Anatomic Boundary-aware Explanation for Convolutional Neural Network in Diagnostic Radiology

Han Yuan[1] YUAN.HAN@U.DUKE.NUS.EDU
Lican Kang[1] KANGLICAN@WHU.EDU.CN
[1] *Duke-NUS Medical School, National University of Singapore*

Abstract

Convolutional neural network (CNN) has achieved remarkable success not only in general computer vision tasks but also in medical image analysis. However, unlike some general 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 on its inference logic, clinicians can evaluate and correct the intervention suggestions provided by CNN. Therefore, various explanation methods have augmented CNN decisions with interpretative explanations. Nevertheless, prior approaches treated medical image tasks akin to general computer vision tasks, adhering to the end-to-end paradigm while overlooking crucial medical domain knowledge that has the potential to enhance the quality of model explanations.

In this study, we propose a plug-and-play module that explicitly integrates anatomic boundary information into the explanation process for CNN-based thoracopathy classification. To generate the anatomic boundary of the lung area, we utilize a lung segmentation model developed on public datasets and deploy it on the target dataset for thoracopathy classification. Assessed by Intersection over Union (IoU) and Dice Similarity Coefficient (DSC) between model-extracted explanations and expert-annotated lesion areas, our method consistently outperformed the baseline devoid of domain knowledge in 47 out of 48 scenarios, encompassing 2 CNN architectures (VGG-11 and ResNet-18), 2 classification settings (binary and multi-label), 3 explanation methods (Saliency Map, Grad-CAM, and Integrated Gradients), and 4 co-occured thoracic diseases (Atelectasis, Fracture, Mass, and Pneumothorax).

The studies underscored the effectiveness of embedding domain knowledge in improving CNN explanations and we hope it could inspire future efforts in integrating clinical domain knowledge into medical image analysis. The code and dataset have been released at GitHub[1] for reproducibility.

Keywords: Diagnostic Radiology, Explainable Artificial Intelligence, Medical Domain Knowledge

1. https://github.com/Han-Yuan-Med/constrained-explanation

1. Introduction

In the last decade, the convolutional neural network (CNN) has reshaped the diagnostic process of thoracopathy (Yasaka et al., 2018). Although CNN features high-fidelity accuracy on various retrospective tasks, clinicians still cannot fully trust their decisions on prospective medical practice due to their black-box characteristics and the high stakes affecting patients' lives (Xie et al., 2022; Liu et al., 2023). Based on model explanations on its inference logic, clinicians can evaluate and correct the intervention suggestions provided by CNN (Van der Velden et al., 2022). Therefore, various explainable artificial intelligence (XAI) methods have been proposed to map the final diagnostic decision back onto the input images by highlighting the important pixels (regions) toward decisions made by CNN models. Such a heatmap will be provided to clinicians to conduct further evaluation on whether the model focuses on the clinically relevant region such as lesion area (Yuan et al., 2023a). To generate the model focus area, researchers have proposed a spectrum of methods such as Saliency Map (Simonyan et al., 2014), Grad-CAM (Selvaraju et al., 2017), and Integrated Gradients (Sundararajan et al., 2017).

Thoracopathy comprises conditions of the heart, lungs, mediastinum, esophagus, chest wall, diaphragm, and great vessels, especially these diseases of the lungs (Baker et al., 2022). In this study, we focus on four co-occurring thoracic diseases: Atelectasis, Fracture, Mass, and Pneumothorax (Harford et al., 2023; Hong et al., 2022; Park et al., 2023). Conventional clinical diagnosis is based on the clinicians manual evaluation of chest radiographs while multiple CNN-based models are recently proposed to automate this process. For example, Chen et al. used two asymmetric CNNs of DenseNet and ResNet 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 the classification and explanation stages and neglected the explicit domain knowledge that thoracic diseases mainly occur in the lung area. Several prior studies have noticed this gap and presented the enhancement of accuracy via the embedding of domain knowledge (Wang et al., 2019; Jung et al., 2023).

To explore the domain knowledge in the explanation of CNN-based thoracopathy classifiers, we present a plug-and-play module to constrain the model explanations within the lung area. To obtain the lung area, we transfer an external lung segmentation model trained on public datasets. Such a transferred model can efficiently alleviate the additional annotation costs on the target dataset for thoracopathy classification. Assessed on 2 CNN architectures, 3 XAI methods, 4 thoracic diseases, and 2 classification settings, the proposed approach consistently outperformed the baseline devoid of domain knowledge. The studies underscored the effectiveness of embedding domain knowledge in improving CNN explanations and we hope it could inspire future efforts in integrating clinical domain knowledge into medical image analysis.

