Investigation on Utility of Multi-source Datasets to Reduce Bias in Chest X-ray Classification

Adanna Vardian
George Mason University
4400 University Dr, Fairfax, VA 22030
avardian@gmu.edu

Abstract

As electronic data has become increasingly available in the healthcare domain, areas to employ automation are being investigated. Examination and classification of chest X-rays is one area that has attempted to leverage the applications and techniques of deep learning on common images. Models developed for this task have shown efficacy, however they have also shown performance bias among various protected subgroups such as religion, race, sex, or age. The TorchXrayVision library [2] has increased the accessibility of images as well as models designed for multi-class classification of chest X-rays. As a result, this research effort attempted to replicate the decrease in model bias as depicted by Seyyed-Kalantari et al. *[8] with datasets available in the TorchXraVision library. While the model developed on a multi-sourced dataset did demonstrate a reduction in bias for some categories, it increased in others. These results along with recent publications [12], [1] illustrate obvious limitations in the ability to provide unbiased and generalized models for chest X-ray classification.*

1. Introduction

As a result of the advancements and acceptance of deep neural networks for a range of tasks, many domains have started to research applicability for their use cases. Medicine is an active application area in which the role of deep learning techniques is still developing. The American Recovery and Reinvestment Act required that in 2014, healthcare providers were required to adopt electronic medical records (EMR). Since then, the amount and type of data in the field has been growing to include patient records and diagnostic readings/results. This increased data availability makes it a conducive area for incorporating deep learning. Processing of X-rays has been a focus area for application, since it is one of the most commonly administered diagnostic evaluations and requires manual evaluation by a trained individual. Automating some aspects of X-ray analysis has the potential to reduce the cost of this diagnostic which in turn could possibly increase its availability.

In 2016 a large publicly available de-identified dataset of chest X-rays and findings was made available Demner-Fushman *et al*. [4]. A few additional datasets Wang *et al*. [10], Bustos *et al*. [9] emerged along with the first applications of image classification and segmentation in the domain. In 2017 the CheXNet model (a version of a Huang *et al*.'s DenseNet [5]) along with a curated dataset of chest X-rays paired with labels from radiology reports proved effective for multi-class classification [7].

A surge of research and development in this area emerged from the outbreak of COVID-19, as chest X-rays were one of the tools being used to detect if the patient had the virus. Simultaneously, impact of the virus was reportedly disproportionate across racial and economic groups in the U.S. [14]. While the number of publicly available datasets and models tuned for medical tasks is growing [2], the assessment and reporting on biases is unfortunately not typically included or provided.

Research from Larrazabal *et al*. [6], and Seyyed-Kalantari *et al*. [8] have specifically highlighted gender, age, and socioeconomic biases in CheXNet's classification results. Seyyed-Kalantari *et al*. provided an example for assessing the model on one metric for fairness - Equal Opportunity measured by True Positive Rate (TPR) Disparity [8] . In addition, the paper presented integrating multiple datasets as one tactic for reducing disparity. Motivated by these results and the recent availability of the TorchXRayVision library [2] for normalizing chest X-ray datasets, this research effort aimed to replicate the reporting and reduction of bias in a version of the CheXNet model trained with three publicly available datasets. The TorchXrayVision library enabled datasets to be selected that shared the 14 pathologies compared to 8 provided in Seyyed-Kalantari *et al*. [8]. The results of the study demonstrated that while AUCs for labels were preserved or slightly improved when training on a merged dataset TPR disparity

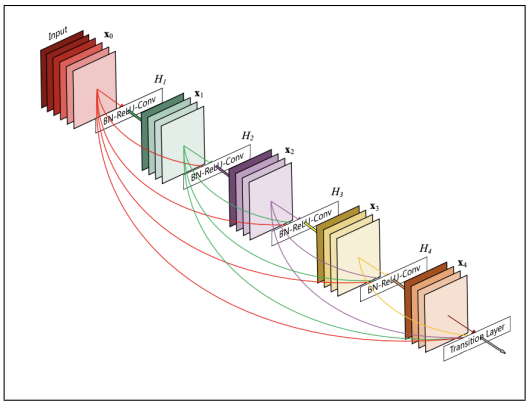

Figure 1. Copy of the Huang *et al*. schematic of a 5-layer dense block with a growth rate of k=4

among male and female patients was not universally improved. For this reason, it seems imperative that additional research and effort is focused on identifying the image features that may be attributing to this disparity. Additional insights from Cohen *et al*. [1] on the limitations of cross-domain generalization and Zech *et al*. [12] on the negative performance of pretrained models on X-rays from different hospitals further suggest that model utility may be limited to the specifics of the dataset leveraged in training.

