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

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

***Abstract*—Chest X-ray interpretation is a critical task in clinical diagnostics, yet deep learning models often exhibit variability in predictive reliability and lack transparent explanations. In this work, we conduct a comprehensive evaluation of four widely used architectures—ResNet-50, DenseNet-121, ConvNeXt-Tiny, and ViT-B/16—trained across three random seeds (999, 123, and 42) to assess robustness. Models are trained and evaluated on the CheXpert dataset, with out-of-distribution (OOD) performance assessed using the NIH ChestX-ray14 dataset. To analyze model explainability, we generate Grad-CAM visualizations and evaluate them using the CheXlocalize dataset, enabling a systematic comparison of localization faithfulness.**

We benchmark all models using a broad suite of classification and explanation metrics and propose the Unified Classification–Explainability Score (UCES), a single metric that jointly captures predictive performance and explanation quality. UCES provides a principled way to identify the most reliable architecture when both accuracy and interpretability are essential. Our results show that ConvNeXt-Tiny consistently achieves the highest UCES across seeds, indicating superior overall performance in both classification and explainability. This unified evaluation framework offers a practical approach for selecting clinically reliable models in medical imaging tasks.

I. Introduction

Chest X-rays (CXRs) are among the most frequently used imaging modalities for screening and diagnosing thoracic diseases, making automated interpretation an important goal in medical AI [1], [2]. Deep learning models have demonstrated impressive performance on large-scale CXR datasets, often surpassing traditional machine learning approaches [3], [4]. However, despite promising accuracy, several challenges continue to limit their clinical deployment.

A key issue is the limited robustness of deep models to real-world distribution shifts. Models trained on a single source dataset frequently suffer performance degradation when evaluated on external datasets due to differences in population characteristics, imaging devices, or acquisition protocols [5], [6]. This raises concerns about generalizability and reliability beyond controlled research settings. Additionally, modern architectures—including convolutional networks and vision transformers—differ significantly in their inductive biases [7], [8], optimization behavior, and sensitivity to random initialization, making it unclear which models deliver stable performance across training seeds or evaluation environments.

Explainability presents another critical barrier. While methods such as Grad-CAM remain widely adopted in medical imaging for visualizing model attention [9], [10], recent studies show that saliency maps can be noisy, inconsistent, or misaligned with clinically relevant regions [11], [12]. This variability complicates efforts to build trust and interpretability into decision-making pipelines. Existing benchmarks often evaluate classification or explanation quality in isolation, overlooking the interplay between predictive accuracy and explanation fidelity. As a result, high-performing models may still generate poor or misleading explanations, posing safety risks in diagnostic applications.

These gaps highlight the need for a comprehensive, multi-dimensional evaluation framework that examines both classification robustness and explainability across diverse architectures, seeds, and dataset settings. Motivated by this need, our work conducts a systematic comparison of four widely used architectures—ResNet-50 [13], DenseNet-121 [14], ConvNeXt-Tiny [8], and ViT-B/16 [7]—trained on the CheXpert dataset [2] and evaluated under both in-distribution (CheXpert) and out-of-distribution (NIH ChestX-ray14) [3] conditions. We further analyze explainability performance using the CheXlocalize dataset, enabling a rigorous assessment of how faithfully different models localize disease-related regions.

By integrating a broad set of classification and explanation metrics, we introduce a unified evaluation perspective designed to identify models that are both accurate and interpretable. This provides a principled foundation for selecting architectures that are more likely to be clinically dependable in real-world chest X-ray analysis systems.

II. Methodology

We conduct 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, 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 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 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.

