The Explainability Value in CNN-Based Medical Imaging Diagnostics: A Comparative Study of Grad-CAM, LIME, and SHAP

Nino Miljkovic
Independent Researcher
nmiljkovic1@my.hpu.edu

November 22, 2025

Abstract

The integration of deep learning into medical imaging has demonstrated considerable diagnostic potential, yet widespread clinical adoption remains constrained by the opaque decision-making processes inherent to convolutional neural networks (CNNs). This paper presents a systematic comparison of three prominent post-hoc Explainable Artificial Intelligence (XAI) methods — Gradient-weighted Class Activation Mapping (Grad-CAM), Locally Interpretable Model-Agnostic Explanations (LIME), and SHapley Additive exPlanations (SHAP) — applied to a ResNet18 classifier trained on the RSNA Pneumonia Detection Challenge dataset ($n = 30{,}227$). The proposed *medxai* experimental pipeline provides a reproducible, end-to-end framework for training, evaluating, and interpreting CNN predictions on chest radiographs. The model achieves an AUC of 0.8716 and a sensitivity of 94.46%, establishing a clinically meaningful performance baseline. Qualitative and cross-method analysis demonstrates that Grad-CAM, LIME, and SHAP produce convergent explanations, consistently highlighting pathologically plausible pulmonary regions across pneumonia-positive cases. This triangulation effect provides stronger interpretive confidence than any single method alone, supporting the argument that multi-method XAI frameworks are essential for transparent and trustworthy clinical AI deployment. The study further situates these findings within the broader trajectory of XAI research in radiology, evidenced by a substantial growth in publication volume since 2018.

Keywords: explainable AI, convolutional neural networks, medical imaging, Grad-CAM, LIME, SHAP, pneumonia detection, interpretability, trustworthy AI

Contents

1 Introduction

Deep learning has emerged as a transformative technology in medical imaging, enabling detection and classification of pathologies with remarkable accuracy. Convolutional neural networks (CNNs), in particular, have achieved strong performance across complex diagnostic tasks including pneumonia detection, breast cancer screening, and diabetic retinopathy classification [1, 2]. These advances carry particular significance for conditions such as pneumonia and lower respiratory infections, which remain among the leading causes of morbidity and mortality worldwide, especially in resource-limited settings where chest X-rays constitute the primary imaging modality [3].

Despite deep learning models approaching or exceeding radiologist-level performance in retrospective evaluations, their routine deployment in clinical workflows remains limited. The primary barrier is no longer predictive performance but rather *trust*, *interpretability*, and *clinical reliability*. In high-stakes environments, the "black-box" nature of deep learning models, where decision pathways are not directly observable, creates substantial challenges for clinicians who must justify and verify diagnostic decisions [4, 5]. Concerns extend beyond accuracy to include robustness to dataset shift, hidden biases from confounded or imbalanced training distributions, and failure modes on clinically important subgroups that aggregate performance metrics may obscure [6, 7].

Explainable Artificial Intelligence (XAI) has emerged as a central response to these concerns. XAI encompasses a suite of methods and frameworks designed to make model behavior transparent and understandable, enabling clinicians to interrogate and contextualize predictions rather than accept them uncritically [4]. In medical contexts, explainability is not merely an ethical ideal but a functional requirement: clinicians must validate model outputs against established diagnostic knowledge, detect spurious correlations, and ensure that predictions are grounded in medically meaningful features rather than imaging artifacts, institutional markers, or annotation noise [5–7].

This study contributes to that ongoing discourse through an applied, reproducible implementation of three leading XAI techniques within a controlled deep learning environment tailored for medical imaging. The experimental framework, referred to throughout as the *medxai* environment, provides a systematic pipeline for training, evaluating, and interpreting a CNN for pneumonia detection using the RSNA Pneumonia Detection dataset [27]. Within this framework, three widely adopted XAI methods are systematically compared: Grad-CAM [8], LIME [9], and SHAP [10]. Each method embodies a distinct conceptual approach to interpretability: Grad-CAM produces class-specific activation maps highlighting salient image regions; LIME constructs local surrogate models through input perturbation; and SHAP assigns feature-level contributions using cooperative game theory.

By applying these techniques to a CNN trained on a large-scale radiographic dataset, this work illustrates how interpretability methods complement predictive accuracy by revealing which image regions and latent features most strongly influence diagnostic decisions. In doing so, it aims to move beyond abstract discussion of XAI principles toward a concrete, clinically relevant demonstration: explaining CNN behavior on chest X-ray images in a manner that can be meaningfully evaluated by radiologists and domain experts.

The remainder of this paper is structured as follows. Section 2 reviews relevant literature on XAI in medical imaging. Section 3 describes the experimental methodology, including model architecture, dataset preprocessing, and XAI implementation. Section 4 presents quantitative performance results and qualitative XAI visualizations. Section 5

interprets these findings within the broader research landscape. Section 6 concludes with limitations and directions for future work.

2 Related Work

2.1 Explainable AI in Medical Imaging

The rapid adoption of deep learning in medical imaging has amplified long-standing concerns about transparency, accountability, and algorithmic trust. While CNNs have demonstrated strong performance across diagnostic tasks [1, 2, 7], their inherently opaque decision processes pose significant challenges for clinical integration. This tension between performance and interpretability is widely recognized in the healthcare AI literature, where explainability is increasingly framed as a prerequisite for trustworthy systems rather than an optional supplement [4, 5, 19].

Markus et al. [19] provide a comprehensive survey emphasizing that trustworthiness depends not only on model performance but on clearly defined terminology, appropriate design choices, and rigorous evaluation of explanations. Their work highlights that clinicians require explanations that are faithful to the underlying model, comprehensible to domain experts, and aligned with clinical reasoning patterns. Similarly, Amann et al. [4] and Tonekaboni et al. [5] stress that clinical users evaluate AI tools not solely on accuracy but on whether they can reliably interrogate model outputs, identify failure modes, and integrate predictions into existing workflows.

