An Adversarial Approach for the Robust Classification of Pneumonia from Chest Radiographs

Joseph D. Janizek
jjanizek@cs.washington.edu
University of Washington

Gabriel Erion
erion@cs.washington.edu
University of Washington

Alex J. DeGrave
degrave@cs.washington.edu
University of Washington

Su-In Lee
suinlee@cs.washington.edu
University of Washington

ABSTRACT

While deep learning has shown promise in the domain of disease classification from medical images, models based on state-of-the-art convolutional neural network architectures often exhibit performance loss due to dataset shift. Models trained using data from one hospital system achieve high predictive performance when tested on data from the same hospital, but perform significantly worse when they are tested in different hospital systems. Furthermore, even within a given hospital system, deep learning models have been shown to depend on hospital- and patient-level confounders rather than meaningful pathology to make classifications. In order for these models to be safely deployed, we would like to ensure that they do not use confounding variables to make their classification, and that they will work well even when tested on images from hospitals that were not included in the training data. We attempt to address this problem in the context of pneumonia classification from chest radiographs. We propose an approach based on adversarial optimization, which allows us to learn more robust models that do not depend on confounders. Specifically, we demonstrate improved out-of-hospital generalization performance of a pneumonia classifier by training a model that is invariant to the view position of chest radiographs (anterior-posterior vs. posterior-anterior). Our approach leads to better predictive performance on external hospital data than both a standard baseline and previously proposed methods to handle confounding, and also suggests a method for identifying models that may rely on confounders.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; **Learning latent representations**; • **Applied computing** → **Life and medical sciences**.

ACM CHIL '20, April 2–4, 2020, Toronto, ON, Canada

ACM ISBN 978-1-4503-7046-2/20/04…$15.00
https://doi.org/10.1145/3368555.3384458

KEYWORDS

Domain Adaptation, Adversarial Training, Robustness, Domain Shift, Covariate Shift, Deep Learning, Distributional Robustness, Chest Radiograph, Pneumonia, Radiology

ACM Reference Format:
Joseph D. Janizek, Gabriel Erion, Alex J. DeGrave, and Su-In Lee. 2020. An Adversarial Approach for the Robust Classification of Pneumonia from Chest Radiographs. In *ACM Conference on Health, Inference, and Learning (ACM CHIL '20), April 2–4, 2020, Toronto, ON, Canada.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3368555.3384458

1 INTRODUCTION

A variety of recent papers have demonstrated the promise of deep learning for medical imaging tasks. From the prediction of diabetic retinopathy using retinal scan images to the diagnosis of melanoma from photographs, machine learning approaches have achieved near-physician level performance [6, 9]. Deep learning classifiers of chest radiographs are not only promising in a research setting, but have also been deployed in clinical practice. For example, an algorithm to detect 4 different thoracic diseases from frontal chest radiographs was evaluated in an emergency medicine setting and was found to increase radiology residents' sensitivity [12].

Despite these major advances, there are still significant limitations for medical deep learning. One of these problems is dataset shift, or the loss in performance when a model is tested on data that is drawn from a different distribution than the data used for training the model [24, 26]. Zech et al. [39] found that a deep learning pneumonia classifier trained on data from two hospital systems exploited differences in the base rate of pneumonia between the two hospitals by learning to identify each radiograph's hospital of origin rather than anatomically-relevant features of pneumonia. While this model apparently had high predictive performance, when the model was tested on radiographs from a third hospital not present in the training data its performance significantly decreased. Furthermore, even within a single hospital system, confounded predictions may be a problem for deep learning. For example, Badgeley et al. [1] demonstrated that a deep learning hip fracture classifier was leveraging patient-level variables (such as age and gender) and process-level variables (such as scanner model and hospital department) in its predictions. After controlling for these variables during model evaluation by rebalancing the test set, they found that the classifier performed no better than random. A recent multi-society statement on the "Ethics of Artificial Intelligence in Radiology" points to the importance of being able to understand and guide the

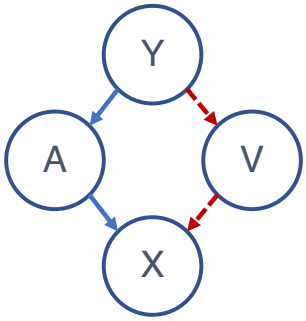

Figure 1: Causal graph showing relationships that form part of one plausible data generating process for chest radiographs: relationships are between pneumonia (Y), radiograph view position (V), anatomically relevant radiographic features (A), and the final chest radiograph (X). Red and dashed edges indicate a view-mediated causal path between the radiograph and pneumonia that may shift between different datasets or hospitals. We emphasize that this does not illustrate the full data generating process, and that many data generating processes are possible.

decision-making process of machine learning algorithms to ensure that these algorithms can be safely and effectively used in clinical practice [8]. While the works above have described the brittleness of deep learning medical imaging classifiers, more work is needed to create robust models.

We propose an approach based on adversarial neural networks to address dataset shift by learning models that are invariant to confounders that may shift across hospitals. In particular, we focus on the problem of pneumonia classification from chest radiographs, as the problem of confounding and dataset shift has been particularly well-documented for this task [39]. We find that (1) potential model confounding can be effectively identified by evaluating how well confounders can be predicted from a model's output, that (2) adversarial training enables pneumonia classification that is independent of radiograph view, and that (3) the adversarially-trained models attain better generalization performance when tested in novel hospital systems. [1]

2 PROBLEM STATEMENT

We first consider some of the causal relationships forming part of one plausible data generating process for chest radiographs, given by the random variable X in Figure 1. A patient's pneumonia status, given by the random variable Y, will lead to a variety of anatomically-relevant features A, such as increased radiopacity or consolidation in the lung fields, that form part of the radiograph. Furthermore, the patient's disease status will lead to a variety of clinical signs and symptoms which will influence which department they are seen in (e.g. in-patient or out-patient). Different departments may use different scanners (portable or fixed) and these scanners may be taken with different views (V). Frontal chest radiographs may be taken with either an anterior-posterior (AP) view where the x-ray source is positioned such that x-rays enter through the front of the chest and exit through the back of the chest, or a posterior-anterior (PA) view where the x-ray source is positioned such that x-rays enter through the back of the chest and exit through the front.

