DeCoDEx: Confounder Detector Guidance for Improved Diffusion-based Counterfactual Explanations

Nima Fathi[*1,2] NIMAFH@CIM.MCGILL.CA

Amar Kumar[*1,2] AMARKR@CIM.MCGILL.CA

Brennan Nichyporuk[1,2] BRENNANN@CIM.MCGILL.CA

Mohammad Havaei[3] MHAVAEI@GOOGLE.COM

[3] *Google Research*

Tal Arbel[1,2] ARBEL@CIM.MCGILL.CA

[1] *Center for Intelligent Machines, McGill University, Canada.*

[2] *MILA (Quebec AI institute), Canada.*

Editors: Under Review for MIDL 2024

Abstract

Deep learning classifiers are prone to latching onto dominant confounders present in a dataset rather than on the causal markers associated with the target class, leading to poor generalization and biased predictions. Although explainability via counterfactual image generation has been successful at exposing the problem, bias mitigation strategies that permit accurate explainability in the presence of dominant and diverse artifacts remain unsolved. In this work, we propose the DeCoDEx framework and show how an external, pre-trained binary artifact detector can be leveraged during inference to guide a diffusion-based counterfactual image generator towards accurate explainability. Experiments on the CheXpert dataset, using both synthetic artifacts and real visual artifacts (support devices), show that the proposed method successfully synthesizes the counterfactual images that change the causal pathology markers associated with Pleural Effusion while preserving or ignoring the visual artifacts. Augmentation of ERM and Group-DRO classifiers with the DeCoDEx generated images substantially improves the results across underrepresented groups that are out of distribution for each class. The code is made publicly available at https://github.com/NimaFathi/DeCoDEx.

Keywords: Bias Mitigation, Causality, Counterfactual Image Synthesis, Diffusion, Explainability, Spurious Correlations

1. Introduction

Deep learning (DL) methods have shown tremendous success in a wide variety of medical image tasks, including disease classification, due to their ability to learn generalizable, discriminative features across subjects. However, DL models are prone to learning shortcuts in order to obtain high overall accuracies, including any prevalent visual artifacts (e.g. marks in the image (DeGrave et al., 2021)) that are correlated with, but not causal of, the target outcome. Models that have not learned the relevant causal visual markers (Jia and Liang, 2017; Zech et al., 2018) are *right for the wrong reasons* (Sun et al., 2023b,a), and fail to generalize across out-of-distribution subgroups (Geirhos et al., 2018, 2020). Explainable

[*] Contributed equally

© 2024 N. Fathi, A. Kumar, B. Nichyporuk, M. Havaei & T. Arbel.

DL models that not only expose these biases but mitigate them are required in order to ensure their trustworthiness for safe clinical deployment.

Counterfactual (CF) image generation methods (e.g. Gifsplanation (Cohen et al., 2021) and Attri-Net (Sun et al., 2023b,a)) have recently been successful at exposing when the classifier is latching onto spurious correlations in order to obtain high performance. These methods employ conditional generation of the counterfactual image when the classifier has the opposing target outcome. Differences between the factual and counterfactual images should reflect the predictive local markers indicative of the class label, but also expose the classifier's reliance on spurious correlations. These methods do not mitigate the biases nor address their poor generalization. A number of debiasing methods (Sagawa et al., 2020; Wang et al., 2020; Sarhan et al., 2020) have recently been successful in several medical imaging contexts. Recent work (Kumar et al., 2023) combined Cycle-GAN counterfactual image generation and a Group-DRO (Sagawa et al., 2020) classifier to expose and mitigate the biases. The results showed improvement for minority subgroups, and classification based on disease-specific features. However, debiasing techniques have a number of known drawbacks, including improved fairness at the expense of a reduction in the performance in the majority subgroup. Integrating them into the counterfactual models requires sub-group labels for each class during training (which is often unavailable), and GANs require retraining the generative model with the pre-trained debiasing classifier in order to provide supervision for the counterfactual synthesis, which makes the process inflexible. Recently, an unconditional DDPM (Denoising Diffusion Probabilistic Model) (Ho et al., 2020), DiME (Jeanneret et al., 2022), has been proposed to generate classifier-guided counterfactual explanations, however, with the classifier latching onto shortcuts present in the dataset. Overall, developing bias mitigation strategies that permit accurate explainability in the presence of dominant and diverse visual artifacts remain open research questions.

