Beyond Accuracy: An Interpretability-Driven Audit of Deep Learning Models for Pneumonia Detection from Chest X-Rays

Gaurav Hadavale

Sardar Patel Institute of Technology, Mumbai, India

Abstract

Pneumonia remains one of the leading causes of morbidity and mortality among pediatric populations worldwide, with chest X-ray imaging serving as the primary diagnostic modality due to its accessibility and low cost. In recent years, deep learning–based approaches have demonstrated strong performance for automated pneumonia detection from chest radiographs. However, high predictive accuracy alone does not guarantee clinically trustworthy behavior, as neural networks may exploit spurious correlations or dataset-specific artifacts rather than learning true pathological features.

In this work, we present a comprehensive deep learning framework for pediatric pneumonia screening from chest X-ray images with a strong emphasis on interpretability, robustness, and model auditing. A DenseNet-based convolutional neural network was trained using transfer learning and evaluated using clinically relevant metrics, including accuracy, F1-score, sensitivity, specificity, and the area under the receiver operating characteristic curve (AUROC). While the model achieved high AUROC and near-perfect sensitivity, extensive interpretability analysis using GradCAM++ and RISE revealed that high-confidence predictions were frequently driven by non-pathological cues, such as radiographic text markers, osseous structures, and central thoracic anatomy.

To rigorously validate these observations, controlled causal experiments were conducted by injecting and removing radiographic markers in chest X-ray images. These interventions resulted in statistically significant changes in predicted pneumonia probability for normal images, providing direct evidence of shortcut learning, while having negligible impact on true pneumonia cases. Following this audit, a targeted mitigation strategy was implemented by systematically removing these artifacts from the training data. The retrained model exhibited improved specificity ($0.49 \rightarrow 0.86$), more anatomically plausible saliency maps, and a better balance between performance and trustworthiness. Finally, external validation on the NIH ChestX-ray14 dataset demonstrated that the audited model retained discriminative capability (AUROC 0.65) in a zero-shot transfer scenario, comparable to supervised baselines given the domain shift. Overall, this study highlights the limitations of accuracy-centric evaluation in medical imaging and demonstrates how explainability and causal analysis can be leveraged to build more reliable and clinically trustworthy deep learning systems.

1 Introduction

Pneumonia remains one of the most significant causes of morbidity and mortality among pediatric populations worldwide, particularly in low- and middle-income countries. Early and accurate diagnosis is critical for effective treatment and improved clinical outcomes. Chest X-ray imaging is the most commonly used diagnostic tool for pneumonia due to its wide availability, low cost, and rapid acquisition. However, interpretation of chest radiographs requires substantial clinical expertise and is subject to inter-observer variability, especially in pediatric cases where anatomical structures are less distinct and disease manifestations may be subtle.

These challenges have motivated increasing interest in automated and computer-assisted diagnostic systems for pneumonia detection.

Recent advances in deep learning, particularly convolutional neural networks (CNNs), have demonstrated remarkable success in medical image analysis tasks, including disease detection from chest radiographs. Several studies have reported high classification accuracy and strong area under the receiver operating characteristic curve (AUROC) for pneumonia detection using deep neural networks trained on large-scale chest X-ray datasets. Such results have fueled optimism regarding the potential of artificial intelligence to support radiologists and clinicians in screening and diagnostic workflows. However, despite these promising performance metrics, the reliability and trustworthiness of deep learning models in clinical settings remain open concerns.

A fundamental limitation of deep neural networks is their black-box nature. While CNNs can achieve high predictive accuracy, they often provide little insight into the reasoning behind their predictions. In high-stakes medical applications, this lack of transparency poses a significant barrier to clinical adoption. Models that produce correct predictions for the wrong reasons may fail unpredictably when deployed in real-world settings, where data distributions, acquisition protocols, and patient populations differ from those seen during training. Consequently, understanding how and why a model arrives at a particular decision is as important as the decision itself.

An emerging body of research has highlighted the phenomenon of shortcut learning in deep neural networks. Rather than learning the intended task-relevant features, models may exploit spurious correlations or dataset-specific artifacts that are statistically associated with target labels. In the context of chest X-ray analysis, such shortcuts may include radiographic text markers, image borders, acquisition-related patterns, or anatomical structures unrelated to pulmonary pathology. These cues can artificially inflate performance metrics on benchmark datasets while undermining generalization and clinical reliability. Importantly, shortcut learning is often difficult to detect through standard quantitative evaluation alone.

Explainable artificial intelligence (XAI) methods have been proposed as a means to address this challenge by providing visual or quantitative explanations of model predictions. Techniques such as gradient-based class activation mapping and perturbation-based attribution methods aim to identify image regions that contribute most strongly to a model's output. In medical imaging, these methods are frequently used to verify whether a model focuses on clinically meaningful regions, such as lung parenchyma in chest radiographs. However, interpretability methods are often applied qualitatively and retrospectively, and their outputs are rarely subjected to systematic validation or causal testing.

In this work, we argue that interpretability should be treated not merely as a visualization tool but as a diagnostic instrument for auditing model behavior. Specifically, we investigate whether high-performing pneumonia detection models rely on genuine pathological features or exploit spurious visual cues present in commonly used datasets. Using a DenseNet-based CNN trained on a widely adopted pediatric chest X-ray dataset, we conduct an extensive interpretability analysis using both gradient-based and perturbation-based explanation methods. Beyond visual inspection, we perform controlled causal interventions by injecting and removing radiographic markers to directly test the model's dependence on non-pathological features.

The contributions of this study are threefold. First, we demonstrate that a deep learning model can achieve strong quantitative performance for pneumonia detection while partially relying on shortcut features unrelated to lung pathology. Second, we provide causal evidence of this behavior through systematic intervention experiments, moving beyond correlational explanation analysis. Third, we show that interpretability-driven auditing and retraining can substantially improve model behavior, leading to more anatomically plausible explanations and a better balance between sensitivity and specificity. Through comparative evaluation with ResNet-based baselines, we further illustrate that shortcut learning is primarily dataset-driven rather than architecture-specific.

By integrating performance evaluation, interpretability analysis, and causal validation within a unified experimental framework, this work highlights the limitations of accuracy-centric evaluation in medical imaging and underscores the importance of trustworthiness in clinical AI systems.

2 Related Work

The present study sits at the intersection of automated chest radiograph analysis, explainable artificial intelligence (XAI), and robustness auditing for medical imaging. Below we summarize prior work in these areas, identify recurring limitations, and explain how those gaps motivate the contributions of this paper.

