

Debiasing Counterfactuals In the Presence of Spurious Correlations

Amar Kumar^{1,2}, Nima Fathi^{1,2}, Raghav Mehta^{1,2}, Brennan Nichyporuk^{1,2},
Jean-Pierre R. Falet^{2,3}, Sotirios Tsaftaris^{4,5}, Tal Arbel^{1,2}

¹Center for Intelligent Machines, McGill University, Canada.

²MILA (Quebec AI institute), Canada.

³Montreal Neurological Institute, McGill University, Canada.

⁴Institute for Digital Communications, School of Engineering, University of
Edinburgh, UK.

⁵The Alan Turing Institute, UK

amarkr@cim.mcgill.ca

Abstract. Deep learning models can perform well in complex medical imaging classification tasks, even when basing their conclusions on spurious correlations (i.e. confounders), should they be prevalent in the training dataset, rather than on the causal image markers of interest. This would thereby limit their ability to generalize across the population. Explainability based on counterfactual image generation can be used to expose the confounders but does not provide a strategy to mitigate the bias. In this work, we introduce the first end-to-end training framework that integrates both (i) popular debiasing classifiers (e.g. distributionally robust optimization (DRO)) to avoid latching onto the spurious correlations and (ii) counterfactual image generation to unveil generalizable imaging markers of relevance to the task. Additionally, we propose a novel metric, *Spurious Correlation Latching Score (SCLS)*, to quantify the extent of the classifier reliance on the spurious correlation as exposed by the counterfactual images. Through comprehensive experiments on two public datasets (with the simulated and real visual artifacts), we demonstrate that the debiasing method: (i) learns generalizable markers across the population, and (ii) successfully ignores spurious correlations and focuses on the underlying disease pathology.

Keywords: Biomarker · Counterfactuals · Debiasing · Explainability

1 Introduction

Deep learning models have shown tremendous success in disease classification-based on medical images, given their ability to learn complex imaging markers across a wide population of subjects. These models can show good performance and still be *biased* as they may focus on spurious correlations in the image that are not causally related to the disease but arise due to confounding factors - should they be common across the majority of samples in the training dataset. As a result, the confounding predictive image markers may not generalize across

the population. For example, a deep learning model was able to accurately detect COVID-19 from chest radiographs, but rather than relying on pathological evidence, the model latched on to spurious correlations such as medical devices or lettering in the image [3]. As a result, these image markers did not generalize across the population.

In order to safely deploy black-box deep learning models in real clinical applications, explainability should be integrated into the framework so as to expose the spurious correlations on which the classifier based its conclusions. Popular post-hoc explainability strategies, such as Grad-CAM [16,6,19], SHAP [10], LIME [11] are not designed to expose the precise predictive markers driving a classifier. Models that integrate counterfactual image generation, along with black-box classifiers [21,2,23], permit exposing the predictive markers used by the classifier. However, should these methods discover that the markers are indeed simply visual artifacts there are no strategies to mitigate the resulting biases. Furthermore, although several debiasing methods have been successfully implemented to account for generalizability [1,26,17,8,27], they do not integrate explainability into the framework in order to provide reasons for improved performance.

Therefore, the important question to be answered is - *Can a model be trained to disregard spurious correlations and identify generalizable predictive disease markers?*

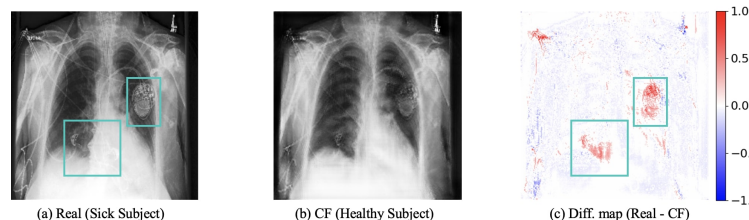


Fig. 1: Counterfactual (CF) image indicating that the classifier latched onto spurious correlations (medical devices) when correctly predicting that subject is sick (class: Pleural Effusion), due to their prevalence in the training dataset for this class. (a) Chest radiograph of a sick subject with several medical devices shown (cyan boxes), (b) Generated (CF) image, (c) Difference heat map shows maximum change around the medical devices, rather than indicating the correct markers for the disease.