This paper introduces DeCoDEx, a diffusion-based (DDPM) counterfactual image generator for debiased classifier explainability in the presence of dominant and diverse visual artifacts. DeCoDEx overcomes a number of limitations of current approaches: Rather than requiring training specialized debiasing classifiers for known subgroups, and then retraining the counterfactual generator that uses them (e.g. GANs), the framework provides debiased explainability of the classifier in question *at inference time* by leveraging the flexibility and explicit inference procedure (Dhariwal and Nichol, 2021; Wang et al., 2022; Kazerouni et al., 2022) of DDPMs, is generalizable to any subgroups, and shows stable training as compared to GANs. The framework can make use of any pre-trained, binary detector trained to indicate the presence or absence of the visual artifact in question (for any classes). During inference, the detector guides the diffusion-based counterfactual image generator towards accurate explainability, as the gradients from the detector counter the gradients from the classifier away from spurious correlations.

Extensive experiments are performed on the publicly available CheXpert dataset (Irvin et al., 2019), using both synthetic artifacts and real visual artifacts (support devices). Qualitative results show that the proposed method successfully synthesizes the counterfactual images by making changes in the pathology associated with Pleural Effusion while preserving or ignoring the visual artifacts. The quality of the counterfactual images was measured via several metrics, such as L1 distance, Counterfactual Prediction Gain (CPG) and Spurious Correlation Latching Score (Kumar et al., 2023). The results indicated the strength of

framework over a baseline (without a detector). Augmentation of the dataset with counterfactual images synthesized with DeCoDEx improves ERM and Group-DRO classification for the minority subgroups.

2. Methodology

The DeCoDEX framework involves a training strategy for explainability via counterfactual image synthesis, while ensuring robustness to spurious correlations. The model consists of a classifier and a trained unconditional denoising diffusion probabilistic model (DDPM). The framework leverages a pre-trained visual artifact detector that guides a DDPM to synthesize counterfactuals while ignoring the artifact.

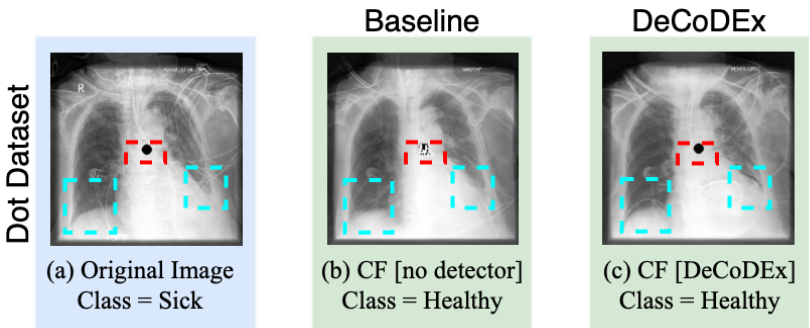

Figure 1: CF explanations for a subject with Pleural Effusion in the presence of an artifact: (a) Chest radiograph of a sick patient: dot artifact, disease pathology; (b) CF image from biased classifier using DDPM (i.e. DeCoDEx without detector) maintains the diseased area but modifies the dot; (c) DeCoDEx CF image modifies the Pleural Effusion area to look healthy as expected (Huang et al., 2022; Wang et al., 2017) while ignoring the dot artifact.

One of the advantages of our approach is the flexibility to use any pre-trained detector that can identify spurious correlations in the input sample. During the counterfactual image generation, when the biased classifier relies on spurious non-causal features in the dataset, the detector's gradient reversal signal readjusts the generation process, steering it back toward focusing on relevant pathological features (Fig. 1). It should be noted here that if there are several spurious correlations in the dataset that are difficult to detect, the detector may only block only some of them. Therefore, for some difficult cases, even after using a detector the counterfactual images may fail to make changes in the area correlated with the target class. An overview of our method is shown in Fig. 2.