In radiology specifically, hidden stratification, dataset shift, and spurious correlations further motivate interpretability. Oakden-Rayner et al. [6] and Zech et al. [7] demonstrate that CNNs can achieve strong aggregate metrics while failing on clinically important subpopulations due to biased or confounded training distributions. These findings underscore the importance of explanations that reveal when models rely on non-causal features, such as laterality markers, scanner artifacts, or institutional signatures, rather than pathological findings.

2.2 Post-Hoc XAI Methods for CNNs

Most XAI research in medical imaging has focused on post-hoc interpretability, where explanations are generated after training a black-box model. For CNNs applied to image data, three classes of methods have proven particularly influential.

Gradient-based methods. Grad-CAM, introduced by Selvaraju et al. [8], computes class-specific localization maps by backpropagating gradients from the target class score to the final convolutional layer, aggregating them to highlight regions most influential to the prediction. In radiology, Grad-CAM heatmaps have been applied to pneumonia detection, COVID-19 classification, and tumor localization, providing intuitive overlays comparable with radiologist annotations.

Local surrogate methods. LIME, proposed by Ribeiro et al. [9], approximates a model's local decision boundary around a specific input through perturbation sampling. For image data, LIME segments radiographs into superpixels and fits a simple interpretable model, typically linear, to mimic CNN behavior in a local neighborhood. This approach has been used in multiple medical imaging studies to identify which regions most strongly support or contradict a given prediction [16, 20].

Game-theoretic attribution. SHAP, introduced by Lundberg and Lee [10], generalizes feature attribution using Shapley values from cooperative game theory. SHAP has been applied at both pixel and superpixel levels to quantify individual feature contributions to predicted disease probability. Its theoretical guarantees of local accuracy and consistency make it particularly suited to high-stakes domains where explanation fidelity is critical [10, 16].

Recent surveys corroborate the centrality of these methods. Sun et al. [20] identify Grad-CAM, LIME, and SHAP as among the most commonly adopted tools for visualizing model behavior across imaging, audio, and multimodal tasks. Chaddad et al. [17] further demonstrate the integration of multiple CAM-based methods with ResNet architectures across several medical datasets, evaluating how different XAI techniques vary in their capacity to highlight clinically meaningful regions.

2.3 Emerging Paradigms: Self-Explainable and Generalizable AI

While post-hoc XAI remains dominant, recent literature has begun to highlight its limitations. Chief concerns include potential mismatches between explanations and true model behavior (low fidelity), instability under small input perturbations, and the risk that visually plausible explanations may still mask biased or non-causal decision pathways [16, 18, 19].

Hou et al. [18] introduce the concept of Self-eXplainable AI (S-XAI), in which models are designed to produce intrinsic explanations as part of their architecture and training process. Through attention mechanisms, concept-based reasoning, and structured prediction, S-XAI embeds interpretability directly into model design. Chaddad et al. [17] argue that generalizability and explainability should be treated as coupled design objectives, showing that different XAI techniques can produce divergent explanations even when model performance is similar, thus emphasizing the need for a standardized quantitative evaluation of explanations alongside conventional accuracy metrics.

3 Methodology

3.1 Experimental Framework

The methodology is structured around the *medxai* experimental environment: a reproducible, end-to-end pipeline for training, evaluating, and interpreting a CNN model for pneumonia detection. The pipeline integrates dataset preparation, model training, and multi-method XAI analysis, producing both quantitative metrics and qualitative visualizations of model decision-making.

3.2 Model Architecture

The backbone architecture is ResNet18 [14], a residual CNN whose skip connections facilitate training of deeper representations while managing gradient degradation. Its moderate depth and computational efficiency make it well-suited to medical imaging tasks requiring a balance between performance and interpretability overhead. The network was initialized with ImageNet pre-trained weights, enabling transfer learning to accelerate convergence and improve feature generalization from limited medical training data. The

final fully connected layer was replaced with a single-output node for binary classification (pneumonia vs. normal).

3.3 Computational Environment

The *medxai* environment was constructed using Anaconda with Python 3.10, PyTorch 2.5.1, and CUDA 12.1 for GPU acceleration on an NVIDIA GeForce RTX 3060 Ti. Python 3.10 was selected to ensure full compatibility with PyTorch 2.5.x and the CUDA 12.1 runtime, as newer Python releases introduce unresolved dependency incompatibilities with several deep learning libraries. All dependencies were pinned to specific versions to guarantee reproducibility, as summarized in Table 1.

Table 1: Core library versions in the *medxai* environment.

Library	**Version**	**Role**
numpy	1.26.4	Numerical operations
scipy	1.11.4	Optimisation and image processing
pandas	2.0.3	Metadata handling
matplotlib	3.8.4	Visualisation
scikit-learn	1.3.2	Preprocessing and metrics
pydicom	3.0.1	DICOM parsing
torch	2.5.1	Deep learning framework [22]
captum	0.8.0	Model interpretability [23]
lime	0.2.0.1	Local interpretability [9]
shap	0.44.1	Shapley-based attribution [10]

NumPy 2.0 and above was explicitly avoided due to known ABI incompatibilities with PyTorch's tensor backend and Scikit-learn compiled extensions [24].

3.4 Dataset

The dataset used is the RSNA Pneumonia Detection Challenge dataset [27], provided by the Radiological Society of North America (RSNA) in collaboration with the National Institutes of Health (NIH). It contains over 30,000 anonymised chest X-ray images in DICOM format, labeled as pneumonia-positive or normal. Representative samples from the dataset are shown in Figure 1. Key characteristics are summarized in Table 2.

Figure 1: Representative chest X-ray samples from the RSNA Pneumonia Detection Challenge dataset, illustrating variability in image quality, patient positioning, and the presence or absence of radiographic opacities.

Table 2: RSNA Pneumonia Detection dataset characteristics.