View directly impacts the appearance of chest radiographs in a variety of ways. Different views cause anatomical structures to have different relative sizes in radiographs since their distance from the radiographic source is altered [25]. Furthermore, AP radiographs are taken on portable scanners, which may place text such as "PORTABLE" or "SEMI-UPRIGHT" directly on the image. For this graph, it is plausible that the relationship between pneumonia and view may not be consistent across hospitals. The AP view position is generally associated with a higher prevalence of disease, as sicker patients are more likely to need to have a portable scanner brought to them [25, 38]. In our source training dataset (described below in subsection 3.1), however, the standard relationship is reversed and the prevalence of pneumonia is 2-fold *higher* in PA view radiographs (2.1% base rate of pneumonia in AP images vs. 3.9% base rate of pneumonia in PA images). As the difference in base rate between the subgroups increases, the worse the generalization performance should be (see Appendix A). Since the relationships between pneumonia and view may not be consistent across hospitals, we hypothesize that by learning a model that is invariant to differences in radiograph view, we can create a model that will be more robust to dataset shift. View is additionally an important confounder to control because commerically-available chest radiograph algorithms are currently designed to accept *both* AP and PA view radiographs as input [12].

We formally state the problem as follows. We are given data from a source distribution $\mathcal{S}$ where each sample (indexed by i) is a 3-tuple consisting of a radiograph $x_i \sim X$, a multi-label classification label $y_i \sim Y$, and a binary indicator of view $v_i \sim V$. We would like to learn a model that outputs a pneumonia score that will generalize well to a target domain $\mathcal{T}$, where the relationship between the nuisance variable and the outcome may be different in the target domain than in the source domain. In our particular problem, we assume that we have no access at all to data from the target distribution, corresponding to what Subbaswamy et al. [33] refer to as a proactive approach to addressing dataset shift. Much of the prior work on adversarial domain adaptation has corresponded to a different problem, in which we assume that we have access to *unlabeled data* from the target distribution, corresponding to what Subbaswamy et al. [33] refer to as a reactive approach to dataset shift [2, 7, 14, 19]. Since we have no data from the target distribution, we instead aim to learn a classifier f that outputs a pneumonia score S such that $S \perp V$. Even though we use all 13 of the different pathologies in Y to train our model, since we only require that our model learns a relationship such that $S \perp V$, and not $Y \perp V$, there is no constraint for the model to learn view-independent scores for any of the other non-pneumonia pathologies.

[1] Code to reproduce this project is available at https://github.com/suinleelab/cxr_adv

3 METHODS

3.1 Data

To assess the robustness of models to dataset shift, we used chest radiographs from two large publicly-available datasets. For our model training source domain, we used the CheXpert dataset from Stanford [13]. This dataset contains 224,316 chest radiographs of 65,240 patients. We considered only the 191,229 frontal radiographs (AP or PA view) in the dataset, excluding all of the lateral radiographs. Since the test split in the original CheXpert dataset only contained 8 radiographs that were positive for pneumonia, all of which were AP radiographs, we moved 92 more positive pneumonia radiographs (for a total of 100 positive pneumonia radiographs) to the test set for the sake of better pneumonia performance evaluation. For our target domain, we used the MIMIC-CXR dataset from Massachusetts Institute of Technology [15]. This dataset includes 371,920 chest radiographs of 65,079 patients. After filtering lateral radiographs, we had 249,995 frontal radiographs remaining. One major advantage of using these two datasets is that they have the same set of 13 labels ("Enlarged Cardiomediastinum," "Cardiomegaly," "Lung Opacity," "Lung Lesion," "Edema," "Consolidation," "Pneumonia," "Atelectasis," "Pneumothorax," "Pleural Effusion," "Pleural Other," "Fracture," and "Support Devices") and are created using the same labeling algorithm. This algorithm takes expert-generated free-text radiological reports associated with each chest radiograph as input and outputs the set of pathology labels. Using data labeled with the same natural language processing algorithm helps to remove the potential effects of dataset shift due to differences in the label generating process.

3.2 Standard Training

To train our baseline models for prediction, we used the architecture and training procedure described in [39] and [28]. The model architecture used was a DenseNet-121 initialized with weights pretrained on ImageNet, which can be downloaded from the PyTorch torchvision models subpackage [11, 22]. While we were primarily interested in pneumonia detection, we found that using all pathology labels available in the CheXpert dataset during training significantly increased pneumonia classification performance. Since the number of classes in the CheXpert dataset is different than the number of classes in the ImageNet dataset, the classification head for the pretrained DenseNet-121 was removed and replaced by a linear layer with output dimensions equal to the number of labels in the CheXpert dataset, followed by a sigmoid activation function. A binary cross-entropy loss was optimized using an SGD optimizer with momentum of 0.9, weight decay of 10^{-4}, and an initial learning rate of 10^{-2}. Early stopping was implemented by monitoring binary cross-entropy loss on a held out split of validation data. Our validation set, representing 5% of the training data, was split on patients rather than radiograph index. If validation loss did not improve over an epoch, the learning rate was decreased by a factor of 10. If validation loss failed to improve for 3 consecutive epochs, training was stopped. Performance was then evaluated on the held out test set. This procedure was repeated three separate times to attain standard deviations of performance.

3.3 Adversarial Deconfounding

To learn more robust models that generalize better to external test data, we propose an approach based on adversarial training. This approach consists of jointly training two neural networks. The first is the classifier, f, which is trained to predict a pneumonia label y from a chest radiograph x. The second is an adversary, g, which is trained to predict the view v from the output score s of the classifier f. The optimization procedure consists of alternating between training the adversary network until it is optimal, then training the classifier to fool the adversary while still predicting pneumonia well.

This approach aims to proactively mitigate the potential effects of domain shift by controlling for known confounders in medical images using adversarial training. In addition to the applications for reactive domain adaptation mentioned above in section 2, adversarial training has been used in a variety of other areas to learn models or representations that are independent of a given variable. For example, there is a significant body of literature in the area of algorithmic fairness where adversarial training has been used to learn representations that are fair with respect to protected classes such as race or gender [4, 18, 37]. In the physical sciences, adversarial training has been used to learn classifiers capable of detecting interesting particle jets in particle colliders that are independent of the presence of nuisance interactions in the collider [17].

We emphasize that one major contribution of our work compared to prior work on deep learning for medical images is that we take advantage of causal domain knowledge to improve generalization performance without needing to use *any* data from the target domain. Where previous approaches to domain adaptation use adversarial training to either learn a score or intermediate representation that are *domain-invariant* by augmenting training with unlabeled data from the target domain, we instead use our domain knowledge about the causal relationships involved in our data to find nuisance variables that potentially will have a different relationship with the outcome in the target domain than in the source domain. We then use an adversarial approach to learn a classifier that is invariant to the nuisance variable, which requires no data whatsoever from the target domain.

To implement our training, we take the approach suggested in Louppe et al. [17] and adapt it for use in the application of radiograph classification. For the notation in the following sections, the parameterization of classifier f will be given as θ_f, while the parameterization of adversary g will be given by θ_g. The classifier's output score for pneumonia is given by $s = f(x)^{pneumo}$ (where the ($pneumo$) superscript indicates the index for pneumonia in the multi-label output vector).

