEE Trans Pattern Anal Mach Intell. 2019;41(9):2131–45. https://doi.org/10.1109/TPAMI.2018.2858759

[4] Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: visualising image classification models and saliency maps. Proc Int Conf Learn Representations. 2013. https://openreview.net/forum?id=cO4ycnpqxKcS9

[5] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks *via* gradient-based localization. Int J Comput Vis. 2020;128(2):336–59. https://doi.org/10.1007/s11263-019-01228-7

[6] Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. Proc Int Conf Machine Learn. 2017. https://proceedings.mlr.press/v70/sundararajan17a.html

[7] Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. 2016 IEEE Conf Comput Vision and Pattern Recogn (CVPR). 2016:2921–9. Las Vegas, NV, USA. https://doi.org/10.1109/CVPR.2016.319

[8] Yuan H, Jiang P-T, Zhao G. Human-guided design to explain deep learning-based pneumothorax classifier. Proc Med Imag with Deep Learn. 2023. https://openreview.net/pdf?id=_kk8KI8MiRE

[9] Leo J. Clinical anatomy and embryology. Luxembourg: Springer Nature; 2023.

[10] Harford P, Tran L, Pollock D, Thiruvenkatarajan V, Munn Z. Effectiveness of erector spinae plane block for rib fracture analgesia: a systematic review protocol. JBI Evidence Synthesis. 2024;22(4):706–12. https://doi.org/10.11124/JBIES-23-00168

[11] Hong W, Hwang EJ, Lee JH, Park J, Goo JM, Park CM. Deep learning for detecting pneumothorax on chest radiographs after needle biopsy: clinical implementation. Radiology. 2022;303(2):433–41. https://doi.org/10.1148/radiol.211706

[12] Park IH, Kim CW, Choi YU, Kang TW, Lim J, Byun CS. Occult pneumothorax in patients with blunt chest trauma: key findings on supine chest radiography. J Thorac Dis. 2023;15(8):4379–86. https://doi.org/10.21037/jtd-23-541

[13] Chen B, Li J, Guo X, Lu G. DualCheXNet: dual asymmetric feature learning for thoracic disease classification in chest X-rays. Biomed Signal Process Control. 2019;53:101554. https://doi.org/10.1016/j.bspc.2019.04.031

[14] Huang G, Liu Z, Pleiss G, Maaten LV, Weinberger KQ. Convolutional networks with dense connectivity. IEEE Trans Pattern Anal Mach Intell. 2022;44(12):8704–16. https://doi.org/10.1109/TPAMI.2019.2918284

[15] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2016 IEEE Conf Comput Vision and Pattern Recogn (CVPR). 2016:770–8. Las Vegas, NV, USA. https://doi.org/10.1109/CVPR.2016.90

[16] Yuan H, Hong C, Jiang PT, Zhao G, Tran NTA, Xu X, et al. Clinical domain knowledge-derived template improves post hoc AI explanations in pneumothorax classification. J Biomed Inf. 2024;156:104673. https://doi.org/10.1016/j.jbi.2024.104673

[17] Liu H, Wang L, Nan Y, Jin F, Wang Q, Pu J. SDFN: segmentation-based deep fusion network for thoracic disease classification in chest X-ray images. Comput Med Imag Graph. 2019;75:66–73. https://doi.org/10.1016/j.compmedimag.2019.05.005

[18] Zaidi SZY, Akram MU, Jameel A, Alghamdi NS. Lung segmentation-based pulmonary disease classification using deep neural networks. IEEE Access. 2021;9:125202–14. https://doi.org/10.1109/ACCESS.2021.3110904

[19] Park B, Park H, Lee SM, Seo JB, Kim N. Lung segmentation on HRCT and volumetric CT for diffuse interstitial lung disease using deep convolutional neural networks. J Digit Imag. 2019;32(6):1019–26. https://doi.org/10.1007/s10278-019-00254-8

[20] Yuan H, Hong C, Tran NTA, Xu X, Liu N. Leveraging anatomical constraints with uncertainty for pneumothorax segmentation. Health Care Sci. 2024;3(6):1–19. https://doi.org/10.1002/hcs2.119