In this paper, we propose the first end-to-end training framework for the explainability of classifier and debiasing via counterfactual image generation. We seek to discover imaging markers that reflect underlying disease pathology and that generalize across subgroups. Extensive experiments are performed on two different publicly available datasets - (i) *RSNA Pneumonia Detection Challenge* and (ii) *CheXpert* [5]. To illustrate the goal, Figure 1 shows an example from the contrived CheXpert dataset, where most of the sick subjects have medical device(s) (e.g. a pacemaker) in their images while most of the healthy subjects do not. As such, there exists a spurious correlation between a confounding visual artifact (the medical devices) and the disease. A classifier based on a standard

optimization technique, empirical risk minimization (ERM), incorrectly indicates the medical device as a disease marker, as depicted by the counterfactual (CF). In this work, we propose replacing ERM with a popular debiasing method, Group-DRO (distributional robust optimization). This permits the classifier to focus on the pathological image markers of the disease rather than on spurious correlation(s). Additionally, we show that Group-DRO ignores the visual artifact when making its decision, and generalizes across subgroups without the spurious correlation. Since standard metrics to evaluate counterfactuals do not indicate the region where the classifier focuses, we also propose a novel metric, the Spurious Correlation Latching Score (SCLS), to measure the degree to which the classifier latches onto spurious correlations. Our experiments indicate an improvement (in terms of differences in classifier outputs) of 0.68 and 0.54 in the SCLS using the Group-DRO classifier over the ERM for each of the two datasets.

2 Methodology

We propose an end-to-end training strategy to explain the output of a classifier. Here, we are considering a scenario where majority of the training data encompasses a spurious correlation with the target label. However, there is also a minority subgroup in the dataset that does not have any spurious correlation with the target label i.e., if the classifier was to rely onto the spurious correlation then the performance on these minority subgroups will be poor. Also, the term ‘majority’ and ‘minority’ is based on the number of samples in these groups. An overview of our approach is shown in Figure 2.

2.1 Classifier Explainability and Debiasing Via Counterfactual Image Generation

Disease Classification Binary (e.g. "sick" or "healthy") classification of the images is performed using either a standard classifier (ERM [24]), or a classifier that mitigates biases across sub-groups (Group-DRO [18]). The ERM classifier (f_{ERM}) is expected to be affected by the spurious correlation present in the training dataset, as it minimizes the loss over the entire training dataset and latching onto spurious correlation is a shortcut to minimize the loss. Thus, it would not generalize across the minority subgroups of the dataset [12,20]. On the contrary, the DRO classifier (f_{DRO}) is not expected to learn the spurious correlation as it considers the majority and minority subgroups separately when optimizing the loss. Thus, it would generalize well across all subgroups.

Generative model for synthesizing counterfactuals We develop an explainability framework that integrates counterfactual image generation together with a classifier during training. We adapted Cycle-GAN [25] as the generative model for counterfactual image generation, chosen for its strong performance across a variety of domains [13,25]. A pre-trained, frozen binary classifier (f_{ERM}

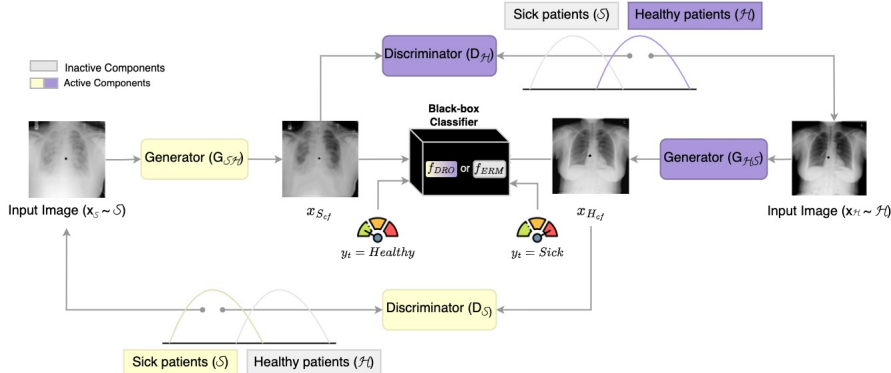


Fig. 2: Training procedure overview: The black-box classifier can be f_{ERM} or f_{DRO} and provides supervision to maintain the correct target class, y_t . Two U-Net generators, G_{SH} and G_{HS} , are employed to synthesize counterfactual images, namely $x_{S_{cf}}$ and $x_{H_{cf}}$. The discriminator D_H and D_S compares the counterfactual images with the domain of healthy \mathcal{H} and sick \mathcal{S} subjects respectively. Note, training a cycle-GAN requires simultaneous use of two input images from the two distributions.