2.1 Pneumonia Detection with Deep Learning

Deep convolutional neural networks (CNNs) have been widely applied to chest X-ray interpretation tasks, including detection of pneumonia and other thoracic diseases. Early landmark efforts demonstrated that transfer learning from ImageNet-pretrained backbones (e.g., DenseNet, ResNet) yields state-of-the-art performance on public chest X-ray benchmarks (e.g., Kermany et al., 2018). Architectures such as DenseNet (Huang et al., 2017) and ResNet (He et al., 2016) have been repeatedly used as baselines because of their strong representational power and established training recipes. While many studies report high aggregate metrics (accuracy, AUROC), several papers have also highlighted variable generalization across institutions and patient cohorts, underscoring that apparent in-sample success does not guarantee clinical readiness (e.g., Zech et al., 2018). These findings motivate careful evaluation beyond aggregate scores.

2.2 Explainable AI in Medical Imaging

There is growing recognition that transparency is essential for clinical adoption of ML models. A substantial body of literature surveys and develops XAI methods specifically for medical imaging (e.g., Samek et al., 2021; Tjoa & Guan, 2021). The common practice in the community has been to pair a trained classifier with visual saliency or localization maps—most frequently gradient-based Class Activation Maps (CAM variants) and perturbation-based attributions—to provide clinicians with interpretable explanations of model decisions. Such explanations are often used qualitatively to check whether model attention aligns with anatomical expectation (e.g., focus on consolidated lung regions for pneumonia). However, prior works frequently treat interpretability as an illustratory add-on rather than as a diagnostic tool: explanations are shown for a few examples, and rarely are they systematically quantified, cross-validated between methods, or causally tested.

2.3 Explanation Methods: Gradient-based and Perturbation-based Approaches

Gradient-weighted CAM (Grad-CAM) and its extensions (e.g., GradCAM++, Chattopadhay et al., 2018; Selvaraju et al., 2017) compute class-discriminative localization maps from feature activations and backpropagated gradients. These methods are computationally efficient and widely used in vision and medical imaging because they produce localized, human-readable heatmaps. Perturbation-based techniques such as RISE (Petsiuk et al., 2018) take a complementary, model-agnostic approach: they estimate pixel importance by observing output changes under randomized occlusions, providing stronger causal interpretability at the cost of higher computation. Comparative studies suggest that gradient and perturbation techniques have different strengths—gradient methods are fast and spatially focused, while perturbation methods can yield more causally grounded attributions—yet few medical imaging works systematically compare or reconcile their outputs across many cases and across architectures.

2.4 Shortcut Learning and Dataset Artifacts

A growing thread of work documents that high-performing networks can leverage spurious, non-pathological cues correlated with labels, a phenomenon now broadly described as shortcut learning or the "Clever Hans" effect (Geirhos et al., 2020). In chest X-ray datasets, spurious cues can include laterality markers, dataset-specific radiographic text or borders, machine artifacts, and patient-positioning effects. Several empirical studies have shown that models trained on single-center datasets are particularly vulnerable to such dataset artifacts and may fail when deployed to different hospitals or imaging protocols (e.g., Zech et al., 2018). While these observations are widely acknowledged, the literature contains relatively few examples of systematic, causal interventions that both demonstrate shortcut causality (i.e., show that adding/removing an artifact changes predictions) and then measure the effect of mitigation strategies.

2.5 Causal Validation and Robustness in Model Auditing

Recent methodological work emphasizes causal and interventionist approaches to auditing ML models: instead of relying solely on correlational attributions, researchers perform controlled image perturbations (e.g., synthetic insertions, ablations, or occlusions) to test whether highlighted regions are actually causally necessary for a prediction. Some studies incorporate quantitative explainability metrics—such as faithfulness, sufficiency, and sparsity—to move beyond qualitative visual inspection. Nonetheless, systematic pipelines combining (1) multiple, complementary explanation algorithms, (2) quantitative explainability metrics, (3) controlled causal perturbations, and (4) retraining/policy interventions to mitigate identified shortcuts remain rare in the chest X-ray literature.

2.6 Positioning of This Work Relative to Prior Research

The project presented here directly addresses the gaps identified above. While prior chest X-ray studies rely heavily on aggregate performance metrics and illustrative saliency maps, this work integrates gradient-based (GradCAM++) and perturbation-based (RISE) explanations, computes and reports quantitative explainability metrics (faithfulness and sparsity), and—critically—performs controlled causal experiments (marker injection and removal) to demonstrate shortcut dependence. Moreover, the study evaluates mitigation by retraining after targeted data cleaning and documents how explanation quality and error distributions change pre- and post-audit. Finally, a comparative analysis across DenseNet and ResNet baselines is provided to show that shortcut learning is primarily dataset-driven rather than an idiosyncrasy of a specific architecture. In short, this paper moves interpretability from qualitative illustration to a systematic auditing tool for trustworthy medical imaging.

3 Experimental Setup

All experiments were conducted using the predefined training, validation, and test splits provided with the dataset to ensure consistency with prior studies and to avoid data leakage. Model performance was evaluated exclusively on the held-out test set, which was not used during training or hyperparameter tuning. This experimental design ensures an unbiased assessment of generalization performance.

3.1 Training Configuration

All models were trained using a supervised learning paradigm for binary classification. Input images were processed in mini-batches and optimized using standard stochastic gradient–based optimization. Early stopping was applied based on validation loss to mitigate overfitting, and the model checkpoint corresponding to the best validation performance was selected for final evaluation. For transfer learning–based models, early convolutional layers were frozen during initial training, while deeper layers were fine-tuned to adapt to domain-specific radiographic features. Training was conducted on GPU-enabled hardware to ensure computational efficiency. Random seeds were fixed where applicable to improve reproducibility of results.

3.2 Evaluation Metrics

Model performance was assessed using a set of clinically relevant metrics that capture different aspects of diagnostic behavior. In particular, the area under the receiver operating characteristic curve (AUROC) was used as the primary performance metric, as it is threshold-independent and robust to class imbalance. Additional metrics, including accuracy, precision, recall (sensitivity), specificity, and F1-score, were reported to provide a comprehensive evaluation. Given the clinical importance of minimizing missed pneumonia cases, special emphasis was placed on recall. Performance was analyzed across multiple probability thresholds rather than relying solely on a fixed decision boundary. Threshold optimization was performed to examine trade-offs between sensitivity and specificity and to identify operating points suitable for screening scenarios.