[21] Yuan H, Kang L, Li Y, Fan Z. Human-in-the-Loop machine learning for healthcare: current progress and future opportunities in electronic health records. Med Adv. 2024;2(3):318–22. https://doi.org/10.1002/med4.70

[22] Jung H-G, Nam W-J, Kim H-W, Lee S-W. Weakly supervised thoracic disease localization *via* disease masks. Neurocomputing. 2023;517:34–43. https://doi.org/10.1016/j.neucom.2022.10.019

[23] Wang H, Jia H, Lu L, Xia Y. *Thorax*-net: an attention regularized deep neural network for classification of thoracic diseases on chest radiography. IEEE J Biomed and Health Informatics. 2020;24(2):475–85. https://doi.org/10.1109/JBHI.2019.2928369

[24] Xie X, Niu J, Liu X, Chen Z, Tang S, Yu S. A survey on incorporating domain knowledge into deep learning for medical image analysis. Med Image Anal. 2021;69:101985. https://doi.org/10.1016/j.media.2021.101985

[25] Sirocchi C, Bogliolo A, Montagna S. Medical-informed machine learning: integrating prior knowledge into medical decision systems. BMC Med Inf Decis Making. 2024;24(Suppl 4):186. https://doi.org/10.1186/s12911-024-02582-4

[26] Yuan H. Clinical decision making: evolving from hypothetico-deductive model to knowledge-enhanced machine learning. Med Adv. 2024;2(4):1–5. https://doi.org/10.1002/med4.83

[27] Lian J, Liu J, Zhang S, Gao K, Liu X, Zhang D, et al. A structure-aware relation network for thoracic diseases detection and segmentation. IEEE Trans Med Imag. 2021;40(8):2042–52. https://doi.org/10.1109/TMI.2021.3070847

[28] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. Proc Int Conf Learn Representations. 2014. https://arxiv.org/abs/1409.1556

[29] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM. 2017;60(6):84–90. https://doi.org/10.1145/3065386

[30] Agliari E, Alemanno F, Barra A, De Marzo G. The emergence of a concept in shallow neural networks. Neural Network. 2022; 148:232–53. https://doi.org/10.1016/j.neunet.2022.01.017

[31] Liao W, Zou B, Zhao R, Chen Y, He Z, Zhou M. Clinical interpretable deep learning model for glaucoma diagnosis. IEEE J Biomed and Health Informatics. 2020;24(5):1405–12. https://doi.org/10.1109/JBHI.2019.2949075

[32] Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature. 1986;323(6088): 533–6. https://doi.org/10.1038/323533a0

[33] Yuan H, Xie F, Ong MEH, Ning Y, Chee ML, Saffari SE, et al. AutoScore-imbalance: an interpretable machine learning tool for development of clinical scores with rare events data. J Biomed Inf. 2022;129:104072. https://doi.org/10.1016/j.jbi.2022.104072

[34] Zhao Y-X, Yuan H, Wu Y. Prediction of adverse drug reaction using machine learning and deep learning based on an imbalanced electronic medical records dataset. Proc 5th Int Conf Med Health Informatics. 2021:17–21. Kyoto, Japan. https://doi.org/10.1145/3472813.3472817

[35] Fu G-H, Yi L-Z, Pan J. Tuning model parameters in class-imbalanced learning with precision-recall curve. Biometrical J Biometrische Zeitschrift. 2019;61(3):652–64. https://doi.org/10.1002/bimj.201800148

[36] Efron B. Better bootstrap confidence intervals. J Am Stat Assoc. 1987;82(397):171–85. https://doi.org/10.2307/2289144

[37] Amorim JP, Abreu PH, Santos J, Cortes M, Vila V. Evaluating the faithfulness of saliency maps in explaining deep learning models using realistic perturbations. Inf Process Manag. 2023; 60(2):103225. https://doi.org/10.1016/j.ipm.2022.103225

[38] Shiraishi J, Katsuragawa S, Ikezoe J, Matsumoto T, Kobayashi T, Komatsu K, et al. Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. AJR Am J Roentgenol. 2000;174(1):71–4. https://doi.org/10.2214/ajr.174.1.1740071

[39] Jaeger S, Candemir S, Antani S, Wáng YJ, Lu P-X, Thoma G. Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. Quant Imag Med Surg. 2014;4(6):475–7. https://doi.org/10.3978/j.issn.2223-4292.2014.11.20

[40] Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention – MICCAI 2015. Cham. Springer International Publishing; 2015. p. 234–41.

