What a Standard ResNet Learns on NIH Chest X-rays— and What We Can (and Can't) Infer From It

Jalil Ahmed

January 2025

Abstract

Currently, Deep learning-based methods are state-of-the-art for medical image analysis. However, the robustness, causal interpretation, and potential for clinical deployment of such methods remain incompletely understood. This post documents a careful evaluation of a ResNet-50 trained for multilabel classification on the NIH ChestX-ray14 dataset. The goal is to demonstrate how to set up a problem rigorously, explain evaluation choices, and measure the interpretation of results. The emphasis is on epistemic humility: understanding the boundaries of what our experiments reveal.

1 Introduction

Deep learning-based methods outperform classical methods for various tasks, including, anatomy segmentation and disease classification. Although saliency maps are used to highlight clinically relevant regions for these models, yet this surface-level success may obscure fundamental questions about what these models have actually learned and under what conditions their predictions remain valid.
This work provides a careful baseline evaluation that makes explicit:

- The dataset characteristics and their implications
- The training procedure and its limitations
- The evaluation methodology and what it measures
- The interpretation of explanations and their scope

The goal of this work is to provide a guide on how to train a standard convolutional architecture on a widely-used medical imaging dataset and interpret its performance.

2 Dataset and Experimental Setup

2.1 The NIH ChestX-ray14 Dataset

The NIH ChestX-ray dataset [1] contains 112,120 frontal-view chest X-ray images of 1024-by-1024 size from 30,805 unique patients, with 14 disease labels extracted from radiological reports by natural language processing. text-mined labels are expected to have more than 90% accuracy.

	Value
Images	112,120
Patients	30,805
Age (mean ± std)	49.6 ± 16.8
Male (%)	56.5
Female (%)	43.5

Table 1: Dataset split following patient-wise partitioning to prevent data leakage.

Each image may have multi-labels. I employ the *official patient-wise split* into training, validation, and test sets.
Patient-wise splitting is important because random image-level splits can leak patient-specific information across the train/test boundaries. The same patient may have multiple X-rays exhibiting similar pathologies, anatomical features, and acquisition characteristics. An image-level split would allow the model to "memorize" patient-specific patterns. This inflates performance estimates and results in models that do not learn generalizable disease representations.

2.2 Dataset Characteristics

Severe class imbalance: is characteristic of real-world clinical datasets and is critical for accurate evaluation of models. The label frequency in the dataset is in a diverse range,

$$\text{Label frequency: } f_{\min} \approx 0.2 \quad \text{to} \quad f_{\max} \approx 0.18 \tag{1}$$

Condition	Count	Prevalence(%)
Atelectasis	11559	10.3
Cardiomegaly	2776	2.4
Effusion	13317	11.8
Infiltration	19894	17.7
Mass	5782	5.1
Nodule	6331	5.6
Pneumonia	1431	1.2
Pneumothorax	5302	4.7
Consolidation	4667	4.1
Edema	2303	2.0
Emphysema	2516	2.2
Fibrosis	1686	1.5
Pleural Thickening	3385	3.0
Hernia	227	0.2

Table 2: Disease prevalence in the dataset

Noisy and weak labels were extracted automatically from radiology reports using natural language processing. This introduces:

- False positives from reports e.g. negations or historical references
- False negatives from incomplete report coverage

- Ambiguous cases are not curated by radiologists

Pathology-dependent ambiguity: Different pathologies have different levels of annotation reliability. For instance, "Cardiomegaly" has relatively clear radiological criteria, while "No Finding" is a broad category that may include other abnormalities which are not explicitly labeled.
These are *not preprocessing failures*. They are inherent properties of the dataset that reflect real-world medical data challenges. Instead of "fixing" these challenges through relabeling or curation, a more robust and generalizable model is achieved by incorporating and tackling these challenges in development.