TABLE I: Per-seed performance on CheXpert test set

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 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 Summary of results (Mean $\pm$ Std across 3 seeds)

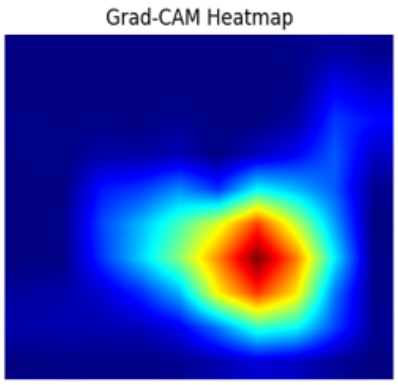
Model	AUROC (↑)	AP (↑)	Macro-F1 (↑)	Micro-F1 (↑)
ResNet-50	**0.759 $\pm$ 0.019**	0.244 $\pm$ 0.017	0.220 $\pm$ 0.010	0.310 $\pm$ 0.012
DenseNet-121	0.779 $\pm$ 0.026	0.257 $\pm$ 0.011	**0.237 $\pm$ 0.011**	**0.338 $\pm$ 0.021**
ViT-B/16	0.746 $\pm$ 0.042	0.223 $\pm$ 0.037	0.218 $\pm$ 0.019	0.320 $\pm$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

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

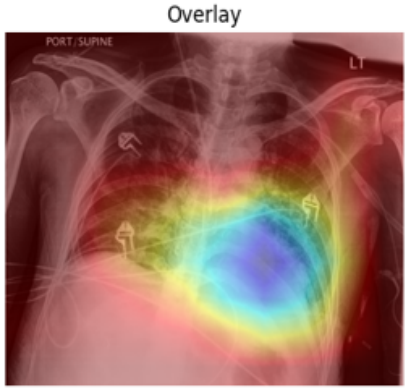
a) Overall Interpretation: Together, these metrics capture both local fidelity (through confidence-based perturbation tests) and structural quality (through sparsity and entropy) of the saliency maps. The combination provides a rigorous framework for comparing explainability performance across models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

F. Explainability Results Summary

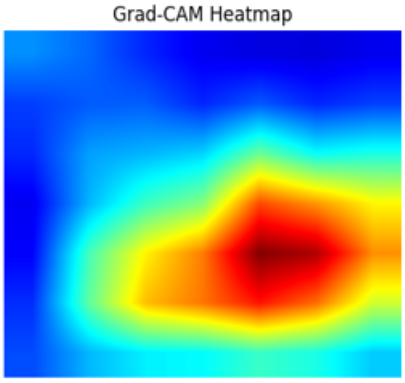
Qualitative explainability can be evaluated from the figure 1 where we find the the predicted Grad-CAM map is close to the original one. Table V reports the per-seed explainability scores for all architectures, while Table VI summarizes the average explainability across seeds. Several distinctive patterns emerge across the metrics.



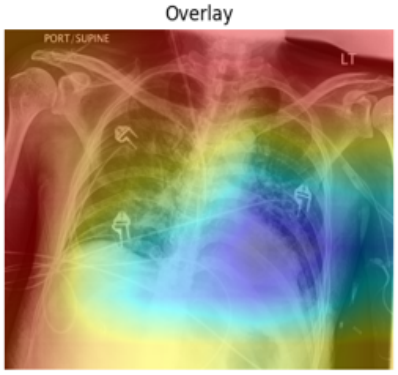
((a)) True-class Grad-CAM heatmap



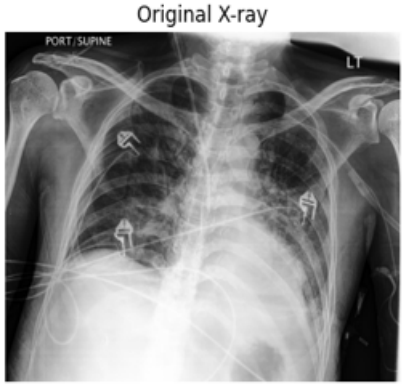
((b)) True-class Grad-CAM overlay



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



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



((e)) Original image

Fig. 1: Comparison between true-class and predicted-class Grad-CAM visualizations using the ResNet model (seed = 999).

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

TABLE V: Per-seed explainability performance mean across models

	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 ± Std 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

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 Unified Classification–Explainability Score

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.

A. Metric Computation

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 involves four conceptual stages. First, we select metrics that capture both classification performance and explanation fidelity. For classification, we use AUROC(↑), Average Precision (AP)(↑), Macro F1(↑), and Micro F1(↑). For explainability, we include Average Drop (↓), Increase in Confidence (↑), Entropy (↓), Sparsity (↑), Deletion AUC (↓), and Insertion AUC (↑).