2. Experiments

2.1. Dataset

We conducted a comprehensive investigation into the association between XAI-based CNN explanations and lesion areas annotated by human experts by utilizing the public dataset

Table 1: An overview of the data split in thoracopathy classification tasks

Thoracopathy				Training set	Validation set	Test set
Atelectasis	Fracture	Mass	Pneumothorax			
✓	✓	✓	✓	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

of ChestX-Det. We extracted 611 healthy samples and 880 samples were diagnosed with at least one of the four-correlated thoracic diseases. 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, ChestX-Det was enriched with pixel-level lesion annotations for each thoracic disease. We randomly split the extracted samples into training, validation, and test datasets at 60: 20: 20 as indicated in Table 1. 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. Two classification scenarios were explored for each disease. The first involved training a multi-label classifier for all 4 diseases, leveraging the entire dataset. The second scenario entailed training individual binary classifiers for each disease, using solely healthy samples and diseased samples specific to the targeted disease.

2.2. CNN-based thoracopathy classification

Considering the middle-scale characteristic of the used dataset, we developed thoracopathy classifiers using three lightweight CNN backbones: AlexNet (Krizhevsky et al., 2012), VGG (VGG-11) (Simonyan et al., 2015), and ResNet (ResNet-18) (He et al., 2016). For model training, we employed Stochastic Gradient Descent (SGD) (Rumelhart et al., 1986) 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 classifier and multi-label classifier to eliminate the impact of dominating classes (Yuan et al., 2022). Each DL 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 performance instead of the Area Under

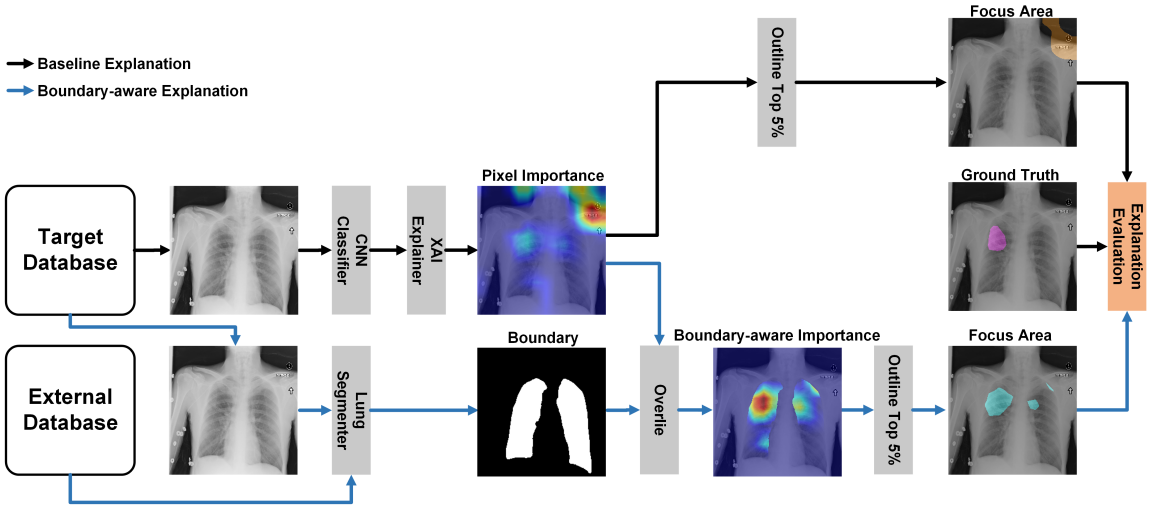

Figure 1: Schematic diagram of the proposed boundary-aware explanation framework.

the Receiver Operating Characteristic Curve (AUROC) (Fu et al., 2019). We also included accuracy, precision, and recall for a comprehensive evaluation. Besides, the standard error (SE) was calculated based on the bootstrapping of samples in the test dataset (Efron, 1987).

2.3. XAI for CNN-based thoracopathy classifiers

To explain the model decision logic behind the trained classifiers, various XAI methods were applied to calculate each pixel's importance towards the model's final decision and we further outlined the focus area consisting of the most significant pixels by selecting the top 5% important pixels. In this study, we utilize three mainstream explanation techniques in the XAI family: Saliency Map (Simonyan et al., 2014), Grad-CAM (Selvaraju et al., 2017), and Integrated Gradients (Sundararajan et al., 2017) to generate the pixel-level importance. The top 5% important pixels are extracted to formulate the model focus area. After that, we compare the model focus area with the concrete lesion area A_i^{test} to calculate Intersection over Union (IoU) and Dice Similarity Coefficient (DSC). For reproduction, the code and dataset are publicly available at GitHub[1].

2.4. Anatomic boundary-aware model explanation

According to the clinical knowledge (Jung et al., 2023), thoracic diseases occur in a potential space within the lung area on a 2D projection of a chest radiograph, and therefore, the model explanations should be constrained within the anatomic boundary of the lung area. As outlined in Figure 1, the proposed method develops an auxiliary lung segmenter using three public lung segmentation datasets, including Japanese Society of Radiological Technology dataset (JSRT) (Shiraishi et al., 2000), Shenzhen dataset (Shenzhen) (Jaeger et al., 2014), and Montgomery County dataset (MC) (Jaeger et al., 2014) and then deploys the lung segmenter on the baseline explanation to constrain the selection of focus area within the predicted lung space.

