Debiasing Counterfactuals in the Presence of Spurious Correlations

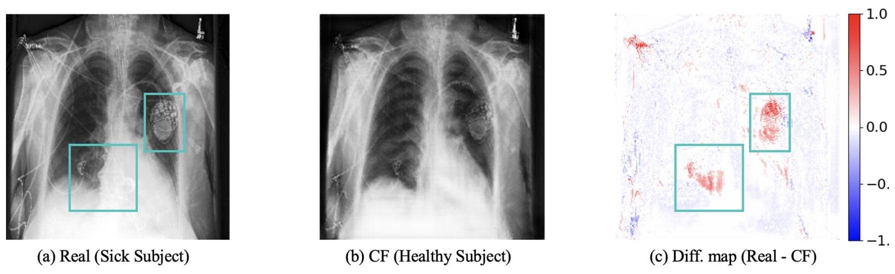
Amar Kumar [1,2], Nima Fathi [1,2], Raghav Mehta [1,2], Brennan Nichyporuk [1,2], Jean-Pierre R. Falet [2,3], Sotirios Tsaftaris [4,5] and Tal Arbel [1,2]

[1] Centre for Intelligent Machines, McGill University, Montreal, Canada
[2] MILA Quebec AI Institute, Montreal, Canada
[3] Montreal Neurological Institute, McGill University, Canada
[4] Institute for Digital Communications, School of Engineering, University of Edinburgh, UK
[5] The Alan Turing Institute, UK

MICCAI 2023 Vancouver CANADA

(1) Introduction

- Deep learning models can take 'shortcut paths to optimization' by latching onto spurious correlation prevalent in the dataset.
- Explainability: verifies model is 'right for right reasons'.



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

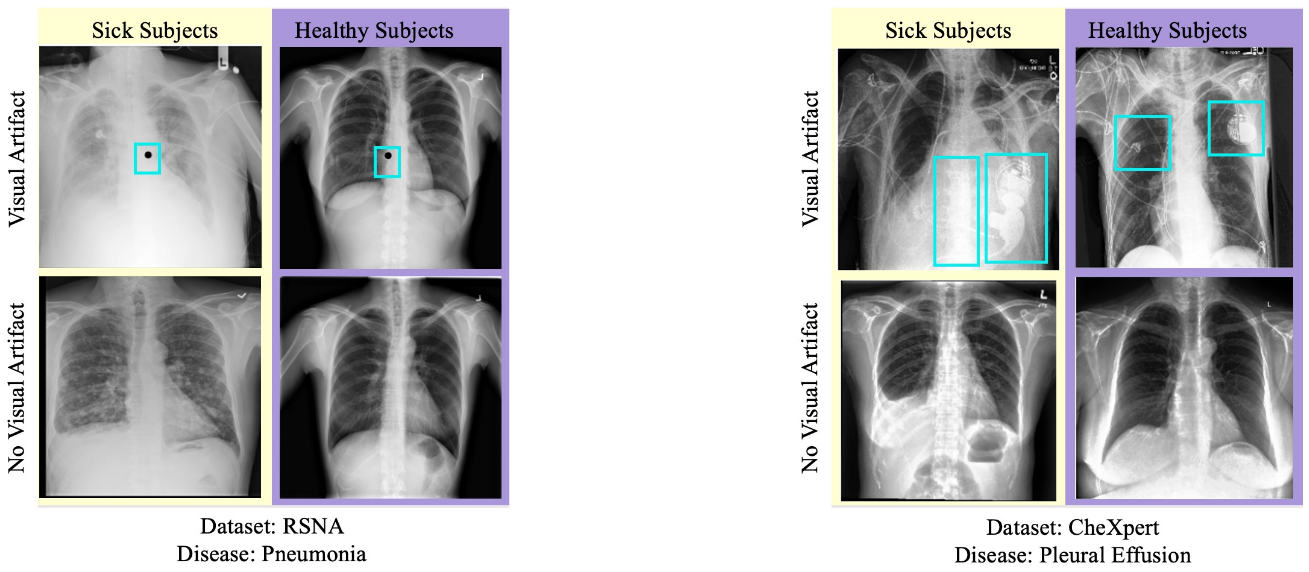
➠ Counterfactual explanation shows 'right for wrong' reasons.

- **Goal**: Develop first end-to-end framework to debias counterfactual explanations in presence of spurious correlations.