or f_{DRO}) provides supervision to the generator. The proposed architecture and optimization objectives (see Figure 2) are designed to generate counterfactual images that adhere to the following common constraints [14,15,7]: (i) *Identity preservation*: The counterfactual images resemble the input images with minimal change; (ii) *Classifier consistency*: Counterfactual images belong to the target class; (iii) *Cycle consistency*: When counterfactual images are fed through the opposing generator, the output reverts to the original image (see Figure 2).

During inference, based on the classifier’s decision (i.e., f_{ERM} or f_{DRO}) for the input image, we generate counterfactual images and analyze the difference heatmap between the factual (input) and counterfactual (synthesized) images. This interpretable heatmap indicates the image markers that contribute the most to changing the classifier’s decision.

2.2 Metrics for Evaluating Counterfactuals: Accounting for Spurious Correlations

Standard counterfactual evaluation metrics are structured so as to ensure that the generated images (a) preserve the subject identity and thus penalize generated counterfactual images that are significantly different from the factual (original) images and (b) result in a maximal change in the class label (e.g. from healthy to sick). Identity preservation is typically measured by *structural similarity index* (SSIM) [4] and *Actionability* [15,14], defined as $\mathbb{E} \left[\|x - x_{cf}\|_{L_1} \right]$ between factual (x) and counterfactual (x_{cf}) images. Here, a higher value for SSIM and a lower value for Actionability would indicate better counterfactuals. The *counterfactual prediction gain* (CPG) [15], defined as $|f(x) - f(x_{cf})|$, indi-

cates the degree of change in the classifier’s decision such that a higher value of CPG indicates better counterfactuals.

While such metrics are required to measure the validity of the generated counterfactuals, they do not assess whether the classifier latched onto spurious correlations. For example, consider an image of a sick subject in the presence of a spurious correlation. If the disease classifier, f_{ERM} , latched onto the spurious correlation when identifying the subject as sick, the corresponding counterfactual image (i.e., depicting a healthy subject) would show changes in the area of the spurious correlation. In this case, all three evaluation metrics mentioned above would determine that this is a valid counterfactual image, based on high SSIM and low Actionability (shows minimal changes made compared to the factual image) and high CPG (due to the classifier decision changing from sick to healthy). However, the counterfactual image shows changes in the area of the spurious correlation rather than depicting the correct predictive image markers for the disease as desired.

In order to indicate that the classifier is correct but for the wrong reasons, we introduce a novel metric called Spurious Correlation Latching Score (SCLS) defined as follows:

$$\text{SCLS} = |d(x) - d(x_{cf})|. \quad (1)$$

Here, $d(\cdot)$ is a separate classifier, trained to identify the presence of spurious correlation in the image. In cases where the counterfactual image makes changes in an area of spurious correlation, SCLS will be high, as the $d(\cdot)$ will show a maximum change in its prediction between factual and counterfactual images. On the other hand, if the counterfactual image does not make changes in the area of the spurious correlation then SCLS will have a low value. As such, this evaluation strategy will validate how well the counterfactuals can help to determine that the classifier latched onto spurious correlations.

