Comparative Evaluation of AI Models for Robust and Explainable Chest X-ray Analysis

Rehna Afroz Shaik(22B3932), Yashaswini K(22B3911)

I. Introduction

Deep learning models have achieved high accuracy in chest X-ray analysis, but their robustness to real-world shifts and explainability remain major barriers to clinical adoption. Existing work often focuses on single datasets and accuracy, overlooking how different architectures generalize across settings or how reliable their explanations are. There is a clear need for a comparative benchmark that evaluates multiple AI models not only for classification performance but also for robustness and explainability on chest X-rays.

II. Methodology

We have conducted a comprehensive study on chest X-ray classification and interpretability using the CheXpert dataset as our primary training source. Four architectures—ResNet-50, DenseNet-121, ConvNeXt-Tiny, and ViT-B/16—are trained across three random seeds (999, 123, and 42) to ensure robustness. To evaluate generalization, all models are tested on both the in-domain CheXpert test set and the out-of-domain NIH ChestX-ray14 dataset. In addition, we assess model interpretability using the CheXLocalize dataset, enabling a quantitative comparison of explanation quality across architectures. Our methodology consists of two components: (i) multi-label classification and (ii) explainability analysis.

Classification. We fine-tune four widely used convolutional and transformer-based architectures for eight-way multi-label classification. All networks start from ImageNet-1K pretrained weights, with the final layer replaced by a linear classifier of dimension eight. The prediction targets follow a canonicalized label space containing *No Finding*, *Atelectasis*, *Cardiomegaly*, *Consolidation*, *Edema*, *Pleural Effusion*, *Pneumonia*, and *Pneumothorax*. Metadata fields in CheXpert are mapped to these unified categories (e.g., "Lung Opacity" $\rightarrow$ "Consolidation") to maintain consistent supervision.

All images are resized to 224×224, and the models are optimized end-to-end using a sigmoid cross-entropy loss. The training pipeline, preprocessing steps, and label definitions are kept identical across architectures to ensure a fair comparison.

Explainability. To evaluate model interpretability, we use Grad-CAM attribution maps provided in the CheXLocalize dataset. Each heatmap is normalized to the range $[0, 1]$, resized to 224×224, and applied as a soft mask on the input image. Using the model's predicted confidence on the original image $f(x)$ and the masked version $f(x \odot h)$, we compute four standard explanation metrics:

- **Average Drop**: proportional reduction in confidence after masking,
- **Increase in Confidence**: cases where masking strengthens the prediction,
- **Entropy**: spatial concentration of the heatmap,
- **Sparsity**: fraction of zero-valued pixels.

We additionally measure *deletion* and *insertion* curves by progressively removing or revealing pixels in descending order of heatmap importance, recording the model confidence at each step. The area under these curves provides a summarized measure of explanation sensitivity. All metrics are computed per image and averaged over the dataset for each model.

III. Experimental Results

A. Classification Evaluation Metrics

This section summarizes the in-domain (CheXpert) and out-of-distribution (NIH ChestXray14) classification performance of all four architectures across three random seeds(42, 123, 999). We analyze both per-seed stability and cross-model trends using AUROC, Average Precision (AP), Macro-F1, and Micro-F1.

B. In-domain CheXpert Performance

Tables I and II report the in-domain CheXpert results. Several consistent trends emerge across architectures. ResNet-50 demonstrates highly stable performance across seeds, achieving an average AUROC of 0.819 ± 0.007 with moderate AP and F1-scores. Its low variance indicates robust convergence behavior and reliable supervised learning dynamics.

DenseNet-121 achieves the strongest classification performance on CheXpert in terms of macro-level metrics. It records the highest Macro-F1 (0.460 ± 0.009) and Micro-F1 (0.512 ± 0.008), highlighting its ability to capture relevant disease patterns more effectively across both rare and common classes. Although its AUROC is comparable to ResNet, DenseNet shows a consistent advantage in calibration and threshold-based metrics.

ViT-B/16 underperforms the convolutional baselines, with noticeably lower AUROC (0.785 ± 0.012) and AP, along with higher variance. This degradation is expected given ViT's comparatively weaker inductive biases and increased data requirements. Despite this, ViT yields competitive F1-scores, suggesting that its errors are evenly distributed across classes rather than dominated by specific pathologies.

