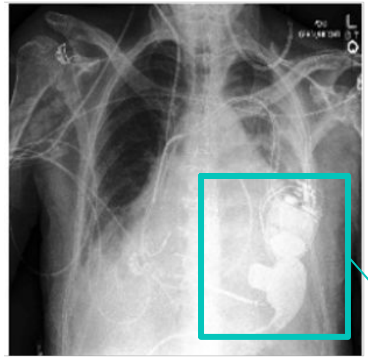
Debiasing Counterfactuals In the Presence of Spurious Correlations

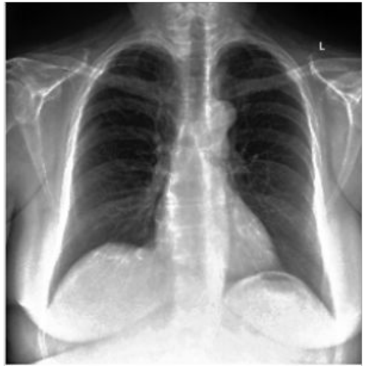
Amar Kumar, Nima Fathi, Raghav Mehta, Brennan Nichyporuk, Jean-Pierre R. Falet, Sotirios Tsaftaris, Tal Arbel

Motivation



Medical Device

Sick patients

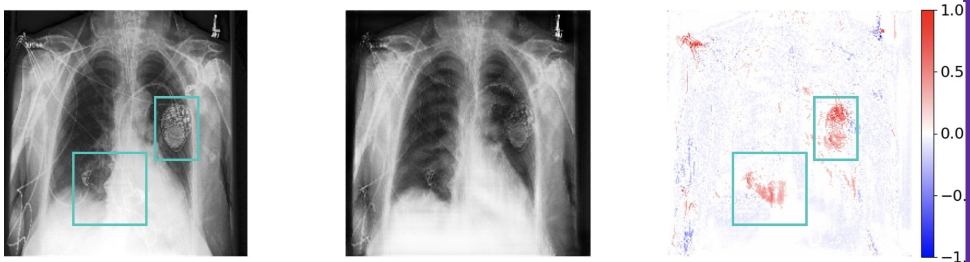


Healthy patients

Deep learning methods learns a ‘shortcut’
Disease = Medical Devices

- Deep learning model optimizes for majority population

Deep learning methods learns a ‘shortcut’
Disease = Medical Devices



(a) Real (Sick Subject) (b) CF (Healthy Subject) (c) Diff. map (Real - CF)

Counterfactual (CF) Explainability: classifier latches onto spurious correlations (prevalent in the training dataset for sick subjects)

- Deep learning model optimizes for majority population
- Explainability - Counterfactual generation shows when the model is ‘right for wrong reasons’

Background

Debiasing and Explainability

- Debiasing
  - Stochastic Weight Averaging Densely (SWAD) [1]
  - Sharpness-Aware Minimization (SAM) [2]
- Explainability
  - Grad-CAM [3], LIME [4], SHAP [5], Gifsplanation [6]
  - Counterfactual Explanations[7]
  - Attrinet[8]
- Ours - debias + explain

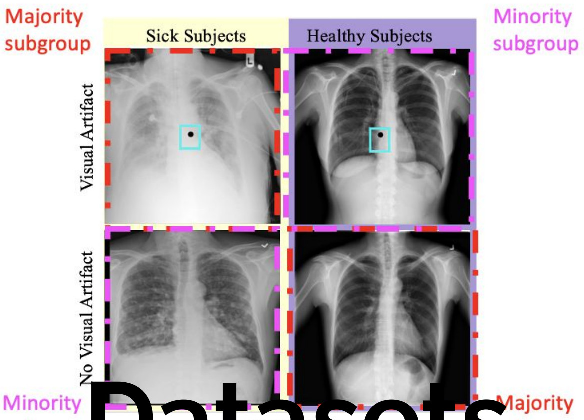
[1] Cha et al.: Swad: Domain generalization by seeking flat minima..Neurips 2021
[2] Foret et al.: Sharpness-aware minimization for efficiently improving generalization. arXiv preprint arXiv:2010.01412
[3] Selvaraju et al.: Grad-cam: Visual explanations from deep networks via gradient-based localization. ICCV 2017
[4] Ribeiro et al.: "Why should i trust you?" explaining the predictions of any classifier. ACM SIGKDD 2016
[5] Lundberg et. al.: A unified approach to interpreting model predictions. Neurips 2017
[6] Cohen et al..: Gifsplanation via latent shift: A simple autoencoder approach to progressive exaggeration on chest x-rays. MIDL 2021
[7] Ribeiro et al..: High Fidelity Image Counterfactuals with Probabilistic Causal Models.
[8] Sun et al..: Inherently Interpretable Multi-Label Classification Using Class-Specific Counterfactuals.. MIDL 2023