3.3.1 Separately Pretraining Classifier and Adversary. The classifier f is first trained using the procedure described in the standard training section above to optimize the negative log-likelihood of $Y|X$ under θ_f:

$$\mathcal{L}_f(\theta_f) = \mathbb{E}_{x\sim X}\mathbb{E}_{y\sim Y|x}[-\log p_{\theta_f}(y|x)]. \quad (1)$$

Then, the parameters of the classifier are fixed and the adversary network is trained. The architecture used for the adversary is a simple feed-forward network with 3 hidden layers of 32 nodes. We used ReLU activation functions between the hidden layers, and

a linear output. This architecture was selected to have sufficient capacity to model non-linear dependency between the score and view while still being lightweight enough for quick optimization. The network is optimized to minimize the following objective:

$$\mathcal{L}_r(\theta_f, \theta_r) = \mathbb{E}_{s\sim f(X;\theta_f)}\mathbb{E}_{v\sim V|s}[-\log p_{\theta_r}(v|s)]. \quad (2)$$

This means that the adversary takes the scalar-valued pneumonia score output by the classifier as its input, and outputs a scalar-valued prediction of view. The adversary was pretrained for a single epoch.

3.3.2 Joint adversarial optimization. After both the classifier and the adversary were pretrained, we began joint adversarial optimization. Each "joint optimization epoch" consisted of first fixing the classifier, then training the adversary for one epoch by minimizing the loss of the batch stochastic gradients for each of $K = N/M$ minibatches present in the entire dataset (where N is the number of total samples in the training data and M is the size of the minibatch):

$$\nabla_{\theta_r} \sum_{k=1}^{K} \sum_{m=1}^{M_k} -\log p_{\theta_r}(v_m|s_m). \quad (3)$$

Then, after the adversary is trained to optimally predict the nuisance variable V from the score output by the classifier, the parameters of the adversarial network θ_r are fixed, and we draw a single minibatch of data and update the model by descending the stochastic gradients of the minibatch

$$\nabla_{\theta_f} \sum_{m=1}^{M} \Big[-\log p_{\theta_f}(y_m|x_m) + \log p_{\theta_r}(v_m|s_m) \Big]. \quad (4)$$

The procedure of an entire epoch of training for the adversary with the classifier fixed, and a single minibatch of training for the classifier with the adversary fixed, is repeated until the model achieves optimal performance while its output is independent of the nuisance variable.

Louppe et al. [17] showed that the optimal solution of this minimax optimization scheme is a classifier f that is optimal with respect to the training data with output S that is independent of V. If no such classifier exists, then the weight of the adversarial loss term given in Equation 2 can be tuned with an additional hyperparameter λ to make a tradeoff between stability (in terms of independence of the classifier from the nuisance variable) and accuracy (in terms of classification performance given the data). For all of our models, we used a value of $\lambda = 1$. Finally, while other approaches have enforced independence between V and some intermediate layer of the network, if we want a pneumonia score S that is independent of V, we observe that it suffices to directly adversarially optimize the prediction of V from S.

3.4 Previous approaches for controlling confounders

Attempting to control for confounding in machine learning models is a well studied problem, and has previously been specifically studied in the domain of medical imaging [29]. In addition to testing the performance of our adversarial approach, we also compared to a variety of previously used approaches for modeling medical images in the presence of confounders.

3.4.1 Instance sampling. One approach to domain adaptation involves re-weighting samples in the training data [16, 23, 31, 34]. We re-implement the approach suggested in Rao et al., called Instance Weighting [29]. In a normal empirical risk minimization framework, we assume that the data the model will be evaluated on will be drawn from the same data generating process as that which the model is trained on, and thus aim to minimize the empirical risk:

$$f^* = \text{argmin}_{f\in\mathcal{F}} \sum_{i=1}^{n} \frac{1}{n} \ell(f(X_i), Y_i). \quad (5)$$

If we assume that we will have test data drawn from a different distribution, we can try to reweight the samples in our training set to minimize the empirical risk in the *target population* instead of the source population:

$$f^* = \text{argmin}_{f\in\mathcal{F}} \sum_{i=1}^{n} \frac{1}{n} \left[\frac{\hat{P}^{\mathcal{T}}(V_i, Y_i)}{\hat{P}^{\mathcal{S}}(V_i, Y_i)} \right] \ell(f(X_i), y_i), \quad (6)$$

where $\hat{P}^{\mathcal{S}}(V_i, Y_i)$ and $\hat{P}^{\mathcal{T}}(V_i, Y_i)$ indicate the joint density of radiograph view and pneumonia in the source and target domains respectively.

Since we do not have any information about the target distribution, we assume the target marginal distributions of the targets and the confounders are identical to the source marginal distributions, which means that the loss function factorizes to the following form:

$$f^* = \text{argmin}_{f\in\mathcal{F}} \sum_{i=1}^{n} \frac{1}{n} \left[\frac{\hat{P}^{\mathcal{S}}(Y_i)}{\hat{P}^{\mathcal{S}}(Y_i|V_i)} \right] \ell(f(X_i), y_i). \quad (7)$$

In order to avoid particular unbalanced batches during optimization, rather than applying the weights as a multiplicative factor during the calculation of the loss function, we instead re-weight the probability of each particular instance in the training data being sampled at each batch.

3.4.2 Matching. In addition to changing the sampling weights of each sample in the training set, the most straight-forward possible approach to handling confounding suggested in [29] is matching the base rate across subgroups in the training data. The drawbacks to this approach are that it either requires deliberately collecting data that is balanced across subgroups in advance, or throwing out data. Since we could not go back and alter the data collection process for our dataset, in order to match the base rate of pneumonia in AP and PA radiographs in the training data, we had to delete 77,117 AP radiographs from the training data. This represented a substantial portion of the total data, amounting to 40% of the samples negative for pneumonia in the CheXpert dataset, and 35% of all samples in the training data.