Counterfactual Image Generation: The counterfactual image generation is designed to adhere to a set of constraints (Mothilal et al., 2020; Nemirovsky et al., 2020) through the following loss functions: (i)*Identity preservation loss,* $\mathcal{L}_{perc}$: counterfactual images preserves the identity of the factual image, $\mathcal{L}_{perc}$; (ii) *Classifier consistency loss*, $\mathcal{L}_{class}$: counterfactual image belongs to the correct target class. An additional *Detector loss* term, $\mathcal{L}_{det}$, is intro-

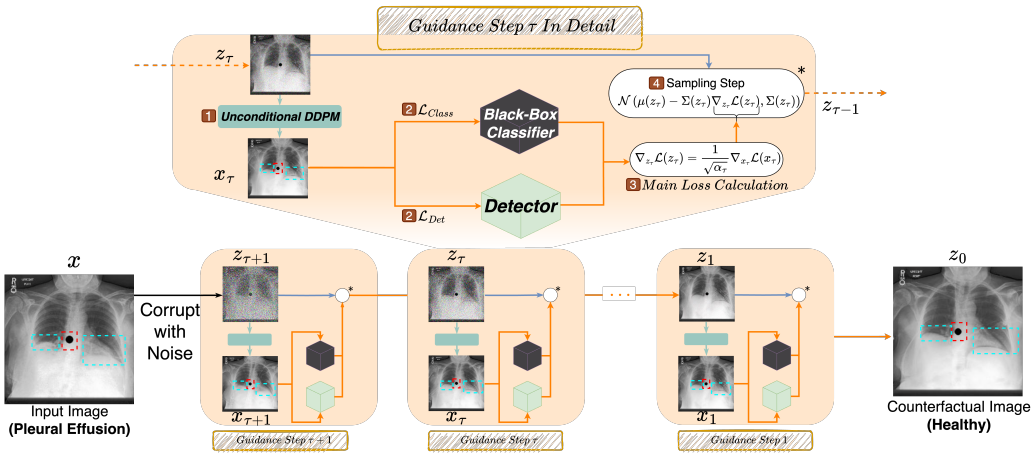

Figure 2: DeCoDEx Framework: Generating the counterfactuals (CFs) involves several inference steps. At each step, there are several components: (1) Denoising via unconditional DDPM, (2) pretrained classifier and detector loss, (3) gradient of the classifier, detector and perceptual loss, (4) counterfactual synthesis via sampling and backpropagating loss from black-box classifier and detector. The classifier, detector and unconditional DDPM are all pre-trained components. The resulting CF makes changes to the disease markers while disregarding visual artifacts.

duced to help guide the generation away from the spurious correlations that the classifier would have otherwise latched onto.

The generative model for synthesizing counterfactual images use a pre-trained DDPM with classifier guidance. DDPMs operate through a forward diffusion process (Eq. 1), which incrementally corrupts the original image x (or z_0) by adding Gaussian noise, culminating in a highly noised image z_T over T timesteps.

$$z_t \sim \mathcal{N}\left(\sqrt{1-\beta_t} z_{t-1}, \beta_t \mathbb{I}\right), \tag{1}$$

where β_t are predefined noise levels. The reverse diffusion process (Eq. 2) aims to recover the original image from z_T, using a neural network to predict and subtract the added noise iteratively.

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)\right), \tag{2}$$

where $\alpha_t := \prod_{k=1}^{t}(1-\beta_k)$ and $\epsilon_\theta(x_t, t)$ is the noise estimated by the network at step t. Guided-diffusion sampling (Dhariwal and Nichol, 2021) is used to denoise the image at any time step t, given by $z_{t-1} \sim \mathcal{N}\left(\mu(z_t) - \Sigma(z_t)\nabla_{z_t}\mathcal{L}(z_t; y_c, y_s, x_t), \Sigma(z_t)\right)$. The complete loss function is given by:

$$\mathcal{L}(x_t; y_c, y_s, x) = \lambda_c \mathcal{L}_{class}\left(C(y_c|x_t)\right) + \lambda_d \mathcal{L}_{det}\left(D(y_s|x_t)\right) + \lambda_p \mathcal{L}_{perc}(x_t, x), \tag{3}$$

where C and D refer to our classifier and detector modules, y_c and y_s refer to the target labels of the class and the presence of spurious correlation and $\lambda_c, \lambda_d, \lambda_p$ are hyperparmeters.