ConvNeXt-Tiny achieves the strongest AUROC among all models (0.831 ± 0.008) and the highest AP (0.559 ± 0.023), demonstrating excellent ranking performance. However, its Macro-F1 remains slightly below DenseNet, indicating that

TABLE I: Chexpert Per-seed performance

Model	Seed	Mean AUROC	Mean AP	Macro F1	Micro F1
ResNet-50	42	0.814	0.572	0.440	0.494
	123	0.816	0.560	0.442	0.506
	999	0.827	0.551	0.430	0.511
DenseNet-121	42	0.819	0.533	0.458	0.508
	123	0.810	0.543	0.452	0.505
	999	0.833	0.554	0.469	0.522
ViT-B/16	42	0.792	0.502	0.447	0.504
	123	0.769	0.512	0.418	0.486
	999	0.794	0.531	0.434	0.486
ConvNeXt-Tiny	42	0.823	0.540	0.455	0.508
	123	0.830	0.553	0.460	0.518
	999	0.839	0.585	0.450	0.511

TABLE II: Chexpert Summary of Results (Mean $\pm$ Std across 3 seeds)

Model	AUROC (↑)	AP (↑)	Macro F1 (↑)	Micro F1 (↑)
ResNet-50	**0.819 $\pm$ 0.007**	0.561 $\pm$ 0.011	0.437 $\pm$ 0.005	0.504 $\pm$ 0.007
DenseNet-121	0.821 $\pm$ 0.012	0.543 $\pm$ 0.011	**0.460 $\pm$ 0.009**	**0.512 $\pm$ 0.008**
ViT-B/16	0.785 $\pm$ 0.012	0.515 $\pm$ 0.015	0.433 $\pm$ 0.012	0.492 $\pm$ 0.010
ConvNeXt-Tiny	**0.831 $\pm$ 0.008**	**0.559 $\pm$ 0.023**	0.455 $\pm$ 0.004	0.512 $\pm$ 0.005

TABLE III: Per-seed performance on NIH OOD test set

Model	Seed	Mean AUROC	Mean AP	Macro-F1	Micro-F1
ResNet-50	42	0.780	0.261	0.232	0.298
	123	0.740	0.228	0.215	0.310
	999	0.757	0.243	0.214	0.322
DenseNet-121	42	0.781	0.262	0.243	0.334
	123	0.753	0.244	0.234	0.319
	999	0.804	0.266	0.233	0.360
ViT-B/16	42	0.785	0.263	0.237	0.334
	123	0.702	0.189	0.194	0.300
	999	0.752	0.216	0.223	0.326
ConvNeXt-Tiny	42	0.797	0.283	0.239	0.330
	123	0.758	0.266	0.214	0.335
	999	0.797	0.273	0.234	0.323

TABLE IV: NIH cross-domain (OOD) test results: Mean $\pm$ Std across 3 seeds

Model	AUROC (↑)	AP (↑)	Macro-F1 (↑)	Micro-F1 (↑)
ResNet-50	**0.759 $\pm$ 0.019**	0.244 $\pm$ 0.017	0.220 $\pm$ 0.010	0.310 $\pm$ 0.012
DenseNet-121	0.779 $\pm$ 0.026	0.257 $\pm$ 0.011	**0.237 $\pm$ 0.011**	**0.338 $\pm$ 0.021**
ViT-B/16	0.746 $\pm$ 0.042	0.223 $\pm$ 0.037	0.218 $\pm$ 0.019	0.320 $\pm$ 0.017
ConvNeXt-Tiny	**0.784 $\pm$ 0.020**	**0.274 $\pm$ 0.009**	0.229 $\pm$ 0.010	0.329 $\pm$ 0.006

while ConvNeXt excels at distinguishing positive from negative cases overall, DenseNet maintains better class-level balance. ConvNeXt displays moderate seed stability, though with notably higher AP variance.

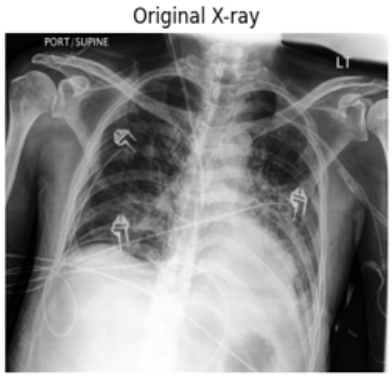
Overall, DenseNet-121 is best for *balanced F1 performance*, ResNet-50 for *training stability*, and ConvNeXt-Tiny for *ranking-based metrics* such as AUROC and AP.

C. Out-of-Distribution (OOD) NIH Performance

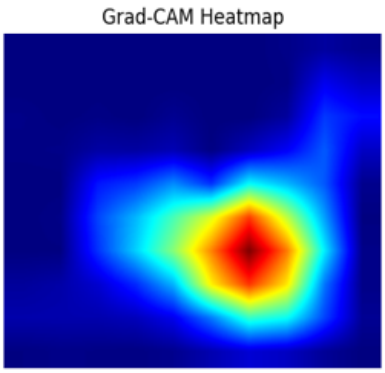
Tables IV and III show model generalization to the NIH ChestXray14 dataset. All models experience performance degradation, confirming the presence of domain shift. Nevertheless, consistent relative performance ranking is observed.