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)
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)
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)
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)

3. Results

First, we quantitatively demonstrated the performance of various CNN architectures in different classification settings in Table 2. Due to data imbalance, AUPRC instead of AUROC was utilized as the primary evaluation metric for model classification performance (Fu et al., 2019). Analyzing the results based on CNN architectures, VGG-11 consistently outperformed ResNet-18 across scenarios concerning AUPRC, accuracy, and in most scenarios, precision. However, its average recall of 0.619 was lower than the 0.676 achieved by ResNet-18. Additionally, we explored different settings of binary and multi-label classification. Binary classifiers consistently showcased superior performance across all four disease classifications in terms of AUPRC and precision compared to multi-label classifiers. Also, binary classifiers achieved an average accuracy of 0.740, outperforming 0.627 by multi-label classifiers. Nonetheless, multi-label classifiers demonstrated an average recall of 0.656, marginally better than the 0.640 achieved by binary classifiers.

Table 3 presents the explanation performance of VGG-11 and ResNet-18 utilizing different XAI methods for thoracopathy classification under binary or multi-label settings. In Atelectasis and Pneumothorax classifications, binary VGG-11 employing boundary-aware Saliency Map achieved the best explanation performance, whereas in Fracture and Mass, binary ResNet-18 utilizing boundary-aware Grad-CAM demonstrated superior explanation performance. Also, anatomic boundary-aware explanations consistently exhibited improvements in IoU and DSC compared to baseline explanations, except for Grad-CAM explanation of VGG-11 in multi-label fracture classification. Additionally, binary classifiers outperformed multi-label classifiers in 17 out of 24 scenarios for baseline explanations and in 18 out of 24 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 explanation performance. Particularly in

Fracture and Pneumothorax, only 2 and 3 classifier-based explanations achieved better IoU and DSC compared to the classifier-free boundary.

Table 3: Thoracopathy explanation performance of CNN models by various XAI methods on the test dataset. Top 5% activated pixels were selected as the model focus area.

Disease	Model	Setting	XAI	IoU	DSC
ATE	VGG-11	Binary	Saliency Map	5.69 (0.51)	10.44 (0.92)
			Saliency Map + B*	5.89 (0.54)	10.69 (0.94)
		Multi-label	Saliency Map	1.92 (0.38)	3.61 (0.66)
			Saliency Map + B	3.84 (0.51)	7.09 (0.89)
		Binary	Grad-CAM	1.63 (0.48)	2.93 (0.82)
			Grad-CAM + B	2.74 (0.59)	4.95 (1.05)
		Multi-label	Grad-CAM	1.49 (0.38)	2.74 (0.69)
			Grad-CAM + B	1.76 (0.64)	3.17 (1.07)
		Binary	Integrated Gradients	3.81 (0.36)	7.18 (0.64)
			Integrated Gradients + B	4.92 (0.51)	9.05 (0.84)
		Multi-label	Integrated Gradients	0.99 (0.26)	1.91 (0.48)
			Integrated Gradients + B	3.29 (0.51)	6.09 (0.89)
	ResNet-18	Binary	Saliency Map	2.28 (0.33)	4.36 (0.61)
			Saliency Map + B	3.56 (0.46)	6.59 (0.77)
		Multi-label	Saliency Map	3.80 (0.46)	7.07 (0.82)
			Saliency Map + B	4.56 (0.56)	8.29 (0.94)
		Binary	Grad-CAM	4.12 (0.77)	7.41 (1.35)
			Grad-CAM + B	4.42 (0.71)	7.91 (1.22)
		Multi-label	Grad-CAM	3.77 (1.07)	6.31 (1.63)
			Grad-CAM + B	5.55 (0.99)	9.73 (1.66)
		Binary	Integrated Gradients	2.82 (0.36)	5.32 (0.64)
			Integrated Gradients + B	3.56 (0.46)	6.61 (0.79)
		Multi-label	Integrated Gradients	4.08 (0.54)	7.51 (0.89)
			Integrated Gradients + B	4.67 (0.56)	8.46 (0.97)
	/		B	3.80 (0.66)	6.89 (1.10)
FRA	VGG-11	Binary	Saliency Map	0.40 (0.05)	0.79 (0.10)
			Saliency Map + B	0.94 (0.15)	1.85 (0.26)
		Multi-label	Saliency Map	0.29 (0.08)	0.58 (0.20)
			Saliency Map + B	0.86 (0.18)	1.67 (0.33)
		Binary	Grad-CAM	0.38 (0.15)	0.73 (0.31)
			Grad-CAM + B	1.40 (0.48)	2.43 (0.79)
		Multi-label	Grad-CAM	0.51 (0.15)	0.98 (0.28)
			Grad-CAM + B	0.46 (0.13)	0.89 (0.23)
		Binary	Integrated Gradients	0.28 (0.05)	0.56 (0.10)
			Integrated Gradients + B	0.85 (0.15)	1.66 (0.28)
		Multi-label	Integrated Gradients	0.15 (0.05)	0.30 (0.10)
			Integrated Gradients + B	0.93 (0.20)	1.80 (0.41)
	ResNet-18	Binary	Saliency Map	0.77 (0.10)	1.51 (0.23)
			Saliency Map + B	1.62 (0.23)	3.10 (0.41)