3 Model Architecture and Training Protocol

3.1 Model Architecture

The model is a standard ResNet-50 [2] backbone from `torchvision`, with the following specifications:

- 50 layers organized into residual blocks
- Approximately 25.6M parameters
- Initialized with random weights

The final fully-connected layer is replaced with:

$$f(\mathbf{x}) = \sigma(W^\top h(\mathbf{x}) + b) \tag{2}$$

where $h(\cdot)$ is the ResNet feature extractor, $W \in \mathbb{R}^{2048\times 15}$, $b \in \mathbb{R}^{15}$, and σ is the sigmoid function applied element-wise for multi-label prediction.
Important Note: Pre-trained ImageNet weights were not used to achieve a more accurate and realistic baseline performance.

3.2 Training Configuration

Training choices were deliberately *standard and minimal*:

Algorithm 1 Training Protocol

1: **Optimizer:** Adam with default parameters ($\beta_1 = 0.9$, $\beta_2 = 0.999$)
2: **Learning rate:** $\alpha = 10^{-3}$ (fixed, no scheduling)
3: **Loss function:** Binary cross-entropy with logits:

$$\mathcal{L} = -\frac{1}{N \cdot c} \sum_{i=1}^{N} \sum_{c=1}^{C} [y_{ic} \log \hat{y}_{ic} + (1 - y_{ic}) \log(1 - \hat{y}_{ic})]$$

4: **Batch size:** 32
5: **Epochs:** 25
6: **Data augmentation:** None
7: **Random seed:** Single run with seed=123
8: **Hyperparameter search:** None performed

The objective for minimal training was to evaluate a *reasonable baseline* and not to optimize for leaderboard performance. Extensive hyperparameter tuning may lead to:

- Overfitting to validation set
- Obscure which design choices matter
- Make results harder to reproduce

3.3 Computational Resources

Training was performed on a single NVIDIA RTX A6000 GPU, requiring approximately 16 hours for 25 epochs. The baseline performance on this dataset does not require extensive resources.

4 Evaluation Methodology

4.1 Performance Metrics

Performance is quantified using the **Area Under the Receiver Operating Characteristic curve (AUROC)**, computed per pathology class. AUROC provides a threshold-independent measure of discrimination ability for highly imbalanced datasets. AUROC is also relatively stable compared to accuracy as it is dominated by majority class predictions. For example, consider a pathology present in 2% of cases. A model that predicts "negative" for all cases achieves 98% accuracy while learning nothing. In summary, AUROC measure how well the model ranks positive cases above negative cases regardless of the class distribution.
For each pathology k:

$$\text{AUROC}_c = \mathbb{P}(\hat{c}_i^{(c)} > \hat{y}_j^{(c)} \mid y_i^{(k)} = 1, y_j^{(c)} = 0) \tag{3}$$

Also,

$$\text{Macro-AUROC} = \frac{1}{C} \sum_{c=1}^{C} \text{AUROC}_c \tag{4}$$

4.2 What Is Not Evaluated

It is equally important to state what this evaluation does *not* include:

- **No out-of-distribution testing:** Performance on other chest X-ray datasets (CheXpert, MIMIC-CXR, PadChest) is not assessed
- **No subgroup analysis:** Performance stratified by patient demographics (age, sex, race) is not reported
- **No calibration evaluation:** Whether predicted probabilities match true frequencies is not examined
- **No temporal validation:** Robustness to temporal drift is not tested
- **No clinical validation:** Agreement with radiologist interpretations is not measured

These omissions define the scope of valid inference. We can characterize performance on the NIH test set under i.i.d. assumptions, but cannot make claims about generalization beyond this distribution.

5 Results: What the Numbers Show

5.1 Per-Class Performance

Table 3 presents AUROC scores for each pathology. Performance varies substantially across classes.

Pathology	**AUROC**	**Prevalence (%)**
Cardiomegaly	0.852	2.4
Edema	0.825	2.0
Pneumothorax	0.816	4.7
Emphysema	0.784	2.2
Hernia	0.786	0.2
Effusion	0.768	11.8
Fibrosis	0.770	1.5
Atelectasis	0.728	10.3
Consolidation	0.710	4.1
Pleural Thickening	0.709	3.0
Mass	0.706	5.1
Infiltration	0.684	17.7
Pneumonia	0.668	1.2
Nodule	0.642	5.6
Macro-Average	**0.744**	—

Table 3: Per-pathology AUROC scores and label prevalence in descending order of AUROC scores. Performance varies from 0.642 to 0.852, highlighting that "model performance" is not a single number.