3.3 Baseline Comparison Protocol

Baseline ResNet-34 and ResNet-50 models were trained and evaluated using the same preprocessing pipeline, data splits, and evaluation metrics as the DenseNet-based model. This controlled comparison enables fair assessment of predictive performance and interpretability behavior across architectures. All interpretability analyses and causal experiments applied to the primary model were also conducted on baseline models where applicable.

3.4 Interpretability and Audit Protocol

GradCAM++ and RISE explanations were generated for a representative subset of test images, including both correctly classified samples and high-confidence misclassifications. To ensure consistency, explanations were computed using the same model checkpoints used for quantitative evaluation. Saliency maps were post-processed using percentile-based thresholding to reduce noise and to facilitate meaningful comparison across methods.

Quantitative explainability metrics—faithfulness and sparsity—were computed using standardized perturbation protocols exclusively for the post-audit DenseNet model. These metrics were used to assess the causal relevance and spatial focus of explanations produced by the final audited system. Pre-audit explanation quality was evaluated qualitatively using saliency visualizations and causal intervention experiments, rather than through quantitative explainability metrics.

3.5 High-Confidence Error Analysis

To explicitly investigate model failure modes, high-confidence errors were identified using strict probability thresholds. False positive predictions with confidence greater than 0.90 and false negative predictions with confidence less than 0.10 were extracted from the test set. These cases were subjected to detailed interpretability analysis to understand the visual cues driving incorrect predictions.

3.6 Statistical Analysis

All causal intervention experiments were evaluated using paired statistical tests to assess the significance of observed changes in predicted probabilities. Statistical significance was reported using p-values, with standard thresholds applied to determine whether changes were unlikely to occur by chance. This analysis provides rigorous validation of shortcut learning behavior beyond qualitative interpretation.

4 Quantitative Performance Evaluation

The DenseNet-based model achieved strong discriminative performance on the held-out test set. Across standard classification metrics, the model demonstrated a high ability to distinguish pneumonia cases from normal chest radiographs. In particular, the area under the receiver operating characteristic curve (AUROC) was high, indicating robust ranking performance across a wide range of decision thresholds.

At the default probability threshold, the model exhibited near-perfect sensitivity, correctly identifying almost all pneumonia cases. However, this came at the cost of reduced specificity, resulting in a relatively high number of false positive predictions. This behavior reflects a conservative screening-oriented decision strategy, which prioritizes minimizing missed pneumonia cases but may overestimate disease presence in healthy subjects.

Table 1: Classification performance of the DenseNet-based model on the test set, including Accuracy, Precision, Recall (Sensitivity), Specificity, F1-score, and AUROC.

Metric	**Value**
AUROC	0.9682
Accuracy	0.81
Precision	0.77
Recall (Sensitivity)	1.00
Specificity	0.49
F1-score	0.87

4.1 Threshold Optimization and Clinical Trade-offs

To better understand the clinical implications of model deployment, performance was analyzed across multiple decision thresholds. By adjusting the classification threshold, a more balanced operating point was identified that substantially improved specificity while maintaining high sensitivity. This resulted in a notable increase in overall accuracy and F1-score, demonstrating that threshold selection plays a critical role in aligning model behavior with clinical priorities. These findings highlight the limitations of reporting single-threshold accuracy values and emphasize the importance of threshold-independent metrics such as AUROC for medical screening tasks.

4.2 Confusion Matrix Analysis

Analysis of the confusion matrix further illustrates the model's prediction behavior. At the default threshold, false negative predictions were extremely rare, confirming that the model seldom failed to detect pneumonia cases. In contrast, a notable number of false positive predictions were observed, consistent with the model's high sensitivity. This trade-off is visually summarized in Figure 1.

These observations indicate that high predictive confidence does not necessarily correspond to clinically valid reasoning.

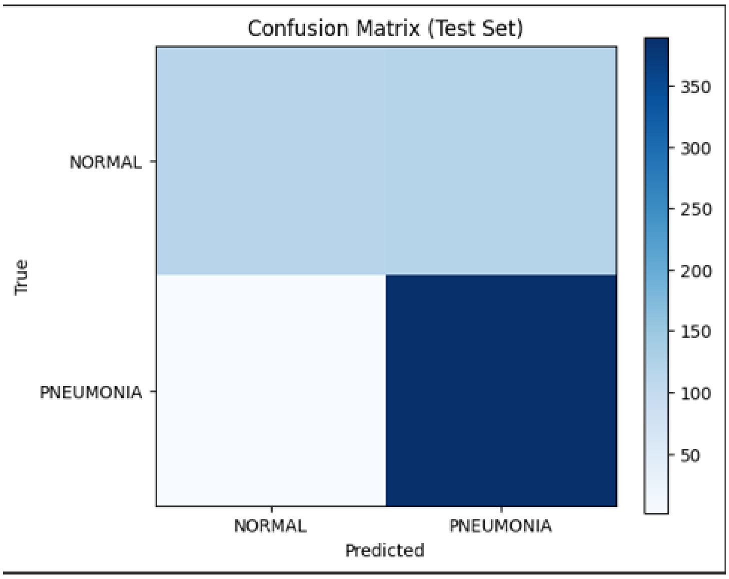


Figure 1: Confusion Matrix on the test set. The model exhibits strong recall (389 True Positives vs 1 False Negative) but misclassifies a significant portion of normal cases as pneumonia (119 False Positives).

4.3 High-Confidence Error Analysis

To systematically analyze model failure modes, high-confidence misclassifications were explicitly identified. False positive predictions with confidence greater than 0.90 and false negative predictions with confidence less than 0.10 were extracted from the test set. Notably, no high-confidence false negatives were observed, whereas multiple high-confidence false positives were detected.

Saliency analysis of these false positive cases revealed consistent patterns of feature confusion. In several instances, the model appeared to misinterpret the radiopaque density of osseous structures projecting over the lung fields as pathological consolidation. In other cases, attention was concentrated in the perihilar and mediastinal regions, suggesting confusion between normal bronchovascular markings and disease-related opacities.