ConvNeXt-Tiny achieves the strongest OOD AUROC (0.784 ± 0.020) and AP (0.274 ± 0.009), suggesting superior robustness to distributional changes. Its per-seed results are also the most stable, reinforcing its ability to generalize effectively.

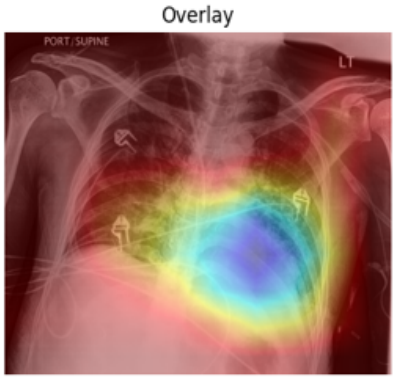
DenseNet-121 and ResNet-50 exhibit competitive performance, with DenseNet again leading in Macro-F1 (0.237 ± 0.011) and Micro-F1 (0.338 ± 0.021). DenseNet's inductive structure appears to aid in retaining class-level discrimination



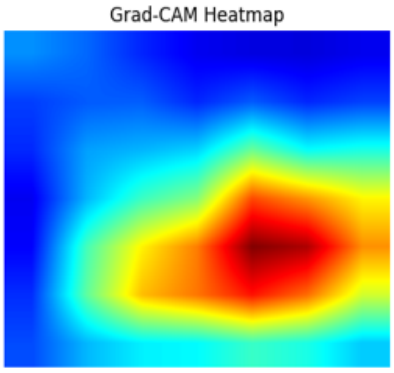
((a)) Original image



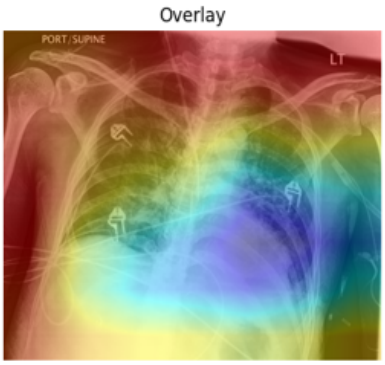
((b)) True-class Grad-CAM overlay



((c)) True-class Grad-CAM heatmap



((d)) Predicted-class Grad-CAM heatmap



((e)) Predicted-class Grad-CAM overlay

Fig. 1: Comparison between true-class and predicted-class Grad-CAM visualizations using the ResNet model (seed = 999). The first row shows the original input and the ground truth heatmap, and the second row shows the predicted heatmap and overlays for both true and predicted classes.

even under shift. ResNet-50, while stable, shows the largest AUROC drop relative to CheXpert.

ViT-B/16 performs the weakest under OOD shift, particularly in AP and F1. This aligns with known vulnerabilities of transformer architectures in low-data medical settings and in cross-domain environments.

D. Overall Findings

Across both datasets, three clear conclusions emerge:

- **DenseNet-121 provides the best overall F1-based performance**, both in-domain and OOD, making it a strong choice for clinical decision support where thresholded predictions matter.
- **ConvNeXt-Tiny achieves the strongest AUROC and AP**, indicating the best ranking capability and the most robust OOD generalization.
- **ResNet-50 remains the most stable across seeds**, offering highly consistent performance suitable for reproducible analysis.
- **ViT-B/16 lags behind CNN-based architectures**, particularly under domain shift, but still produces competitive F1 scores due to balanced error distribution.

E. Explainability Evaluation Metrics

To assess the quality of Grad-CAM heatmaps, we compute four standard perturbation- and distribution-based metrics, along with deletion and insertion curves.

Average Drop quantifies the proportional decrease in the model's confidence when only salient regions are retained. Higher values indicate that the highlighted regions are crucial for the prediction.

Increase in Confidence (IncConf) measures cases where removing non-salient regions increases the model's confidence, indicating that the explanation may be suppressing irrelevant or noisy areas.

Entropy evaluates the spatial concentration of the normalized saliency map. Lower entropy reflects more focused, interpretable explanations, whereas higher entropy indicates diffuse attributions.

Sparsity captures the fraction of pixels assigned near-zero importance, with high sparsity suggesting compact and noise-free explanations.

