# Jack of All Trades? Generalization Challenges in Neural Networks for Chest X-Ray Classification

**Charles Zhou**
Computer Science
Harvard University
czhou@college.harvard.edu

## Abstract

Deep learning methods have achieved highly accurate results on chest X-ray classification when training and test data share the same source. However, real-world deployment of AI models requires models to generalize across datasets beyond similar distributions with varying formats and labeling methods. In this study, we train convolutional neural networks on the VinBigData chest X-ray dataset and evaluate their performance on both in-distribution data and the out-of-distribution dataset CheXpert. We compare single model and ensemble results for both, and use GradCAM for visual interpretability of model behavior. While ensemble models achieve high levels of accuracy (over 95%) on in-distribution data, their performance significantly drops (to 26%) on out-of-distribution examples, highlighting generalization challenges. Our results suggest that domain shift remains a core limitation for medical imaging AI and emphasizes the need for broader evaluation and cross-domain adaptation strategies.

## 1 Introduction

Medical imaging is one of the most effective and widely used tools in modern health-related diagnostics, especially with radiology. Among the various applications that this can be applied to, chest X-rays are one of the most commonly performed examinations, with over 70 million being conducted each year within the U.S. alone. [1]. Analyzing them is useful as they allow for diagnostics of diseases such as cardiomelagy, lung opacities, and other torso-centered illnesses. Naturally, as artificial intelligence has progressed, applying these tools to automate chest X-ray interpretation has become a point of interest. After being trained on large annotated datasets, deep learning models such as those based on convolutional neural network (CNN) architecture have had promising results: being capable of improving physician accuracy for these datasets [2].

Despite these recent advances, however, lies a core challenge: generalizing beyond just the dataset across patients and imaging domains. Many studies have found the application of AI-image detection tools to be successful for in-distribution data [3]. While demonstrating high performance for test sets sampled from the same source as the training data could be important for validation, there remains a significant issue of generalization across sources. When models are applied to out-of-distribution (OOD) images, such as other slight variation in imaging techniques, hospitals, or countries, there is a drop in capability [4]. Despite similar protocols these models often struggle to perform, and this failure to generalize can lead to real-world consequences, especially since its primary usage is in clinical settings where data often various greatly between patients, imaging practices, and equipment [5, 6].

This problem is especially apparent in chest X-ray classification tasks where large datasets like VinBigData, CheXpert, and MIMIC-CXR contain similar disease labels but differ in exact methods, imaging practices, photo resolution, file types, and context. For example, VinBigData uses bounding

Preprint. Under review.

boxes for abnormalities from radiologists [7]. In contrast, CheXpert uses an automated rule-based labeler that extracts observations from reports via natural language processing, turning them into structured labels [8]. However, despite similarities in "labels" from the outside perspective, there are large differences in prevalence, noise, and even practices that differ between datasets. Thus, a model trained on one of the datasets could potentially overfit to in-distribution artifacts or not pick up the correct signals to allow for generalization to OOD data.

In this paper, we study the extent to which a neural network trained on one chest X-ray dataset can generalize to another. More specifically, we trained a ResNet-18 model on the VinBigData's VinDr-CXR dataset and evaluated the model on Stanford AIMI's CheXpert dataset, experimenting with different training splits and ensemble methods to determine how generalization is affected. Our results show that, as expected, models perform well on their in-distribution validation data, but their performance significantly depreciates on OOD data, underscoring the importance and challenge of generalization in medical imaging AI.

## 2 Background

Deep learning has become a dominant approach in medical image analysis, especially using CNN architecture. In this study, we use deep residual learning (ResNets) which has been shown to perform well for image recognition tasks [9]. This approach of using ResNet architecture has been extensively explored and is common practice, and we choose to use a smaller 18-layer network instead of a larger network like the 121-layer DenseNet due to the capabilities of achieving the same results with less compute and complexity [10].