		Multi-label	Saliency Map	0.58 (0.10)	1.15 (0.15)
			Saliency Map + B	1.29 (0.18)	2.51 (0.36)
		Binary	Grad-CAM	1.52 (0.33)	2.82 (0.59)
			Grad-CAM + B	2.27 (0.46)	4.19 (0.84)
		Multi-label	Grad-CAM	0.35 (0.15)	0.66 (0.28)
			Grad-CAM + B	1.18 (0.31)	2.20 (0.56)
		Binary	Integrated Gradients	0.54 (0.08)	1.06 (0.15)
			Integrated Gradients + B	1.24 (0.20)	2.38 (0.38)
		Multi-label	Integrated Gradients	0.37 (0.05)	0.73 (0.13)
			Integrated Gradients + B	0.96 (0.18)	1.87 (0.36)
		/	B	1.42 (0.18)	2.75 (0.33)
MAS	VGG-11	Binary	Saliency Map	1.46 (0.48)	2.75 (0.87)
			Saliency Map + B	3.64 (0.79)	6.76 (1.38)
		Multi-label	Saliency Map	1.65 (0.59)	3.05 (1.10)
			Saliency Map + B	4.23 (1.07)	7.60 (1.79)
		Binary	Grad-CAM	0.62 (0.41)	1.12 (0.71)
			Grad-CAM + B	5.75 (1.94)	9.20 (2.83)
		Multi-label	Grad-CAM	1.10 (0.74)	1.85 (1.20)
			Grad-CAM + B	4.28 (1.58)	6.89 (2.30)
		Binary	Integrated Gradients	1.79 (0.43)	3.42 (0.84)
			Integrated Gradients + B	5.26 (0.99)	9.51 (1.66)
		Multi-label	Integrated Gradients	0.45 (0.15)	0.88 (0.31)
			Integrated Gradients + B	4.26 (0.82)	7.86 (1.45)
	ResNet-18	Binary	Saliency Map	3.99 (0.56)	7.47 (0.99)
			Saliency Map + B	6.30 (0.94)	11.35 (1.68)
		Multi-label	Saliency Map	3.68 (0.48)	6.97 (0.89)
			Saliency Map + B	6.32 (1.07)	11.37 (1.81)
		Binary	Grad-CAM	6.06 (1.86)	10.11 (2.98)
			Grad-CAM + B	9.36 (2.40)	15.08 (3.78)
		Multi-label	Grad-CAM	3.27 (1.05)	5.76 (1.79)
			Grad-CAM + B	5.71 (1.66)	9.73 (2.70)
		Binary	Integrated Gradients	5.21 (0.77)	9.55 (1.30)
			Integrated Gradients + B	7.48 (1.15)	13.29 (1.89)
		Multi-label	Integrated Gradients	3.99 (0.64)	7.44 (1.15)
			Integrated Gradients + B	7.21 (1.07)	12.81 (1.86)
		/	B	5.51 (0.74)	10.07 (1.30)
PTX	VGG-11	Binary	Saliency Map	1.71 (0.28)	3.30 (0.51)
			Saliency Map + B	2.65 (0.31)	5.09 (0.56)
		Multi-label	Saliency Map	0.46 (0.15)	0.89 (0.28)
			Saliency Map + B	1.89 (0.38)	3.59 (0.71)
		Binary	Grad-CAM	1.17 (0.43)	2.14 (0.77)
			Grad-CAM + B	1.59 (0.46)	2.97 (0.82)
		Multi-label	Grad-CAM	0.37 (0.18)	0.72 (0.36)
			Grad-CAM + B	1.82 (0.82)	3.18 (1.33)

ResNet-18	Binary	Integrated Gradients	1.34 (0.41)	2.56 (0.74)
		Integrated Gradients + B	2.64 (0.54)	4.98 (0.97)
	Multi-label	Integrated Gradients	0.22 (0.13)	0.44 (0.23)
		Integrated Gradients + B	2.02 (0.41)	3.86 (0.77)
	Binary	Saliency Map	0.45 (0.10)	0.89 (0.20)
		Saliency Map + B	1.61 (0.20)	3.12 (0.41)
	Multi-label	Saliency Map	0.51 (0.13)	1.02 (0.23)
		Saliency Map + B	1.32 (0.20)	2.58 (0.43)
	Binary	Grad-CAM	1.45 (0.71)	2.63 (1.25)
		Grad-CAM + B	1.81 (0.74)	3.32 (1.35)
	Multi-label	Grad-CAM	0.69 (0.36)	1.28 (0.66)
		Grad-CAM + B	1.25 (0.56)	2.25 (0.99)
	Binary	Integrated Gradients	0.68 (0.20)	1.32 (0.41)
		Integrated Gradients + B	1.59 (0.31)	3.04 (0.56)
	Multi-label	Integrated Gradients	0.72 (0.28)	1.35 (0.54)
		Integrated Gradients + B	1.48 (0.38)	2.82 (0.69)
/		B	2.39 (0.31)	4.61 (0.56)