Since these metrics differ in scale and interpretability, 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), we apply

$$x_{norm} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}.$$

For metrics where lower values are preferred (Average Drop, Entropy, Deletion AUC), we use

$$x_{norm} = \frac{x_{\max} - x}{x_{\max} - x_{\min}}.$$

After normalization, we compute the classification component C and the explainability component E as

$$C = \frac{1}{4}\Big(\text{AUROC}_n + \text{AP}_n + \text{MacroF1}_n + \text{MicroF1}_n\Big)$$

and

$$E = \frac{1}{6}\Big(\text{AvgDrop}_n + \text{IncConf}_n + \text{Entropy}_n + \text{Sparsity}_n + \text{DeletionAUC}_n + \text{InsertionAUC}_n\Big).$$

The final unified score is obtained using a weighted sum,

$$\text{UCES} = \alpha C + (1 - \alpha)E,$$

where we set $\alpha = 0.7$, placing slightly higher importance on classification performance while still incorporating explainability.

B. UCES Scores and Model Ranking

A summary of the UCES metric across all architectures is provided in Table VII.

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

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

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.

V. Conclusion

This work presented a comprehensive evaluation of four widely used deep learning architectures for chest X-ray analysis, examining not only their classification performance but also their robustness across training seeds, distribution shifts, and explainability quality. By training all models on the CheXpert dataset and evaluating them under both in-distribution and out-of-distribution conditions, we highlighted substantial variability in model performance that cannot be captured by accuracy alone. Our analysis of Grad-CAM explanations using the CheXlocalize dataset further demonstrated that explanation quality is highly model-dependent and does not always correlate with predictive performance.

To address the need for a unified assessment framework, we introduced the Unified Classification–Explainability Score (UCES), which integrates normalized classification and explanation metrics into a single, interpretable score. UCES provides a principled approach for identifying models that are simultaneously accurate and trustworthy. According to this metric, ConvNeXt-Tiny consistently achieved the highest overall performance across all random seeds, indicating strong robustness, reliable generalization, and superior explainability compared to the other evaluated architectures.

These findings emphasize the importance of jointly evaluating accuracy and interpretability when selecting models for clinical deployment. Future work may explore extending UCES to other tasks such as multi-label localization, integrating additional explainability methods, or evaluating more diverse datasets to further strengthen model reliability in real-world medical imaging settings.

References

[1] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Y. Ng, "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," *arXiv preprint arXiv:1711.05225*, 2017.

[2] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya *et al.*, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 590–597.

[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)*, 2017, pp. 2097–2106.

[4] T. Ozturk, M. Talo, E. A. Yildirim, U. B. Baloglu, O. Yildirim, and U. R. Acharya, "Automated detection of covid-19 cases using deep neural networks with x-ray images," *Computers in Biology and Medicine*, vol. 121, p. 103792, 2020.

[5] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study," *PLOS Medicine*, vol. 15, no. 11, p. e1002683, 2018.

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

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2021.

[8] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 11 976–11 986.

[9] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.

[10] A. Saporta, X. Gui, A. Agrawal, A. Pareek, S. Q. Truong, C. D. Nguyen, V.-D. Ngo, J. Seekins, F. G. Blankenberg, A. Y. Ng, M. P. Lungren, and P. Rajpurkar, "Benchmarking saliency methods for chest x-ray interpretation," *Nature Machine Intelligence*, vol. 4, no. 10, pp. 867–878, 2022.

[11] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 31, 2018.

[12] M. Reyes, R. Meier, S. Pereira, C. A. Silva, F.-M. Dahlweid, H. von Tengg-Kobligk, R. M. Summers, and R. Wiest, "On the interpretability of artificial intelligence in radiology: Challenges and opportunities," *Radiology: Artificial Intelligence*, vol. 2, no. 3, p. e190043, 2020.

[13] 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)*, 2016, pp. 770–778.

[14] 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)*, 2017, pp. 2261–2269.