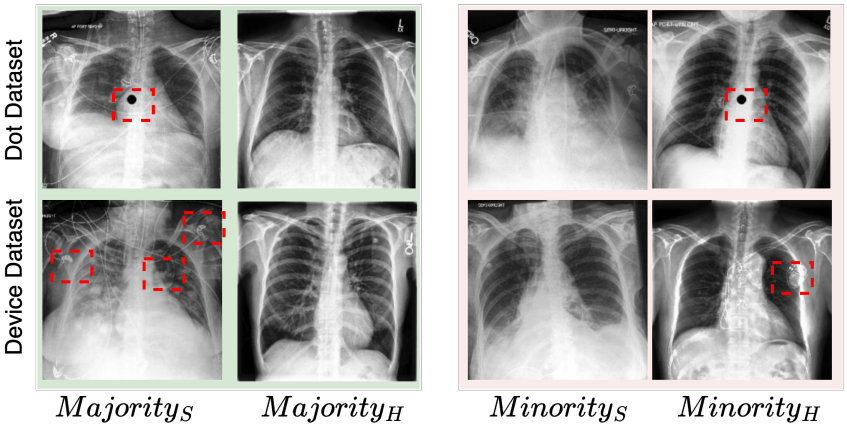

Figure 3: Majority and Minority subgroup samples from the Dot dataset (top row) and Device dataset (bottom row). Red boxes show the location of the artifacts.

Finally, the gradient of the complete loss function can be expressed as follows:

$$\nabla_{z_t}\mathcal{L}(z_t; y_c, y_s, x_t) = \frac{1}{\sqrt{\alpha_t}}\nabla_{x_t}\mathcal{L}(x_t; y_c, y_s, x). \tag{4}$$

3. Experiments and Results

3.1. Dataset and Implementation Details

We perform experiments on the publicly available CheXpert dataset (Irvin et al., 2019) that contains over 200,000 chest X-ray images, with binary labels for 14 diseases (e.g. Pleural Effusion, Cardiomegaly, Pneumonia), as well as binary labels indicating the presence of support devices (visual artifacts). We create two variants of the CheXpert dataset: (i) ***Dot dataset***: We introduce a synthetic artifact, a black dot of radius 9 pixels, in the center of the image for evaluating the quality of counterfactual images in the presence of spurious correlations. This also helps to compare the behaviour of counterfactual images synthesized from the baseline method: DeCoDEx without a detector and the proposed DeCoDEx. (ii) ***Device dataset:*** A subset of the CheXpert data is used to demonstrate the performance of DeCoDEx in the presence of real artifacts (Support devices). In both datasets, the artifacts are present in the majority of images of subjects with Pleural Effusion and in the minority of images of healthy subjects. In contrast, the majority of the images of healthy subjects and the minority of the images of subjects with Pleural Effusion do not contain these artifacts. Fig. 3 shows the sample images from both the datasets and all four subgroups. For both datasets, the ratio of the number of samples in majority to minority is 90:10 and the dataset is divided into training/validation/testing with a 70/15/15 random split. The details of number of samples in different split is included in Appendix A. The DenseNet-121 (Huang et al., 2017) architecture is used to train the classifier and detector. The classifier is trained separately on Dot and Device datasets. We use the standard Empirical Risk Minimization

