An Empirical Evaluation of Reducing Spurious Signals in Chest X-Rays via Non-ROI Masking

Rhys Compton
Courant Institute
New York University
New York City, NY
rc4499@nyu.edu

Abstract

Deep learning has exhibited strong performance over a range of computer vision tasks with an especially large research area focusing on the medical imaging domain, given the high cost of human prediction and large potential upside of automating this procedure. However, generalizability of these models is often underwhelming and remains a significant barrier to widespread adoption. Often this poor performance on new (out-of-distribution) data is due to hidden spurious signals in the original dataset which the model leverages during training. In chest x-rays, previous work has indicated these signals to be outside the lung boundaries (non-ROI) where left/right markers and other auxiliary information is placed. To examine the significance of these artifacts on model performance, lung boundary masks are acquired for two classification datasets and applied to hide the non-ROI region within each image. Classification performance on the resulting datasets show mixed improvement in model generalizability when predicting pneumonia on an out-of-distribution dataset, indicating that masking the non-ROI region may force the model to learn more from physiological signals over spurious background artifacts. Despite this promising result, when trained *directly* to leverage these spurious artifacts (by predicting hospital system), the model is able to predict hospital systems with near perfect accuracy under a heavily data-constrained scenario, implying that significant noise exists *within* the ROI, requiring more advanced techniques than simple masking to remove it. Code can be found at `https://github.com/basedrhys/non-roi-masking`.

1 Introduction

Computer-aided prediction of radiographic images is a task well suited for the application of deep learning, given the large size of datasets and high cost of human prediction (requiring expert radiologists). Approaches that tackle this task commonly report performance metrics on a held-out test set, which although is unseen during training, is from the same distribution as the original training data and so gives an overestimate of real-world performance when applied to new datasets due to distribution shift. Unfortunately, model performance often drops under these conditions [14] which has created a barrier to real-world adoption of this technology; if model performance degrades significantly when applied to images from a new hospital, the model is learning from a hidden/unwanted signal in the data (assuming that a given disease presents identically wherever imaged). In this report I employ the term *nuisance* to refer to this signal [8]. Due to the often-insidious nature of these nuisances (difficult/impossible for humans to recognize), if it changes imperceptibly in production, the model performance can suffer without a clear indication.

Preprint. Under review.

One area identified in previous work is the area outside the ROI (referred to as non-ROI), which can include information such as the L/R marker, x-ray position description, and other patient information [14]. This project performs an empirical investigation into this nuisance signal and proposes a candidate solution: *non-ROI masking*. If the primary spurious signals are in the non-ROI region, then masking this out should force the model to learn from more physiologically-based signals; as a consequence, generalization (as measured by performance on an out-of-distribution dataset) should improve. In the case of no change in performance, these experiments provide evidence for a counter-statement: the primary (or significant) nuisances exist within the image ROI, meaning they are non-trivial to remove. In particular, I am asking two research questions:

- *Does masking the non-ROI improve model generalization (by forcing the model to learn from more physiologically-based signal over nuisance)?*
- *How strong and deeply embedded can this nuisance info be? Can it be mitigated significantly via non-ROI masking?*

To investigate the strength/location of nuisance data in Chest X-rays, I perform an experimentation of non-ROI masking: hiding everything in the image outside the lung boundaries (which shouldn't be used for disease prediction). To achieve this, a UNET model is trained to segment lung boundaries in PA chest x-rays and applied to images in two curated classification tasks: *pneumonia prediction* and *hospital prediction*. A separate classification model is then trained and evaluated on variations of each dataset, with the resulting performance acting as a proxy answer to the research questions above.

The experiments indicate that non-ROI masking has promise as a method for improving model generalizability, shown by a significant improvement in classification performance when applying a trained pneumonia predictor to an unseen dataset from a separate hospital. This trend does not hold in all out-of-distribution cases, however, so further experimentation on other datasets is required to validate these findings.