For **deletion** and **insertion** tests, pixels are progressively removed or revealed according to saliency ranking. The model confidence at each step produces a monotonic curve. Steep deletion curves and steep insertion curves correspond to faith-

ful and informative explanations. The area under each curve (AUC) is used as the final summary metric.

a) Overall Interpretation: Together, these metrics capture both local fidelity (through confidence-based perturbation tests) and structural quality (through sparsity and entropy) of the saliency maps. The combination provides a rigorous framework for comparing explainability performance across models and seeds, revealing how effectively each method identifies the features that drive model predictions.

F. Explainability Results Summary

Table V presents the per-seed explainability evaluation for all architectures. Several distinctive patterns emerge across the metrics.

ResNet-50 offers the most *faithful and stable* explanations. It consistently achieves low Avg Drop values (0.003–0.19) and positive Increase in Confidence across most seeds, indicating strong alignment between its saliency maps and decision process. Its deletion and insertion AUC scores (4–5) further confirm moderate causal agreement. Overall, ResNet-50 demonstrates the most reliable and seed-stable explainability behavior among the evaluated models.

DenseNet-121 shows the weakest gradient-based faithfulness, with the highest Avg Drop (0.24–0.45) and near-zero Increase in Confidence for all seeds. However, it achieves the strongest perturbation-based performance: the best insertion AUC (8–9.3) and high deletion AUC (7.3–7.7). This indicates that DenseNet explanations excel under reconstruction-based causal tests even when gradient metrics appear weak.

ViT-B/16 displays sparse and structured explanations, consistent with transformer-based attention mechanisms. It achieves intermediate Avg Drop values (0.09–0.23) and strong deletion/insertion AUC performance (6.2–8.1). Its entropy and sparsity metrics remain stable across seeds, showing compact and interpretable attribution patterns. While moderately faithful, ViT exhibits some seed sensitivity, especially in Increase in Confidence.

ConvNeXt-Tiny shows the highest variability across seeds. While seed 999 yields very high Increase in Confidence (1.089), other seeds perform poorly or show zero values. Avg Drop ranges from excellent (0.000) to very poor (0.541), indicating unstable gradient behavior. Deletion and insertion AUC values (5.8–8.0) lag behind DenseNet and ViT, suggesting weaker causal alignment. Thus, ConvNeXt explanations are sharp but unreliable.

Overall, the results highlight that: (1) ResNet-50 yields the most stable and faithful saliency maps; (2) DenseNet-121 excels in causal reconstruction metrics; (3) ViT-B/16 provides sparse, structured, moderately faithful explanations; (4) ConvNeXt-Tiny shows strong local sensitivity but poor robustness across seeds. These findings underscore the necessity of multi-metric evaluation when assessing explainability.

IV. Proposed Metric

A. Metric definition

As we have seen, different models excel in different aspects—some offer stronger predictive performance, while others provide better interpretability. To capture these complementary strengths, we introduce a unified metric that integrates both dimensions.

B. Metric definition

The **Unified Classification–Explainability Score (UCES)** provides a single, coherent measure that jointly reflects a model's predictive accuracy and the quality of its explanations. The computation consists of four steps:

1) **Metric selection:** We include metrics that capture both classification quality and explanation fidelity:
   - Classification: AUROC, Average Precision (AP), Macro F1, and Micro F1.
   - Explainability: Average Drop (↓), Increase in Confidence (↑), Entropy (↓), Sparsity (↑), Deletion AUC (↓), and Insertion AUC (↑).
2) **Normalization:** Since these metrics differ in scale and interpretation, each is normalized using min–max normalization across all evaluated models:
   - For metrics where higher values indicate better performance (AUROC, AP, Macro F1, Micro F1, Increase in Confidence, Sparsity, Insertion AUC):
   $$x_{norm} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$
   - For metrics where lower values are preferred (Average Drop, Entropy, Deletion AUC):
   $$x_{norm} = \frac{x_{\max} - x}{x_{\max} - x_{\min}}$$
3) **Aggregating classification and explainability components:** After normalization, we compute the classification component C and the explainability component E as:
   $$C = \frac{1}{4}\Big(\text{AUROC}_n + \text{AP}_n + \text{MacroF1}_n + \text{MicroF1}_n\Big)$$
   $$E = \frac{1}{6}\Big(\text{AvgDrop}_n + \text{IncConf}_n + \text{Entropy}_n + \text{Sparsity}_n + \text{DeletionAUC}_n + \text{InsertionAUC}_n\Big)$$
4) **Final UCES computation:** The final unified score is obtained using a weighted sum:
   $$\text{UCES} = \alpha C + (1 - \alpha)E$$
   Following standard practice, we set $\alpha = 0.7$, giving slightly higher importance to classification while still incorporating explainability.