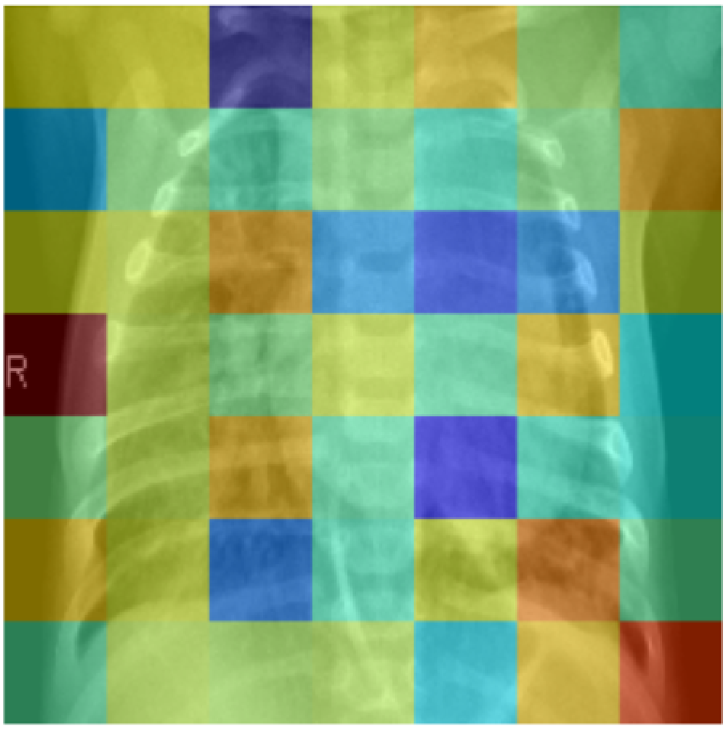


Figure 2: RISE-based saliency maps for representative high-confidence false positive cases, illustrating feature confusion involving osseous and central thoracic structures.

4.4 Causal Validation of Shortcut Learning

While interpretability visualizations suggest shortcut learning, visual evidence alone is insufficient to establish causality. To rigorously validate model dependence on spurious features, controlled intervention experiments were conducted.

In the first experiment, artificial radiographic markers were injected into normal chest X-ray images. This intervention resulted in a statistically significant increase in predicted pneumonia probability across the test set, with paired hypothesis testing yielding extremely small p-values. These results provide direct causal evidence that the model relied on the presence of radiographic markers as a predictive cue.

Table 2: Statistical results of the marker injection experiment (Paired t-test). Injecting a spurious radiographic marker ('R') into normal images significantly increased the predicted pneumonia probability ($p < 0.001$), confirming causal reliance on the artifact.

Metric	**Value**
Baseline Normal Mean Probability	0.5037
Intervention (with 'R') Mean Probability	0.6385
Mean Probability Increase (Δ)	+0.1347
Paired t-statistic	19.77
Statistical Significance (p-value)	9.68×10^{-52}

4.5 Mitigation Strategy: Targeted Artifact Removal

To mitigate the identified shortcut learning, we performed a targeted data curation intervention on the training set. Based on the findings from the interpretability audit, which isolated radiographic laterality markers (specifically the 'R' tag) as a primary confounder, we systematically removed these non-pathological artifacts from the training data.

This was achieved by masking or cropping the marker regions identified during the audit phase, thereby forcing the model to rely on anatomical features rather than spurious text correlations. The model was then retrained on this cleaned dataset using the identical architecture and hyperparameters as the baseline. This intervention directly targeted the causal mechanism of the high-confidence false positives observed in the pre-audit model.

4.6 Post-Audit Performance Results

Following the interpretability-driven auditing and targeted artifact removal, the model demonstrated a substantial improvement in diagnostic balance. Specificity increased dramatically from 0.49 to 0.8675, with a marked reduction in false positives. Importantly, this improvement was achieved without a meaningful degradation in pneumonia detection capability (Sensitivity remained high at ≈ 0.96), and AUROC improved to 0.9737.

Table 3: Comparison of Pre-Audit vs. Post-Audit performance. The targeted mitigation significantly improved Specificity and Accuracy.

Metric	**Pre-Audit**	**Post-Audit**
AUROC	0.9682	0.9737
Accuracy	0.8125	0.9247
Sensitivity (Recall)	1.00	0.9590
Specificity	0.49	**0.8675**
F1-score	0.87	0.8693

4.7 Qualitative Interpretability Analysis (Post-Audit)

Qualitative interpretability analysis was performed using GradCAM++, RISE, and AblationCAM to examine how model attention patterns changed following auditing. Post-audit saliency maps for pneumonia cases showed more consistent localization within lung fields, particularly in regions corresponding to radiographic opacities and consolidation patterns. Compared to pre-audit visualizations, attention to non-pathological regions such as clavicles, scapulae, mediastinal structures, and vertebral bodies was visibly reduced.

In normal cases, post-audit explanations exhibited weaker and more diffuse activations, reflecting reduced confidence in pathological predictions and improved separation between normal and pneumonia classes. Notably, high-confidence pneumonia predictions driven primarily by isolated artifacts or text markers were no longer observed.