5.2 Interpreting the Variation

The spread in AUROC values (Figure-1)reflects substantive differences in task difficulty w.r.t each pathology that persist despite identical architecture and training:
Highest performance (AUROC $>$ 0.80):

- *Cardiomegaly* (0.852): Cardiomegaly is the condition characterized abnormally thick or overly stretched heart causing an enlarge heart sihouette in x-ray imaging. This change in geometric feature (cardiothoracic ratio) of a major organ make it relatively easy for a CNN to identify distinct outlines.

- *Edema* (0.825): Pulmonary Edema is the accumulation of fluid in the lung tissue. This condition results in widespread, hazy, white opacities in the x-ray images that are visually distinct from normal, dark, air-filled lung images. These large areas produce a strong contrast making them highly detectable.

- *Pneumothorax* (0.816): Pneumothorax is the condition of presence of air in the pleural space causing partial or complete collapse of the lung. In the x-ray images of Pneumothorax patients, the lung edge has a clear demarcation, and complete blackness (absence of lung markings) in the surrounding area provide a very high-contrast which is easily identifiable visual cue for a CNN model.

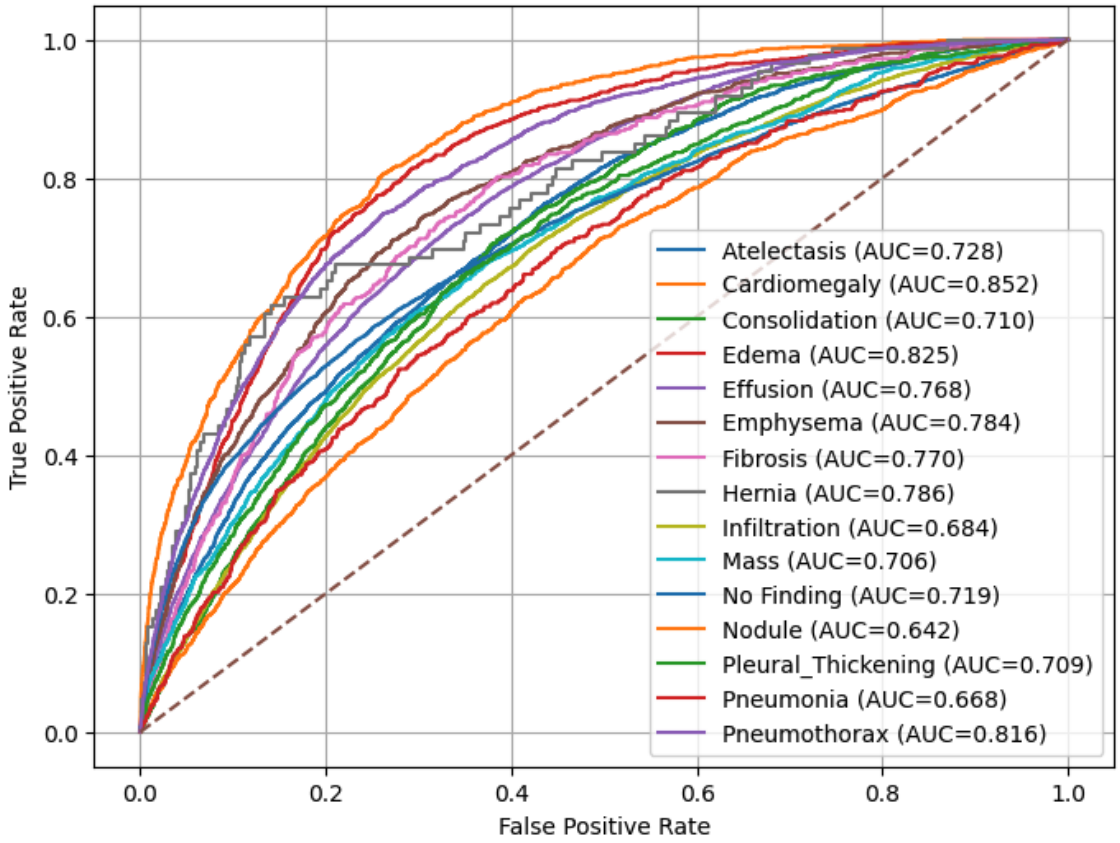

Figure 1: AUROC per pathlogy. The Macro-average AUROC is 0.744

Lower performance (AUROC < 0.70):