C. UCES Scores and Model Ranking

- ConvNeXt-Tiny achieves the highest UCES score, reflecting a strong balance between classification and explainability.

TABLE V: Per-seed explainability performance across models (mean values only).

	Seed	Avg Drop (↓)	Inc Conf (↑)	Entropy (↓)	Sparsity (↑)	Deletion AUC (↓)	Insertion AUC (↑)
ResNet-50	42	0.0513	0.0717	10.02	0.207	**4.22**	5.40
	123	0.1925	0.00	10.02	0.207	5.37	5.19
	999	**0.0031**	0.14	10.02	0.207	4.36	4.75
DenseNet-121	42	0.2409	0.0014	10.02	0.207	7.69	**9.32**
	123	0.3008	0.00	10.02	0.207	7.58	8.15
	999	0.4494	0.00	10.02	0.207	7.32	8.00
ViT-B/16	42	0.094	0.019	9.637	**0.411**	7.30	7.27
	123	0.165	0.031	9.637	**0.411**	8.09	8.02
	999	0.233	0.0003	**9.637**	**0.411**	6.24	6.57
ConvNeXt-Tiny	42	0.0601	0.0658	9.64	0.411	**6.02**	**5.89**
	123	0.541	0.00	9.64	0.411	8.07	7.55
	999	0.000	**1.089**	9.592	0.409	7.82	7.89

TABLE VI: Explainability performance across models (mean ± standard deviation across 3 seeds).

Model	**Avg Drop (↓)**	**Inc Conf (↑)**	**Entropy (↓)**	**Sparsity (↑)**	**Deletion AUC (↓)**	**Insertion AUC (↑)**
ResNet-50	0.0823 ± 0.0951	0.0706 ± 0.0835	10.02 ± 0.00	0.207 ± 0.00	4.65 ± 0.58	5.11 ± 0.33
DenseNet-121	0.3304 ± 0.1047	0.00047 ± 0.00081	10.02 ± 0.00	0.207 ± 0.00	7.53 ± 0.19	8.49 ± 0.72
ViT-B/16	0.1640 ± 0.0695	0.0168 ± 0.0155	9.637 ± 0.00	0.411 ± 0.00	7.21 ± 0.93	7.29 ± 0.73
ConvNeXt-Tiny	0.2004 ± 0.2965	0.385 ± 0.6106	9.624 ± 0.028	0.410 ± 0.0012	7.30 ± 1.12	7.11 ± 1.07

TABLE VII: Model-wise strengths, weaknesses, and recommended use-cases based on updated explainability results.

Model	**Strengths**	**Weaknesses**	**Best Use-Cases**
ResNet-50	Most stable across seeds; lowest Avg Drop; consistently positive Increase in Confidence; moderate deletion/insertion AUCs	Slightly weaker reconstruction performance than DenseNet	Applications requiring highly *faithful, stable, and interpretable* saliency maps
DenseNet-121	Best insertion AUC; strong deletion AUC; excellent causal recovery performance	Highest Avg Drop; weak gradient-based faithfulness; near-zero Increase in Confidence	Settings where *perturbation-based causal reconstruction* is more important than gradient alignment
ViT-B/16	Sparse and structured explanations; low entropy; strong AUC values	Moderate faithfulness; some seed sensitivity in Increase in Confidence	Scenarios requiring *compact, token-like, structured explanations* as in medical or fine-grained tasks
ConvNeXt-Tiny	High Increase in Confidence for some seeds; sensitive to salient regions	Very high seed sensitivity; inconsistent Avg Drop; weaker causal alignment	Useful for *local sensitivity analysis*; not recommended for robust global explainability

Model	Normalized C (↑)	Normalized E (↑)	UCES (↑)	Rank
ConvNeXt-Tiny	0.943	0.698	0.870	1
DenseNet-121	0.848	0.167	0.643	2
ResNet-50	0.622	0.364	0.544	3
ViT-B/16	0.000	0.572	0.172	4

TABLE VIII: Normalized classification and explainability scores, UCES values, and final ranking for $\alpha = 0.7$.

- DenseNet-121 performs well in classification but obtains weaker explainability scores, reducing its final ranking.
- ViT-B/16 shows comparatively good explainability but low classification performance, resulting in the lowest UCES value.
- ResNet-50 yields moderate scores in both categories and ranks third overall.
- The UCES framework provides a single quantitative criterion for selecting models in settings where both accuracy and interpretability are important.

V. Conclusion

The UCES metric offers a unified and quantitative framework for evaluating classification models using both predictive performance and explainability. In our experiments, ConvNeXt-Tiny attains the highest overall score and is the recommended model for robust and interpretable chest X-ray classification.