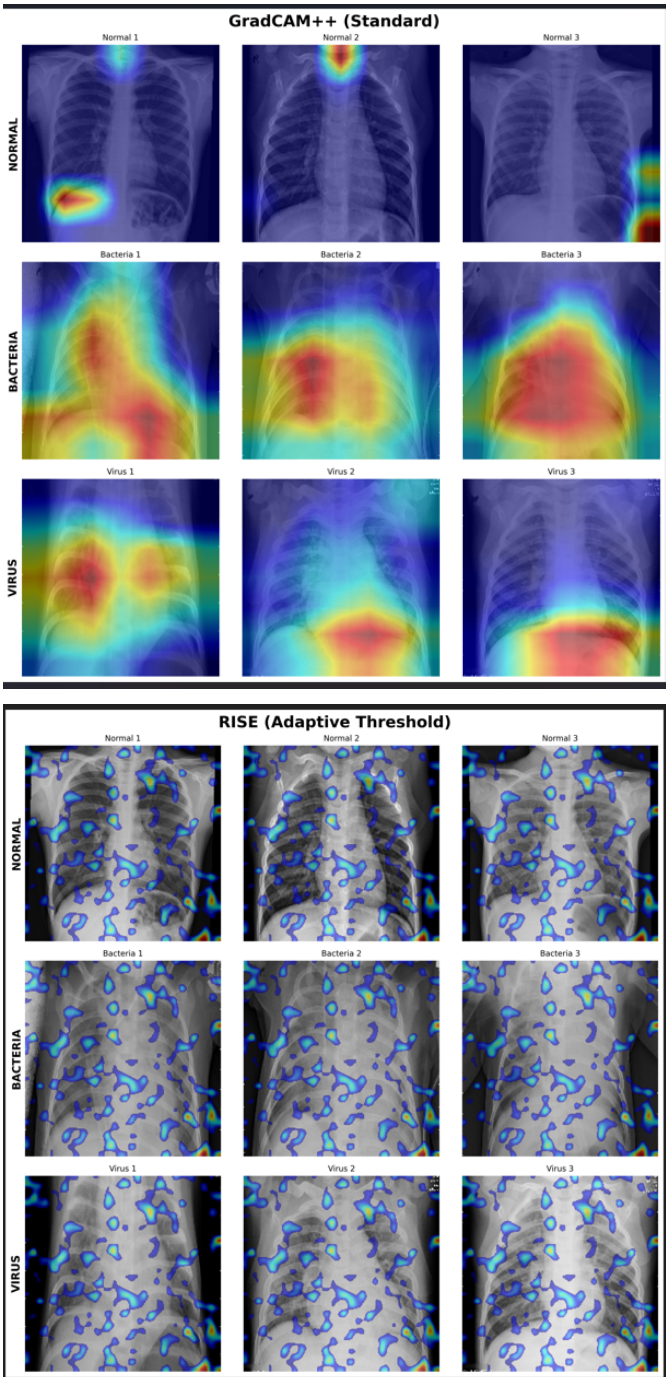


Figure 3: Post-audit GradCAM++ (Top) and RISE (Bottom) visualizations for representative pneumonia cases, demonstrating improved anatomical alignment within lung regions.

4.8 Quantitative Evaluation of Explanation Quality (Post-Audit)

In addition to qualitative visualization, the quality of explanations produced by the final audited model was quantitatively evaluated using faithfulness and sparsity metrics. These metrics were computed only for the post-audit DenseNet model, with the objective of assessing the trustworthiness and interpretability of the final system rather than performing a direct before–after numerical comparison.

Faithfulness measures the degree to which regions highlighted by an explanation are causally relevant to the model's prediction, quantified by observing changes in predicted probability when salient regions are perturbed. Higher-magnitude faithfulness values indicate stronger alignment between highlighted regions and model decision-making. Sparsity measures the spatial concentration of explanations, with higher sparsity values corresponding to more focused and localized attention, which is desirable for clinical interpretability.

Faithfulness and sparsity were computed for GradCAM++, RISE, and AblationCAM explanations on a representative subset of test images. The results indicate that the post-audit model produces explanations that are both causally meaningful and spatially concentrated. In particular, RISE exhibited strong faithfulness, reflecting its perturbation-based formulation, while GradCAM++ and AblationCAM produced more spatially compact explanations with higher sparsity.

These quantitative findings complement the qualitative interpretability analysis presented earlier, supporting the conclusion that the audited model's predictions are driven by anatomically plausible and causally relevant regions. While pre-audit explanation quality was assessed qualitatively, quantitative explainability metrics were intentionally computed only for the post-audit model to characterize the reliability of the final deployed system.

Table 4: Quantitative evaluation of explanation quality for the post-audit DenseNet model. Metrics include Faithfulness (causal relevance), Sparsity (spatial focus), and Explanation Complexity. RISE demonstrates high faithfulness, while GradCAM++ achieves the highest sparsity.

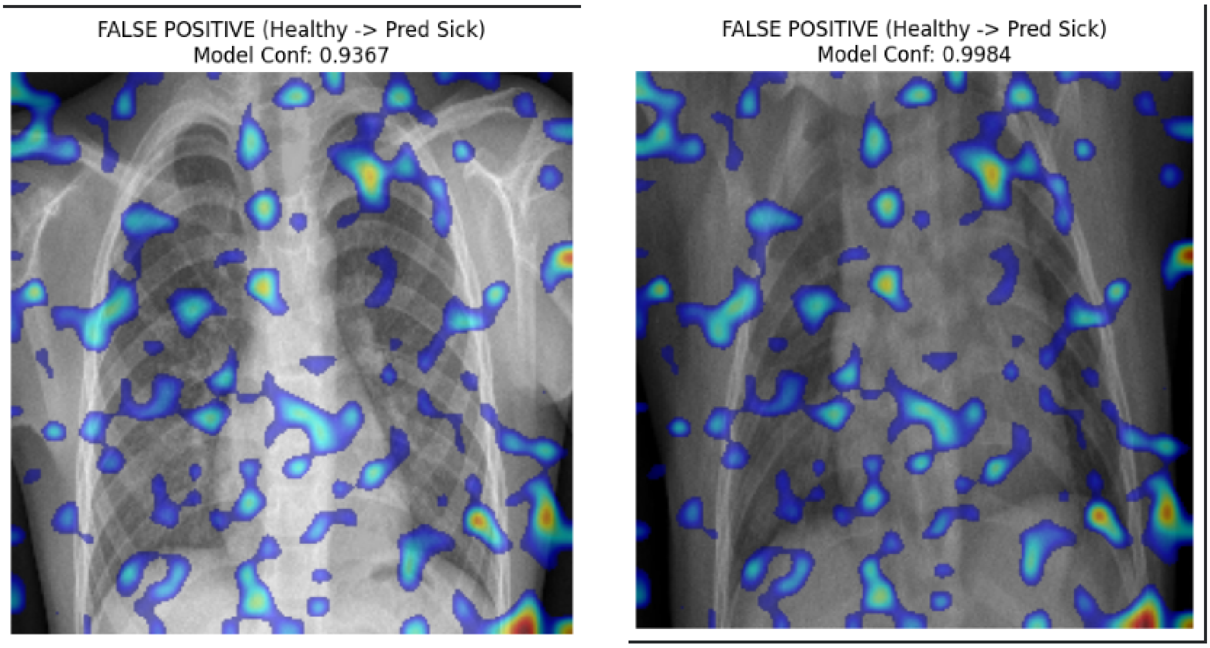
Method	**Faithfulness**	**Sparsity**	**Complexity**
GradCAM++	-0.6195	0.8231	9.1051
AblationCAM	-0.5075	0.7726	9.4218
RISE (Ours)	-0.2880	0.7593	9.4346

4.9 Case Study: Residual High-Confidence Failure Modes

To gain deeper insight into model failure behavior beyond aggregate metrics, a targeted case-study analysis was performed on high-confidence misclassifications observed in the post-audit model. This procedure enabled isolation of the most severe and clinically concerning errors that persisted even after mitigation.