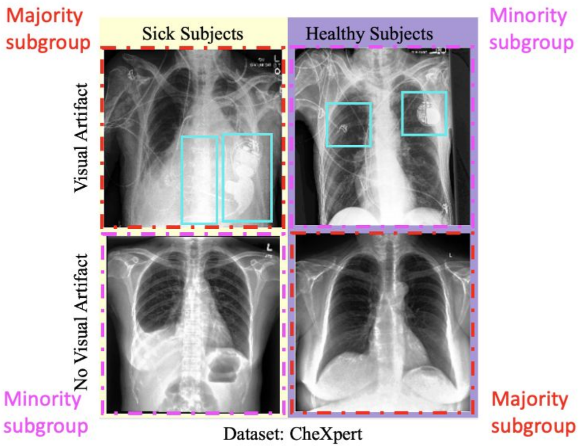
Explainability via Counterfactual Images

…. Debiasing the results

Can a model be trained to disregard spurious correlations and identify generalizable predictive disease markers?

Datasets





Dataset: CheXpert
Disease: Pleural Effusion

Experiments are performed on two publicly available datasets:

(i) RSNA Pneumonia Detection Challenge

… with synthetic artifacts

(ii) CheXpert

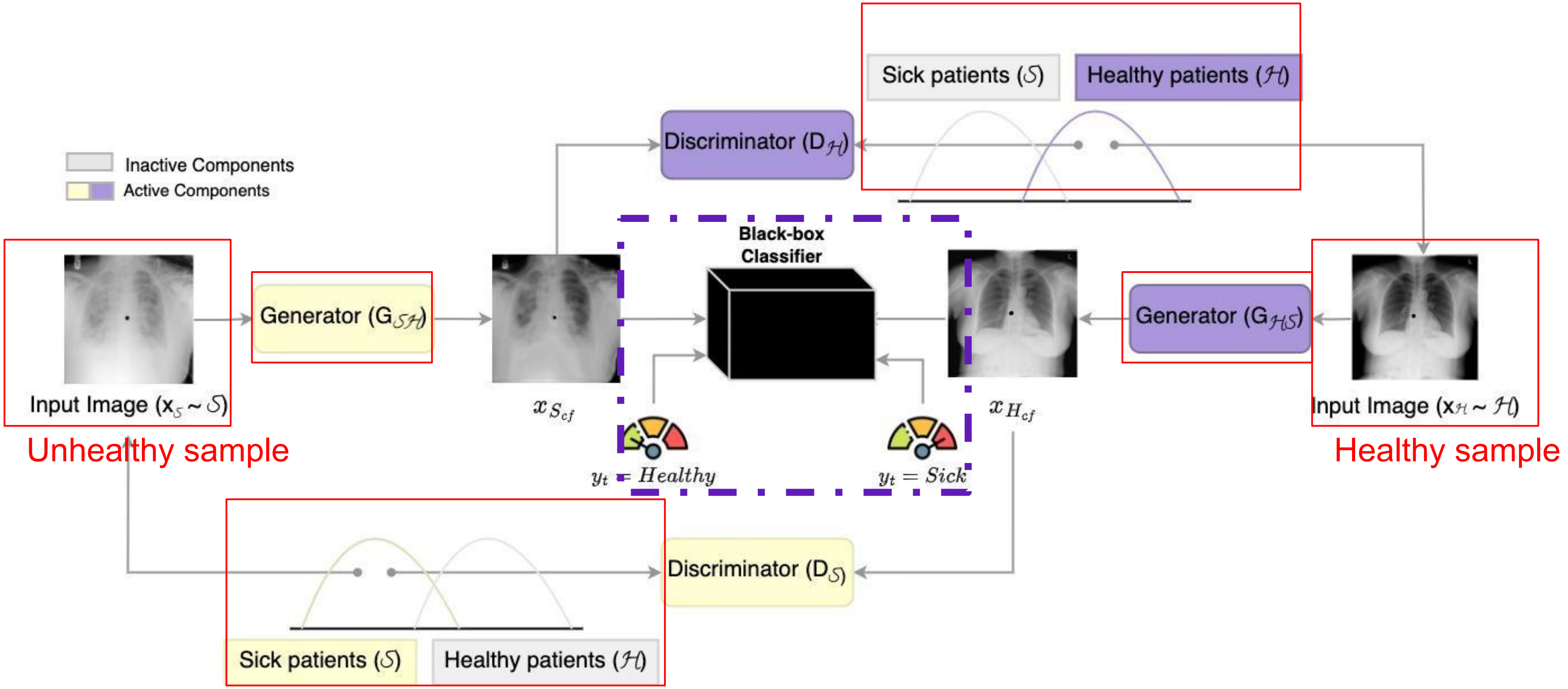
… with real artifacts (medical devices)