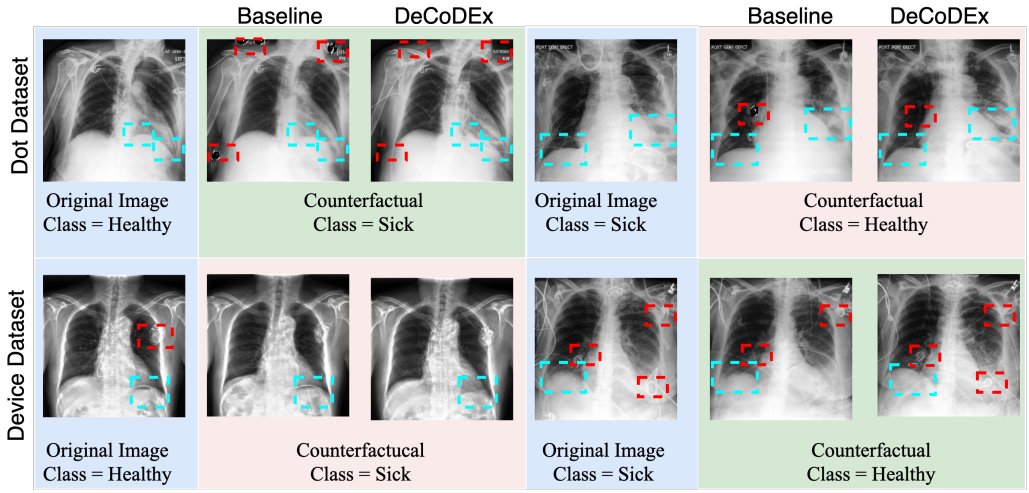

Figure 4: Qualitative comparison of counterfactual images synthesized via Baseline (i.e. DeCoDEx without detector) and DeCoDEx. For the baseline, most of the changes were made to the spurious correlation but for DeCoDEx visual artifacts were ignored and changes pertained to disease pathology.

(ERM) (Sagawa et al., 2020) as the optimization method. The detector for the support devices is pre-trained on the entire CheXpert dataset (except the held out test set). The detector is trained with binary target labels indicating the presence/absence of support devices. DeCoDEx is capable of handling images with multiple artifacts. These can vary in type, shape, size, location and intensities. An analysis showing the performance of the detector via counterfactual image generation is discussed in Appendix B.

3.2. Metrics and Experiments to evaluate counterfactuals

Several metrics are used to evaluate the quality of counterfactual images: (i) *Subject Identity Preservation: L1 Score* as depicted by the L1 distance between the counterfactual and the factual (original) image (as in (Mothilal et al., 2020; Nemirovsky et al., 2020)); (ii) *Counterfactual Prediction Gain (CPG)* (Nemirovsky et al., 2020) which measures the absolute value of the difference in the prediction of the classifier on the factual and counterfactual images (a higher score indicates a maximal change in the classifier decision boundary); (iii) *Spurious Correlation Latching Score (SCLS)*) (Kumar et al., 2023) assesses whether the spurious correlation was preserved in the counterfactual image (a lower SCLS score is desirable); (iv) *Classifier Flip Rate (CFR)* which represent the number of samples that flipped their class as per the classifier and (v) *Detector Robustness Rate (DRR)* showing number of samples that were robust to the detector.

In order to show that the synthesized counterfactuals learn useful features associated with the disease, the training data for the original classifier was augmented with synthesized images. The ERM classifier is retrained with the augmented data. An increase in perfor-

0. Only the training dataset is augmented while the validation and test split remains the same.

	CFR↑	DRR↑	L1↓	CPG↑	SCLS↓
Dot dataset					
Baseline	0.8975	0.2575	0.038	**0.592**	0.7394
DeCoDEx	**1**	**0.9775**	**0.036**	0.559	**0.058**
Device dataset					
Baseline	**0.97**	0.7625	0.040	0.377	0.201
DeCoDEx	0.89	**0.99**	**0.035**	**0.529**	**0.068**

Table 1: Quantitative results comparing the scores for the counterfactuals generated by the Baseline and DeCoDEx on both datasets. Notice the high DRR and low SCLS values for DeCoDEx showing the spurious correlation were ignored in the counterfactual images.

mance indicates that synthesized images learned discriminative features generalizable to the subgroups. We augment the Dot dataset with 200 counterfactual images and the Device dataset with 600 counterfactual images synthesized using basline method and DeCoDEx. For completeness, the augmentation experiments are repeated for a debiasing Group-DRO classifier.