When predicting hospital system from chest x-rays, the model is able to predict with perfect accuracy under all mask variations (variations outlined in §2.3), and with near-perfect accuracy under a heavily data-constrained scenario (limited training instances, tiny image region). These results confirm findings found in previous work [1], that medical image datasets contain deeply embedded spurious artifacts *within the ROI*, and further processing should be applied to mitigate these artifacts before training deep learning models for prediction.

2 Evaluation Datasets

I employ two binary classification tasks to evaluate the effectiveness of non-ROI masking, namely *Pneumonia Prediction* and *Hospital Prediction*.

2.1 Pneumonia Prediction

Pneumonia is a common lung disease, seriously effecting pediatric patients and causing an estimated 2 million deaths in children under 5 years old every year making it the estimated leading cause of childhood mortality [10]. This disease is commonly diagnosed by radiological analysis of chest x-rays, so is a prime candidate for deep learning.

To evaluate the domain-generalizability improvements caused by non-ROI masking, I use two pneumonia classification datasets; one for training/validation and the other for testing. Using a completely distinct test dataset (rather than simply reserving some portion of the original dataset) is important as this scenario (referred to as *out-of-distribution*) is where deep learning models commonly fail [14].

Refer to Appendix A for dataset sizes. The first pneumonia dataset is the *Pediatric* Pneumonia Detection dataset [7], comprised of 5k chest X-rays from children.

The second pneumonia prediction dataset is the *CheXpert* Pneumonia Detection [5] dataset. CheXpert is a large multi-label chest X-ray classification dataset (224,316 images) with each instance containing 0 or more of 14 different pathologies. To make this comparable to the Pediatric dataset, I modified the dataset into a binary classification task by labelling an instance as `NORMAL` if the `No Finding` label

was positive (and no other labels positive), and labelled `PNEUMONIA` if the `Pneumonia` label was positive. Other labels were occasionally also positive with `Pneumonia`, but these instances were left in due to their strong correlation/indication of pneumonia (e.g., `Pneumonia` causes `Lung Consolidation`, so one would expect these two labels to co-occur frequently in the dataset). The resulting dataset contains 14228 instances (9553 `NORMAL`, 4675 `PNEUMONIA`).

2.2 Hospital Prediction

I design a binary hospital prediction task to test the research question directly: *how strong and deeply embedded can this nuisance information be?*; a model tasked with predicting which hospital an x-ray came from will utilise all spurious/nuisance features available in the data, so training in this manner will learn from the nuisance signals clearly. It should be noted that this is an "anti-task", where I am aiming for the lowest prediction accuracy possible; this result would imply that images from each hospital are visually homogenous, reducing/eliminating the effect of nuisance signals and the downstream problems they cause.

The dataset for this task is collated from two open-access chest x-ray datasets:

- **CheXpert** [5]: 224k images, sourced from Stanford Hospital.
- **ChestX-ray14** [13]: 112k images, compiled by the NIH from the NIH Clinical Center.

CheXpert contains both PA (front view) and Lateral (side view) x-rays, while **ChestX-ray14** only contains PA images. All lateral x-rays (32,419) were removed to induce further homogeneity and avoid any *obvious* spurious signals in the data that may contribute to predicting hospital system. The *Indiana University* [2] dataset which was used in previous work on generalizability [14] is open-access, however the lateral images are not labelled clearly and cannot be trivially removed, so this dataset was left out of experimentation.

Refer to Appendix A for exact dataset sizes.

2.3 Dataset / Mask Processing

Once fully trained, the *lung segmentation* model (§3.1) was used to generate lung masks for all images in the evaluation datasets. Predicted segmentation maps are commonly noisy with irregular borders, so basic image processing methods were applied to smooth the masks out. A morphological open was applied to remove small noisy blobs while retaining the larger lung regions, before applying a median blur to smooth the edges of the mask. I also removed instances where the smoothed mask was too small ($< 20\%$ of the total image area) as this was usually due to a poorly predicted mask (e.g., missed a lung); these post-processing steps aim to limit any confounding effects of applying the masks themselves. Appendix A details the evaluation datasets used, with the final *postprocessed* size in parentheses. From these smoothed masks, four versions of each dataset were created to better identify the behaviour of models trained under non-ROI masking (refer to Figure 1 for example images):