## 3 Methods

### 3.1 Dataset

We used two large public datasets of chest X-rays: VinBigData and CheXpert. Original format of the VinBigData dataset was in DICOM, which is a standard file format for medical images but is not typically processed by most computers. These were large in resolution and greyscale. To train a model that had to deal with multiple data formats and resolutions, we opted to create a standardized method of processing the data so that there would be no bias towards one specific format for the model. To achieve this, we converted each image of the VinBigData to greyscaled JPEGs and applied equalization due to different coloring scales. Afterward, all the images were resized to 224x224 pixels, matching the input resolution that our model architecture expected through this streamlined process. Similar methods were done for CheXpert, though they did not originally come in DICOM format. Ultimately, the pipeline was the same, applying necessary methods to achieve the same input representation across datasets. This was done for both training and validation data splits for both datasets. Figures 1a and 1b show representative examples from both datasets post-processing.

After processing the data, we took a subset of the top six most common diseases from the VinBigData dataset that had overlap with those in CheXpert due to compute constraints: No finding, Cardiomelagy, Nodule/Mass, Lung Opacity, Pleural Effusion, and Consolidation. By only taking this subset for training and validation for both datasets, we avoid introducing unnecessary noise while also preserving the exploration of how the model generalizes by having many classes. We mapped these consistently to both datasets to numbers (0-5). From this we took a further randomly sampled subset to have a maximum amount of data for a class as to not overwhelm the model in training and also to allow for reasonable training times. In total, we had a subset of 3,204 images from VinBigData to train our model on.

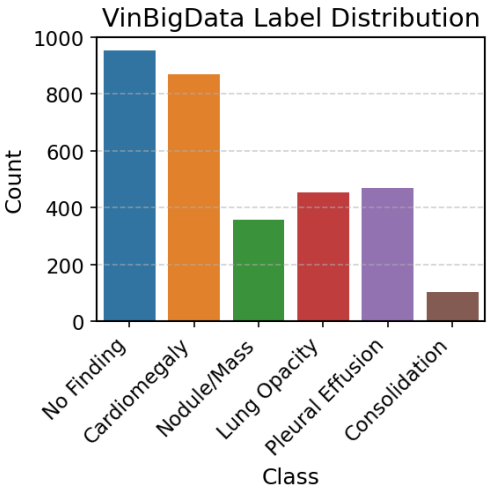
Naturally, class imbalance could impact our models. To better understand this imbalance within the datasets, we visualized the distribution of the size labels that we selected:

As evidence above, "No Finding" is the most common class, followed by Cardiomelagy in the VinBigData dataset which we used for training. Other labels such as Consolidation are underrepresented in comparison, which potentially posed as challenges for training balanced classifiers. This variance likely carried over to our results of applying the model on CheXpert's OOD data.

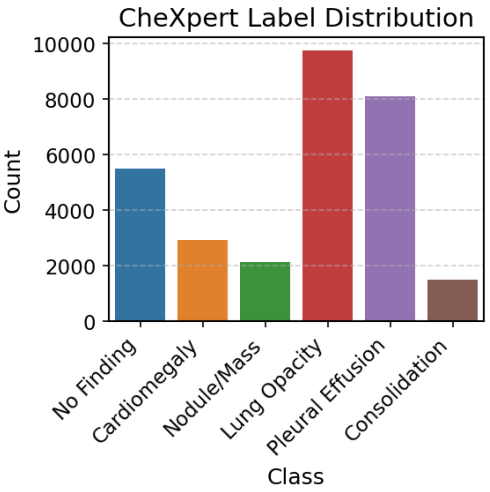

(a) Example of VinBigData after reformatting


(b) Example of CheXpert after reformatting

Figure 1: Reformatted examples from the VinBigData and CheXpert datasets.




(a) Class Distribution for VinBigData




(b) Class Distribution for CheXpert

Figure 2: Comparison of class distributions between VinBigData and CheXpert datasets.