Property	**Value**
Total studies	30,227
Unique patients	26,684
Normal cases	20,672 (68.4%)
Pneumonia cases	9,555 (31.6%)
Image format	DICOM (.dcm)
Input resolution	224×224 (standardised)

DICOM files were decoded using the `pydicom` library with the `apply_voi_lut()` method to apply DICOM windowing functions [25]. Images in `MONOCHROME1` format were inverted to maintain diagnostic consistency. Pixel values were scaled to $[0, 255]$, cast to 8-bit unsigned integers, and replicated across three channels to produce RGB tensors compatible with ImageNet-pretrained architectures.

3.5 Data Augmentation and Sampling

Image normalization followed ImageNet standards (mean $= [0.485, 0.456, 0.406]$; std $= [0.229, 0.224, 0.225]$). Training augmentations were applied using the `torchvision.transforms` API [26] and are summarized in Table 3.

Table 3: Training augmentation pipeline.

Transformation	**Parameter**	**Rationale**
Resize	224×224	Standardise variable-resolution inputs
Random rotation	$\pm 15°$	Simulate patient repositioning
Horizontal flip	$p = 0.5$	Mimic lateral orientation variability
Color jitter	brightness/contrast ± 0.1	Compensate for scanner variation
Normalisation	ImageNet mean/std	Align with pretrained feature space

To address class imbalance (68.4% normal, 31.6% pneumonia), a Weighted Random Sampler was applied during training, assigning each instance a sampling probability inversely proportional to its class frequency. The dataset was partitioned into patient-independent splits to prevent data leakage: 70% training (18,678 images), 15% validation (4,003 images), and 15% test (4,003 images).

3.6 Training Configuration

The model was trained using AdamW (learning rate $= 1 \times 10^{-4}$; $\beta_1 = 0.9$, $\beta_2 = 0.999$) with binary cross-entropy loss (`BCEWithLogitsLoss`). Training proceeded for five epochs with a batch size of 32; the model checkpoint with the highest validation AUC was retained for all subsequent evaluation and XAI analyses. Each epoch completed in approximately 15 minutes on the RTX 3060 Ti.

3.7 Explainability Methods

Three XAI methods were implemented and compared within the *medxai* pipeline:

Grad-CAM [8] uses gradients of the target class score flowing into the final convolutional layer to produce class-discriminative localization maps. Backward hooks were registered on the final convolutional block; the weighted average of gradient-scaled feature maps was computed and normalized to $[0, 1]$ to produce a heatmap overlay on the original radiograph. Warm colors in the overlay indicate regions with greatest positive influence on the pneumonia prediction.

LIME [9] approximates the model's local decision boundary around each instance through perturbation sampling. Radiographs were segmented into superpixels using the Quickshift algorithm (kernel size = 4; max_dist = 10). Randomized masking of superpixel subsets generated 500 perturbed samples per image; a linear surrogate model was fitted to approximate CNN behavior locally. The resulting binary masks identify superpixels that most strongly support or contradict the predicted class.

SHAP [10] extends cooperative game theory to model interpretation by assigning each feature a Shapley value representing its marginal contribution to the model output. The `GradientExplainer` variant was applied, computing gradients relative to a background distribution. Positive contributions (red) denote regions increasing predicted pneumonia probability; negative contributions (blue) denote suppressive evidence.

Together, these three methods provide a multi-angle interpretability framework: Grad-CAM reveals coarse spatial attention, LIME delineates specific contributing superpixel regions, and SHAP quantifies pixel-level attributions within a theoretically grounded additive framework.

4 Results

4.1 Model Performance

The best-performing checkpoint was selected based on validation AUC across five training epochs. Figure 2 presents the per-epoch training and validation logs, showing steady improvement in AUC and loss. Table 4 reports test-set performance, and Figure 3 shows the full classification report.

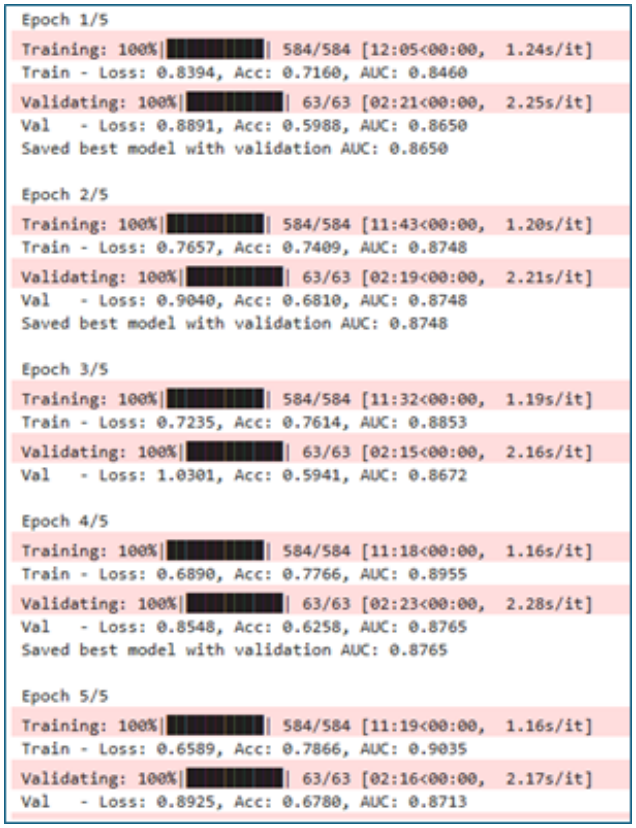


Figure 2: Per-epoch training and validation logs across 5 epochs. Each epoch reports training loss, accuracy, and AUC, followed by validation metrics. The best checkpoint (highest validation AUC = 0.8765) was saved at epoch 4.

Table 4: Test set performance metrics of the ResNet18 pneumonia classifier.