- `None` (baseline): No masking applied. This leaves all information visible to the model to learn from, so is expected that maximal learning from nuisance signals will occur.
- `Raw`: Apply the smoothed masks as-is to the original image. The segmentation masks for each lung trace tightly around the lung boundary, excluding the heart outline and space between left/right lungs, so this version of the dataset masks out as much of the image as possible while retaining the regions where lung disease can be present.
- `Convex Hull`: Create a convex hull from the smoothed masks. This includes more anatomical information than `Raw` (space between lungs, heart, oesophagus) while still excluding non-ROI information.
- `Crop`: Naive crop strategy - ignore the predicted mask and only keep the center third square of the image. Although hiding most/all of the non-ROI, this also hides much of the lung area so is expected to decrease disease prediction performance. Compared to the masking approaches above, the `Crop` masked area is held constant across images, ensuring that applying the mask is not adding any more spurious signals; for example, some images may have poorly predicted lung masks, so applying these back to the image could sway the model's predictions even further.

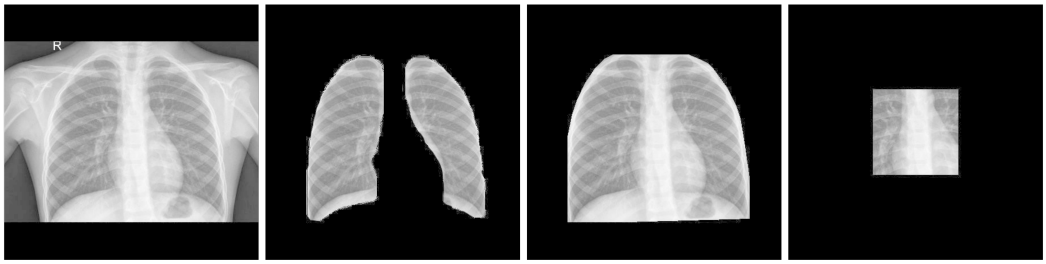

Figure 1: Dataset versions

Data Variation	CheXpert → Pediatric F1	CheXpert → Pediatric AUROC	Pediatric → CheXpert F1	Pediatric → CheXpert AUROC
None	0.729 (0.02)	0.876 (0.01)	0.533 (0.01)	0.715 (0.01)
Crop	0.707 (0.02)	0.767 (0.01)	0.413 (0.01)	0.617 (0.01)
Raw	0.524 (0.01)	0.753 (0.02)	0.526 (0.01)	0.696 (0.01)
Convex Hull	0.615 (0.03)	0.798 (0.02)	0.576 (0.01)	0.777 (0.01)

Table 1: Out of distribution results for pneumonia classification. Left columns represent model trained on CheXpert and evaluated on Pediatric, vice versa for right columns. SEM in parentheses. Color coding is applied column-wise from green-red.

3 Models

3.1 Lung Segmentation Model

To study the spurious information present in the non-ROI region of chest x-rays, we need accurate lung masks to mask this region. Unfortunately, the evaluation datasets used do not have ground truth lung/chest masks included so I trained a *lung segmentation* model to acquire these. The lung segmentation model training dataset consists of two public chest X-ray datasets compiled by the U.S. National Library of Medicine designed for the diagnosis of pulmonary tuberculosis [6], containing PA chest x-rays from Montgomery County Hospital and Shenzen No.3 People's Hospital. There are 704 images with expert-labelled lung boundary masks which were used as the ground-truth label for the lung segmentation model.

I use the UNET [9] architecture with an ImageNet pretrained ResNet50 [3] as the backbone. The model is trained for 8 epochs with a learning rate of $1e^{-5}$ using the fastai[4] library, reaching a Sørensen–Dice coefficient (Dice score) of 0.965.

3.2 Image Classification Model

I trained an image classification model on each dataset version to evaluate the effect that non-ROI masking has on downstream model performance. Images were resized to a square 448x448 size with zero-padding. An ImageNet pretrained ResNet50 model is used as the classification model for the evaluation datasets, trained until AUROC saturates (~12 epochs) with a batch size of 32 and learning rate of $3e^{-4}$.