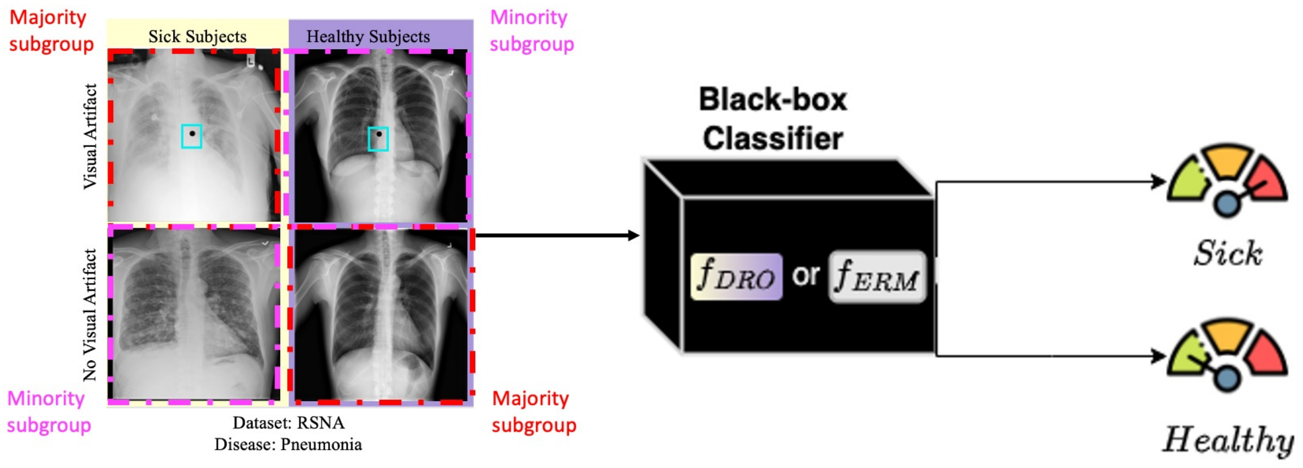
(2) Proposed Framework

Dataset Preparation:

Spurious Correlation (visual artifact) is prevalent in majority of patients.