3.4.3 Include Nuisance Covariate in Regression. Another potential approach to handle confounding suggested in [29] is to "regress out" the effect of view on the outcome. We make use of the fact that the classification head of the DenseNet-121 is a logistic regression with the learned features (nodes of the last hidden layer, H^{n-1}) as covariates. Therefore, we simply append an extra feature for our covariate V to the last layer $H^{appended} = [H_0^{n-1}, H_1^{n-1}, \cdots, H_i^{n-1}, V]$. We can then model the data using a standard logistic regression:

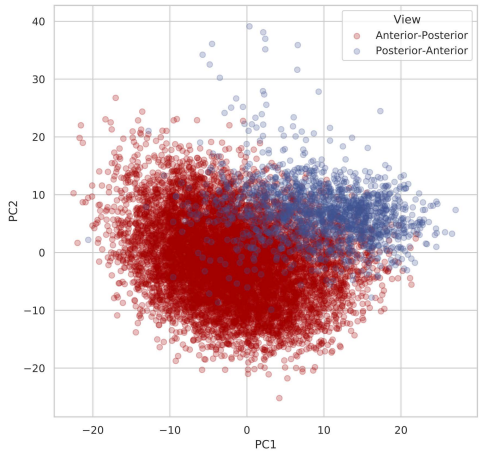

Figure 2: A pretrained CNN with no task-specific supervision represents radiographs in a manner easily separable by view.

Table 1: Pneumonia classifier performance (AUROC ± st. dev.) on held out test data from CheXpert dataset (Source), and held out test data from the external MIMIC dataset (Target). Standard deviations reported across three independent re-initializations of the training procedure. Best performance on external test data highlighted in bold and red.

Method	Source (Internal)	Target (External)
Standard	0.791 ± 0.016	0.703 ± 0.016
Adversarial (Ours)	0.747 ± 0.013	**0.739 ± 0.001**
Instance Weighting	0.685 ± 0.049	0.648 ± 0.038
Covariate	0.793 ± 0.008	0.715 ± 0.016
Matching	0.684 ± 0.036	0.689 ± 0.024

$$Y = \sigma(H^{appended} w + \beta), \tag{8}$$

where $w \in \mathbb{R}^{h+|V|}$ is the vector of weights of the classification head, $\beta \in \mathbb{R}$ is the bias term for the classification head, and $\sigma(t) = \frac{1}{1+e^{-t}}$.

We then train the modified DenseNet-121 following the exact same procedure as described in subsection 3.2. When evaluating our model in the external target domain, we remove the effect of the confounding variable by setting it equal to the mean across all samples.

4 RESULTS

4.1 CNN pneumonia classifiers fail to generalize to external health datasets

To assess the generalization performance of standard deep learning approaches to pneumonia classification, we trained a classifier using the procedure described in subsection 3.2 on data from the Stanford CheXpert dataset, then evaluated the model on both held-out patients from the same dataset (source performance) and held-out patients from the external MIMIC dataset (target performance). We evaluated performance using area under the ROC curve (AUROC), which evaluates the true positive rate and false positive rate attainable by the model across all possible thresholds.

We found that this model was able to achieve an AUROC of pneumonia classification of 0.791 ± 0.016 (see Table 1). When we tested this same model on data from the PhysioNet MIMIC dataset, we found a substantial drop in performance, with the model only able to achieve an AUROC for pneumonia classification of 0.703 ± 0.016 (see Table 1). This result again confirms the concerns raised in [24] and [39], that state-of-the-art training and model architectures for deep learning medical imaging classifiers lead to models that do not generalize well to external datasets.

4.2 Adversarial predictions improve model interpretability by identifying potentially confounded models

In this section we first show that state-of-the-art convolutional neural network (CNN) architectures are capable of detecting potential confounders given only pixel-level data. We then show how current approaches to model interpretability are of limited usefulness in determining whether or not a particular trained model depends on a potential confounder. We finally propose an approach based on training a neural network to predict the confounder from the model output, and show that it does a better job of identifying potential confounding.

4.2.1 Pretrained networks separate radiographs on basis of view and sex without any supervision. To assess how easily CNNs separate radiographs on the basis of features other than pathology, we examined the features extracted by a DenseNet-121 pretrained on ImageNet before *any* training on chest radiographs (Figure 2). We randomly sampled 10,000 radiographs and applied the DenseNet-121 features submodule to them (i.e. the entire model except the classification head). We then average pooled over the last two dimensions to get 1024 features for each sample. To visualize how different sorts of radiographs were spread over these pretrained features, we performed principal components analysis on the resulting matrix, and compared the distributions of different subsets of the data along the principal components. We found that the ImageNet-pretrained DenseNet-121 easily separates chest radiographs on the basis of their view, as AP and PA radiographs are embedded in different parts of the last layer.

4.2.2 Pretraining may alleviate some of the effects of confounding. Since pretraining alone was so easily able to separate confounders, we wanted to evaluate what sort of impact pretraining on ImageNet had in terms of the extent of the confounding and generalizability we observed in subsection 4.1, especially in light of recent results on the potentially limited benefit of transfer learning for medical imaging [27]. Therefore, we also tried training a deep CNN architecture from randomly initialized weights using the same training approach and same CheXpert data (see Appendix B). While we found

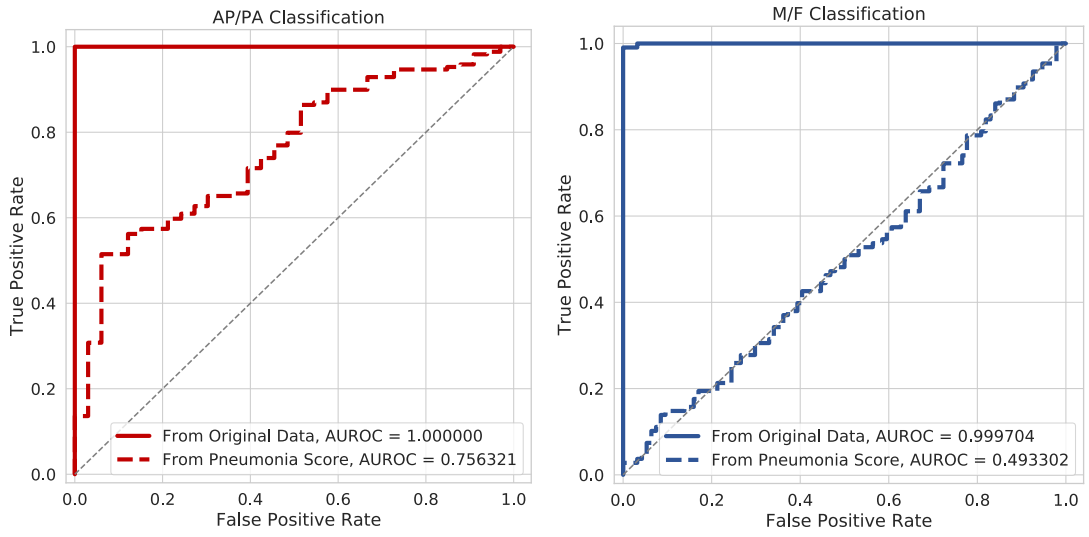

Figure 3: The DenseNet-121 model architecture (solid lines) is capable of near-perfect prediction of the potential nuisance variables radiograph view and patient sex from the original image data. After training a DenseNet-121 to predict pneumonia from the original image data, we see that a simple feed-forward classifier is capable of predicting radiograph view using only the scalar-valued score output by the pneumonia model as input (dashed line, left). However, a neural network classifier fails to attain better than random performance at predicting patient sex from the same scalar-valued score (dashed line, right). This indicates that the pneumonia classification score is independent of patient sex, but not of radiograph view.

that while this model was able to achieve comparable classification performance on held-out test data from the CheXpert dataset, it generalized significantly worse to the external target domain MIMIC data. We therefore are able to conclude that the network trained from scratch on medical data seems to learn domain-specific confounders more effectively when compared to an ImageNet pre-trained model. This is somewhat intuitive, as the scratch-trained model can more easily adapt to the space of input data, including potential confounders.