Metric	**Value**
Accuracy	0.6295
Precision	0.3729
Recall (Sensitivity)	0.9446
F1-Score	0.5347
AUC	0.8716
Specificity	0.5379

```
Loaded best model from resnet18_pneumonia_best.pt

Evaluating model on test set...
Testing: 100%|██████████| 63/63 [02:17<00:00,  2.17s/it]
================================================
TEST SET PERFORMANCE METRICS
================================================
Accuracy  : 0.6295
Precision : 0.3729
Recall    : 0.9446
F1-Score  : 0.5347
AUC       : 0.8716

================================================
CLASSIFICATION REPORT
================================================
              precision    recall  f1-score   support

      Normal     0.9709    0.5379    0.6923      3101
   Pneumonia     0.3729    0.9446    0.5347       902

    accuracy                         0.6295      4003
   macro avg     0.6719    0.7412    0.6135      4003
weighted avg     0.8361    0.6295    0.6568      4003
```

Figure 3: Test set classification report showing per-class precision, recall, and F1-score for Normal and Pneumonia classes, alongside macro and weighted averages.

The confusion matrix (Figure 4) yields: True Negatives (TN) = 1,668; False Positives (FP) = 1,433; False Negatives (FN) = 50; True Positives (TP) = 852. The predominance of false positives is consistent with classifiers trained to minimize missed positive cases, reflecting the asymmetric clinical cost of false negatives in pneumonia screening.

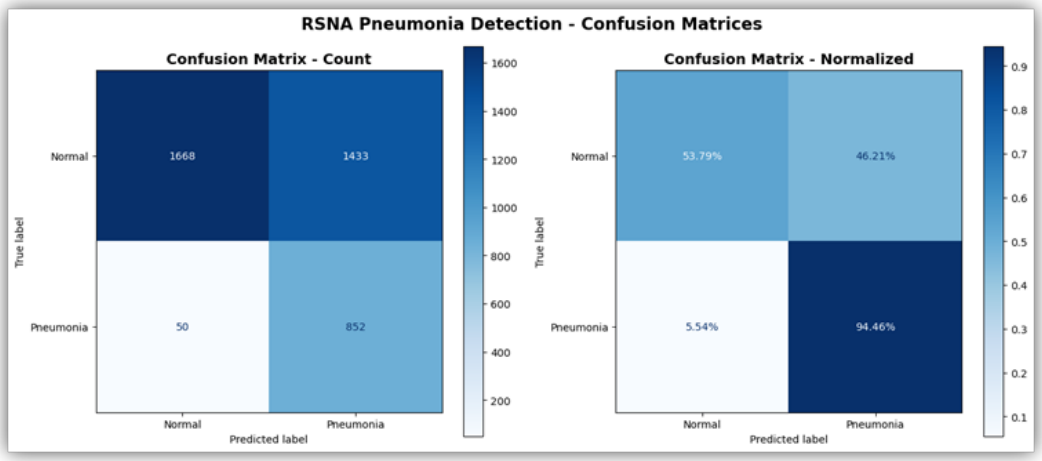


Figure 4: Confusion matrices for the ResNet18 classifier on the test set. *Left:* raw counts. *Right:* normalized proportions. The high false-positive rate reflects the model's sensitivity-optimized behavior under class imbalance.

ROC analysis identified an optimal decision threshold of 0.808 (Youden's J statistic), yielding TPR = 0.808 and FPR = 0.222. The ROC curve, Precision-Recall curve, and

prediction probability distributions are shown in Figure 5.

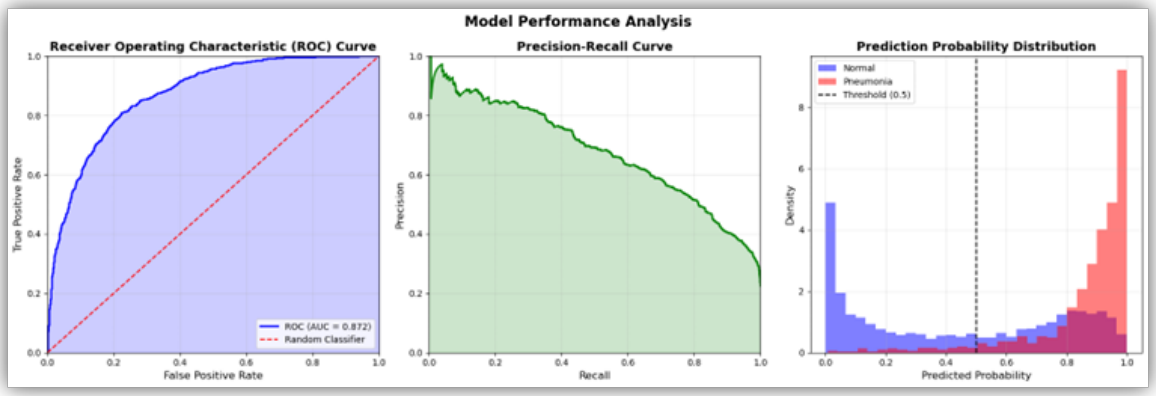


Figure 5: Model performance analysis. *Left:* ROC curve (AUC = 0.872) with random classifier baseline. *Centre:* Precision-Recall curve. *Right:* Predicted probability distributions for Normal and Pneumonia classes, with the default decision threshold at 0.5 indicated.

4.2 XAI Methods: Cross-Patient Comparison

Figure 6 presents a side-by-side comparison of all three XAI methods applied to two representative patients: one with a high pneumonia probability ($P(\text{Pneumonia}) = 0.999$) and one classified as normal ($P(\text{Pneumonia}) = 0.000$). The four columns show the original radiograph, the Grad-CAM heatmap, the LIME superpixel mask, and the SHAP attribution map.