- *Infiltration* (0.684): Infiltration is a highly ambiguous label with non-specific patterns and diverse locations. This makes it harder to define a consistent visual signature for a CNN to learn across all images.
- *Pneumonia* (0.668): Pneumonia is difficult to detect by a model because it manifests in a wide variety of ways. It may have diverse patterns such as, lobar consolidation, patchy infiltrates, or interstitial patterns. Moreover, the visual presentations overlap visually with other conditions like edema, atelectasis, or even non-infectious inflammation.
- *Nodule* (0.642): Nodules are localized lesions that can be small and easily obscured by overlapping structures, such as, ribs or the heart shadow. This makes consistent detection challenging for a model.

Key observation: "Model performance" is *not a single scalar*. Even within a single dataset, with identical architecture and training procedure, different labels exhibit drastically different learnability. This heterogeneity is important to fully understand the capability of the model.

6 Grad-CAM Analysis:

6.1 Gradient-weighted Class Activation Mapping

Gradient-weighted Class Activation Mapping (Grad-CAM) [3] was used to qualitatively investigate the behavior of the model. For a target class c and convolutional feature map A^k in layer l, Grad-CAM computes:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{5}$$

The class-discriminative localization map is then:

$$L_{\text{Grad-CAM}}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k \right) \tag{6}$$

This highlights regions where changes in activation most influence the prediction for class c. We generated class-averaged Grad-CAM visualizations by:

1. Selecting all test set images with positive labels for each pathology
2. Computing Grad-CAM for each image
3. Averaging the attention maps spatially

6.2 Observed Patterns

Anatomically plausible attention (some cases) (Figure-2):

- *Cardiomegaly*: The average Grad-CAM for cardiomegaly show that the model's attention is consistently concentrated over the central thorarcic region corresponding to the cardiac silhouette. The mean and standard deviation maps demonstrates that the model relies on spatially consistent cues across patients rather than diffuse or image-peripheral features. This indicates that the model has learned to focus on the expected anatomical location of pathology. However, the lack of attention maps localized to the cardiac borders implies that the model may be responding to global heart-region cues rather than explicitly encoding boundary-based size relationships that are clinicaly relevant features for diagnosing cardiomegaly.
- *Edema*: The average Grad-CAM for edema reveals a diffuse attention pattern spanning the central and bilateral lung fields. This distribution is anatomically plausible, as pulmonary edema typically presents with widespread interstitial and alveolar involvement rather than a focal lesion. The observed spatial spread and increased variability in the standard deviation maps are consistent with the heterogeneous radiographic appearance of edema. However, the central emphasis and recurrent activation in lower-peripheral regions suggest that the model may also be attending to non-pathological cues. Edema-related visual patterns on X-ray images include perihilar bat-wing opacities and basal interstitial markings. Overall, the attention maps partially correspond to the lung-centered distribution of edema but do not demonstrate clear localization to clinically meaningful radiographic features.