3 Experiments and Results

3.1 Dataset and Implementation Details

We perform experiments on two publicly available datasets. The absence of ground truth makes the validation of counterfactual images particularly challenging. Therefore, to directly evaluate the quality of the generated counterfactual images in the presence of spurious correlations, we modify a publicly available dataset (*RSNA Pneumonia Detection Challenge*) by adding a synthetic artifact to the majority of the sick images (90%). The majority of the sick and few of the healthy subjects have an artifact in the image, whereas the majority of the healthy and a few sick subjects do not have this artifact. The spurious correlation (artifact) is a black dot of radius 9 pixels at the center of the image. Thus, there are a total of four subgroups ($majority_S$, $majority_H$, $minority_S$ and $minority_H$) in the dataset with varying number of images: $majority_S$ and $majority_H$ are majority subgroups (sick with artifact and healthy without artifact), while $minority_S$ and $minority_H$ are minority subgroups (sick without

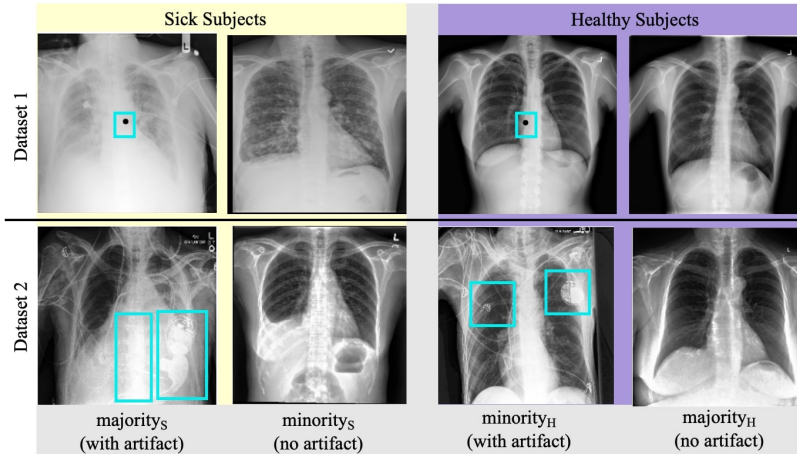


Fig. 3: Datasets 1 and 2 group division: The majority of the sick subjects [$majority_S$] and the minority of healthy subjects [$minority_H$] have visual artifacts (shown in cyan boxes). The majority of healthy subjects [$majority_H$] and the minority of sick subjects [$minority_S$] do not have visual artifacts. Top row: Simulated artifacts (black dots); Bottom row: Real artifacts (medical devices).

artifact and healthy with artifact). Henceforth, this dataset will be referred to as Dataset 1.

	Disease	Image size	Classifier	# samples [$majority_S$, $minority_S$, $minority_H$, $majority_H$]
Dataset 1	Pneumonia	512 x 512	AlexNet [13]	5413, 1526, 883, 7968
Dataset 2	Pleural Effusion	224 x 224	Resnet-50 [22] (pre-trained)	2600, 260, 350, 3456

Table 1: Implementation details for the two datasets

We also show experiments on a subset of a publicly available dataset (*CheXpert* [5]) with medical devices (visual artifacts), spuriously correlated with the disease. Specifically, we extract the subset of images that have labels “healthy” or “pleural effusion” (subjects with the presence of other diseases are removed from the dataset). This dataset will be referred to as Dataset 2. More details about both datasets are provided in Table 1. Note that both the datasets are divided into training/validation/testing with 70/10/20 random split. Example images for both datasets and all four subgroups are shown in Figure 3.

3.2 Results

Classifier Evaluation For both datasets (Figure 4), the DRO-based classifier (f_{DRO}) performs better for the minority subgroups ($minority_S$ and $minority_H$);

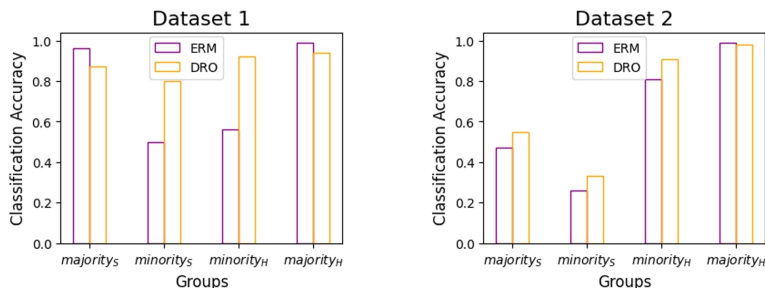


Fig. 4: Performance of ERM (f_{ERM}) and DRO (f_{DRO}) based classifier on a held out test set across all subgroups for both datasets. Notice that DRO has improved performance on minority subgroup [$minority_S$ and $minority_H$] showing improved generalizability across all subgroups.

indicating that it can better generalize to sub-populations that do not have the same visual artifact as the majority subgroups. Both classifiers perform similarly for the majority subgroups ($majority_S$ and $majority_H$).