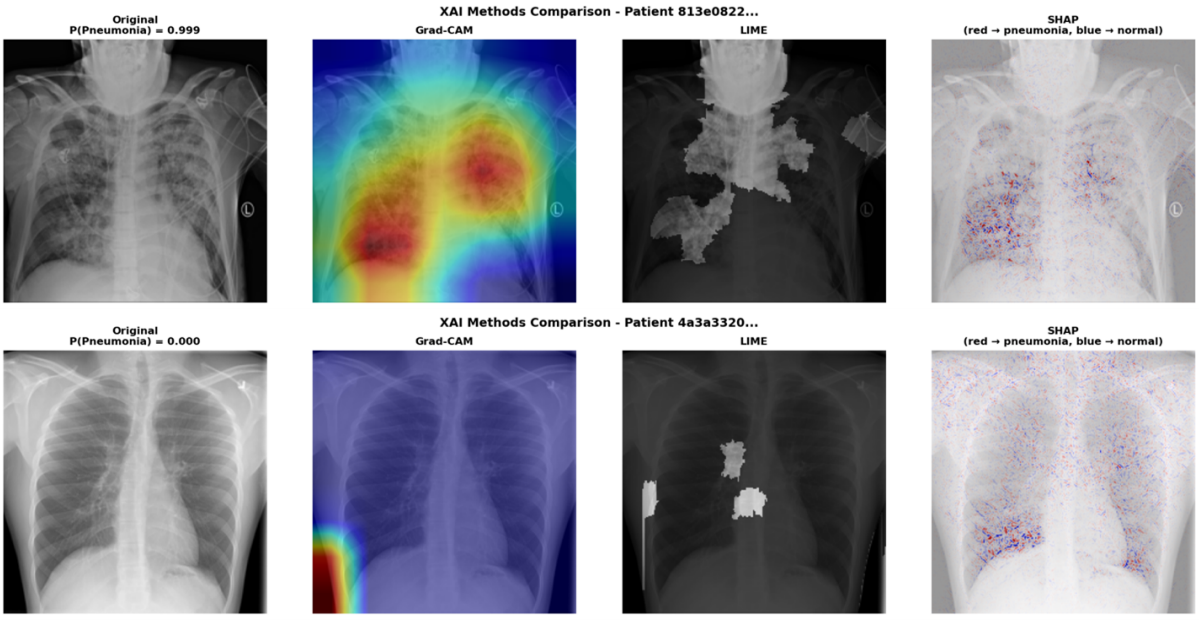


Figure 6: XAI method comparison for two representative patients. *Top row:* Patient 813e0822 — pneumonia positive ($P = 0.999$). *Bottom row:* Patient 4a3a3320 — normal ($P = 0.000$). Columns from left to right: original radiograph, Grad-CAM heatmap, LIME superpixel mask, and SHAP attribution map (red → pneumonia evidence; blue → normal evidence).

4.3 Grad-CAM Visualizations

For pneumonia-positive samples, Grad-CAM heatmaps produced concentrated activations within the lower and central lung zones, overlapping radiographic opacities consistent with pneumonic consolidation (Figure 7). This spatial correspondence between model activations and pathologically plausible regions supports the inference that the network learned diagnostically meaningful representations rather than spurious correlations. For pneumonia-negative cases, activations were diffuse and peripheral, reflecting the absence of strong pathological cues.

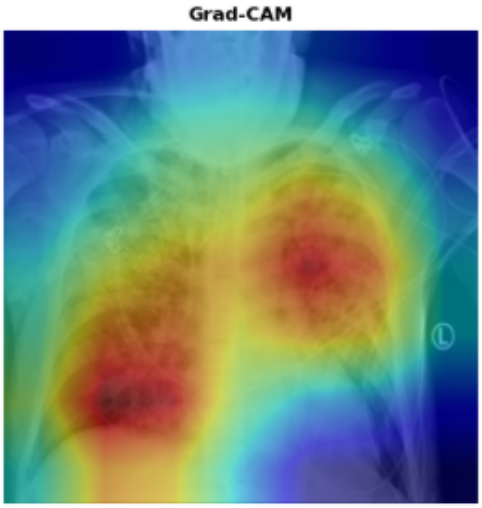


Figure 7: Grad-CAM heatmap for a pneumonia-positive chest radiograph. Warm colors (red/yellow) indicate regions exerting the greatest positive influence on the pneumonia prediction, concentrated in the lower and central lung zones consistent with radiographic consolidation.

These findings are consistent with prior applications of Grad-CAM to chest radiograph classification [8,17], where warm-color overlays concentrated in lung parenchyma serve as a proxy for anatomical plausibility of learned representations. A limitation of Grad-CAM is its coarse spatial resolution, which occasionally blends signals from anatomically adjacent regions.

4.4 LIME Visualizations

LIME superpixel masks provided sharper spatial boundaries than Grad-CAM, offering finer-grained delineation of contributing image regions (Figure 8). In pneumonia-positive cases, masks frequently emphasized lung opacities and regions of consolidation, corroborating Grad-CAM findings at higher spatial granularity. Pneumonia-negative cases exhibited fewer and less intense active superpixels.

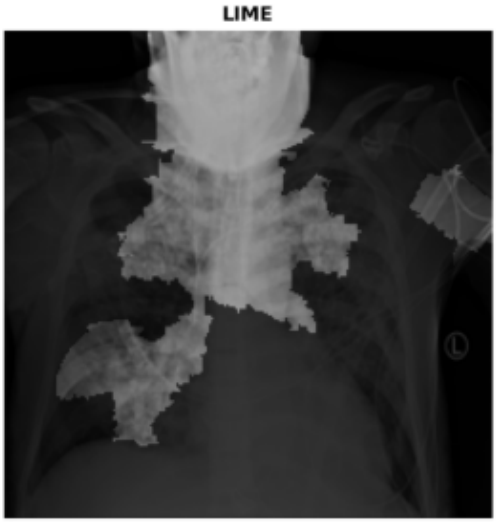


Figure 8: LIME superpixel mask for a pneumonia-positive chest radiograph. Highlighted regions (white) identify superpixels most strongly contributing to the positive classification, offering sharper spatial delineation than Grad-CAM at the cost of stochastic run-to-run variability.

A known limitation of LIME is its stochastic nature: repeated runs with different random seeds occasionally produced variations in mask shape or activation strength. Nevertheless, LIME outputs were consistently aligned with the model's predicted confidence levels, reinforcing the spatial coherence of Grad-CAM explanations.