B: Boundary, ATE: Atelectasis, FRA: Fracture, MAS: Mass, PTX: Pneumothorax

To provide a comparative visualization of model explanations with and without anatomic boundary, Figure 2, 3, and 4 portrays a comparative visualization of Saliency Map, Grad-CAM, and Integrated Gradients, respectively. 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 of the lung area constrained model explanations within the lung area and therefore enhanced the overlapping between the lesion region and model focus area.

4. Discussion

In this study, we evaluated the performance of three popular XAI methods for CNN explanations and proposed a plug-and-play module using the anatomic boundary of the lung area for explanation improvements. Based on various combinations of CNN architectures, XAI methods, and classification settings, the proposed method consistently improved baseline explanations for thoracopathy classifiers.

In our experiments, VGG-11 outperformed ResNet-18, a more sophisticated architecture, in terms of AUPRC, accuracy, and precision. The phenomenon of "The deeper is not the better" has been reported in various medical applications (Santos-Bustos et al., 2022; Ikechukwu et al., 2021) and was verified again by our experimental results on thoracopathy classification, implying that the conventional VGG-11 is still capable of handling middle scale datasets. However, the success of VGG-11 in classification failed to guarantee it is better than ResNet-18 in thoracic disease explanations. Additionally, in our experimentation with VGG-11 and ResNet-18, 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 exper-

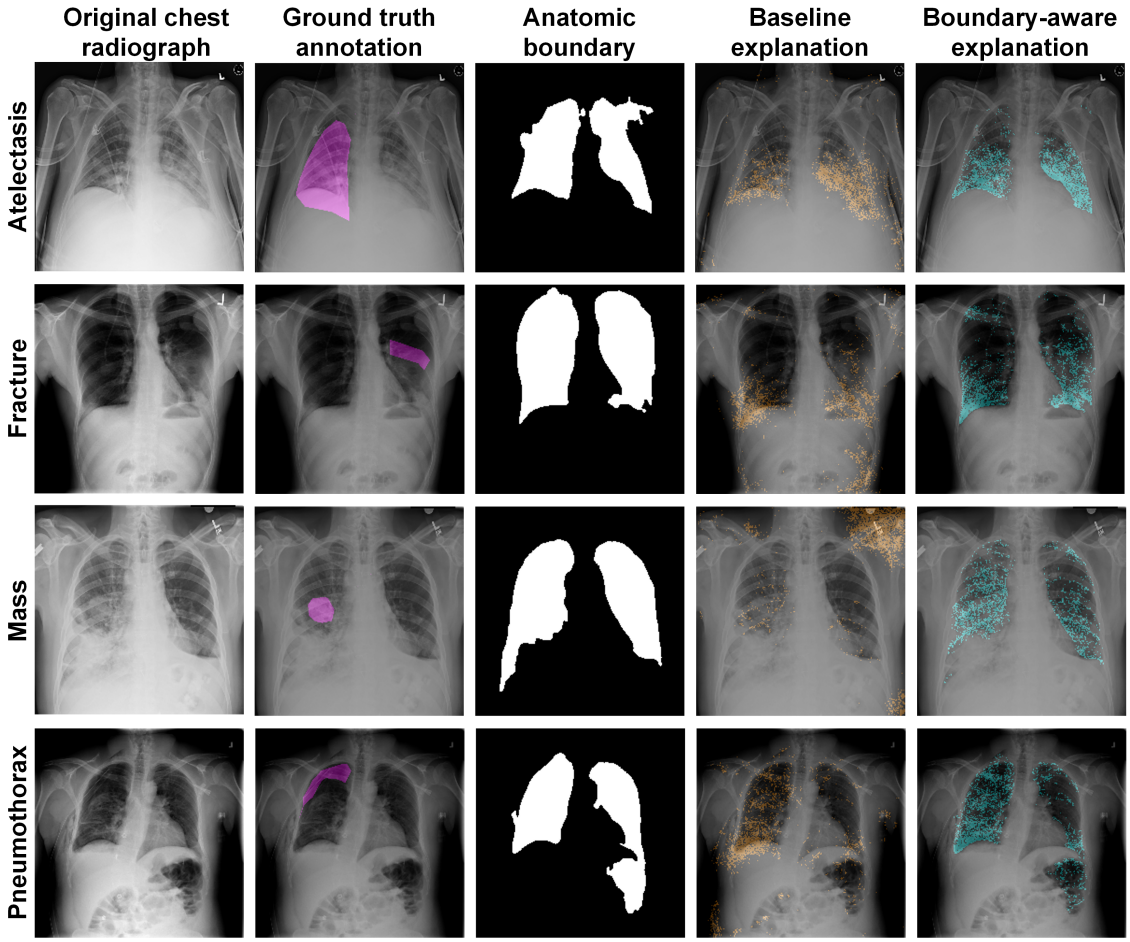

Figure 2: Visualization comparison of thoracopathy radiograph explained by baseline and boundary-aware Saliency Map.