2. Approach

Versions of Huang *et al*.'s DenseNet Figure 1 have been used in multiple studies including one of the primary leading papers that emerged along with the first publicly available datasets for chest X-ray classification [7]. The original DenseNet paper and subsequent references attribute its effectiveness to the retention of information captured in feature maps of previous layers assisting in the prevention of vanishing gradients. DenseNet also serves as the basis for the various pretrained state of the art models included in the TorchXRayVision library. Rajpurkar *et al*. modified the output layer of DenseNet with the softmax over the 14 class labels included in the data. This work was replicated with the latest PyTorch libraries and publicly available by John Zech [11]. This publicly available version of CheXNet is provided with starter code for the NIH-ChestX-ray8 data [11]. Since Seyyed-Kalantari *et al*. [8] leveraged the NIH database, it along with two other datasets that included the same pathologies were used to evaluate the impacts on model fairness with regard to patient's indicator for Male or Female.

Methods for measuring fairness of deep learning models are primarily focused on post-processing results. Typically fairness is discussed in one or more of the following categories: demographic parity, equal opportunity, equal accuracy, or group unaware [3]. In this case, equal opportunity is the most applicable as it evaluates positive predictions are equivalent for each group. True positive rates (TPR) are used as the metric for comparison. As in Seyyed-Kalantari *et al*. [8] for a given group g TPR:

$$TPR_{g,i} = p[\hat{Y}_i = y_i | G = g, Y_i = y_i]$$

In this research, there are only two groups of interest and as a result the gap of TPR between groups is calculated as:

$$Gap_{g,i} = TPR_{g,i} - TPR_{\neg g,i}$$

. For completeness and potentially future work that includes more than two groups, the gap is calculated as follows if g is the group of interest and additional groups represented as $S_1, .., S_N$:

$$Gap_{g,i} = TPR_{g,i} - Median(TPR_{S_1,i}, .., TPR_{S_N})$$

Ideally, TPR-disparities for all subgroups would be 0. In the event that the disparities are not zero, the subgroup with the negative disparity has less likelihood of being labeled with the category when it should be. Which in the domain of healthcare could mean missing a pathology or diagnosis.

In this study, three datasets (NIH-ChestX-ray8, PadChest, and OpenI) were pulled from the TorchXRayVision and normalized for the following 14 pathologies listed in the NIH-ChestX-ray8 dataset: Atelectasis, Cardiomegaly, Consolidation, Edema, Effusion, Emphysema, Fibrosis, Hernia, Infiltration, Mass, Nodule, Pleural Thickening, Pneumonia, Pneumothorax. A DenseNet-121 representative of the CheXNet model was trained and evaluated on each dataset separately. A fourth model was trained using all three datasets. The test sets from each of the datasources were then re-evaluated on the fourth model. TPR disparities were tabulated based on patient's sex. Note: the OpenI did not contain labels for the patient's sex and could not be included in the evaluation of TPR disparity.

3. Results

3.1. Datasets

As previously mentioned, three datasets available through TorchXRayVision were used in this study, additional details regarding composition of training set as well as notes are included below:

- NIH-ChestX-ray8: 112,120 frontal-view X-ray images that were automatically labeled by mining associated radiography reports; labels are reported to be $> 90\%$ accurate
- OpenI: 7,470 chest x-rays with 3,955 radiology reports; objective for dataset was to provide a deidentified dataset; note that it does not contain any M/F labels or equivalent

- PadChest: 160,000 images 27% manually labeled, reamining images auto labeled with text-mining methods (validated by independent test set). Reports were originally in Spanish and mapped to diagnosis in the Unified Medical Language System (UMLS)

In all datasets negative examples are more predominate than positive examples. Table 1 provides the number of positive examples along with its percent composition in the training set.

Table 1. Positive Represenations in Training Sets

Label	NIH		OpenI		PadChest	
Atelectasis	1467	6.39%	150	8.04%	1811	3.83%
Cardiomegaly	597	2.60%	112	6.00%	4136	8.76%
Consolidation	355	1.55%	-	-	411	0.87%
Edema	56	0.24%	33	1.77%	87	0.18%
Effusion	1432	6.23%	71	3.80%	1277	2.70%
Emphysema	279	1.21%	42	2.25%	418	0.89%
Fibrosis	444	1.93%	9	0.48%	266	0.56%
Hernia	56	0.24%	24	1.29%	739	1.56%
Infiltration	2888	12.57%	55	2.95%	3352	7.10%
Mass	989	4.30%	3	0.16%	375	0.79%
Nodule	1270	5.53%	30	1.61%	1706	3.61%
Pleural Thickening	647	2.82%	15	0.80%	1503	3.18%
Pneumonia	163	0.71%	37	1.98%	1541	3.26%
Pneumothorax	405	1.76%	8	0.43%	71	0.15%

Datasets were divided into 70-10-20 for training, validation, and testing. The following transforms were applied to the images: RandomHorizontalFlip(Training only), Resize(224), CenterCrop(224), Normalization(ImageNet mean, ImageNet std). TorchXRayVision datasets were extended to read images in RGB instead of the single grayscale provided by the library in order to align with the given model architecture.