Methodology & Contributions

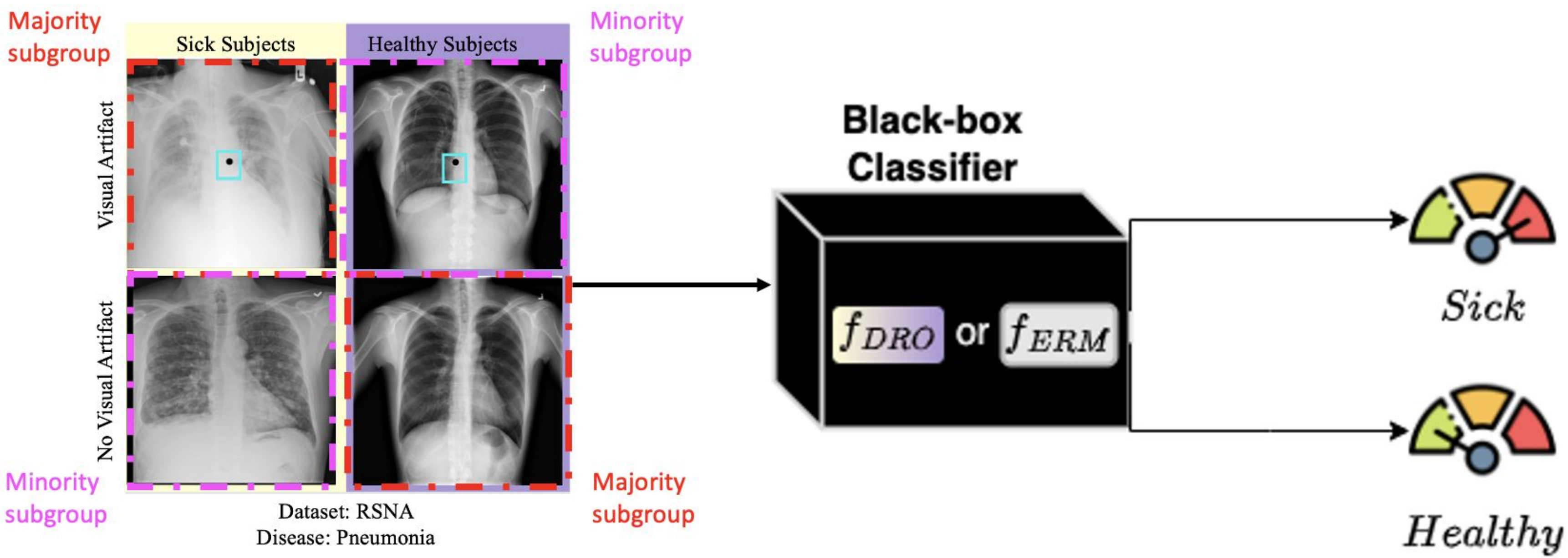
End-to-end training of a generative model to (i) debias and (ii) explain the classifier decision.