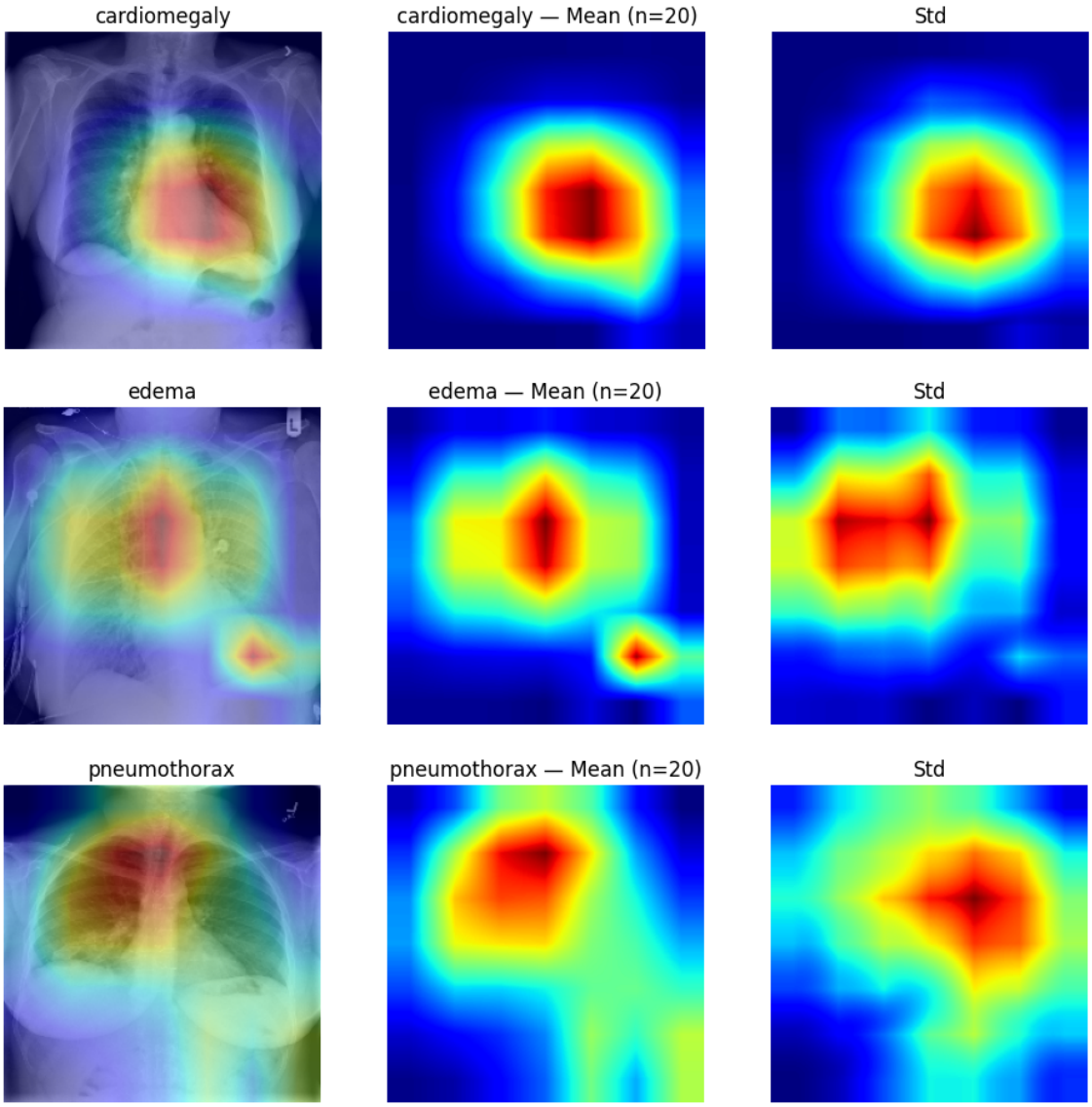

Figure 2: Class-averaged Grad-CAM visualizations for pathologies with top-3 AUROC

- *Pneumothorax*: The average Grad-CAMs show that the model's attention is mostly in the upper chest, not spread out across the lungs. This is only partly accurate, since pneumothorax often appears at the top of the lungs in upright X-rays. However, the attention is broad and centered, not sharply focused on the pleural line or lung edges. The mean map shows a steady focus area, but the standard deviation indicates significant variation, suggesting the model's attention is inconsistent. This suggests the model may be using general regional cues, such as differences in brightness or symmetry, rather than subtle edge features that are important for diagnosing pneumothorax.