4.5 SHAP Visualisations

SHAP attribution maps extended interpretability into a quantitative dimension by assigning pixel-level Shapley values (Figure 9). Positive contributions (red) denote pneumonia-supporting evidence; negative contributions (blue) denote suppressive evidence. Across pneumonia-positive cases, SHAP maps exhibited strong spatial alignment with Grad-CAM heatmaps: regions identified as salient by gradient-based localization also tended to carry high positive Shapley values.

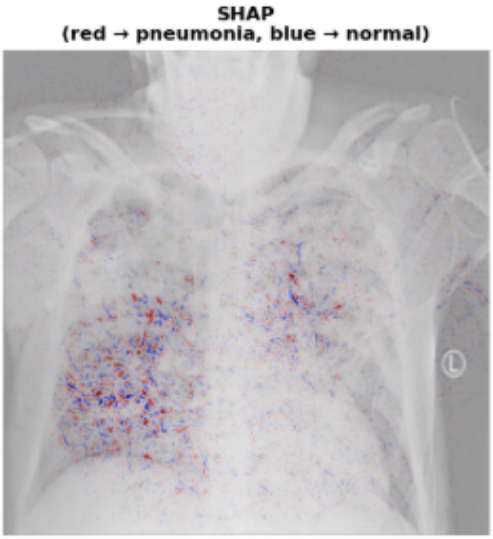


Figure 9: SHAP (GradientExplainer) attribution map for a pneumonia-positive chest radiograph. Red pixels indicate positive contributions to the pneumonia prediction; blue pixels indicate suppressive (normal-supporting) evidence. The bilateral distribution of positive attributions across the lung fields is consistent with the diffuse nature of the pneumonic infiltrates in this case.

SHAP's bidirectional encoding enabled detection of both supporting and contradictory evidence within the same image, offering richer interpretive context for uncertain or borderline predictions. SHAP yielded stable and reproducible explanations across test samples, though at greater computational cost relative to Grad-CAM.

4.6 Cross-Method Convergence

The comparative evaluation revealed substantial complementarity across all three XAI methods. Despite operating on distinct theoretical foundations (gradient propagation, local surrogate modeling, and cooperative game theory) Grad-CAM, LIME, and SHAP consistently highlighted overlapping pulmonary regions as the most discriminative features for pneumonia classification. Specifically:

- Grad-CAM identified coarse spatial attention patterns within the lung parenchyma.
- LIME localized specific superpixel clusters at finer spatial granularity.
- SHAP confirmed these regions through positive Shapley attribution, while additionally identifying suppressive peripheral regions.

When visualized side-by-side (Figure 6), the overlapping attention zones form an interpretable consensus map of the model's diagnostic logic. This triangulation effect mitigates the individual limitations of each method: Grad-CAM's coarse resolution, LIME's stochastic variability, and SHAP's computational intensity are counterbalanced when their outputs are combined, yielding an ensemble of explanations with both visual interpretability and mathematical rigor.

5 Discussion

The results highlight a dual reality in contemporary medical AI: CNNs can achieve high diagnostic sensitivity in chest X-ray analysis, yet clinical value depends equally on the transparency of decision-making. The ResNet18 model achieved a strong AUC (0.8716) and exceptional sensitivity (94.46%), yet the comparatively low specificity (53.79%) and high false-positive rate underscore persistent challenges related to dataset heterogeneity and class imbalance, broadly documented in radiology AI research [6, 7]. These findings reinforce interpretability not as an auxiliary feature but as an essential component of any clinically deployable AI system.

The XAI analysis demonstrates how Grad-CAM, LIME, and SHAP each contribute a distinct interpretive function. Grad-CAM provided spatially coherent heatmaps aligned with radiographically relevant regions, consistent with prior literature [8, 17]. LIME generated superpixel-based explanations capturing local texture and intensity variations; while more spatially precise, it exhibited greater run-to-run variability [9, 18]. SHAP offered theoretically grounded quantification of pixel-level contributions, producing nuanced attribution maps revealing both positive and suppressive evidence. Its axiomatic properties [10] make it particularly suitable for high-risk settings where explanation fidelity is non-negotiable.

Critically, all three methods demonstrated partial convergence across representative test images, consistent with recommendations from Tjoa and Guan [16] and Markus et al.'s [19] argument that cross-method agreement serves as an informal indicator of explanation reliability.

The broader publication landscape corroborates the increasing prominence of XAI in radiology. Bibliometric analysis [21] reveals a substantial increase in radiology- focused XAI publications from 2018 onward, peaking in 2021–2022 (Figures 10 and 11). This is further reflected in Google Trends data (Figures 12), with both *XAI in Healthcare* and *Explainable AI in Healthcare* showing steep upward trends from 2022 to 2025.

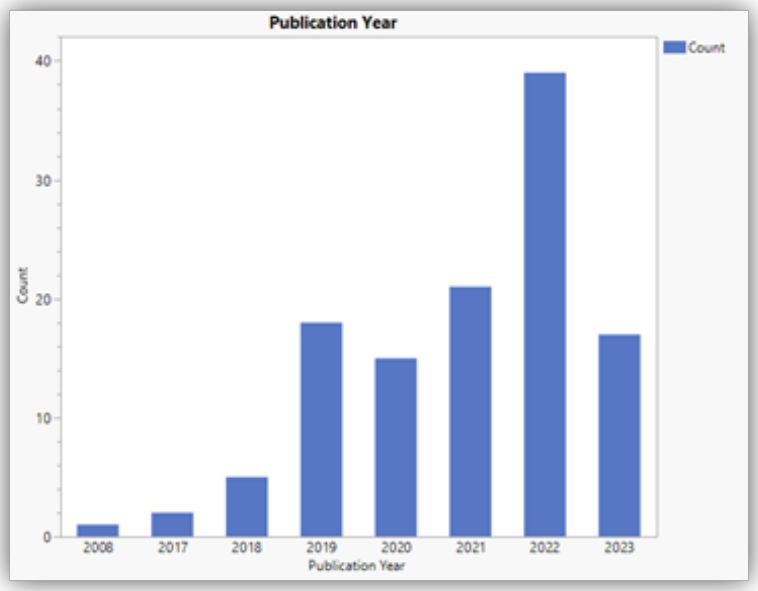


Figure 10: Annual publication counts for XAI research in Radiology, Nuclear Medicine, and Medical Imaging (2008–2023). A marked acceleration is visible from 2018, with a peak of 39 publications in 2022. Data source: [21].