Evaluation of the counterfactual image using a new proposed - Spurious Correlation Latching Score (SCLS)

Cycle-GAN for Counterfactual Image

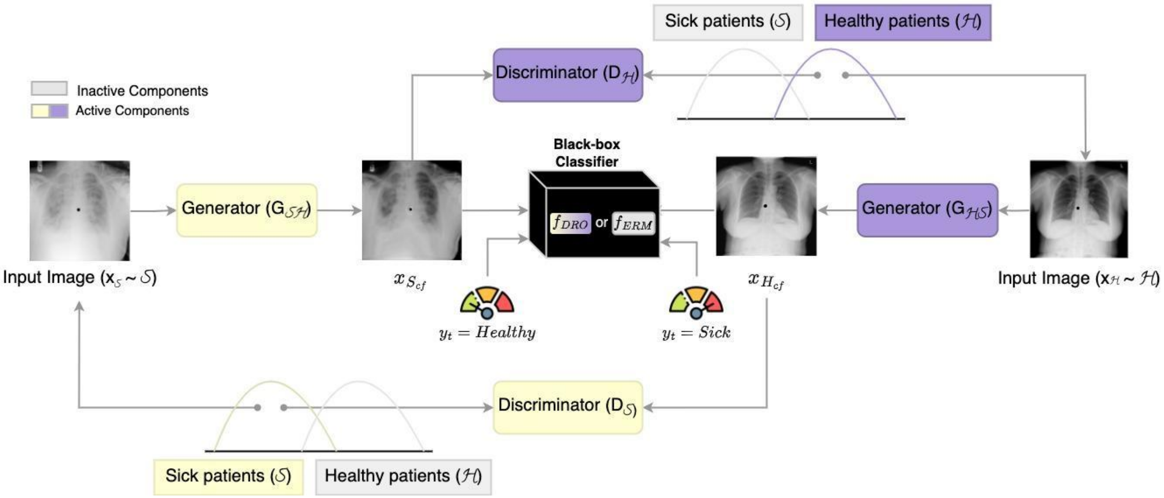


Debiasing Classifier - DRO



ERM: Empirical Risk Minimization; DRO: Distributionally Robust Optimization

Counterfactuals Image Synthesis



Constraints on Counterfactual images:

1. Identity Preservation
2. Classifier consistency
3. Cycle consistency

Evaluation of the counterfactual images

1. **Identity Preservation :** Structural Similarity Index (SSIM) and Actionability to ensure counterfactual images look similar to factual images

1. **Counterfactual Prediction Gain (CPG):** Ensures the counterfactual images belong to the correct target class.

Standard Metrics

1. **Spurious Correlation Latching Score (SCLS):** Identifies the presence of spurious correlation in the image