Diffuse or inconsistent attention (other cases) (Figure-3):

- *Infiltration*: The Grad-CAM results for the infiltration prediction indicate difficulty with anatomical focus and localization on the primary lung fields, where pulmonary infiltration should be identified. The attention map (red zones) is concentrated in the lower right corner

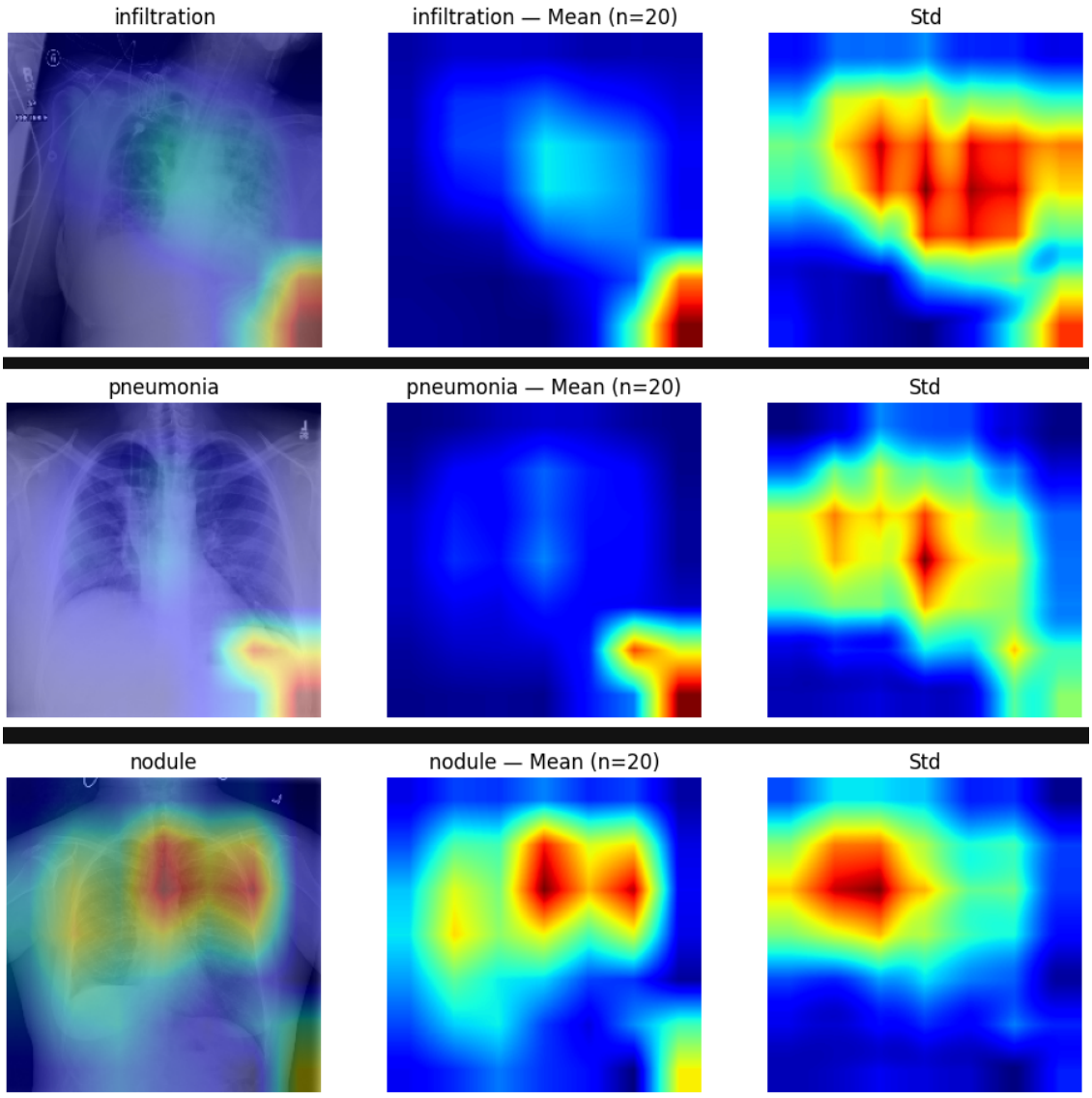

Figure 3: Class-averaged Grad-CAM visualizations for pathologies with lowest-3 AUROC