3.3. Results

Classifier and Detector evaluation The classifier's performance on both datasets is shown in Table 2 on the row labeled 'ERM'. Note that the performance on the minority subgroup samples is significantly lower than the majority for both datasets. The dot and device detectors perform very well with subgroups [$majority_S$, $minority_S$, $minority_H$, $majority_H$] accuracies of [100, 99.9, 100, 99.8] and [91.8, 88.8, 79.2, 88.7]. Perfect accuracies for the dot detector can be attributed to the fact that the position and size of the dot remain fixed for all the subjects. However, the variability including size, position, intensity of medical devices are large, thus making detection challenging.

Qualitative evaluation Pleural effusion (PE) is characterized by the rounding of the costophrenic angle, augmented lung opacity, and reduced clarity of the diaphragm and lung fissures (Light, 2002). It is observed in the lower corner of the lungs (Wang et al., 2017; Huang et al., 2022). Qualitative results for counterfactual generation from the models can be seen in Fig. 4. For the baseline method on the Dot dataset, the counterfactual for a healthy subject simply added dots to the image, indicating that the classifier has latched onto the dot artifact. However, DeCoDEx ignored the dot artifact (and maintained the original text artifact) and made changes reflective of PE disease. For the Device dataset, the baseline generated a counterfactual for a sick patient from the majority subgroup by simply removing the medical device(s) while DeCoDEx makes (correct) changes in the area associated with the disease. Therefore, DeCoDEx indeed ignores the spurious correlation.

Quantitative evaluation In Table 1, CF images generated by the baseline and DeCoDEx show similar results for the metric L1 indicating CF images are similar to the factual image. CF synthesized by DeCoDEx have SCLS score close to zero indicating that the artifact was preserved in the counterfactual images. Table 2 shows the results of the augmentation

Dot dataset				
	Pleural Effusion		**Healthy**	
	Dot	No Dot	Dot	No Dot
ERM	97	2.2	8	100
ERM augmented with Baseline CFs	**98.5**	0	16.6	100
ERM augmented with DeCoDEx CFs	90.6	**12.1**	**53.1**	98.0
Group-DRO	90	61.8	56.0	88.0
Group-DRO augmented with Baseline CFs	**97**	8.0	33.0	**98.9**
Group-DRO augmented with DeCoDEx CFs	91.0	**70.2**	**60.9**	80.3
Device dataset				
	Pleural Effusion		**Healthy**	
	Support Device	No Support Device	Support Device	No Support Device
ERM	**92.7**	75.2	84.9	87.0
ERM augmented with Baseline CFs	92.5	74.8	83.4	87.0
ERM augmented with DeCoDEx CFs	92.6	**76.3**	**85.9**	86.8
Group-DRO	92.7	83.4	77.3	88.4
Group-DRO augmented with Baseline CFs	**97.8**	**85.4**	55.9	72.0
Group-DRO augmented with DeCoDEx CFs	93.2	84.7	**79.0**	88.4

Table 2: Augmented Classifier Accuracies: CFs generated by DeCoDEx and the Baseline are used to augment the imbalanced datasets. Both ERM and Group-DRO are retrained on these augmented datasets and the effects are examined. The accuracies (percentages) of all classifiers are shown on the held out test set. Green indicate majority subgroups (90%) and red are the minority subgroups (10%). Best results are in bold. Note the increase in performance for the minority classes when both the methods, ERM and Group-DRO are augmented with DeCoDEx counterfactual images, illustrating the power of our method on extracting disease pathology and generating better counterfactual explanations.

experiments. Augmented training significantly improves the performance over the minority groups for the Dot dataset, indicating that CF samples have learned discriminative features common to this subgroup. Among both ERM and Group-DRO based techniques, our method outperforms the Baseline along with augmented CFs. CF synthesized from DeCoDEx when augmented with ERM performs as well as Group-DRO for some minority classes which is a positive result, particularly when debiasing cannot be performed due to the lack of annotations.