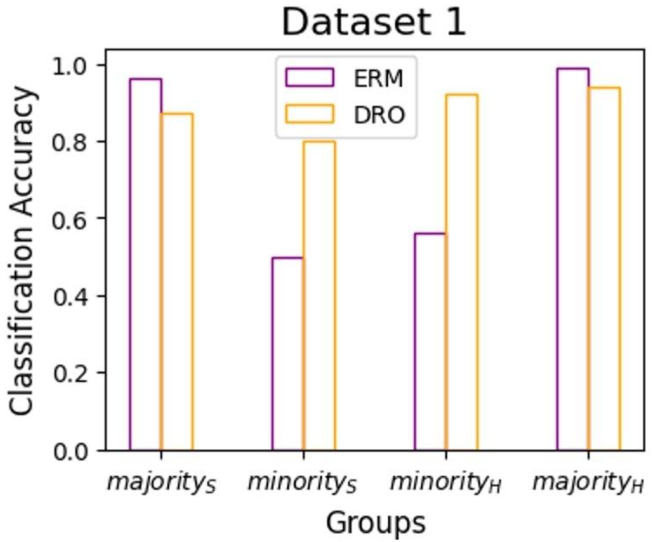
New Proposed Metri

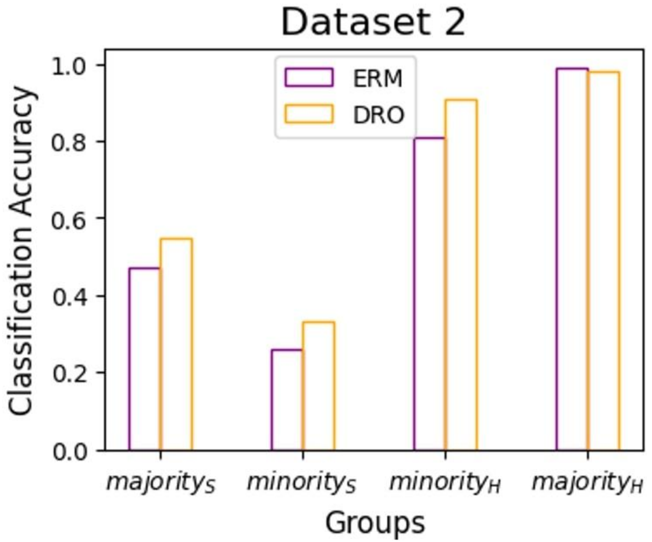
Results

We evaluate the performance of

1. Classifier
2. Counterfactuals
   a. Qualitatively
   b. Quantitatively

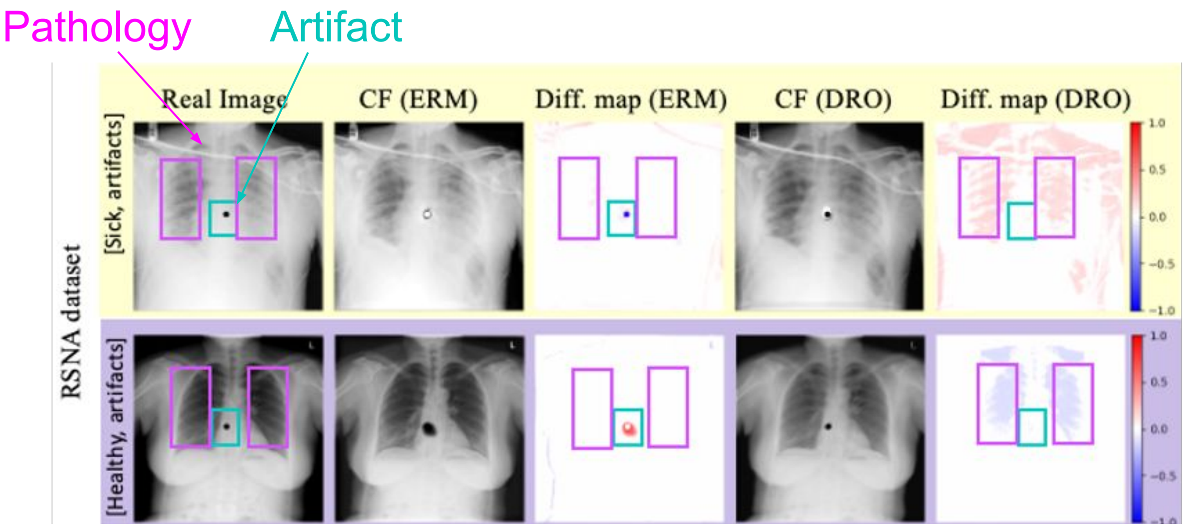
Classifier Evaluation





DRO performs better indicating generalization on the underrepresented classes.

Counterfactual Evaluation [Qualitative]



- ERM : Significant changes in artifact; DRO: No change in artifact
- ERM : No changes in disease pathology; DRO: Significant changes in disease pathology

Counterfactual Evaluation [Quantitative]

	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**

Lower SCLS score indicates that DRO based classifier does not latch onto the spurious correlation.

Conclusion

- Safe deployment of DL models in medical imaging -> Explainability
  - To expose and mitigate spurious correlation/ biases
- First integrated end-to-end training strategy for generating unbiased counterfactual images
  - DRO classifier to enhance generalization

Mila

Microsoft Research

McGill UNIVERSITY

THE UNIVERSITY of EDINBURGH

NSERC CRSNG

Royal Academy of Engineering

CIFAR

Canon
CANON MEDICAL

Digital Research Alliance of Canada | Alliance de recherche numérique du Canada

Thank you!

amarkr@cim.mcgill.ca