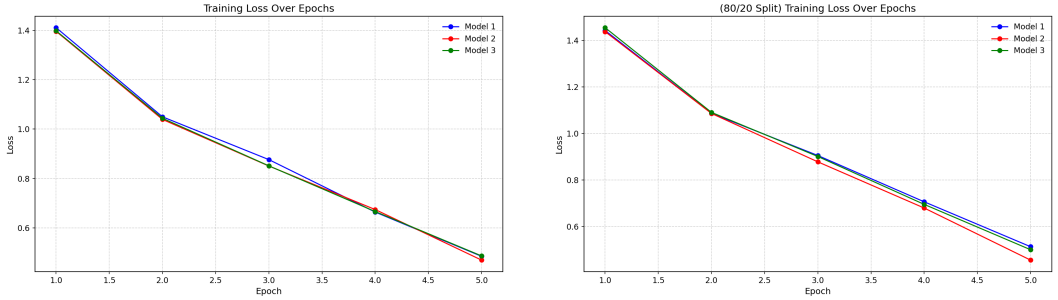
### 3.2 Modeling

As mentioned prior, we used ResNet 18 model architecture in our training as CNNs are widely used for image recognition and the 18-layer depth has been shown have sufficient performance. The marginal improvement of a deeper network like ResNet 50 or DenseNet 121 is computationally not worthwhile and is also much more prone to overfitting.

There were two training datasets that we used of the subset mentioned above: the full VinBigData train dataset and a 80/20 training and validation split dataset. We did this for a few reasons. First, since these are fairly deep CNNs, more training images would likely improve the model's performance. Second, we hoped to achieve a semblance of validation of our initial model before testing it against the dataset itself or OOD data. To achieve this, we trained models on a 80/20 split to somewhat replicate a five fold cross validation split. Due to training constraints, we replicated one fold. We hypothesized that the large size of the dataset would reduce the variance in outcome if any were to

exist. The 20% validation data was withheld during training. We created a data loader that could handle our formatted data which used batch sizes of 32 and shuffled our data.

For these two different partitions, we trained three models each, introducing randomness with data shuffling in an attempt to avoid local minima. All models were trained for 5 epochs using stochastic gradient descent with momentum of 0.9, learning rate of $1e-3$, and L2 penalty without weight decay. Cross-entropy loss was applied as our objective function due to the multiclass classification setting we are working in. Thus, the models output a probability distribution which we can use later to ensemble. Training was performed based off whichever device was available to CUDA at time of training, though a majority of this training was done in a CPU environment which limited complexity. The loss was recorded in Figure 3.




(a) Training loss curve for full dataset model

(b) Training loss curve for split dataset model

Figure 3: Training loss curves for full vs. split dataset models

## 4 Results

We evaluated our models using two methods: single model output and an ensemble output. Thus, we define four total model configurations based on training regime and output method: (1) full-data single model, (2) full-data ensemble, (3) val-split single model, and (4) val-split ensemble. The "ensembled" model represents the output of an ensemble prediction of three underlying independently trained ResNet-18 models. The "val-split" configuration refers to models trained on only 80% of the VinBigData training set, with 20% held out. The "full-data" models were trained on the entire VinBigData subset.

Given three independently trained models $\{f_1, f_2, f_3\}$, we define the ensemble prediction $\hat{y}$ for an input image $x$ as:

$$\hat{y} = \frac{1}{3}\sum_{i=1}^{3} \text{softmax}(f_i(x)), \quad \hat{c} = \arg\max_j \hat{y}_j$$

In the rare event of a tie between classes, we select the class $\hat{c}$ with the lower index, as reflected by NumPy's argmax function. However, this is unprobable to occur due to floating point precision. Overall, this corresponds to averaging the softmax probability vectors output by each of the three models. This technique of ensembling is a standard approach that typically reduces variance within predictions and improves robustness by "marginalizing" away the risk [11].