Qualitative Counterfactual Evaluation Pneumonia in chest radiograph manifests as increased brightness in some regions of the lungs. In dataset 1, when examining the majority subgroup of sick subjects, the ERM-based classifier latches onto the spurious correlation, as seen by the difference maps. On the other hand, a DRO-based classifier focuses on the pathology of the disease, indicated by darker intensity regions over the lungs, as shown in Figure 5. The behavior of f_{ERM} is also evident in the minority subgroup, where the counterfactual for a healthy subject exhibits an enlarged artifact, wrongly suggesting that the visual artifact serves as a disease marker. Pleural effusion is characterized by the rounding of the costophrenic angle, augmented lung opacity, and reduced clarity of the diaphragm and lung fissures [9]. For the majority subgroup of sick subjects in Dataset 2, the counterfactual images based on ERM remove the medical device rather than focusing on the disease. In addition, for healthy subjects from the minority subgroup, maximum changes are observed around the medical device. On the other hand, for the majority subgroup, the DRO-based counterfactuals show changes around the expected areas while preserving the medical device.

Quantitative Counterfactual Evaluation In Table 2, counterfactual images generated by ERM and DRO show similar scores according to standard metrics: SSIM, Actionability and CPG. As these metrics are not designed to quantify whether the generated counterfactuals are affected by spurious correlations (see Section 2.2), the quality of the counterfactuals is now examined based on the proposed SCLS metric. The AUC of the classifier, d , trained to detect the presence of artifacts is 1.0 and 0.82 for Dataset 1 and Dataset 2, respectively. As indicated by the last row of Table 2, the ERM-based classifier shows a high value (poor performance) for SCLS for both datasets. On the other hand, the DRO-based classifier has a low value (good performance) for SCLS for both datasets. These results corroborate the finding made by visual comparison of the counterfactual images generated by the ERM and DRO classifiers. Overall, both qualitative

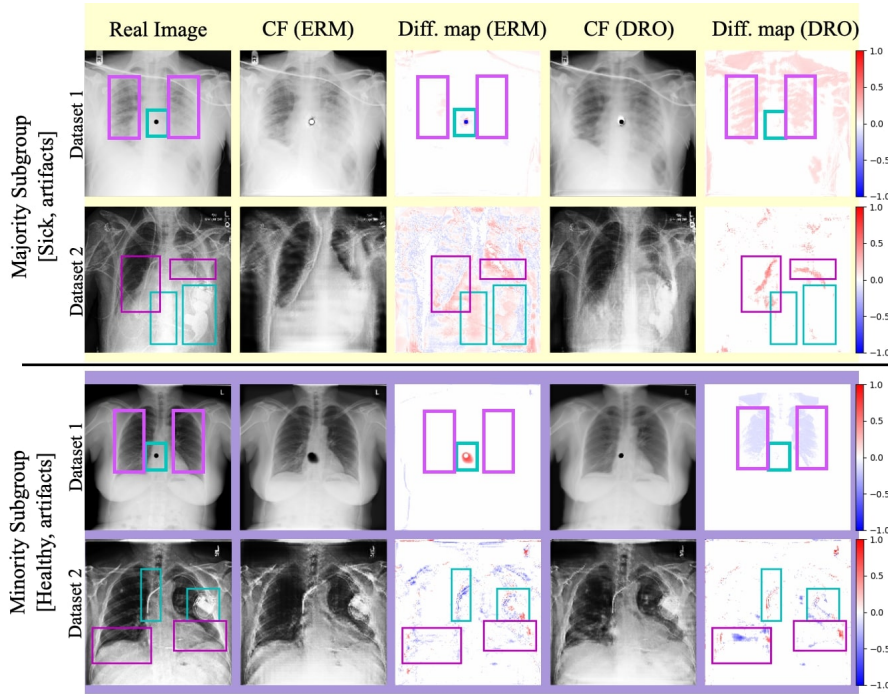


Fig. 5: Qualitative comparison of counterfactual (CF) images generated with ERM and DRO classifiers for both majority (top row) and minority (bottom row) subgroups. The ERM-based CFs show significant changes in the areas of spurious correlation (cyan boxes), whereas the DRO-based CFs show almost no changes in the same areas. In contrast, significant changes can be seen in the expected area of disease pathology (magenta boxes) in DRO-based CFs, while the ERM-based CFs show little to no changes in these areas.

and quantitative evaluations indicate that an ERM-optimized classifier latches on to the spurious correlation prevalent in the dataset, while a DRO-optimized classifier can be trained to successfully ignore the spurious correlation.