[41] Rajpurkar P, Irvin J, Ball RL, Zhu K, Yang B, Mehta H, et al. Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS Med. 2018;15(11):e1002686. https://doi.org/10.1371/journal.pmed.1002686

[42] Santos-Bustos DF, Nguyen BM, Espitia HE. Towards automated eye cancer classification *via* VGG and ResNet networks using transfer learning. Eng Sci Technol an Int J. 2022;35:101214. https://doi.org/10.1016/j.jestch.2022.101214

[43] Victor Ikechukwu A, Murali S, Deepu R, Shivamurthy RC. ResNet-50 vs vgg-19 vs training from scratch: a comparative analysis of the segmentation and classification of pneumonia from chest X-ray images. Glob Transitions Proc. 2021;2(2): 375–81. https://doi.org/10.1016/j.gltp.2021.08.027

[44] Bau D, Zhou B, Khosla A, Oliva A, Torralba A. Network dissection: quantifying interpretability of deep visual representations. 2017 IEEE Conf Comput Vision and Pattern Recognition (CVPR), Honolulu, HI, USA. 2017:3319–27. https://doi.org/10.1109/CVPR.2017.354

[45] Chen H, Miao S, Xu D, Hager GD, Harrison AP. Deep hiearchical multi-label classification applied to chest X-ray abnormality taxonomies. Med Image Anal. 2020;66:101811. https://doi.org/10.1016/j.media.2020.101811

[46] Shiraishi Y, Fukumizu K. Statistical approaches to combining binary classifiers for multi-class classification. Neurocomputing. 2011;74(5):680–8. https://doi.org/10.1016/j.neucom.2010.09.004

[47] Lee SY, Ha S, Jeon MG, Li H, Choi H, Kim HP, et al. Localization-adjusted diagnostic performance and assistance

effect of a computer-aided detection system for pneumothorax and consolidation. NPJ Digit Med. 2022;5(1):107. https://doi.org/10.1038/s41746-022-00658-x

[48] Yuan H. Toward real-world deployment of machine learning for health care: external validation, continual monitoring, and randomized clinical trials. Health Care Sci. 2024;3(5):360–4. https://doi.org/10.1002/hcs2.114

[49] de Vries BM, Zwezerijnen GJC, Burchell GL, van Velden FHP, Menke-van der Houven van Oordt CW, Boellaard R. Explainable artificial intelligence (XAI) in radiology and nuclear medicine: a literature review. Front Med. 2023;10:1180773. https://doi.org/10.3389/fmed.2023.1180773

[50] Loh HW, Ooi CP, Seoni S, Barua PD, Molinari F, Acharya UR. Application of explainable artificial intelligence for healthcare: a systematic review of the last decade (2011–2022). Comput Methods Progr Biomed. 2022;226:107161. https://doi.org/10.1016/j.cmpb.2022.107161

[51] Ouyang X, Karanam S, Wu Z, Chen T, Huo J, Zhou XS, et al. Learning hierarchical attention for weakly-supervised chest X-ray abnormality localization and diagnosis. IEEE Trans Med Imag. 2021;40(10):2698–710. https://doi.org/10.1109/TMI.2020.3042773

[52] Geirhos R, Jacobsen J-H, Michaelis C, Zemel R, Brendel W, Bethge M, et al. Shortcut learning in deep neural networks. Nat Mach Intell. 2020;2(11):665–73. https://doi.org/10.1038/s42256-020-00257-z