of the image, which appears to be an artifact or an area outside the thoracic cavity rather than the lung parenchyma. The mean and standard deviation maps show that the model is highly inconsistent and likely relies on peripheral features or image edges rather than stable anatomical markers. Therefore, the resulting maps are largely anatomically implausible. This behavior is characteristic of suboptimal models that learn using clinically irrelevant features, such as image borders, medical equipment, or patient positioning, rather than identifying the pathology itself.

- *Pneumonia*: The Grad-CAM results indicate Extrathoracic Bias because the heatmaps show high intensity in the image corners. The model is likely relying on "shortcuts", such as hospital-specific tags or patient positioning artifacts, instead of clinical pneumonia markers. Anatomically plausible pneumonia features show focused activations over the pulmonary zones. The activation maps fail to isolate the actual parenchymal infections, making them

anatomically implausible, as shown by the diffuse standard deviation map.

- *Nodule*: The Grad-CAM results show a disconnect between the model's focus and the actual clinical presentation of lung nodules. Anatomically, a lung nodule is a small, well-defined lesion inside the lung parenchyma. Therefore, an effective model should demonstrate highly localized, "spot-like" attention on specific regions of the lung fields, as nodules are typically less than 3cm in diameter. The mean map shows a central mediastinal bias. The model appears to be incorrectly associating the dense structures of the heart or great vessels with the presence of a nole. Also, the model is making a "global" decision based on general image texture or patient positioning, as evidenced by a diffused mean attention map. The standard deviation map shows high variability in the lateral lung zones and the shoulders. This suggests that the model's predictions are unstable and highly sensitive to peripheral noise or anatomical structures. In summary, the Grad-CAM confirms that the model has not learned the "concept" of a nodule in a pathological sense.

6.3 Critical Limitations of Grad-CAM

Grad-CAM serves as a tool for visualizing the relationship between a model's complex mathematical predictions and human interpretation. The spatial patterns and regions where feature activations are most strongly correlated with the model's prediction are identified by calculating gradient flow into the final convolutional layer for a specific class. These heatmaps demonstrate where the model focuses under its learned representation. In summary, these heatmaps reveal the distinguishing characteristics of the training distribution that the models use to classify an image as a certain class. This, in turn, helps researchers verify whether the model is focusing on legitimate anatomical or pathological identifiers or is relying on non-pathological artifacts to make a prediction.

What Grad-CAM does NOT establish: The Grad-CAM heatmaps demonstrate regions that are statistically associated with the label in the training set. This only demonstrates correlation, not causation, so the highlighted regions might not be the underlying biological cause of the pathology. Also, anatomically plausible attention maps do not equate to medically sound reasoning because the model might be focusing on correct anatomy but relying on flawed logic or irrelevant regions to classify. Another frequent issue with Grad-CAM is fragility and lack of robustness as attention patterns shift significantly under distribution shifts. In summary, Grad-CAM represents only a single component of a complex, non-linear decision process and provides a simplified view that does not capture the full scope of the model's internal reasoning.

6.4 The Interpretability Trap

Grad-CAM and similar explanation methods often produce *compelling visualizations* that appear to validate model reasoning often leading researchers into what is known as the "Interpretability Trap." This represents a flawed chain of inference where visual plausibility is mistaken for logical correctness:

Model produces plausible-looking attention map

↓

Human observer recognizes anatomically relevant region

↓

Observer concludes model has learned "correct" reasoning

↓

Observer's confidence in model increases (**The Trap**)

This phenomenon occurs because of ignoring a critical fact that *plausible attention is necessary but not sufficient for establishing robust or clinically valid logic*. It happens when a human observer instinctively recognizes anatomically plausible attention maps and prematurely concludes that the model has learned medically relevant reasoning. It is important to know, that a model may highlight the correct anatomy while actually exploiting spurious correlations, dataset-specific artifacts, or confounding variables which correlated with the pathology in the training distribution.
In order to avoid it, Grad-CAM should be strictly used as a diagnostic tool for understanding model behaviour to effectively identify potential failure models, generating hypotheses about learned features and guiding further investigations e.g. ablation studies. It should not be used as a validation for model correctness to justify clinical deployment and claim human-like model reasoning.