4 Conclusion

Safe deployment of black-box models requires explainability to disclose when the classifier is basing its predictions on spurious correlations and is therefore not generalizable. In this paper, we presented the first integrated end-to-end training strategy for generating unbiased counterfactual images, capitalizing on a DRO classifier to enhance generalization. Our experiments based on two datasets demonstrate that, unlike standard ERM classifiers which are susceptible to latching onto spurious correlations, the unbiased DRO classifier performs significantly better for minority subgroups in terms of- (a) the classifier performance and (b) the novel SCLS metric, which quantifies the degree to which the classifier latches

	Dataset 1		Dataset 2	
	ERM	DRO	ERM	DRO
Actionability ↓	7.68 ± 0.01	7.86 ± 0.01	4.93 ± 0.01	5.68 ± 0.04
SSIM ↑	98.03 ± 0.00	98.44 ± 0.01	98.21 ± 0.01	98.36 ± 0.01
CPG ↑	0.91 ± 0.04	0.96 ± 0.03	0.88 ± 0.07	0.89 ± 0.04
SCLS ↓	0.80 ± 0.08	0.12 ± 0.07	0.76 ± 0.09	0.22 ± 0.06

Table 2: Quantitative results to compare counterfactual images generated for both datasets. A low SCLS value implies that the model (f_{DRO} in this case) did not latch onto the spurious correlation.

on to the spurious correlation as depicted by the generated counterfactual images.

Current datasets typically do not provide the ground truth predictive markers of interest. Future work will require localizing the predictive markers (e.g. with bounding boxes) and determining the degree of overlap with the discovered markers. Further, we intend to explore the power of alternative debiasing techniques and their potential contribution to discovering generalizable image markers.

Acknowledgements The authors are grateful for funding provided by the Natural Sciences and Engineering Research Council of Canada, the Canadian Institute for Advanced Research (CIFAR) Artificial Intelligence Chairs program, the Mila - Quebec AI Institute technology transfer program, Microsoft Research, Calcul Quebec, and the Digital Research Alliance of Canada. S.A. Tsafaris acknowledges the support of Canon Medical and the Royal Academy of Engineering and the Research Chairs and Senior Research Fellowships scheme (grant RCSR1819 / 8 / 25), and the UK’s Engineering and Physical Sciences Research Council (EPSRC) support via grant EP/X017680/1.