For the VinBigData dataset, the probability and thus resultant expected accuracy for naively guessing can be found. There are two methods of naive guessing: uniform random guessing and proportional random guessing. Under uniform random guessing, we naively output a random guess between the possible classes for each datapoint, which in this case is $K = 6$. Thus, our expected accuracy from this would be:

$$\mathbb{E}[\text{Accuracy}] = \sum_{i=1}^{K} P(\hat{y} = y_i) \cdot P(y = y_i) = \frac{1}{K} = \frac{1}{6} \approx 0.167$$

Alternatively, under proportionally random guessing, we would use the distribution from the training set and aim to assign $P(\hat{y} = y_i) = P(y = y_i)$. This would work best if the validation data/test data

follows a similar proportional distribution. At best, this would have improved accuracy but would still not be sufficient in the real-world. For our subset of data, the expected accuracy becomes:

$$\mathbb{E}[\text{Accuracy}] = \sum_{i=1}^{K} P(\hat{y} = y_i) \cdot P(y = y_i) = \sum_{i=1}^{K} [P(y = y_i)]^2 \approx 0.217$$

This is better than uniformly random guessing and will serve as our baseline for evaluating our models.

We tested our models on in-distribution and out-of-distribution data. For the in-distribution, we had the following test datasets: the full dataset for full-data models and the withheld validation set for the val-split models. The OOD data was the entire CheXpert dataset containing only labels from the six overlapping categories. The results of both is in Table 1.

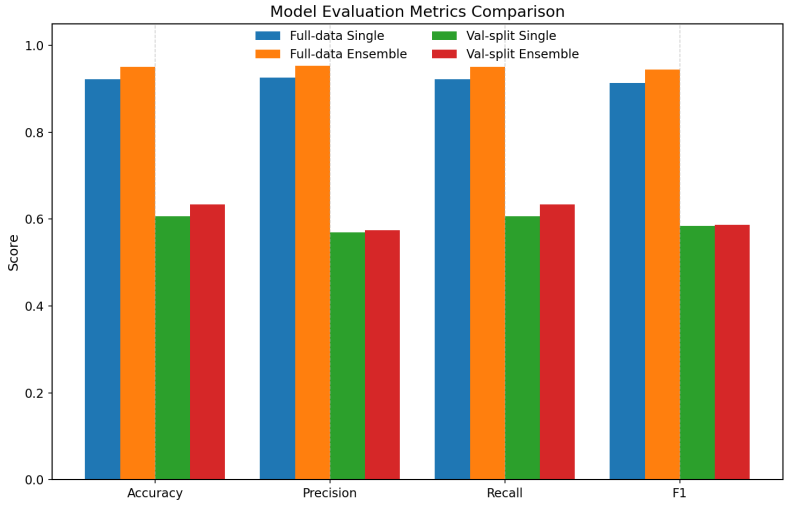



Figure 4: Model Evaluation Metrics for In-distribution

We evaluated the models on in-distribution data and plotted the outcomes to better visualize results. As shown in Figure 4, models trained on the full dataset consistently outperformed those trained on a validation split across all metrics. Additionally, ensemble models yielded slightly higher scores than their single-model counterparts, highlighting the benefits of model aggregation. These observations support the importance of using both full data availability and ensemble methods for achieving robust and high-performing models. We also caution overfitting in this step as it is possible that the full-data models have too much variance and too little bias. Since our ultimate test data is the CheXpert dataset, we attempt to validate our models with that dataset and observe the differences in results. Due to the relatively low performances of the val-split configurations, we anticipate that they are not overfitting, setting their performance as a benchmark for comparison.

Table 1: Accuracies (recall) of all trained models across datasets

| **Model (both ensembled)** | **VinBigData** | **CheXpert** |
|---|---|---|
| Full-data Single Model | 92.26% | – |
| Full-data Ensembled | 95.04% | 26.58% |
| Val-split Single Model | 60.69% | – |
| Val-split Ensembled | 63.34% | 25.30% |

From the OOD dataset evaluations, we observe that the performance of all the models significantly drops when changing away from the source material. While this is expected behavior, the extent to which performance falls is notable. For the best configuration, this is a 95.04% to 26.58% decrease, with the val-split configuration dropping by a less severe difference. This suggests severe limitations in model generalization, even when controlling for label overlap.