4.2.3 CNNs can detect potential confounders from image data with high accuracy. A previously proposed approach for detecting potential confounders has been to evaluate how well that confounder can be predicted from the original data [39]. When we train the same architecture CNN using the same training procedure to predict nuisance variables like sex or radiographic view from the chest radiograph data, we find that our models are capable of predicting these variables with incredibly high accuracy (Figure 3). For example, we see that a model can predict view with an AUROC of 1.0, perfectly classifying every example from the held out test data. Similarly, we see that this same model architecture is capable of predicting patient sex with an AUROC of 0.9997, again on held out test data. This result establishes that even if potentially confounding nuisance variables like radiograph view or patient sex are not explicitly included in the input features of CNN classifiers, deep CNN architectures are able to extract them with high accuracy from the pixel-level features of the radiographs, allowing them to still be used in classification.

While this result indicates that CNNs can detect potential confounders from just the radiograph data, it does give us any way to tell whether or not a particular model is invariant to a particular confounder. For example, in the CheXpert dataset, the base rate of pneumonia in male patients is 2.39% while the base rate of pneumonia in female patients is 2.42%. Therefore, even though we have seen that a CNN *can* identify whether a radiograph is from a male or female patient with high accuracy, it seems likely that a model would already be invariant to a feature that does not have an association with the outcome of interest.

Saliency maps are another previously proposed approach for understanding model behavior [30, 32, 35]. These methods highlight the pixels or regions that were most important for the classifier in a given image. We therefore used Expected Gradients, a pixel-level feature attribution method [5], to generate saliency maps to help understand which pixels were important for classifying view from radiographs (see Figure 4 and Appendix section Appendix C for more details). We observe that there is no specific region in the image that is indicative of PA vs. AP view. While both the laterality marker and text marker on the image are important for classification of view, pixels throughout the entire image, including within the lung fields, are also important for this prediction. Therefore, saliency map-based approaches are also not necessarily useful for identifying whether a model is invariant to a confounder or not.

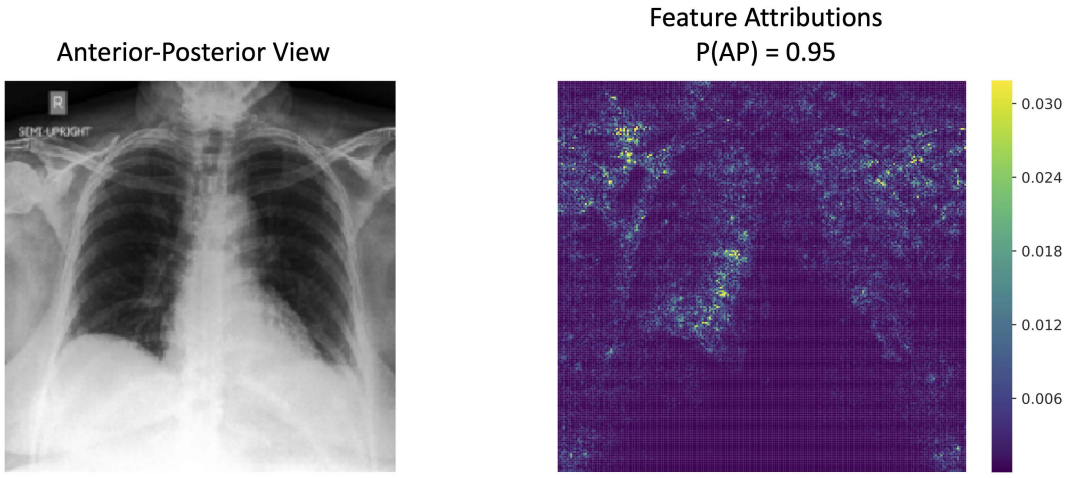

Figure 4: Expected Gradients feature attributions for a DenseNet-121 classifier trained to predict radiograph view position (AP vs. PA). We see that while parts of the image like the laterality markers are important (and have previously been shown to be important confounders for identifying source hospital from chest radiographs [39]), the most important pixels for identifying confounders are spread throughout the entire image, *including* within the lung fields.

4.2.4 Confounders can be detected directly from score. While seeing if nuisance variables can be predicted from the images can help understand if a confounded model *could* be learned from some data, it does not help identify how much a *particular model* is actually invariant to confounders. To assess this, we instead evaluate how well a neural network model (adversary) can classify the confounder of interest using only the scalar output score of the model we care about as input. In our case, this quantifies the dependence between the output score for pneumonia S and the confounding variable V by measuring the difference between the two distributions $p(S|X, V = \text{AP})$ and $p(S|X, V = \text{PA})$, and is well-justified as an empirical approximation to the $\mathcal{H}$-divergence discussed in [4, 7]. For a classifier where the output with respect to our class of interest, S is independent of V, prediction of V from S should be random, while S not independent of V will lead to better than random prediction. As an adversary, we trained a simple feed-forward network with 3 hidden layers of 32 nodes.

While both view and sex were nearly perfectly classified from the original data, when we first train a DenseNet-121 classifier to predict pneumonia from chest radiographs, then try to predict view and sex from the predicted probability of pneumonia, we see that our model attains far greater performance at predicting radiograph view than sex, and that patient sex is not predicted better than random (Figure 3). Therefore, we can conclude that while the pneumonia classifier is likely independent of sex, it is potentially not invariant to view.

4.3 Adversarial training can increase model robustness by controlling for confounders

4.3.1 Adversarial framework learns view-independent classifier. Following the insight of the previous section, we can directly optimize for a classifier that learns a score for our class of interest S that is independent of view using an adversarial framework. Prior to adversarial training, an adversary neural network could predict the counfounder with relatively high accuracy given only the score (Figure 3). Following our adversarial optimization procedure (subsection 3.3), a neural network is not able to predict the confounder any better than random accuracy (see Figure 5, left). Furthermore, when we look at the actual score distributions output by our model, we find that they are more closely matched within the two different view subgroups (see Figure 5, right top and right bottom). While we mainly present results for the binary view variable, one strength of our approach is that it can be applied to any sort of nuisance variable, including continuous-valued variables like age (see Appendix E).