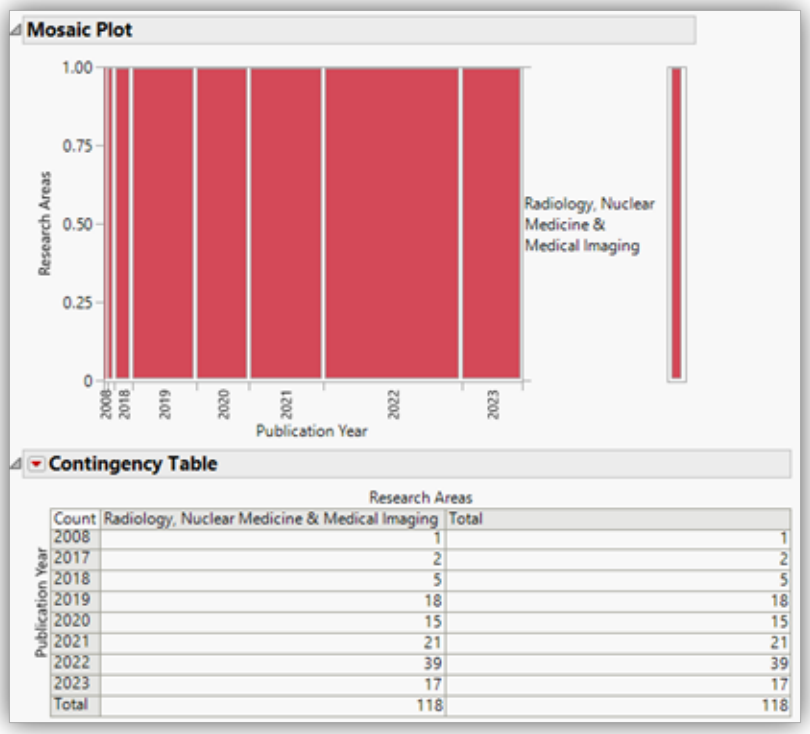
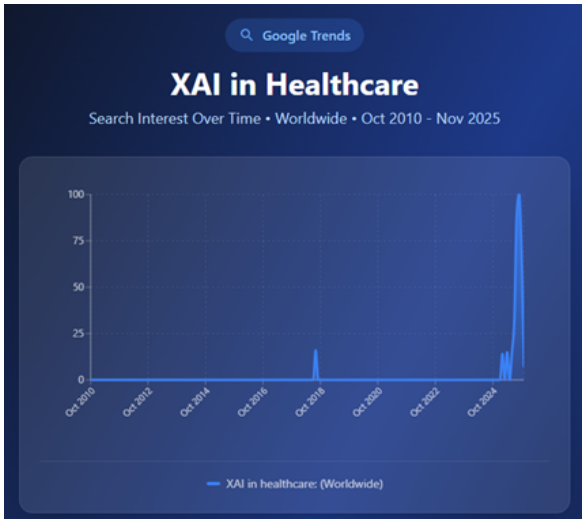
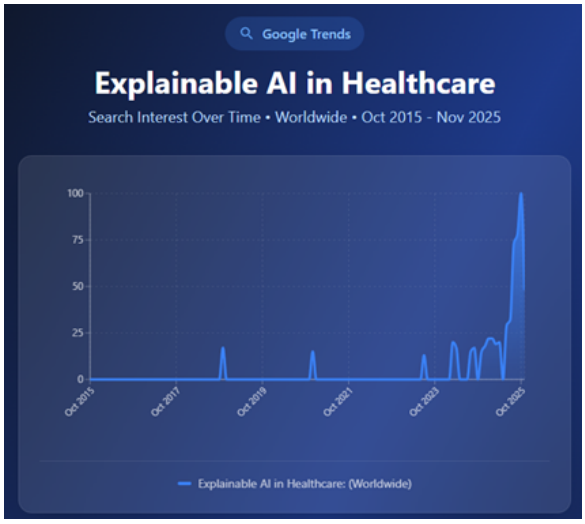


Figure 11: Mosaic plot and contingency table of XAI publications by year in the Radiology, Nuclear Medicine, and Medical Imaging research area. Total: 118 publications. Data source: [21].



(a) "XAI in Healthcare" (Oct 2010 – Nov 2025).



(b) "Explainable AI in Healthcare" (Oct 2015 – Nov 2025).

Figure 12: Google Trends worldwide search interest for XAI-related healthcare queries. Both terms exhibit a dramatic surge from 2023 onward, corroborating the bibliometric evidence of rapidly growing community interest in medical AI explainability.

Limitations. Reliance on a single CNN architecture (ResNet18) limits generalizability to other model families. The binary classification task does not fully capture the complexity of multi-label or multi-pathology interpretation in clinical practice. The RSNA dataset may contain hidden biases attributable to variability across institutions, imaging devices, and patient populations. Finally, Grad-CAM, LIME, and SHAP remain post-hoc techniques subject to documented concerns regarding fidelity, stability, and potential for misleading interpretations [16, 18, 19].

6 Conclusion

This paper presented a systematic, reproducible comparison of three post-hoc XAI methods (Grad-CAM, LIME, and SHAP) applied to a ResNet18 classifier trained for pneumonia detection on the RSNA chest radiograph dataset. Despite operating on distinct theoretical foundations, all three methods exhibit meaningful convergence in their identification of clinically plausible pulmonary regions. This multi-method triangulation provides greater interpretive confidence than any single technique alone, offering empirical support for the growing consensus that robust clinical interpretability requires complementary explanation frameworks.