Additionally, it is worth considering the bias-variance tradeoff within these models. Considering that the full-data ensemble model still performed better than the val-split ensemble model on the large OOD dataset, we are inclined to believe that the full-data model is not extremely overfit. While this does not suggest that the local minima between the tradeoff was found, it does provide information in regards to general performance capabilities. In future studies, a model that is trained on full-data for less epochs could provide a greater understanding of how optimized our models truly are.

We can calculate the expected accuracy from naively guessing based on proportional random guesses for the CheXpert dataset as a fundamental baseline. From the label distribution in Figure 2, we can find the expected accuracy to be:

$$\mathbb{E}[\text{Accuracy}] = \sum_{i=1}^{K} [P(y = y_i)]^2 \approx 0.231$$

With the entire dataset size of 29,842 images, we see that both of our ensemble models perform better than this baseline accuracy. This performance, however, is distant from the models' capabilities for in-distribution data, further cementing difficulties in generalization beyond concerns for overfitting.

In Table 2, we further examine the per-class performance on the OOD CheXpert dataset for the two ensemble models. The "No Finding" class achieves the greatest accuracy. Although this class is overrepresented in the training VinBigData dataset, it is not the most common label within CheXpert, suggesting that the model may have a lower confidence threshold in guessing "No Finding" or may have picked up important patters. However, rarer cases in the trainin data such as Nodule/Mass and Consolidation show significantly lower accuracy, which further highlights the class imbalances and the challenges from it.

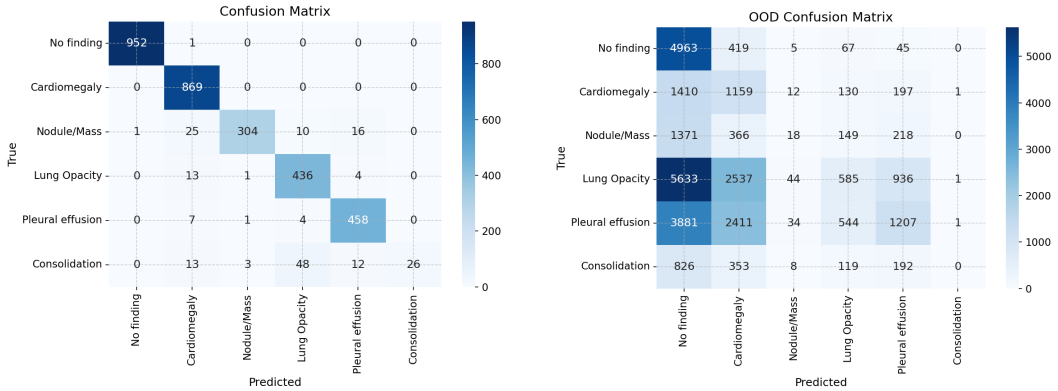
Table 2: Per-class accuracy of ensembled models on CheXpert dataset[1]

| **Model (Ensembled)** | **Overall** | **0** | **1** | **2** | **3** | **4** | **5** |
|---|---|---|---|---|---|---|---|
| Full VinBigData | 27% | 90% | 40% | 2% | 10% | 22% | 0% |
| Split VinBigData | 25% | 91% | 48% | 1% | 3% | 10% | 0% |

Figure 5a shows the confusion matrix for the full-data ensemble model evaluated on VinBigData, while Figure 5b shows the same model evaluated on CheXpert. The in-distribution matrix shows strong diagonal dominance, which signals that a majority of the chest X-rays are being correctly classified. In contrast, the OOD matrix reveals much more widespread misclassifications. Particularly for rarer diseases like Nodule/Mass and Consolidation, the model correctly classifies little to none of the images. Instead, a majority of the images for all the labels are classified directly as "No Finding," with many others being incorrectly classified as "Cardiomegaly." This behavior is reflected in the val-split configuration as well, and could potentially come from bias regarding class imbalances.