We also find that looking at the predictive performance of the adversarial classifier is far more indicative of model behavior than saliency map-based approaches in this case. When we plot saliency maps (see Figure 8 in the Appendix, as described in Appendix C) we can see that there are definite differences in the pixel-level attributions. Furthermore, it appears that the important pixels are more localized to the lung fields in the adversarially-trained model than in the standard model. However, it is difficult to quantitatively assess to what extent that is the case, and since we have shown that pixels throughout the entire image are important for view

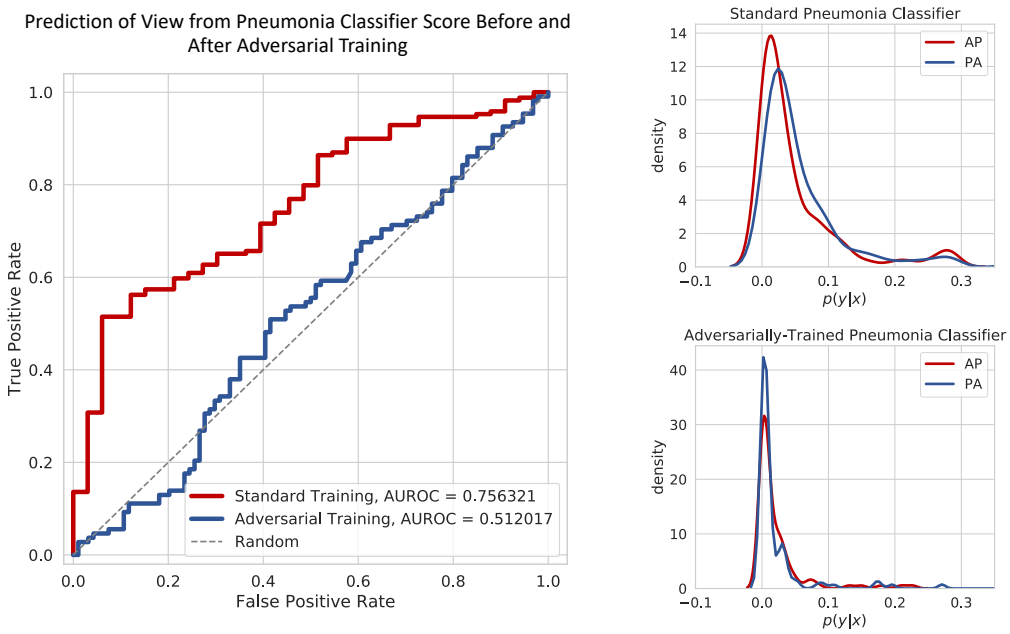

Figure 5: Adversarial training learns a pneumonia score that is independent of view. LEFT: ROC curves for the prediction of radiograph view (AP vs. PA) from a classifier's pneumonia score for a classifier trained with a standard approach (red) and for a classifier trained with our adversarial approach (blue). View can be predicted with relatively high accuracy just using the pneumonia score from the standard classifier, indicating that this model's output and view are not independent. After adversarial training, view can no longer be predicted with better-than-random accuracy, indicating that the output of this classifier is independent of view. RIGHT: When we look at the distribution of pneumonia scores actually output by the two models (Top: Standard, Bottom: Adversarial), we see that the distributions are not identical between AP and PA subgroups in the standard training model, but are much more closely matched betweeen the AP and PA subgroups in the adversarially-trained model.

classification by CNNs, it is very difficult to answer whether or not a model is confounded by view or not based only on its pixel-level feature attributions.

4.3.2 View-independent classifier generalizes better to unseen target domain. In addition to being able to learn a classifier that is independent of view, we find that adversarial training also is able to learn a model that generalizes better to external target domain test data (see Table 1). When we compare the performance of the adversarial model to the standard model, we find that while the adversarial model attains slightly worse performance on the source domain (AUROC = 0.747 ± 0.013 vs. AUROC = 0.791 ± 0.016), it attains better performance on the target domain (AUROC = 0.739 ± 0.001 vs. AUROC = 0.703 ± 0.016).

When we compare to the other baseline methods for controlling for confounding (instance weighting, including the covariate, and matching), we find that adversarial training also outperforms these methods. Of these other methods, we find that including the confounder as an additional covariate in modeling is the most effective, followed by matching, then the instance weighting resampling scheme.

4.3.3 Adversarial training learns a representation where pathology is independent of view. While our adversarial approach only explicitly constrains the final output score to be independent of the nuisance variable V representing view, we wanted to see how the earlier representations in the DenseNet-121 were impacted by this approach. We therefore take the output of the last dense layer before the classification head and average pool over the last two dimensions, and then perform principal components analysis in the same way as we did for the "unsupervised" ImageNet-pretrained classifier in section subsubsection 4.2.1.

The representation in the last layer of our adversarially-trained classifier is interesting, in that it is able to learn an embedding where the axis of differentiation separating the two views seems to be orthogonal to the axis representing pathology (see Figure 7). When we plot the first two principal components of the radiograph embedding and color by view, we see that the views are separated from the bottom left of the plot towards the top right. When we color the same embedding by pathology, the images with no findings are separated to the bottom right, while images containing pathology separate to the top left.

To quantify if this adversarially-trained representation has a more orthogonal relationship between view and pathology than

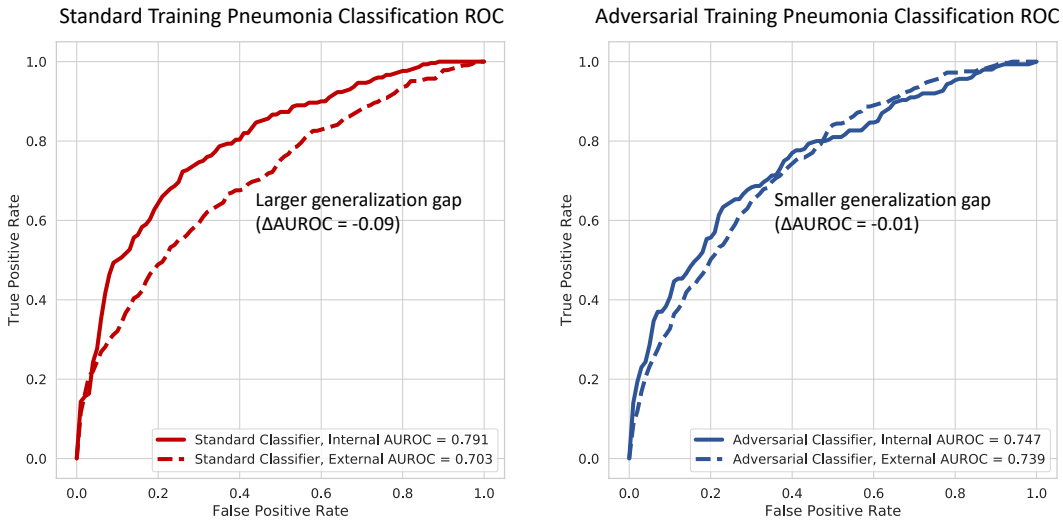

Figure 6: Adversarial training leads to less performance drop and significantly better performance when classifier is tested on data from a hospital system external to the one the training data comes from.