[53] DeGrave AJ, Janizek JD, Lee S-I. AI for radiographic COVID-19 detection selects shortcuts over signal. Nat Mach Intell. 2021;3(7):610–9. https://doi.org/10.1038/s42256-021-00338-7

[54] Ben Ahmed K, Hall LO, Goldgof DB, Fogarty R. Achieving multisite generalization for CNN-based disease diagnosis models by mitigating shortcut learning. IEEE Access. 2022;10:78726–38. https://doi.org/10.1109/ACCESS.2022.3193700

[55] Bateson M, Dolz J, Kervadec H, Lombaert H, Ben Ayed I. Constrained domain adaptation for image segmentation. IEEE Trans Med Imag. 2021;40(7):1875–87. https://doi.org/10.1109/TMI.2021.3067688

[56] Kervadec H, Dolz J, Tang M, Granger E, Boykov Y, Ben Ayed I. Constrained-CNN losses for weakly supervised segmentation. Med Image Anal. 2019;54:88–99. https://doi.org/10.1016/j.media.2019.02.009

[57] Gibson E, Giganti F, Hu Y, Bonmati E, Bandula S, Gurusamy K, et al. Automatic multi-organ segmentation on abdominal CT with dense V-networks. IEEE Trans Med Imag. 2018;37(8):1822–34. https://doi.org/10.1109/tmi.2018.2806309

[58] Yuan H, Liu M, Kang L, Miao C, Wu Y. An empirical study of the effect of background data size on the stability of SHapley Additive exPlanations (SHAP) for deep learning models. Proc Int Conf Learn Representations. 2023. https://openreview.net/forum?id=L38bbHmRKx

[59] Jiang P-T, Zhang C-B, Hou Q, Cheng M-M, Wei Y. Layer-CAM: exploring hierarchical class activation maps for localization. IEEE Trans Image Process. 2021;30:5875–88. https://doi.org/10.1109/TIP.2021.3089943

[60] Han K, Wang Y, Chen H, Chen X, Guo J, Liu Z, et al. A survey on vision transformer. IEEE Trans Pattern Anal Mach Intell. 2022;45(1):87–110. https://doi.org/10.1109/TPAMI.2022.3152247

[61] Ma J, He Y, Li F, Han L, You C, Wang B. Segment anything in medical images. Nat Commun. 2024;15(1):654. https://doi.org/10.1038/s41467-024-44824-z

[62] Xue Z, Jaeger S, Antani SK, Long R, Karagyris A, Siegelman J, et al. Localizing tuberculosis in chest radiographs with deep learning. SPIE Med Imag. 2018;10579:28. https://doi.org/10.1117/12.2293022

[63] Kang L, Liu Y, Luo Y, Yang JZ, Yuan H, Zhu C. Approximate policy iteration with deep minimax average bellman error minimization. IEEE Trans Neural Netw Learn Syst. 2024;1–12, https://doi.org/10.1109/TNNLS.2023.3346992

[64] Zhou SK, Le HN, Luu K, V Nguyen H, Ayache N. Deep reinforcement learning in medical imaging: a literature review. Med Image Anal. 2021;73:102193. https://doi.org/10.1016/j.media.2021.102193

[65] Saporta A, Gui X, Agrawal A, Pareek A, Truong SQH, Nguyen CDT, et al. Benchmarking saliency methods for chest X-ray interpretation. Nat Mach Intell. 2022;4(10):867–78. https://doi.org/10.1038/s42256-022-00536-x

[66] Yuan H, Yu K, Xie F, Liu M, Sun S. Automated machine learning with interpretation: a systematic review of methodologies and applications in healthcare. Med Adv. 2024;2(3):205–37. https://doi.org/10.1002/med4.75

[67] Allgaier J, Mulansky L, Draelos RL, Pryss R. How does the model make predictions? A systematic literature review on the explainability power of machine learning in healthcare. Artif Intell Med. 2023;143:102616. https://doi.org/10.1016/j.artmed.2023.102616

How to cite this article: Yuan H. Anatomic boundary-aware explanation for convolutional neural networks in diagnostic radiology. iRADIOLOGY. 2024;1–16. https://doi.org/10.1002/ird3.113