However, we note that certain diseases are still classified with some success despite class imbalances. For example, Pleural Effusion is not a common class in VinBigData but has greater per-class performance than other labels. Other more diseases such as Lung Opacity are commonly misclassified on a basis which seems to be related to proportionality. In general, this behavior seems to indicate that the extent to which models can generalize on OOD data is influenced by the composition of the in-distribution data that it is trained on, even when controlling for discrepancies.

[1]Class mapping: 0 = No Finding, 1 = Cardiomegaly, 2 = Nodule/Mass, 3 = Lung Opacity, 4 = Pleural Effusion, 5 = Consolidation



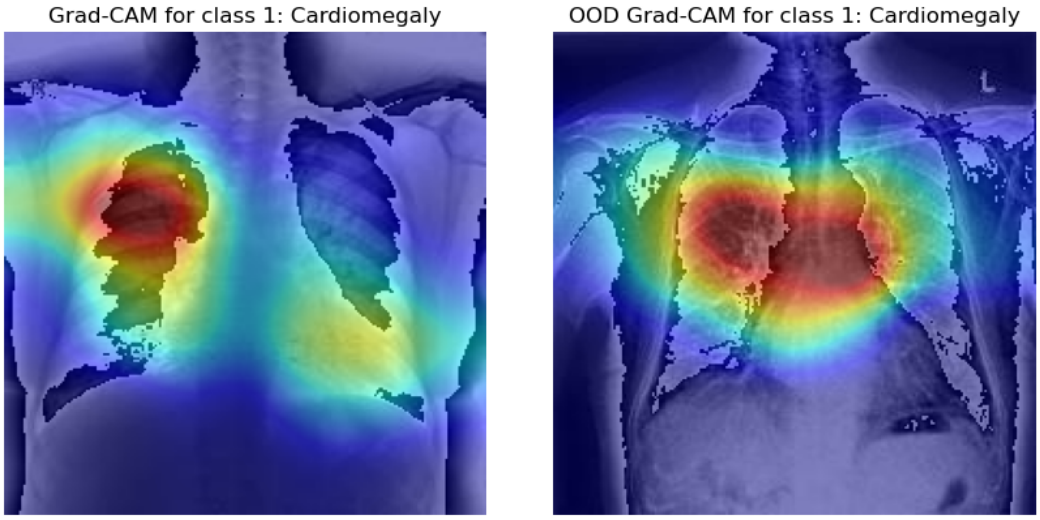

(a) In-distribution confusion matrix

(b) Out-of-distribution confusion matrix

Figure 5: Comparison of confusion matrices for in-distribution and out-of-distribution model predictions.

### 4.1 Visual Explainability

To learn more regarding model behavior beyond just accuracy, we employed GradCAM to generate saliency maps to highlight regions of interest used in classification [12]. This technique has been used in other medical imaging studies for interpretation, and we aim to examine them as well to provide insight [13]. In both VinBigData and CheXpert, our model highlighted meaningful regions depending on disease. For example, we include in Figure 6 cases for Cardiomegaly.