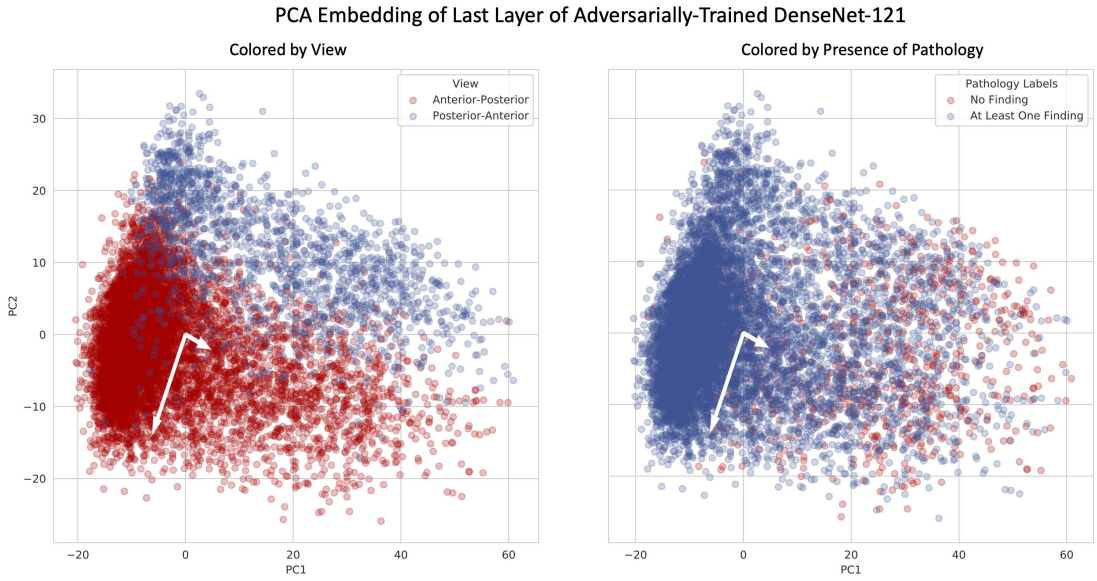

Figure 7: Adversarial training leads to a final representation where general pathology is orthogonal to view. White arrows indicate magnitude and direction of view and pathology classification weight vectors.

a classifier with standard training, we learned a simple logisitic regression classifier using the first two principal components of the last layer embeddings of the standard and adversarially trained classifiers as input. The output for prediction was either view or pathology. We then measured the linear correlation (r) between the weight vectors of the two linear classifiers.

We found over a 10-fold decrease in the correlation between the view and pathology vectors in the final embeddings from the standard to adversarially-trained models (decreasing from $r = 0.1974$ to $r = 0.008$), indicating that the view-axis was substantially more orthogonal to the pathology-axis in the adversarially-trained classifier. Again, this was particularly remarkable in that there was

no constraint to learn a more independent representation in the hidden layers of the model.

5 DISCUSSION

Our results demonstrate that an approach based on adversarial optimization is capable of learning more robust medical imaging classifiers. For the specific case of chest radiographs, we show that a pneumonia classifier trained to be independent of view is more stable to dataset shift, attaining better generalization performance when tested on radiographs from an external dataset. Finally, our results show that attempting to predict potential nuisance variables directly from a model's output score can be a valuable tool for model interpretability, indicating whether or not a particular model is independent of potential confounders. While any measure of the difference in the distributions of the model's output conditional on potential confounders is likely to work well, we believe that our approach is well-suited in that it also lends itself naturally to a technique to create confounder-invariant models.

Examination of the causal diagram relating chest radiographs to pneumonia points to important future research directions. Our experiments showed increased stability to dataset shift at the expense of decreased performance on new samples from the same hospital system as the training data. Given the causal diagram, where view mediates the relationship between the presence of pneumonia and the pixel features of the chest radiograph, it is not surprising that controlling for view should decrease performance. We note, however, that pneumonia is a diagnosis that is made in the context of clinical evidence of disease, and a disease where there is not necessarily perfect concordance between severity of symptoms and radiographic evidence of infiltrate [3, 20, 21, 36]. In the description of the creation of the "Pneumonia" label in the CheXpert dataset, the authors note that while pneumonia is a clinical diagnosis, "Pneumonia... was included as a label in order to represent the images that suggested primary infection as the diagnosis," suggesting that clinical information may play a role in labeling [13]. Disentangling the relationship between radiographic evidence of consolidation, the clinical presence of pneumonia symptoms, and the influence of the latter on the labeling of the former in these datasets could be helpful.

Finally, while we showed results from controlling radiograph view (and patient age), we expect that future work could show even more benefits from applying our approach to a wider variety of variables, both individually and in combination. However, it would be required for these variables to be recorded as metadata in datasets. As more and more additional variables are recorded in medical imaging datasets, and the causal relationships between these variables are better explicated, we expect the potential benefit of our approach to further increase.

ACKNOWLEDGMENTS

This work was funded by the National Science Foundation [CAREER DBI-1552309, and DBI-1759487]; American Cancer Society [127332-RSG-15-097-01-TBG]; and National Institutes of Health [R35 GM 128638, and R01 NIA AG 061132].

We would like to thank Samantha Gilbert, Pascal Sturmfels, Nicasia Beebe-Wang, and Hugh Chen for their feedback on the manuscript. We would also like to thank all of the members of Prof. Su-In Lee's lab for their valuable general feedback on the project.

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. 2019. Deep learning predicts hip fracture using confounding patient and healthcare variables. *npj Digital Medicine* 2, 1 (2019), 31. https://doi.org/10.1038/s41746-019-0105-1

[2] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010. A theory of learning from different domains. *Machine learning* 79, 1-2 (2010), 151–175.