3.2. Model

As briefly mentioned in the approach a DenseNet-121 model represented of the published CheXNet [7] provided by by John Zech [11] was used in the research project. The model hyper parameters were maintained as provided in the repository with the values as listed below:

- Pretrained: ImageNet
- Loss: Binary Cross Entropy Loss
- Optimizer: SGD+momentum

3.3. Threshold Determination

The validation dataset was used to determine threshold values for predicting the label based on the highest F1 measure. Threshold values were then applied to the test results to obtain binary predictions.

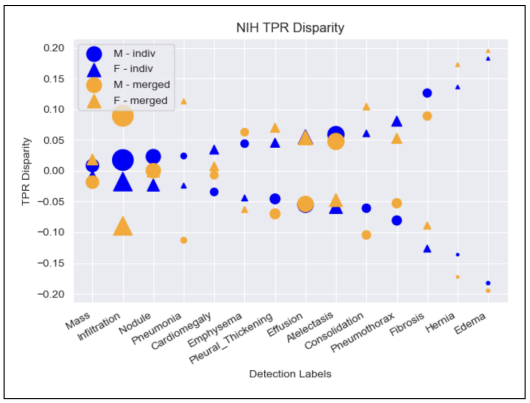

Figure 2. TPR comparison for NIH dataset.

3.4. Empirical Results

A comparison of AUCs among the published CheXNet [7], updated reproduction [11], and models/datasets are provided in Table 2.

The NIH dataset is the reference set common to both SOTA models. Comparing the individual dataset AUCs for the NIH dataset, all but the following three categories fall within the range defined in Zech *et al*. [13] : Edema, Fibrosis, and Hernia - two of which were the least represented labels in the training dataset. The AUC's for OpenI on the merged model were the most improved when compared to the individual dataset. This might be attributed to the overall increase in the number of images available for training. It is notable that neither model performed greater than $> 80\%$ classification on any label in the PadChest dataset. This is particularly intriguing given that this dataset contains 28 different labels. It maybe that the pretrained architecture or selected hyperparmeters are not adequately capturing the dataset specific features. Perhaps customizing either or both would lead to a more effective classifier.

TPR disparity gaps between the patient's record designated with 'M' or 'F' were calculated for individual datasets using the models trained on a single and the merged datasets. Only the NIH and PadChest datasets included the designations of interest. The TPR gaps are plotted together for comparison Figure 2 and 3. Points are scaled based on the category's number of positive labels in the test set.

Results from [8] suggested multi-source datasets as one mechanism to mitigate model bias. While this may be true for some specific labels and datasets, the results above indicate this is not universally true as there are well-represented classes that show greater disparity between selected subgroups when evaluated with the model trained on the merged dataset. The scaling of points highlights that for the NIH dataset disparity for Hernia and Edema may actually be attributed to the overall availability of positive labels than the model itself. In fact, it may bring to question if this

Table 2. Generated by Spread-LaTeX

Label	SOTA		NIH		OpenI		PadChest	
	repo - retrained	CheXNet	indiv	merged	indiv	merged	indiv	merged
Atelectasis	0.818	0.809	**0.822**	0.821	0.768	0.783	0.643	0.630
Cardiomegaly	0.909	**0.925**	0.912	0.903	0.823	0.850	0.716	0.717
Consolidation	**0.800**	0.790	0.747	0.754	-	-	0.710	0.710
Edema	**0.895**	0.888	0.812	0.795	0.885	0.883	0.744	0.784
Effusion	0.883	0.864	0.900	**0.903**	0.832	0.889	0.709	0.712
Emphysema	0.932	**0.937**	0.827	0.826	0.862	0.870	0.661	0.671
Fibrosis	0.825	0.805	0.742	0.731	0.840	**0.869**	0.687	0.667
Hernia	**0.918**	0.916	0.816	0.797	0.774	0.601	0.725	0.730
Infiltration	0.716	**0.735**	0.663	0.654	0.735	0.600	0.651	0.648
Mass	0.838	0.868	0.840	0.841	0.505	**0.904**	0.606	0.613
Nodule	0.776	0.780	0.745	0.745	0.624	**0.810**	0.584	0.575
Pleural Thickening	0.789	0.806	0.757	0.755	0.845	**0.878**	0.610	0.614
Pneumonia	0.762	**0.768**	0.702	0.704	0.651	0.729	0.684	0.694
Pneumothorax	0.878	0.889	0.859	0.847	0.886	**0.977**	0.753	0.772
Difference > 0.01			4		11		5	