7 Limitations: Explicit and Bounded

This work has clear, explicit limitations that define the scope of valid inference:

7.1 Experimental Limitations

- Only single training run
- No distribution shift evaluation
- Class imbalance was not addressed

7.2 Explanation Limitations

- Grad-CAM was not stress-tested
- No ground-truth localization of attention patterns

Grad-CAM not stress-tested:

- Stability under small input perturbations not evaluated
- Consistency across model checkpoints not assessed
- Robustness to adversarial manipulations not tested

No ground-truth localization:

- Attention patterns not compared to radiologist annotations
- Cannot validate whether "correct" regions are used
- Plausibility is subjective, not quantified

7.3 Dataset Limitations

The NIH ChestX-ray14 dataset itself has known issues:

- Labels extracted via automated NLP, not manually curated
- Potential biases in patient population and imaging protocols
- Label noise and ambiguity present but not quantified
- No pixel-level annotations for localization tasks

7.4 Scope of Inference

Given these limitations, we can conclude:
Valid inferences:

- A standard ResNet-50 can achieve AUROC 0.64–0.85 on NIH ChestX-ray14 test set
- Performance varies substantially across pathologies
- Some attention patterns appear anatomically plausible
- Minimal tuning is sufficient for reasonable baseline performance

Invalid inferences:

- The model has learned clinically valid reasoning
- Performance will generalize to other datasets or populations
- Grad-CAM attention validates model correctness
- The model is robust to distribution shift

8 Conclusion

While training a deep learning model for classification is relatively straightforward, comprehending the model's learned representations, identifying its strengths and weaknesses, and assessing the trustworthiness of its predictions present significant challenges.
This study establishes a clear baseline using a standard architecture, minimal hyperparameter tuning, a rigorous evaluation protocol, and transparent acknowledgment of limitations. Its purposes are as follows:

- A demonstration of appropriate evaluation practices
- A reminder that metrics and model explanations are meaningful only within their defined scope

Importantly, this work is a practical guide towards initial steps for training reliable and robust AI models. It emphasizes that questions of robustness, fairness, interpretability, and clinical validity are more important than performance metrics alone. Effective deployment of AI in healthcare requires not only high AUROC scores but also rigorous evidence of reliability in real-world settings.

Acknowledgments

The NIH ChestX-ray14 dataset was created by Wang et al. and made publicly available for research purposes. This work uses standard open-source tools (PyTorch, torchvision) and follows established best practices from the machine learning and medical imaging communities.

References

[1] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). *ChestX-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases.* IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2097–2106.

[2] He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep residual learning for image recognition.* IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778.

[3] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). *Grad-CAM: Visual explanations from deep networks via gradient-based localization.* IEEE International Conference on Computer Vision (ICCV), 618–626.

[4] 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 preprint arXiv:1711.05225.

[5] Oakden-Rayner, L., Dunnmon, J., Carneiro, G., & Ré, C. (2020). *Hidden stratification causes clinically meaningful failures in machine learning for medical imaging.* Proceedings of ACM Conference on Health, Inference, and Learning, 151–159.

[6] Finlayson, S. G., Subbaswamy, A., Singh, K., Bowers, J., Kupke, A., Zittrain, J., ... & Saria, S. (2021). *The clinician and dataset shift in artificial intelligence.* New England Journal of Medicine, 385(3), 283–286.

[7] Rudin, C. (2019). *Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead.* Nature Machine Intelligence, 1(5), 206–215.

[8] DeGrave, A. J., Janizek, J. D., & Lee, S. I. (2021). *AI for radiographic COVID-19 detection selects shortcuts over signal.* Nature Machine Intelligence, 3(7), 610–619.