imental results cannot conclusively demonstrate the superiority of binary classifiers over multi-label ones. The diagnostic process in clinical settings is often complex and likely to involve multi-label tasks rather than simplified binary tasks (Chen et al., 2020). Lastly, it is noteworthy that the best-achieved explanation performance by anatomic boundary-aware Grad-CAM on Mass still failed to meet the administrative regulation standards with a minimum DSC of 20% for clinically relevant XAI explanations (Lee et al., 2022), demonstrating that there still exists a gap for current explanation methods to be deployed in the real world.

In contrast to the previous introduction of domain knowledge in medical image analysis methods (Bateson et al., 2021; Kervadec et al., 2019) that required additional annotation of clinical knowledge on the target dataset, our method employed an external lung area segmenter to generate the anatomic boundary and demonstrated its effectiveness through the consistent improvements on explanation performance. However, we acknowledge the value of additional annotations like the delineation of the lung area by human experts on the target dataset. Given the prevalent domain shift indicated by the anatomic boundary in Figure 2, 3, and 4, the external segmenter could be fine-tuned to match the representation

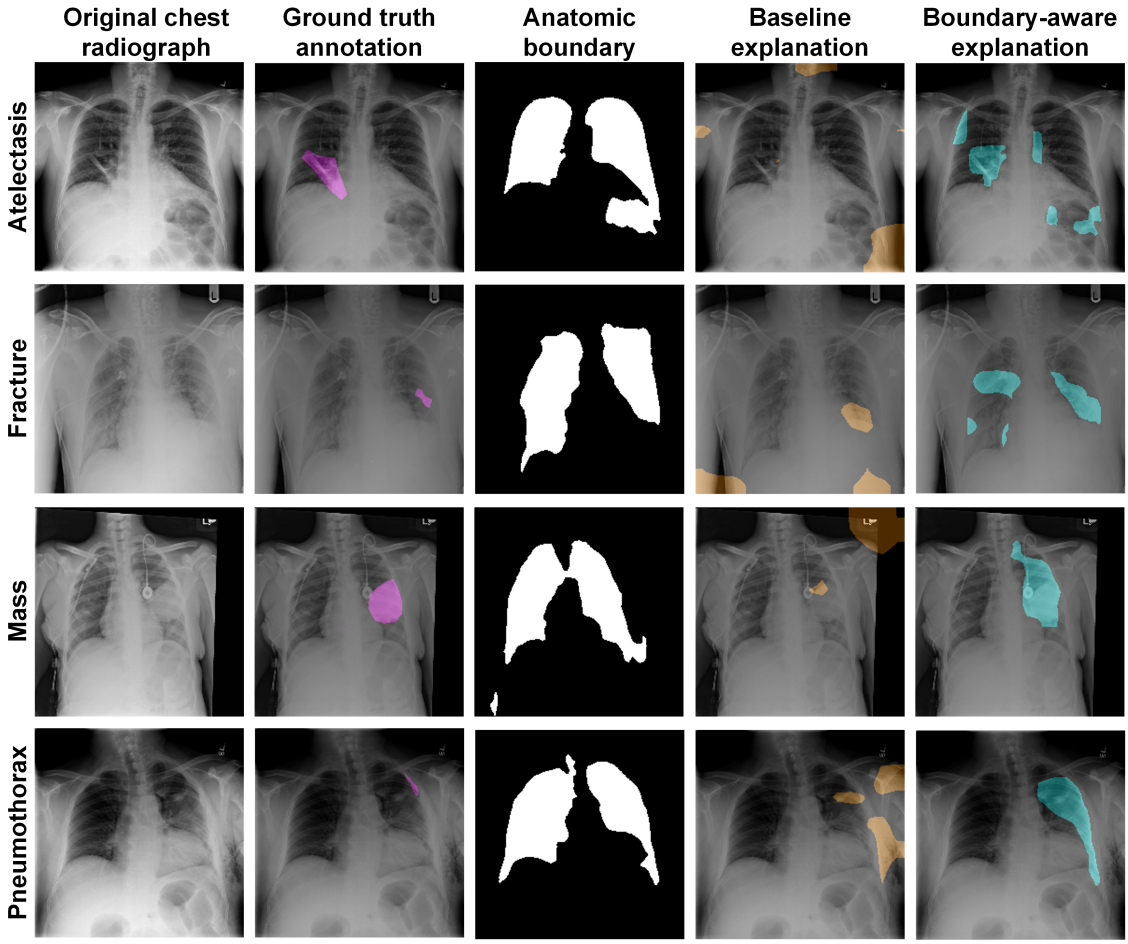

Figure 3: Visualization comparison of thoracopathy radiograph explained by baseline and boundary-aware Grad-CAM.

of the target dataset, potentially offering improved constraints for XAI methods (Yuan et al., 2023b). Also, the proposed method relied on a rough boundary of the lung area for upgrading model explanations. Future research would tailor fine-grained constraints by considering the characteristics of different thoracic diseases.