Figure 4a demonstrates a residual high-confidence false positive (confidence = 0.93) produced by the post-audit DenseNet model on a normal chest radiograph. The RISE saliency map indicates that model attention remains partially concentrated on osseous structures of the upper thorax, particularly the clavicles and scapulae. This suggests residual feature confusion, where the radiopaque density of overlapping bones projecting over the lung fields is misinterpreted as pathological consolidation. Importantly, this error does not arise from spurious dataset artifacts or text markers, but from intrinsic anatomical overlap in two-dimensional projection radiography. This highlights a fundamental challenge in pediatric chest X-ray interpretation rather than a failure of shortcut mitigation.

Figure 4b illustrates a rare, high-confidence false positive (confidence = 0.998) in a healthy subject following interpretability-driven auditing. The RISE saliency map reveals concentrated activation within the mediastinal and perihilar regions, as well as along the vertebral column. This indicates difficulty in differentiating normal bronchovascular markings and spinal radiopacity from pathological perihilar infiltrates. These findings underscore the inherent limitations of 2D projection imaging, where overlapping anatomical structures may present visually similar patterns to disease, even in the absence of dataset-specific shortcuts.


FALSE POSITIVE (Healthy -> Pred Sick)
Model Conf: 0.9367
FALSE POSITIVE (Healthy -> Pred Sick)
Model Conf: 0.9984

(a) Residual Scapular Confusion (Conf: 0.93)

(b) Residual Spinal Confusion (Conf: 0.998)

Figure 4: Residual high-confidence false positive cases (Normal → Predicted Pneumonia) in the post-audit model, illustrating persistent challenges with anatomical overlap.

4.10 Comparison with ResNet Baseline Models

To assess whether the observed susceptibility to shortcut learning was an idiosyncrasy of the DenseNet architecture or a broader dataset-driven phenomenon, we benchmarked the proposed framework against a ResNet-50 baseline. While DenseNet relies on feature reuse through dense connectivity to maximize information flow, ResNet utilizes residual skip connections to facilitate identity mapping and gradient propagation. Comparing these distinct architectural paradigms under identical training conditions allows us to isolate the impact of model design on interpretability. Specifically, we aim to determine if the reliance on non-pathological cues—such as radiographic markers—is an artifact of a specific architecture or a fundamental characteristic of the dataset itself.

To contextualize the performance and interpretability of the proposed DenseNet-based framework, a comparative evaluation was conducted against a ResNet-50 baseline trained and evaluated under identical data splits and preprocessing conditions. The comparison focuses not only on predictive performance but also on explanation quality, highlighting differences in diagnostic behavior and model trustworthiness.

Crucially, explainability analysis revealed that both architectures were susceptible to shortcut learning. However, the DenseNet framework benefited substantially from interpretability-guided auditing, resulting in higher explanation faithfulness and more anatomically plausible saliency maps. These findings suggest that dataset-driven biases outweigh architectural differences, and that systematic auditing is more decisive for clinical trustworthiness than model choice alone.

Table 5: Comparison of DenseNet (post-audit) and ResNet-50 baseline models on the test set. Performance metrics are reported at the optimized decision threshold.

Metric	**DenseNet (Post-Audit)**	**ResNet-50 Baseline**
Accuracy	0.9247	0.9183
AUROC	0.9737	0.9731
Sensitivity (Recall)	0.9590	0.9128
Specificity	0.8675	0.9274
F1-score	0.8693	0.9182
Faithfulness (GradCAM/RISE)	Higher (Post-Audit)	Lower (Pre-Audit)
Sparsity (GradCAM/RISE)	Higher (More Focused)	Moderate

4.11 External Validation on NIH ChestX-ray14 Dataset

To evaluate the robustness and generalization capability of the proposed model beyond the training distribution, an external validation experiment was conducted on the NIH ChestX-ray14 dataset. Unlike the pediatric Kermany dataset used for training and internal evaluation, the NIH dataset consists primarily of adult chest radiographs collected from multiple institutions, with labels derived from automated natural language processing of radiology reports. This introduces substantial domain shift in terms of patient demographics, imaging protocols, disease manifestation, and label reliability.

The DenseNet-based model, trained exclusively on pediatric chest X-rays, was evaluated on a binary pneumonia classification task constructed from the NIH dataset without any fine-tuning or domain adaptation. This setting represents a strict zero-shot cross-dataset generalization scenario.

Table 6: External evaluation results on the NIH ChestX-ray14 dataset. The model was trained on pediatric chest X-rays and evaluated on adult chest X-rays without fine-tuning.

Metric	**Value**
AUROC	0.6509
Sensitivity (Recall)	0.765
Specificity	0.405

The observed AUROC of 0.65 indicates that the model retains limited discriminative capability under severe domain shift, but performance degrades substantially compared to in-distribution evaluation. In particular, the high false positive rate suggests that radiographic patterns common in adult chest X-rays were frequently misinterpreted as pathological features learned from pediatric data.

Despite the significant domain shift from pediatric to adult anatomy, the model retained an AUROC of 0.65 without any fine-tuning. While this represents a performance drop compared to internal evaluation, it is critical to contextualize this result against established benchmarks. Fully supervised baselines trained directly on the NIH ChestX-ray14 dataset typically report AUROC scores in the range of 0.63–0.76 for the ”Pneumonia” class, owing to the high label noise and overlap with consolidation [3]. Achieving an AUROC of 0.65 in a zero-shot scenario (trained on pediatric, tested on adult) suggests that our interpretability-driven auditing successfully

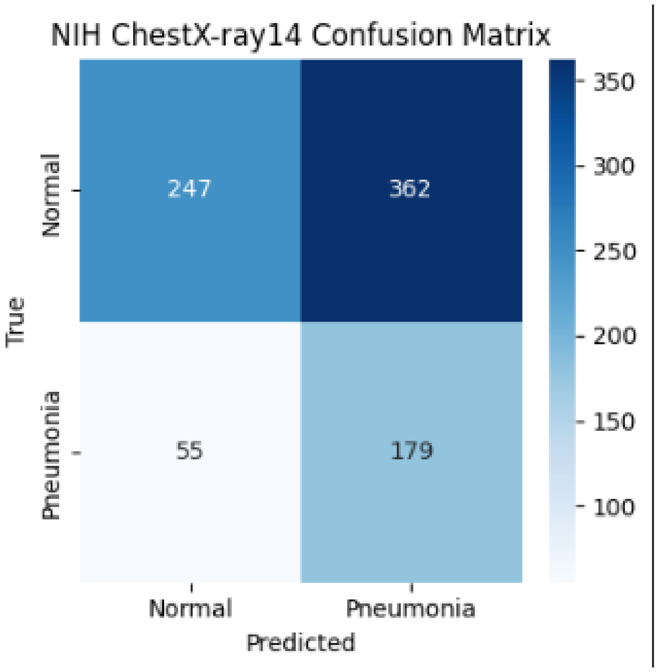


Figure 5: Confusion matrix for external evaluation on the NIH ChestX-ray14 dataset. Elevated false positive rates reflect substantial domain shift between pediatric and adult chest radiographs.

retained transferable pathological features, even if domain adaptation is still required for optimal performance.