Future work may productively explore: (1) intrinsic or self-explainable architectures that embed interpretability into model design [18]; (2) extension to multi-label and multi-pathology detection tasks; (3) standardized quantitative metrics for explanation fidelity and stability [19]; and (4) clinician-focused user studies evaluating the practical utility of XAI explanations in diagnostic workflows. As AI continues to integrate into healthcare, multi-method explainability will remain central to ensuring that diagnostic models are not only accurate but interpretable, reliable, and ethically grounded.

Acknowledgements

The author thanks Professor Chon Ho Alex Yu (Ph.D., D.Phil.) at Hawaii Pacific University for supervision and guidance during the original capstone project from which this work derives. The RSNA Pneumonia Detection Challenge dataset was made available through Kaggle under the terms of the RSNA challenge.

Data and Code Availability

The RSNA Pneumonia Detection Challenge dataset is publicly available at https://www.kaggle.com/competitions/rsna-pneumonia-detection-challenge. The *medxai* experimental codebase is available upon reasonable request to the author.

References

[1] Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., ... Ng, A. Y. (2017). CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. *arXiv*:1711.05225. https://arxiv.org/abs/1711.05225

[2] Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., ... Webster, D. R. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22), 2402–2410. https://pubmed.ncbi.nlm.nih.gov/27898976/

[3] Troeger, C., Blacker, B., Khalil, I. A., Rao, P. C., Cao, J., Zimsen, S. R., ... Reiner, R. C. (2018). Estimates of the global, regional, and national burden of lower respiratory infections, 1990–2016. *The Lancet Infectious Diseases*, 18(11), 1191–1210.

[4] Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020). Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Medical Informatics and Decision Making*, 20(1), 310. https://pubmed.ncbi.nlm.nih.gov/33256715/

[5] Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). What clinicians want: Contextualizing explainable machine learning for clinical end use. *Proceedings of the 4th Machine Learning for Healthcare Conference*. https://proceedings.mlr.press/v106/tonekaboni19a/tonekaboni19a.pdf

[6] Oakden-Rayner, L., Dunnmon, J., Carneiro, G., & Ré, C. (2020). Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. *Proceedings of the ACM Conference on Health, Inference, and Learning*. https://dl.acm.org/doi/pdf/10.1145/3368555.3384468

[7] Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs. *PLOS Medicine*, 15(11), e1002683.

[8] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-CAM: Visual explanations from deep networks via gradient-based localization.

International Journal of Computer Vision, 128, 336–359. https://arxiv.org/pdf/1610.02391

[9] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://arxiv.org/pdf/1602.04938

[10] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems* (NeurIPS), 30. https://arxiv.org/pdf/1705.07874

[11] Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. *Communications of the ACM*, 63(12), 54–63. https://dl.acm.org/doi/pdf/10.1145/3381831

[12] Patterson, D., et al. (2021). Carbon emissions and large neural network training. *arXiv*:2104.10350. https://arxiv.org/pdf/2104.10350

[13] Luccioni, A. S., Viguier, S., & Ligozat, A.-L. (2023). Counting carbon: A survey of factors influencing the emissions of machine learning. *arXiv*:2302.08476. https://arxiv.org/pdf/2302.08476

[14] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. https://arxiv.org/pdf/1512.03385

[15] Yang, S., Xiao, W., Zhang, M., Guo, S., Zhao, J., & Shen, F. (2022). Image data augmentation for deep learning: A survey. *arXiv*:2204.08610. https://arxiv.org/pdf/2204.08610

[16] Tjoa, E., & Guan, C. (2021). A survey on explainable artificial intelligence (XAI): Toward medical XAI. *IEEE Transactions on Neural Networks and Learning Systems*, 32(11), 4793–4813. https://ieeexplore.ieee.org/document/9233366

[17] Chaddad, A., Hu, Y., Wu, Y., Wen, B., & Kateb, R. (2025). Generalizable and explainable deep learning for medical image computing: An overview. *Current Opinion in Biomedical Engineering*, 100567. https://arxiv.org/pdf/2503.08420

[18] Hou, J., Liu, S., Bie, Y., Wang, H., Tan, A., Luo, L., & Chen, H. (2024). Self-explainable AI for medical image analysis: A survey and new outlooks. *arXiv*:2410.02331. https://arxiv.org/html/2410.02331v1

[19] Markus, A. F., Kors, J. A., & Rijnbeek, P. R. (2021). The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics*, 113, 103655. https://doi.org/10.1016/j.jbi.2020.103655

[20] Sun, Q., Akman, A., & Schuller, B. W. (2025). Explainable artificial intelligence for medical applications: A review. *ACM Transactions on Computing for Healthcare*. https://arxiv.org/pdf/2412.01829

[21] Alghamdi, S., Mehmood, R., Alqurashi, F., & Alzahrani, A. (2025). Explainable AI (XAI) and Interpretable Machine Learning (IML) in Healthcare Dataset. *Mendeley Data*, V1. https://doi.org/10.17632/5tcdzzsmx8.1

[22] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems* (NeurIPS), 32. https://arxiv.org/abs/1912.01703

[23] Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., ... Reblitz-Richardson, O. (2020). Captum: A unified and generic model interpretability library for PyTorch. *arXiv*:2009.07896. https://arxiv.org/abs/2009.07896

[24] NumPy Developers. (2024). For downstream package authors — NumPy 2.0-specific advice. *NumPy Documentation*. https://numpy.org/devdocs/dev/depending_on_numpy.html

[25] pydicom Developers. (2024). `pydicom.pixels.apply_voi_lut`: Apply a VOI LUT or windowing operation. *pydicom API Reference*. https://pydicom.github.io/pydicom/dev/reference/generated/pydicom.pixels.apply_voi_lut.html

[26] PyTorch Contributors. (2024). Torchvision models — Pre-trained weights & input normalisation. *PyTorch Documentation*. https://docs.pytorch.org/vision/stable/models.html

[27] Radiological Society of North America. (2018). RSNA Pneumonia Detection Challenge. *Kaggle*. https://www.kaggle.com/competitions/rsna-pneumonia-detection-challenge