There are several other limitations of our work. First, the explanation methods utilized in this study were confined to three XAI methods of Saliency Map, Grad-CAM, and Integrated Gradients, and the DL models were limited to 2 lightweight architectures of VGG-11 and ResNet-18. Additional XAI methods such as LayerCAM (Jiang et al., 2021) and DL models like Vision Transformer (Dosovitskiy et al., 2021) would offer a more comprehensive analysis. Second, our experiments revealed that the anatomic boundary consistently improved CNN explanations and suggested the potential of anatomic information as a reward in improving other tasks (Kang et al., 2023, 2024). Furthermore, this research exclusively explored 4 thoracic diseases and other diseases such as pleural effusion, edema, and consolidation were underexplored (Saporta et al., 2022). Finally, rigorous statistical tests can be implemented to explore the association between XAI performance and geometric features

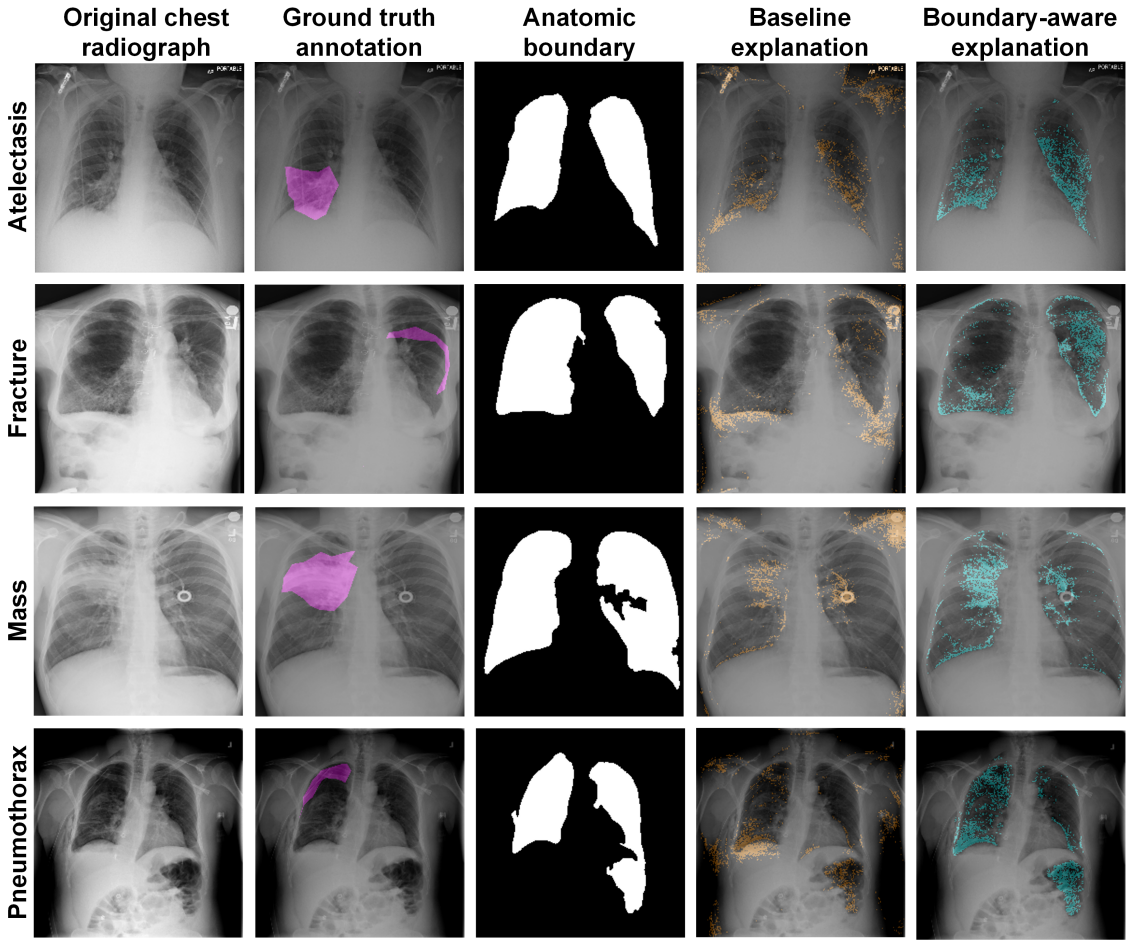

Figure 4: Visualization comparison of thoracopathy radiograph explained by baseline and boundary-aware Integrated Gradients.

of pathologies (Saporta et al., 2022) and the association between XAI performance and classification performance to offer deeper insights into model behaviors and inference logics (Yuan et al., 2024).