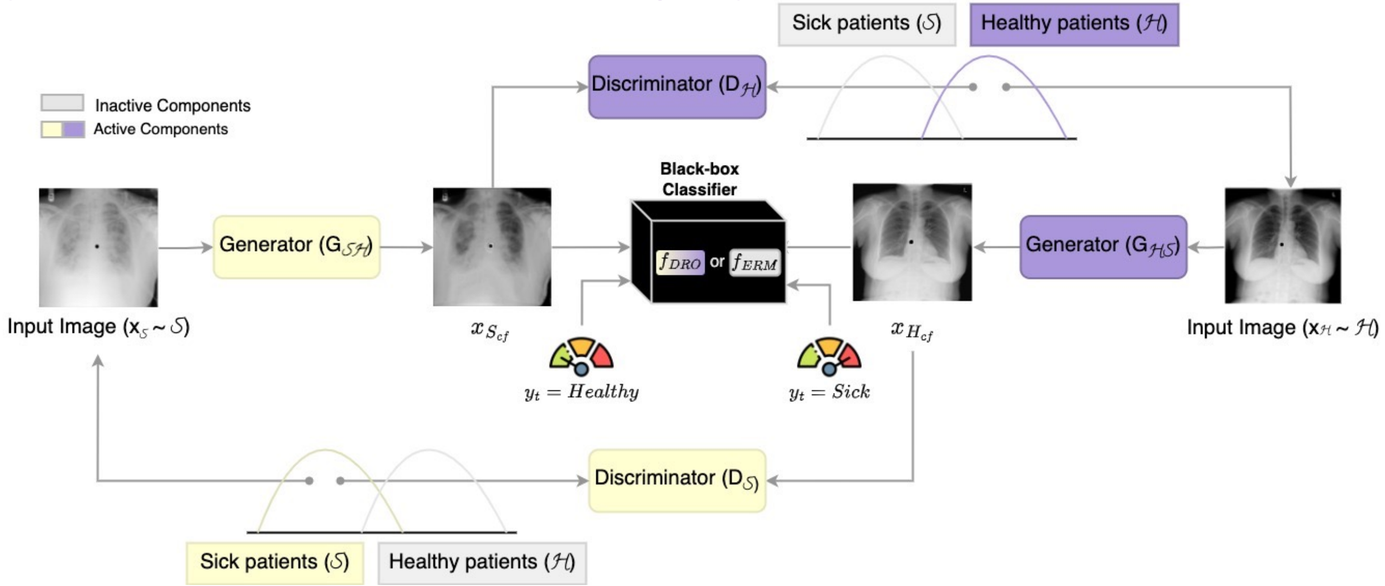


Debiased Classifier (DRO) to Overcome Spurious Correlations



ERM: Empirical Risk Minimization; DRO: Distributionally Robust Optimization

Cycle-GAN for Counterfactual Image Synthesis


Sick patients ($\mathcal{S}$)
Healthy patients ($\mathcal{H}$)
Inactive Components
Active Components
Discriminator ($D_{\mathcal{H}}$)
Black-box Classifier
f_{DRO} or f_{ERM}
Generator ($G_{\mathcal{SH}}$)
Generator ($G_{\mathcal{HS}}$)
Input Image ($x_{\mathcal{S}} \sim \mathcal{S}$)
$x_{S_{cf}}$
$y_t = Healthy$
$y_t = Sick$
$x_{H_{cf}}$
Input Image ($x_{\mathcal{H}} \sim \mathcal{H}$)
Discriminator ($D_{\mathcal{S}}$)
Sick patients ($\mathcal{S}$)
Healthy patients ($\mathcal{H}$)

Evaluating Counterfactual Images

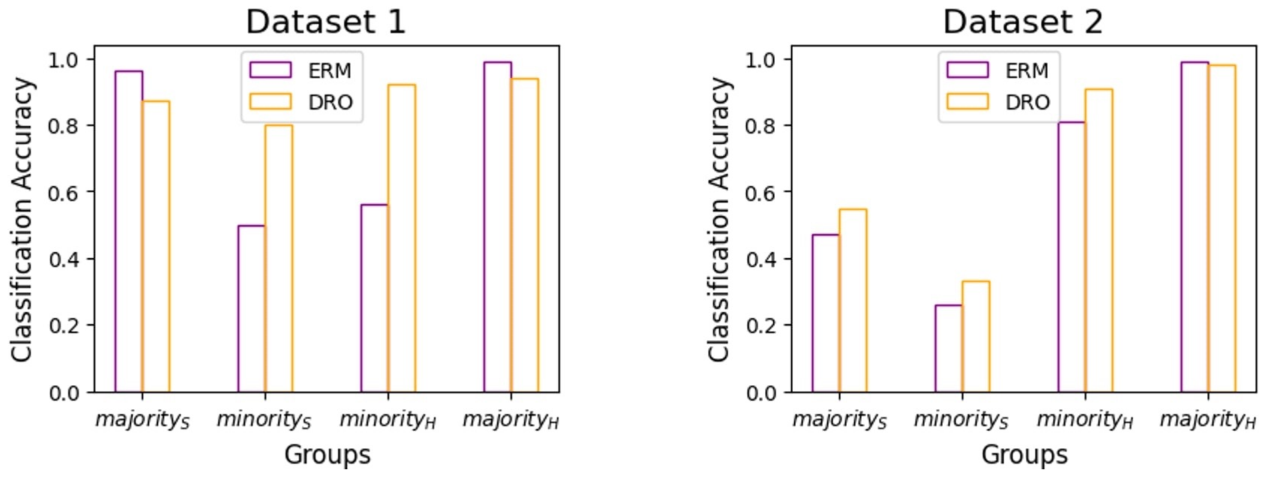
Standard Metrics: Structural Similarly Index Measure (SSIM), Actionability and Counterfactual Prediction Gain (CPG)

New Proposed Metric: Spurious Correlation Latching Score (SCLS) measures the presence of spurious correlation in the synthesized image using a detector, d.