4. Conclusions

In medical image analysis, explainable models are needed to expose and mitigate the bias to improve the trustworthiness of complex models. This paper presents DeCoDEx, an explainability framework that leverages a pre-trained classifier and detector to guide diffusion-based counterfactual synthesis towards accurate disease markers while ignoring spurious correlations. Qualitative and quantitative analysis of our extensive experiments indicate that the proposed method outperforms the baseline model that does not use a detector. Furthermore, the flexibility of our method allows it to be used with any pre-trained detector, does not require retraining a debiasing classifier and associated generative architecture, and provides guidance during inference. One of the current limitations of the model is resulting minor changes throughout the generated CF images. Future work will explore conditional score-based generative models.

References

Joseph Paul Cohen, Rupert Brooks, Sovann En, Evan Zucker, Anuj Pareek, Matthew P Lungren, and Akshay Chaudhari. Gifsplanation via latent shift: a simple autoencoder approach to counterfactual generation for chest x-rays. In *Medical Imaging with Deep Learning*, pages 74–104. PMLR, 2021.

Alex J DeGrave, Joseph D Janizek, and Su-In Lee. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, 2021.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

Tao Huang, Rui Yang, Longbin Shen, Aozi Feng, Li Li, Ningxia He, Shuna Li, Liying Huang, and Jun Lyu. Deep transfer learning to quantify pleural effusion severity in chest x-rays. *BMC Medical Imaging*, 22(1):1–11, 2022.

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

Guillaume Jeanneret, Loïc Simon, and Frédéric Jurie. Diffusion models for counterfactual explanations. In *Proceedings of the Asian Conference on Computer Vision*, pages 858–876, 2022.

Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328*, 2017.

Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Moein Heidari, Reza Azad, Mohsen Fayyaz, Ilker Hacihaliloglu, and Dorit Merhof. Diffusion models for medical image analysis: A comprehensive survey. *arXiv preprint arXiv:2211.07804*, 2022.

Amar Kumar, Nima Fathi, Raghav Mehta, Brennan Nichyporuk, Jean-Pierre R Falet, Sotirios Tsaftaris, and Tal Arbel. Debiasing counterfactuals in the presence of spurious correlations. In *Workshop on Clinical Image-Based Procedures*, pages 276–286. Springer, 2023.

Richard W Light. Pleural effusion. *New England Journal of Medicine*, 346(25):1971–1977, 2002.

Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 607–617, 2020.

Daniel Nemirovsky, Nicolas Thiebaut, Ye Xu, and Abhishek Gupta. Countergan: Generating realistic counterfactuals with residual generative adversarial nets. *arXiv preprint arXiv:2009.05199*, 2020.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization, 2020.

Mhd Hasan Sarhan, Nassir Navab, Abouzar Eslami, and Shadi Albarqouni. Fairness by learning orthogonal disentangled representations. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*, pages 746–761. Springer, 2020.

Susu Sun, Lisa M Koch, and Christian F Baumgartner. Right for the wrong reason: Can interpretable ml techniques detect spurious correlations? In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 425–434. Springer, 2023a.

Susu Sun, Stefano Woerner, Andreas Maier, Lisa M Koch, and Christian F Baumgartner. Inherently interpretable multi-label classification using class-specific counterfactuals. *arXiv preprint arXiv:2303.00500*, 2023b.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.

Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8919–8928, 2020.

Zhendong Wang, Huangjie Zheng, Pengcheng He, Weizhu Chen, and Mingyuan Zhou. Diffusion-gan: Training gans with diffusion. *arXiv preprint arXiv:2206.02262*, 2022.

John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to

detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11): e1002683, 2018.

Appendix A. Detailed dataset description

We elaborate on the datasets and their variants used in our experiments in this appendix, complemented by statistical data on participant distribution across groups as detailed in Table 3.