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 module for improving XAI methods in CNN explanations. We hope that the consistent improvements in model explanations could inspire future efforts in integrating clinical domain knowledge into medical image analysis.

References

Qassim Baker et al. Anatomy of the thorax. In *Anatomy*. 2022.

Mathilde Bateson et al. Constrained domain adaptation for image segmentation. *IEEE Transactions on Medical Imaging*, 2021.

Bingzhi Chen et al. Dualchexnet: dual asymmetric feature learning for thoracic disease classification in chest x-rays. *Biomedical Signal Processing and Control*, 2019.

Haomin Chen et al. Deep hiearchical multi-label classification applied to chest x-ray abnormality taxonomies. *Medical Image Analysis*, 2020.

Alexey Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *the International Conference on Learning Representations*, 2021.

Bradley Efron. Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 1987.

Guanghui Fu et al. Tuning model parameters in class-imbalanced learning with precision-recall curve. *Biometrical Journal*, 2019.

Philip Harford et al. Effectiveness of erector spinae plane block for rib fracture analgesia: a systematic review protocol. *JBI Evidence Synthesis*, 2023.

Kaiming He et al. Deep residual learning for image recognition. In *the Computer Vision and Pattern Recognition Conference*, 2016.

Wonju Hong et al. Deep learning for detecting pneumothorax on chest radiographs after needle biopsy: clinical implementation. *Radiology*, 2022.

Victor Ikechukwu et al. Resnet-50 vs vgg-19 vs training from scratch: A comparative analysis of the segmentation and classification of pneumonia from chest x-ray images. *Global Transitions Proceedings*, 2021.

Stefan Jaeger et al. Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative Imaging in Medicine and Surgery*, 2014.

Pengtao Jiang et al. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 2021.

Honggyu Jung et al. Weakly supervised thoracic disease localization via disease masks. *Neurocomputing*, 2023.

Lican Kang et al. Error analysis of fitted q-iteration with relu-activated deep neural networks. In *the International Conference on Learning Representations*, 2023.

Lican Kang et al. Approximate policy iteration with deep minimax average bellman error minimization. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

Hoel Kervadec et al. Constrained-cnn losses for weakly supervised segmentation. *Medical Image Analysis*, 2019.

Alex Krizhevsky et al. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Sunyeop Lee et al. Localization-adjusted diagnostic performance and assistance effect of a computer-aided detection system for pneumothorax and consolidation. *npj Digital Medicine*, 2022.

Mingxuan Liu et al. Handling missing values in healthcare data: A systematic review of deep learning-based imputation techniques. *Artificial Intelligence in Medicine*, 2023.

Ilhwan Park et al. Occult pneumothorax in patients with blunt chest trauma: key findings on supine chest radiography. *Journal of Thoracic Disease*, 2023.

David Rumelhart et al. Learning representations by back-propagating errors. *Nature*, 1986.

Daniel Fernando Santos-Bustos et al. Towards automated eye cancer classification via vgg and resnet networks using transfer learning. *Engineering Science and Technology*, 2022.

Adriel Saporta et al. Benchmarking saliency methods for chest x-ray interpretation. *Nature Machine Intelligence*, 2022.

Ramprasaath Selvaraju et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *the International Conference on Computer Vision*, 2017.

Junji Shiraishi 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. *American Journal of Roentgenology*, 2000.

Karen Simonyan et al. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *the International Conference on Learning Representations*, 2014.

Karen Simonyan et al. Very deep convolutional networks for large-scale image recognition. In *the International Conference on Learning Representations*, 2015.

Mukund Sundararajan et al. Axiomatic attribution for deep networks. In *the International Conference on Machine Learning*, 2017.

Bas HM Van der Velden et al. Explainable artificial intelligence (xai) in deep learning-based medical image analysis. *Medical Image Analysis*, 2022.

Hongyu Wang et al. Thorax-net: an attention regularized deep neural network for classification of thoracic diseases on chest radiography. *IEEE Journal of Biomedical and Health Informatics*, 2019.

Feng Xie et al. Deep learning for temporal data representation in electronic health records: A systematic review of challenges and methodologies. *Journal of Biomedical Informatics*, 2022.

Koichiro Yasaka et al. Deep learning and artificial intelligence in radiology: Current applications and future directions. *PLoS Medicine*, 2018.

Han Yuan et al. Autoscore-imbalance: An interpretable machine learning tool for development of clinical scores with rare events data. *Journal of Biomedical Informatics*, 2022.

Han Yuan et al. Human-guided design to explain deep learning-based pneumothorax classifier. In *the International Conference on Medical Imaging with Deep Learning*, 2023a.

Han Yuan et al. Leveraging anatomical constraints with uncertainty for pneumothorax segmentation. *arXiv*, 2023b.

Han Yuan et al. Opening the black box of deep learning: Validating the statistically significant association between class activation map (cam) and clinical domain knowledge in fundus image-based glaucoma diagnosis. *arXiv*, 2024.