$$SCLS = |d(x) - d(x_{cf})|$$

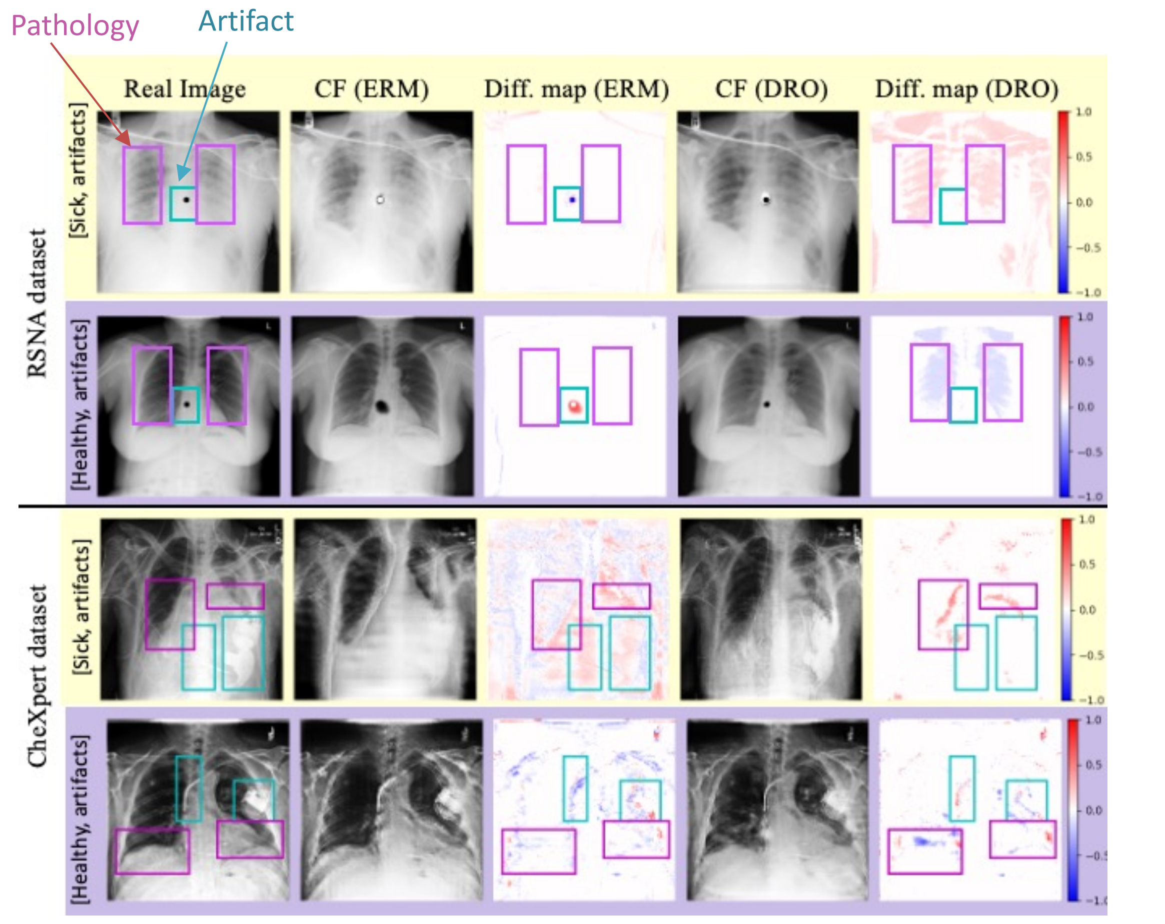
(3) Experiments and Results

Performance of ERM and DRO based classifiers across all subgroups



➠ DRO performs better across the underrepresented subgroups.

Qualitative Comparison of Counterfactuals with ERM and DRO classifiers



- 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 DRO based classifier does not latch onto 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
- First integrated end-to-end training strategy for generating unbiased counterfactual images, leveraging a DRO classifier to enhance generalization

References:

1) DeGrave, A.J, Janizek, J.D. et. al: AI for radiographic COVID-19 detection selects shortcuts over signal. Nature Machine Intelligence 3(7), 610–619 (2021)
2) Kumar A., et. al: Counterfactual Image Synthesis for Discovery of Personalized Predictive Image Markers
3) Irvin, J., Rajpurkar, P., Ko, M., 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)
4) 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)
5) Vapnik, V.: Principles of risk minimization for learning theory. Advances in neural information processing systems 4 (1991)

Acknowledgement : 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. Tsaftaris acknowledges the support of Canon Medical and the Royal Academy of Engineering and the Research Chairs and Senior Research Fellowships scheme, and the UK's Engineering and Physical Sciences Research Council (EPSRC)

Paper

✉ amarkr@cim.mcgill.ca