Dot dataset						
Spurious Correlation	**Pleural Effusion**			**Healthy**		
	Train	Validation	Test	Train	Validation	Test
Dot	1359	191	242	56	5	49
No Dot	166	28	242	340	51	49
Device dataset						
Spurious Correlation	**Pleural Effusion**			**Healthy**		
	Train	Validation	Test	Train	Validation	Test
Support Device(s)	6653	1432	1389	665	143	138
No Support Device(s)	665	143	138	6653	1432	1389

Table 3: Summary of the number of samples for both dataset variants

Appendix B. Explaining the detector

Since we show that the gradients from the detector can help avoid spurious correlation, it is important to understand the features important to the detector. Fig. 5 shows how medical devices can be removed from the original image based on the gradients received from a binary classifier.

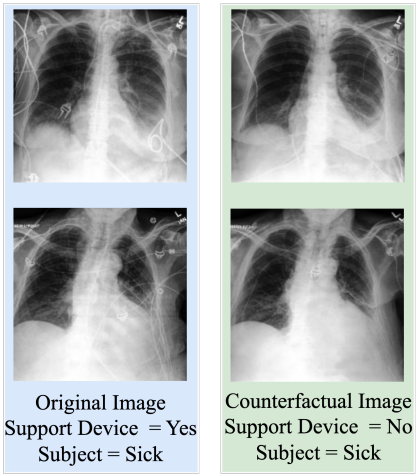

Figure 5: CF explanations for the detector: Removing medical devices from the original images while explaining the detector

Appendix C. Extensive augmentation of the minority subgroup with synthesized CFs

Dot dataset				
	Pleural Effusion		**Healthy**	
	Dot	No Dot	Dot	No Dot
ERM augmented with DeCoDEx CFs [200]	90.6	12.1	53.1	98.0
ERM augmented with DeCoDEx CFs [400]	**93.8**	**26.5**	**61.2**	98.3
Group-DRO augmented with DeCoDEx CFs [200]	91.0	70.2	60.9	80.3
Group-DRO augmented with DeCoDEx CFs [400]	**93.1**	**81.6**	**65.3**	85.9
Device dataset				
	Pleural Effusion		**Healthy**	
	Support Device	No Support Device	Support Device	No Support Device
ERM augmented with DeCoDEx CFs [600]	92.6	76.3	85.9	86.8
ERM augmented with DeCoDEx CFs [1600]	**92.8**	**76.8**	**86.8**	82.3
Group-DRO augmented with DeCoDEx CFs [600]	93.2	84.7	79.0	88.4
Group-DRO augmented with DeCoDEx CFs [1600]	**93.5**	**86.5**	**79.9**	**90.1**

Table 4: Improved accuracy at subgroup level through extensive classifier augmentation: Using DeCoDEx, we expanded on counterfactual generation to demonstrate improvement in the subgroup accuracy. We synthesise 400 counterfactual samples for the Dot dataset and 1600 for the Device dataset (the number in the square bracket refers to the total number of augmentation samples added to both majority and minority subgroups). First row is the original results discussed in Table 2 and the second row shows the result after extensive augmentation. Notice the improvement in the accuracy of minority subgroups across both Dot and Device datasets. Notably, 90% of these counterfactuals represent minority subgroups, thereby achieving a more equalized distribution in the dataset. Our findings indicate improvement in the accuracy of minority groups across all scenarios.

Appendix D. Validating Sex-Preservation for Counterfactual Images

The identity preservation loss in Equation 3 does not guarantee that all the other attributes of the patients are maintained in the counterfactual images. A quick experiment was performed to validate that the sex of the patients is maintained in the counterfactual images generated by DeCoDEx. To this end, a sex classifier, $\mathcal{G}$, is trained on the real images, and then tested on real and synthesized counterfactual images. The sex classifier had an AUC-ROC of 0.98 on the real (factual) images. The differences in the sex classifier results based on the factual (F) and the counterfactual (CF) images, $|\mathcal{G}(\text{F}) - \mathcal{G}(\text{CF})|$, were 0.08 on average, indicating that the sex attribute was maintained in the counterfactual images.