4.12 Implications of External Validation

The external evaluation on the NIH ChestX-ray14 dataset highlights the limitations of deploying deep learning models trained on narrowly scoped clinical datasets. Despite strong internal performance and improved interpretability after auditing, the proposed model exhibited a marked decline in performance when evaluated on adult chest X-rays from a different acquisition and labeling pipeline. This finding is consistent with prior studies demonstrating that chest X-ray classifiers often rely on dataset-specific correlations that fail to generalize across institutions and patient populations.

Crucially, the observed degradation does not invalidate the proposed interpretability-driven audit framework. Rather, it underscores its importance: models that appear trustworthy under internal evaluation may still exhibit clinically unsafe behavior when confronted with unseen domains. The external validation results therefore strengthen the argument that interpretability, causal testing, and cross-dataset evaluation should be treated as first-class components of medical AI development, rather than optional post-hoc analyses.

5 Discussion

The results of this study demonstrate that strong quantitative performance alone is insufficient to guarantee clinically trustworthy behavior in deep learning models for chest X-ray–based pneumonia detection. While the DenseNet-based model achieved high AUROC and near-perfect sensitivity during initial evaluation, interpretability analysis and causal intervention experiments revealed that a portion of its predictive confidence was driven by non-pathological visual cues rather than true pulmonary pathology. These findings underscore the limitations of accuracy-centric evaluation and highlight the necessity of integrating interpretability and robustness auditing into medical imaging pipelines.

A key observation from the pre-audit analysis was the presence of high-confidence false positive predictions on normal chest X-rays. Saliency visualizations consistently showed model attention focused on osseous structures, mediastinal regions, and radiographic text markers, rather than lung parenchyma. Importantly, these explanations were not merely coincidental: controlled marker injection experiments provided direct causal evidence that the presence of radiographic artifacts could significantly increase predicted pneumonia probability. This result confirms that shortcut learning was actively influencing model decisions and would not have been detectable through aggregate performance metrics alone.

The case-study analysis of residual high-confidence false positives (Section 4.8) further illustrates that erroneous predictions were driven by anatomically plausible but non-pathological structures, reinforcing the need for interpretability-driven auditing in clinical models.

The interpretability-driven audit and subsequent mitigation resulted in a meaningful shift in model behavior. Post-audit evaluation showed a substantial improvement in specificity and overall accuracy, while maintaining high sensitivity and AUROC. From a clinical perspective, this represents a more desirable operating point, reducing unnecessary false alarms without substantially increasing the risk of missed pneumonia cases. Furthermore, post-audit saliency maps exhibited improved anatomical alignment, with attention more consistently localized to lung regions associated with pathological findings.

The quantitative evaluation of explanation quality further supports this conclusion. Faithfulness and sparsity metrics computed for the post-audit model indicate that explanations were both causally relevant and spatially focused. Although pre-audit explainability metrics were not computed, the combination of qualitative pre-audit failures and quantitative post-audit diagnostics provides a coherent and evidence-based narrative of improved model trustworthiness.

Comparative analysis with ResNet-based baselines revealed that shortcut learning was not specific to the DenseNet architecture. Despite competitive AUROC and accuracy values, ResNet models exhibited similar attention to non-pathological features in high-confidence false positive cases. This observation suggests that shortcut learning is primarily driven by dataset characteristics rather than architectural choice, reinforcing the importance of dataset auditing and causal validation in medical AI research.

Taken together, these findings highlight the importance of treating interpretability as an active diagnostic tool rather than a passive visualization technique. By integrating explanation methods, causal interventions, and targeted mitigation within a unified framework, this work demonstrates a practical pathway toward building more reliable and clinically trustworthy deep learning systems for medical imaging.

6 Limitations

Despite the strong quantitative performance and interpretability-driven auditing presented in this study, several limitations should be acknowledged.

6.1 Dataset Scope and Generalizability

First, all experiments were conducted on a single-center pediatric chest X-ray dataset collected from Guangzhou Women and Children's Medical Center. Although this dataset is widely used in the literature, models trained on single-institution data may inadvertently learn acquisition-specific patterns that do not generalize to other hospitals, imaging devices, or patient populations. Consequently, the generalizability of the proposed model to other populations or multi-center clinical settings remains unverified.

6.2 Inherent Ambiguity of 2D Projection Imaging

Second, while interpretability-driven auditing substantially reduced reliance on spurious visual cues such as radiographic text markers, residual failure cases persist due to intrinsic anatomical ambiguity in two-dimensional projection radiography. Overlapping structures, including clavicles, scapulae, vertebral bodies, and mediastinal anatomy, can produce radiopaque patterns that resemble pathological consolidation. These ambiguities represent a fundamental limitation of 2D imaging rather than a failure of the proposed method.

6.3 Evaluation Scope of Explainability Metrics

Third, quantitative explainability metrics (faithfulness and sparsity) were computed only for the post-audit model. While this choice was intentional to characterize the trustworthiness of the final deployed system, it limits direct numerical comparison between pre-audit and post-audit explanation quality. Pre-audit explanation quality was instead evaluated qualitatively and through causal intervention experiments. Future work could incorporate systematic quantitative explainability evaluation across both stages.

6.4 Computational Constraints

Fourth, perturbation-based explanation methods such as RISE are computationally expensive, requiring multiple forward passes per image. As a result, explainability analysis was performed on a representative subset of the test data rather than the full dataset. Although this subset captured both correct predictions and high-confidence errors, large-scale deployment of such methods in real-time clinical workflows may require additional optimization.