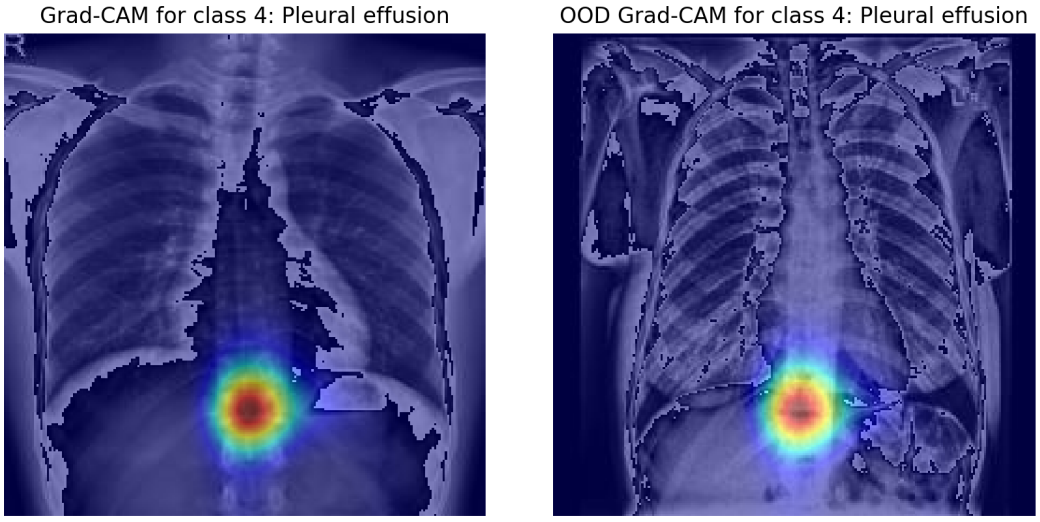



(a) Grad-CAM on VinBigData for cardiomegaly

(b) Grad-CAM on CheXpert for cardiomegaly

Figure 6: Grad-CAM visualizations highlighting regions associated with cardiomegaly predictions in VinBigData and CheXpert datasets.

Due to the type of disease, this GradCAM heatmap emphasized a central cardiac silhouette. However, in other diseases, the points of interest are different, depending on the different aspects and image features the model uses. For example, in Figure 7, we present overlays for a different disease. Pleural effusion is an illness in which liquid builds up inside the lungs of a person, typically toward the bottom. We can see that the highlighted area corresponds to the real-world presence of the disease in both images, suggesting that the model is understanding it. In general, from qualitatively comparing GradCAM highlights across diseases, we observe that the same relevant regions are considered for both datasets. However, some images from the CheXpert dataset had apparent inconsistencies in saliency regions. These could be a result of different reasons such as image textures, imaging methods




(a) Grad-CAM on VinBigData for pleural effusion

(b) Grad-CAM on CheXpert for pleural effusion

Figure 7: Grad-CAM visualizations for Pleural Effusion predictions from models trained on VinBigData and CheXpert.

and framing, and poor generalization as a whole, which impacts the model's confidence on the images. Overall, we see from the GradCAM that generalizing across datasets is a difficult task due to lower confidence in the predictions, potentially arising from different framing, characteristics, and features being important.

## 5 Conclusion

Our study highlights a fundamental challenge in deploying deep learning systems for medical imaging: generalization. While models trained on VinBigData achieved high in-distribution performance (surpassing 95% accuracy when ensembled), their effectiveness dropped significantly when applied to out-of-distribution data from CheXpert, declining to near-baseline levels. This drop in performance reflects the underlying lack of robustness that models have when exposed to even subtle domain shifts, despite consistent labeling methods.

We found that ensembling helped stabilize predictions and marginally improved accuracy but was insufficient in overcoming the domain shift. Additionally, GradCAM visualizations revealed that models trained on one dataset had similar regions of image relevance but occasionally experienced drifts in confidence, which may be due to data discrepancies (such as anatomical framing of the images), offering a visual explanation for some of the performance decay.

Ultimately, these results reiterate that high in-distribution accuracy alone is not a sufficient criteria for real-world medical AI. In the future, we plan on investigating domain adaptation techniques (e.g. optimal transport), improved data reformatting methods, and pretraining strategies (e.g. contrastive learning) to improve generalization. For trustworthy models that can be used in a clinical setting, a broader evaluation using external datasets is necessary to push beyond simply overfitting to specific benchmarks. Through these methods, we aim to develop more robust and accurate models that can be applied to real-world settings.