4 Results

4.1 Pneumonia Prediction

Each model is evaluated in an out-of-distribution manner: trained on one pneumonia classification dataset, and tested on another. This testing regime presents a more accurate view of real-world model performance than simple train/val/test splitting, due to the distribution shift between the two datasets simulating real-world applications.

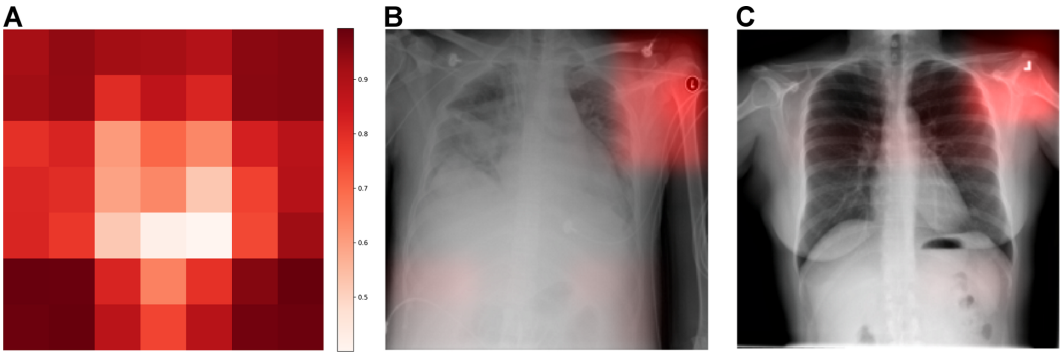

Figure 2: Saliency Map from [14], showing strong reliance on non-ROI signal

Refer to table 1 for results. The `Raw` mask version performed worst overall, likely due to the minimal information that's kept (only that strictly within eacn individual lung boundary), as well as the potential for confounding information with the detailed lung mask.

CheXpert → Pediatric: When training on *CheXpert* and evaluating on *Pediatric*, the highest performance was obtained by the `None` model, achieving an F1 and AUROC of **0.729** and **0.876**, respectively. Interestingly, the `Crop` model achieved the second highest F1 score (even beating `Convex Hull` which can see the entire chest cavity), indicating that cropping to the center third may maintain some physiological signal (or at least enough to predict pneumonia). The `Raw` model performed the worst.

Pediatric → CheXpert: Swapping the train/test datasets (model trained on *Pediatric* was evaluated on the *CheXpert*), the `Convex Hull` model generalised best, achieving an AUROC of 0.777, a significant improvement over the `None` model which only achieved an AUROC of 0.715. The `None` model achieved the highest AUROC of 0.99 on the internal validation set while `Convex Hull` achieved a slightly lower AUROC of 0.972; this indicates that during training, the `None` model may have been leveraging spurious signal outside the ROI to reach higher performance on the validation set, however this led its performance to degrade on a new dataset when these spurious signals change.

4.2 Hospital Prediction

For this question, I reproduce and extend previous work [14] which found that predicting hospital systems could be done with 99% accuracy. In the binary hospital classification task, there was no variability in model performance, with **all data variations** achieving perfect F1 and AUROC scores of **1.0**. Because of the perfect accuracy, it is highly unlikely that the model is using physiologically-based signals for the task, but rather the spurious artifacts, showing the strength of these.

To test this result further, I subsampled the datasets down to only 1000 images per class (~1% of the original data) and took a random 50px crop from the `Crop` dataset (i.e., the 50px crop occured randomly within the center third of the image). Even in this heavily constrained scenario, the model still achieved an AUROC of **0.976**.

These results indicate that medical imaging datasets can contain *strong, deeply embedded* spurious artifacts that are non-trivial to remove; even a minute crop (which kept little more than the width of the spine) had a minimal effect on hospital prediction performance, with the model still predicting highly accurately; therefore, merely hiding the non-ROI region via masking will have limited effect on nuisance signals within the image.