[3] Bálint Botz. 2017. A Few Thoughts About CheXNet — And The Way Human Performance Should (And Should Not) Be Measured. *Web* (2017), 1–4. https://medium.com/@BalintBotz/a-few-thoughts-about-chexnet-and-the-way-human-performance-should-and-should-not-be-measured-68031dca7bf

[4] Harrison Edwards and Amos Storkey. 2015. Censoring Representations with an Adversary. (11 2015). https://arxiv.org/abs/1511.05897

[5] Gabriel Erion, Joseph D Janizek, Pascal Sturmfels, Scott Lundberg, and Su-In Lee. 2019. Learning Explainable Models Using Attribution Priors. *arXiv preprint arXiv:1906.10670* (2019). https://arxiv.org/abs/1906.10670

[6] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542, 7639 (2017), 115.

[7] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research* 17, 1 (2016), 2030–2096.

[8] J Raymond Geis, Adrian P Brady, Carol C Wu, Jack Spencer, Erik Ranschaert, Jacob L Jaremko, Steve G Langer, Andrea Borondy Kitts, Judy Birch, William F Shields, Robert van den Hoven van Genderen, Elmar Kotter, Judy Wawira Gichoya, Tessa S Cook, Matthew B Morgan, An Tang, Nabile M Safdar, and Marc Kohli. 2019. Ethics of Artificial Intelligence in Radiology: Summary of the Joint European and North American Multisociety Statement. *Radiology* 293, 2 (10 2019), 436–440. https://doi.org/10.1148/radiol.2019191586

[9] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, and Jorge Cuadros. 2016. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama* 316, 22 (2016), 2402–2410.

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

[11] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. 2016. Densely Connected Convolutional Networks. (8 2016). http://arxiv.org/abs/1608.06993

[12] Eui Jin Hwang, Ju Gang Nam, Woo Hyeon Lim, Sae Jin Park, Yun Soo Jeong, Ji Hee Kang, Eun Kyoung Hong, Taek Min Kim, Jin Mo Goo, Sunggyun Park, Ki Hwan Kim, and Chang Min Park. 2019. Deep Learning for Chest Radiograph Diagnosis in the Emergency Department. *Radiology* 293, 3 (10 2019), 573–580. https://doi.org/10.1148/radiol.2019191225

[13] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, and Katie Shpanskaya. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *arXiv preprint arXiv:1901.07031* (2019).

[14] Mehran Javanmardi and Tolga Tasdizen. 2018. Domain adaptation for biomedical image segmentation using adversarial training. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 554–558.

[15] Alistair E W Johnson, Tom J Pollard, Seth Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. MIMIC-CXR: A large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042* (2019).

[16] Max A Little and Reham Badawy. 2019. Causal bootstrapping. *arXiv preprint arXiv:1910.09648* (2019).

[17] Gilles Louppe, Michael Kagan, and Kyle Cranmer. 2017. Learning to pivot with adversarial networks. In *Advances in neural information processing systems*. 981–990.

[18] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. 2018. Learning adversarially fair and transferable representations. *arXiv preprint arXiv:1802.06309* (2018).

[19] Faisal Mahmood, Richard Chen, and Nicholas J Durr. 2018. Unsupervised reverse domain adaptation for synthetic medical images via adversarial training. *IEEE transactions on medical imaging* 37, 12 (2018), 2572–2581.

[20] M S Niederman, J B Bass, G Douglas Campbell, A M Fein, R F Grossman, L A Mandell, T J Marrie, A Torres, and V L Yu. 1993. Guidelines for the initial management of adults with community-acquired pneumonia: diagnosis, assessment of severity, and initial antimicrobial therapy. *American Review of Respiratory Disease* 148, 5 (1993), 1418–1426.

[21] Luke Oakden-Rayner. 2018. CheXNet: an in-depth review. *URL: https://lukeoakdenrayner. wordpress. com/2018/01/24/chexnetan-in-depth-review* (2018).

[22] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. (2017).

[23] Judea Pearl and Elias Bareinboim. 2011. *Transportability across studies: A formal approach.* Technical Report.

[24] Eduardo H. P. Pooch, Pedro L. Ballester, and Rodrigo C. Barros. 2019. Can we trust deep learning models diagnosis? The impact of domain shift in chest radiograph classification. *arXiv preprint arXiv:1909.01940* (2019). http://arxiv.org/abs/1909.01940

[25] Elizabeth Puddy and Catherine Hill. 2007. Interpretation of the chest radiograph. *Continuing Education in Anaesthesia, Critical Care and Pain* 7, 3 (2007), 71–75.

[26] Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. 2009. *Dataset shift in machine learning.* The MIT Press.

[27] Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. 2019. Transfusion: Understanding transfer learning with applications to medical imaging. *arXiv preprint arXiv:1902.07208* (2019).

[28] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. 2017. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. (11 2017). http://arxiv.org/abs/1711.05225

[29] Anil Rao, Joao M Monteiro, Janaina Mourao-Miranda, and Alzheimer's Disease Initiative. 2017. Predictive modelling using neuroimaging data in the presence of confounds. *NeuroImage* 150 (2017), 23–49.

[30] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2016. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. (10 2016). https://doi.org/10.1007/s11263-019-01228-7

[31] Hidetoshi Shimodaira. 2000. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference* 90, 2 (2000), 227–244.

[32] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. 2017. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825* (2017).

[33] Adarsh Subbaswamy and Suchi Saria. 2019. From development to deployment: dataset shift, causality, and shift-stable models in health AI. *Biostatistics* (11 2019). https://doi.org/10.1093/biostatistics/kxz041

[34] Masashi Sugiyama, Taiji Suzuki, Shinichi Nakajima, Hisashi Kashima, Paul von Bünau, and Motoaki Kawanabe. 2008. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics* 60, 4 (2008), 699–746.

[35] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 3319–3328.

[36] Saskia F van Vugt, Theo J M Verheij, Pim A de Jong, Chris C Butler, Kerenza Hood, Samuel Coenen, Herman Goossens, Paul Little, and Berna D L Broekhuizen. 2013. Diagnosing pneumonia in patients with acute cough: clinical judgment compared to chest radiography. *European Respiratory Journal* 42, 4 (10 2013), 1076 LP – 1082. https://doi.org/10.1183/09031936.00111012

[37] Christina Wadsworth, Francesca Vera, and Chris Piech. 2018. Achieving fairness through adversarial learning: an application to recidivism prediction. *arXiv preprint arXiv:1807.00199* (2018).

[38] John R. Zech. 2018. What are radiological deep learning models actually learning? *medium.com* (2018).

[39] John R. Zech, Marcus A. Badgeley, Manway Liu, Anthony B. Costa, Joseph J. Titano, and Eric Karl Oermann. 2018. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine* 15, 11 (11 2018), e1002683. https://doi.org/10.1371/journal.pmed.1002683