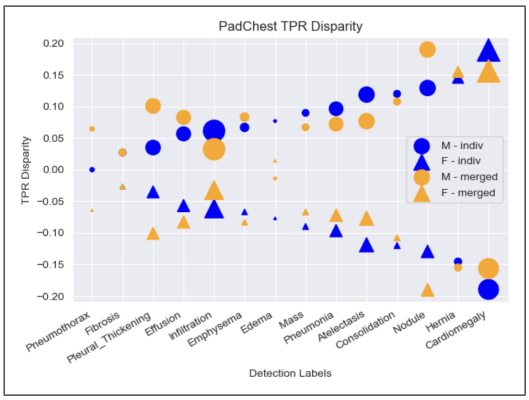

Figure 3. TPR comparison for PadChest dataset.

model should even be considered in the prediction of these labels.

4. Related Work

This work suggests that there are nuances in working with multi-sourced datasets yet to be researched and documented. While incorporating external or additional datasets may incur unintended side effects, the variation in AUCs across the multiple datasets suggest that there maybe differences between datasets in this domain that are yet to be fully characterized.

Research on model generalization for the Chest X-ray classification task has indicated that significant limitations exist on the ability to transfer among tasks[1] and even datasets[12]. The work conducted by Zech *et al*. debunked the notion that models trained on internal X-ray images would have similar prediction rates on images collected from external venues. Additionally, the models trained for Pneumonia detection could also reliably predict an image's origin from a given hospital and even department. As a result, it seems that the applicability of a model trained for Chest X-ray prediction across multi-labels may in reality be more narrow than expected or intended.

5. Summary And Discussion

Reliable automatic Chest X-ray classification is one area in healthcare that could enable accessibility to that diagnostic tool to a broader audience. Previous work [2] as well as this effort indicated that the underlying models used to make predictions contained biases. It was shown that leveraging a multi-source dataset for training minimized the bias across some categories and increased it across others. These results along with those from additional studies that have brought to light limitations on generalizing the chest X-ray task convey a breadth of research remaining in this area before developed models could be broadly used. One aspect open for further investigation is identifying limitations and application scopes for various publicly available datasets and models. Another area of consideration may be the understanding of the specific features in the images of subgroups to which the model has demonstrated bias.

References

[1] J. P. Cohen, M. Hashir, R. Brooks, and H. Bertrand. On the limits of cross-domain generalization in automated x-ray prediction. In *Medical Imaging with Deep Learning*, pages 136–155. PMLR, 2020.

[2] J. P. Cohen, J. D. Viviano, P. Bertin, P. Morrison, P. Torabian, M. Guarrera, M. P. Lungren, A. Chaudhari, R. Brooks, M. Hashir, et al. Torchxrayvision: A library of chest x-ray datasets and models. *arXiv preprint arXiv:2111.00595*, 2021.

[3] A. Cook. Tutorial: Ai fairness, 2022.

[4] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma, and C. J. McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016.

[5] 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*, pages 4700–4708, 2017.

[6] A. J. Larrazabal, N. Nieto, V. Peterson, D. H. Milone, and E. Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23):12592–12594, 2020.

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

[8] L. Seyyed-Kalantari, G. Liu, M. McDermott, I. Y. Chen, and M. Ghassemi. Chexclusion: Fairness gaps in deep chest x-ray classifiers. In *BIOCOMPUTING 2021: proceedings of the Pacific symposium*, pages 232–243. World Scientific, 2020.

[9] G. Shih, C. C. Wu, S. S. Halabi, M. D. Kohli, L. M. Prevedello, T. S. Cook, A. Sharma, J. K. Amorosa, V. Arteaga, M. Galperin-Aizenberg, et al. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology. Artificial intelligence*, 1(1), 2019.

[10] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3462–3471, 2017.

[11] J. Zech. reproduce-chexnet, 2018.

[12] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann. Confounding variables can degrade generalization performance of radiological deep learning models. *arXiv preprint arXiv:1807.00431*, 2018.

[13] J. R. Zech, J. Z. Forde, and M. L. Littman. Individual predictions matter: Assessing the effect of data ordering in training fine-tuned cnns for medical imaging. *arXiv preprint arXiv:1912.03606*, 2019.

[14] A. Zieda, S. Sbardella, M. Patel, and R. W. Smith. Diagnostic bias in the covid-19 pandemic: A series of short cases. *European Journal of Case Reports in Internal Medicine*, 8(5), 2021.