## References


[1] Catherine M Jones, Quinlan D Buchlak, Luke Oakden-Rayner, Michael Milne, Jarrel Seah, Nazanin Esmaili, and Ben Hachey. Chest radiographs and machine learning – Past, present and future. *Journal of Medical Imaging and Radiation Oncology*, 65(5):538–544, August 2021.

[2] Pamela G. Anderson, Hannah Tarder-Stoll, Mehmet Alpaslan, Nora Keathley, David L. Levin, Srivas Venkatesh, Elliot Bartel, Serge Sicular, Scott Howell, Robert V. Lindsey, and Rebecca M. Jones. Deep learning improves physician accuracy in the comprehensive detection of abnormalities on chest X-rays. *Scientific Reports*, 14(1):25151, October 2024. Publisher: Nature Publishing Group.

[3] Preetham Putha, Manoj Tadepalli, Bhargava Reddy, Tarun Raj, Justy Antony Chiramal, Shalini Govil, Namita Sinha, Manjunath KS, Sundeep Reddivari, Ammar Jagirdar, Pooja Rao, and Prashant Warier. Can Artificial Intelligence Reliably Report Chest X-Rays?: Radiologist Validation of an Algorithm trained on 2.3 Million X-Rays, June 2019. arXiv:1807.07455 [cs].

[4] Thomas Eche, Lawrence H. Schwartz, Fatima-Zohra Mokrane, and Laurent Dercle. Toward Generalizability in the Deployment of Artificial Intelligence in Radiology: Role of Computation Stress Testing to Overcome Underspecification. *Radiology: Artificial Intelligence*, 3(6):e210097, October 2021.

[5] Yuzhe Yang, Haoran Zhang, Judy W. Gichoya, Dina Katabi, and Marzyeh Ghassemi. The limits of fair medical imaging AI in real-world generalization. *Nature Medicine*, 30(10):2838–2848, October 2024. Publisher: Nature Publishing Group.

[6] L. Zhang, X. Wang, D. Yang, H. Roth, A. Myronenko, D. Xu, Z. Xu, T. Sanford, B. Turkbey, B. J. Wood, and S. Harmon. Generalizing Deep Learning for Medical Image Segmentation to Unseen Domains via Deep Stacked Transformation. *IEEE transactions on medical imaging*, 39(7):2531–2540, July 2020.

[7] Ha Q. Nguyen, Khanh Lam, Linh T. Le, Hieu H. Pham, Dat Q. Tran, Dung B. Nguyen, Dung D. Le, Chi M. Pham, Hang T. T. Tong, Diep H. Dinh, Cuong D. Do, Luu T. Doan, Cuong N. Nguyen, Binh T. Nguyen, Que V. Nguyen, Au D. Hoang, Hien N. Phan, Anh T. Nguyen, Phuong H. Ho, Dat T. Ngo, Nghia T. Nguyen, Nhan T. Nguyen, Minh Dao, and Van Vu. VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations. *Scientific Data*, 9(1):429, July 2022. Publisher: Nature Publishing Group.

[8] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison, January 2019. arXiv:1901.07031 [cs].

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition, December 2015. arXiv:1512.03385 [cs].

[10] Keno K. Bressem, Lisa C. Adams, Christoph Erxleben, Bernd Hamm, Stefan M. Niehues, and Janis L. Vahldiek. Comparing different deep learning architectures for classification of chest radiographs. *Scientific Reports*, 10:13590, August 2020.

[11] Thomas G. Dietterich. Ensemble Methods in Machine Learning. In *Multiple Classifier Systems*, pages 1–15, Berlin, Heidelberg, 2000. Springer.

[12] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *International Journal of Computer Vision*, 128(2):336–359, February 2020. arXiv:1610.02391 [cs].

[13] Yunyan Zhang, Daphne Hong, Daniel McClement, Olayinka Oladosu, Glen Pridham, and Garth Slaney. Grad-CAM helps interpret the deep learning models trained to classify multiple sclerosis types using clinical brain magnetic resonance imaging. *Journal of Neuroscience Methods*, 353:109098, April 2021.