4.2.1 Validity of Saliency Maps

Although commonly used for post-hoc model explanations, saliency map methods (e.g., Grad-CAM [12]) have been criticized for producing misleading or otherwise inaccurate instance-level model explanations [11]. Figure 2 was shown in previous work looking at model generalizability [14], which implies that for the same task as discussed above (hospital prediction), the model is heavily leveraging non-ROI signals. It should be noted that slightly different datasets were used in our work

so the results discussed above aren't perfectly comparable, however, the results presented in this work show that a model can distinguish hospitals perfectly with these supposedly high-influence regions hidden.

This does not imply that the saliency maps presented are technically incorrect but rather that they are misleading, due to the normalised nature of the heatmaps; saliency maps are normalised to a 0-1 interval, so even if the center part of the image is highly influential in model predictions (which is shown to be the case), if the non-ROI region is more influential, *that* is what will be highlighted in the saliency map.

5 Conclusion

I perform an empirical study of *non-ROI* masking in PA chest x-rays, evaluating it under the goal of generalizability in disease prediction. Through testing on out-of-distribution datasets, we find an improvement in performance when masking images with a convex hull constructed from individual lung masks, keeping visual information within the chest cavity but masking signal outside the ROI. Despite this, the trend is not consistent across all test cases indicating that some datasets contain more non-ROI-based nuisance than others, so further research is required (i.e., by training/evaluating on more datasets).

I also study the influence of spurious signals directly by training a model to predict hospital system from chest x-ray. The results confirm prior work, showing this task can be achieved with perfect accuracy. I also show that this task can be completed with near-perfect accuracy under a severely data-limited scenario, indicating the predictive influence spurious image artifacts can have when trained on directly.

The results show that image datasets contain deeply embedded nuisance correlations that are non-trivial to remove, so more advanced processing should be applied before training and deploying deep models in a healthcare setting.

References

[1] Marcus A Badgeley, John R Zech, Luke Oakden-Rayner, Benjamin S Glicksberg, Manway Liu, William Gale, Michael V McConnell, Bethany Percha, Thomas M Snyder, and Joel T Dudley. Deep learning predicts hip fracture using confounding patient and healthcare variables. *NPJ digital medicine*, 2(1):1–10, 2019.

[2] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[4] Jeremy Howard and Sylvain Gugger. Fastai: a layered api for deep learning. *Information*, 11(2):108, 2020.

[5] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.

[6] Stefan Jaeger, Sema Candemir, Sameer Antani, Yì-Xiáng J Wáng, Pu-Xuan Lu, and George Thoma. Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative imaging in medicine and surgery*, 4(6):475, 2014.

[7] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131, 2018.

[8] Aahlad Puli, Lily H Zhang, Eric K Oermann, and Rajesh Ranganath. Predictive modeling in the presence of nuisance-induced spurious correlations. *arXiv preprint arXiv:2107.00520*, 2021.

[9] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015.

[10] Igor Rudan, Cynthia Boschi-Pinto, Zrinka Biloglav, Kim Mulholland, and Harry Campbell. Epidemiology and etiology of childhood pneumonia. *Bulletin of the world health organization*, 86:408–416B, 2008.

[11] Adriel Saporta, Xiaotong Gui, Ashwin Agrawal, Anuj Pareek, Steven QH Truong, Chanh DT Nguyen, Van-Doan Ngo, Jayne Seekins, Francis G Blankenberg, Andrew Ng, et al. Deep learning saliency maps do not accurately highlight diagnostically relevant regions for medical image interpretation. *medRxiv*, 2021.

[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. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[13] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald 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*, pages 2097–2106, 2017.

[14] John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11):e1002683, 2018.

A Dataset Sizes

Dataset Name	+ve	# -ve
Pediatric	3883 (1193)	1349 (1074)
CheXpert	4675 (3730)	9553 (8550)

Dataset Name	# CheXpert	# ChestX-Ray14
Hospital Prediction	224,316 (156,787)	112,120 (81,930)

Table 2: Instances for pneumonia prediction task (left) and hospital prediction (right). Values in parentheses denote dataset size after cleaning.