References

1. Burlina, P., Joshi, N., Paul, W., Pacheco, K.D., Bressler, N.M.: Addressing artificial intelligence bias in retinal diagnostics. *Translational Vision Science & Technology* **10**(2), 13–13 (2021)
2. Cohen, J.P., Brooks, R., En, S., Zucker, E., Pareek, A., Lungren, M.P., Chaudhari, A.: Gifspianation via latent shift: a simple autoencoder approach to counterfactual generation for chest X-rays. In: *Medical Imaging with Deep Learning*. pp. 74–104. PMLR (2021)
3. DeGrave, A.J., Janizek, J.D., Lee, S.I.: AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence* **3**(7), 610–619 (2021)
4. Hore, A., Ziou, D.: Image quality metrics: PSNR vs. SSIM. In: *2010 20th international conference on pattern recognition*. pp. 2366–2369. IEEE (2010)
5. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghighi, B., Ball, R., Shpanskaya, K., et al.: CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: *Proceedings of the AAAI conference on artificial intelligence*. vol. 33, pp. 590–597 (2019)
6. Jiang, H., Xu, J., Shi, R., Yang, K., Zhang, D., Gao, M., Ma, H., Qian, W.: A multi-label deep learning model with interpretable grad-cam for diabetic retinopathy classification. In: *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. pp. 1560–1563. IEEE (2020)
7. Kumar, A., Hu, A., Nichyporuk, B., Falet, J.P.R., Arnold, D.L., Tsaftaris, S., Arbel, T.: Counterfactual image synthesis for discovery of personalized predictive image markers. In: *Artificial Intelligence over Infrared Images for Medical Applications and Medical Image Assisted Biomarker Discovery: First MICCAI Workshop, AII-IMA 2022, and First MICCAI Workshop, MIABID 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18 and 22, 2022, Proceedings*. pp. 113–124. Springer (2022)
8. Larrazabal, A.J., Nieto, N., Peterson, V., Milone, D.H., Ferrante, E.: 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)
9. Light, R.W.: Pleural effusion. *New England Journal of Medicine* **346**(25), 1971–1977 (2002)
10. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. *Advances in neural information processing systems* **30** (2017)
11. Magesh, P.R., Myloth, R.D., Tom, R.J.: An explainable machine learning model for early detection of parkinson’s disease using lime on datscan imagery. *Computers in Biology and Medicine* **126**, 104041 (2020)
12. Mehta, R., Shui, C., Arbel, T.: Evaluating the fairness of deep learning uncertainty estimates in medical image analysis. In: *Medical Imaging with Deep Learning* (2023)
13. Mertes, S., Huber, T., Weitz, K., Heimerl, A., André, E.: GANterfactual—counterfactual explanations for medical non-experts using generative adversarial learning. *Frontiers in artificial intelligence* **5**, 825565 (2022)
14. Mothilal, R.K., Sharma, A., Tan, C.: Explaining machine learning classifiers through diverse counterfactual explanations. In: *Proceedings of the 2020 conference on fairness, accountability, and transparency*. pp. 607–617 (2020)
15. Nemirovsky, D., Thiebaut, N., Xu, Y., Gupta, A.: CounteRGAN: Generating realistic counterfactuals with residual generative adversarial nets. *arXiv preprint arXiv:2009.05199* (2020)

16. Panwar, H., Gupta, P., Siddiqui, M.K., Morales-Menendez, R., Bhardwaj, P., Singh, V.: A deep learning and grad-CAM based color visualization approach for fast detection of COVID-19 cases using chest X-ray and CT-scan images. *Chaos, Solitons & Fractals* **140**, 110190 (2020)
17. Ricci Lara, M.A., Echeveste, R., Ferrante, E.: Addressing fairness in artificial intelligence for medical imaging. *Nature communications* **13**(1), 4581 (2022)
18. Sagawa, S., Koh, P.W., Hashimoto, T.B., Liang, P.: Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In: *International Conference on Learning Representations* (2019)
19. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: *Proceedings of the IEEE international conference on computer vision*. pp. 618–626 (2017)
20. Shui, C., Szeto, J., Mehta, R., Arnold, D., Arbel, T.: Mitigating calibration bias without fixed attribute grouping for improved fairness in medical imaging analysis. *arXiv preprint arXiv:2307.01738* (2023)
21. Singla, S., Eslami, M., Pollack, B., Wallace, S., Batmanghelich, K.: Explaining the black-box smoothly—a counterfactual approach. *Medical Image Analysis* **84**, 102721 (2023)
22. Targ, S., Almeida, D., Lyman, K.: Resnet in resnet: Generalizing residual architectures. *arXiv preprint arXiv:1603.08029* (2016)
23. Thiagarajan, J.J., Thopalli, K., Rajan, D., Turaga, P.: Training calibration-based counterfactual explainers for deep learning models in medical image analysis. *Scientific reports* **12**(1), 597 (2022)
24. Vapnik, V.: Principles of risk minimization for learning theory. *Advances in neural information processing systems* **4** (1991)
25. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: *Proceedings of the IEEE international conference on computer vision*. pp. 2223–2232 (2017)
26. Zong, Y., Yang, Y., Hospedales, T.: MEDFAIR: Benchmarking fairness for medical imaging. In: *International Conference on Learning Representations* (2023)
27. Zou, J., Schiebinger, L.: AI can be sexist and racist—it’s time to make it fair (2018)