6.5 Clinical Scope (Binary Classification)

Finally, this study focuses exclusively on binary classification (Normal vs Pneumonia). The framework does not distinguish between pneumonia subtypes (e.g., viral vs bacterial) or other thoracic pathologies. Extending the model to multi-class disease differentiation or severity estimation would be necessary for broader clinical applicability.

7 Conclusion and Future Work

In this study, an interpretability-driven auditing pipeline was developed and applied to the problem of pediatric pneumonia detection from chest radiographs. A DenseNet-based classifier trained with transfer learning achieved high discriminative performance (AUROC ≈ 0.97) and near-perfect sensitivity at the default operating point, but interpretability analyses (Grad-CAM++ and RISE) and controlled causal interventions revealed that a portion of the model's high-confidence predictions depended on non-pathological visual cues. The causal marker-injection experiment provided direct evidence that spurious radiographic artifacts (e.g., laterality markers and printed labels) could substantially increase predicted pneumonia probability for otherwise normal images, while marker removal from true pneumonia cases produced negligible changes—indicating that the model mixed genuine disease signal with dataset artifacts. Following interpretability-driven auditing (targeted data cleaning, selective augmentation, and threshold optimization), the model retained strong discriminative ability while achieving a substantially more balanced operating point (post-audit accuracy ≈ 0.925, sensitivity ≈ 0.959, specificity ≈ 0.8675). Post-audit saliency maps were observed to be more anatomically plausible and quantitative explainability diagnostics (faithfulness, sparsity) indicated acceptable explanation quality for the final model.

These results demonstrate three principal conclusions. First, high aggregate metrics such as AUROC and accuracy can mask clinically important failure modes; interpretability tools must therefore be integrated into the standard evaluation pipeline for medical imaging models. Second, causal interventions (insertion and removal of suspected artifacts) provide strong, actionable evidence of shortcut learning and should be considered a necessary complement to qualitative saliency inspection. Third, interpretability-driven remediation—combined with conservative operating-point selection—can meaningfully reduce erroneous high-confidence predictions without destroying the model's discriminatory power, yielding a more deployment-ready system.

Several concrete directions are recommended for future work to increase the clinical robustness, scalability, and utility of the approach presented here:

- **Multi-center and cross-dataset validation.** The present experiments were performed on a single pediatric dataset. External validation on multi-center, multi-vendor, and adult datasets is required to assess generalizability and to detect additional dataset-specific shortcuts that may not be present in the training cohort.

- **Explicit anatomical pre-processing.** Incorporating lung-field segmentation or an anatomically constrained attention mechanism prior to classification can reduce the model's opportunity to attend to irrelevant image regions (e.g., borders, labels, osseous projections).

- **Uncertainty estimation and calibration.** Formal calibration metrics (e.g., Expected Calibration Error, Brier score) and probabilistic uncertainty estimation (Monte Carlo dropout, deep ensembles, or Bayesian approximations) should be computed and reported.

- **Efficient and scalable interpretability.** Perturbation-based methods such as RISE produce causally grounded explanations but are computationally expensive. Future work should explore hybrid approaches or approximation strategies.

- **Robustness-oriented training**. Adversarial or distributionally robust training techniques, domain-adaptation methods, and explicit debiasing objectives could be evaluated to reduce sensitivity to acquisition artifacts and dataset shortcuts.

- **Human-in-the-loop evaluation**. The clinical value of saliency maps should be assessed with radiologists through reader studies that measure whether explanations improve diagnostic accuracy, reduce time-to-decision, or increase user trust.

- **Fine-grained diagnostic extension**. Extending the binary classifier to multi-class tasks (e.g., bacterial vs viral pneumonia, other thoracic pathologies) or to severity scoring would increase clinical applicability but will require carefully curated labels and possibly multimodal clinical context.

- **Open and reproducible auditing**. To facilitate community scrutiny and follow-up work, code, trained checkpoints (where allowed by data licenses), and the exact intervention scripts used for marker injection/removal should be released alongside clear documentation and reproducible evaluation pipelines.

In closing, this work shows that combining explanation methods with causal validation and targeted mitigation yields practical improvements in model trustworthiness that are directly relevant for clinical screening applications. Interpretability must be treated as an active component of model development and validation, not merely as a visualization step. With additional external validation, calibrated uncertainty estimates, and clinician-centered evaluation, the proposed audit pipeline can form the basis for safer and more transparent AI tools in radiology.

References

[1] World Health Organization, *Pneumonia*, WHO Fact Sheets, 2023. https://www.who.int/news-room/fact-sheets/detail/pneumonia

[2] D. S. Kermany, M. Goldbaum, W. Cai, et al., "Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018. https://doi.org/10.1016/j.cell.2018.02.010

[3] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2097–2106, 2017. https://doi.org/10.1109/CVPR.2017.226

[4] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4700–4708, 2017. https://doi.org/10.1109/CVPR.2017.243

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. https://doi.org/10.1109/CVPR.2016.90

[6] R. R. Selvaraju, M. Cogswell, A. Das, et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017. https://doi.org/10.1109/ICCV.2017.74

[7] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 339–346, 2018. https://doi.org/10.1109/WACV.2018.00042

[8] V. Petsiuk, A. Das, and K. Saenko, "RISE: Randomized Input Sampling for Explanation of Black-box Models," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2018. https://arxiv.org/abs/1806.07421

[9] J. R. Zech, M. A. Badgeley, M. Liu, et al., "Variable Generalization Performance of a Deep Learning Model to Detect Pneumonia in Chest Radiographs," *PLoS Medicine*, vol. 15, no. 11, e1002683, 2018. https://doi.org/10.1371/journal.pmed.1002683

[10] R. Geirhos, J. H. Jacobsen, C. Michaelis, et al., "Shortcut Learning in Deep Neural Networks," *Nature Machine Intelligence*, vol. 2, pp. 665–673, 2020. https://doi.org/10.1038/s42256-020-00257-z

[11] W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller (Eds.), *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, Springer, 2019. https://doi.org/10.1007/978-3-030-28954-6

[12] E. Tjoa and C. Guan, "A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 11, pp. 4793–4805, 2021. https://doi.org/10.1109/TNNLS.2020.3027314

[13] Guangzhou Women and Children's Medical Center, "Chest X-Ray Images (Pneumonia) Dataset," *Mendeley Data*, V2, 2018. https://data.mendeley.com/datasets/rscbjbr9sj/2