Investigating the Relationship Between Debiasing and Artifact Removal using Saliency Maps

Lukasz Sztukiewicz[1,†], Ignacy Stępka[1,†], Michał Wiliński[1,†] and Jerzy Stefanowski[1,*]

[1] *Institute of Computing Science, Poznan University of Technology, Poland*

Abstract

The widespread adoption of machine learning systems has raised critical concerns about fairness and bias, making mitigating harmful biases essential for AI development. In this paper, we investigate the relationship between debiasing and removing artifacts in neural networks for computer vision tasks. First, we introduce a set of novel XAI-based metrics that analyze saliency maps to assess shifts in a model's decision-making process. Then, we demonstrate that successful debiasing methods systematically redirect model focus away from protected attributes. Finally, we show that techniques originally developed for artifact removal can be effectively repurposed for improving fairness. These findings provide evidence for the existence of a bidirectional connection between ensuring fairness and removing artifacts corresponding to protected attributes.

Keywords

Deep learning, Fairness, Debiasing, Saliency maps

1. Introduction

Machine learning (ML) systems are becoming widespread across numerous application domains. However, their adoption raises concerns about perpetuating harmful biases and creating discriminatory systems [1]. This problem has been noticed not only by practitioners but also policymakers, resulting in regulatory efforts [2], which underscore the importance of fairness in ML. Machine learning fairness refers to the principle of ensuring that algorithmic decisions do not produce biased or discriminatory outcomes across different groups. Neural networks, especially in computer vision applications, present unique challenges for fairness assessment and bias mitigation [3]. Unlike tabular data, where features are explicitly defined, images lack semantic meaning at the raw pixel level. To gain predictive power, models learn to extract high-level semantic features. This learned featurization becomes problematic when dealing with protected attributes – high-level features (concepts), such as gender, hair color or race, consisting of various pixel combinations. Neural networks are known to develop internal representations that encode not only useful high-level features but also harmful biases [4]. For example, in the CelebA dataset [5], wearing a necktie is highly correlated with the male gender, and can be used as a proxy feature to infer gender, thus creating potential unintended pathways for discrimination.

To address discrimination and ensure fairness in ML models, various approaches have been proposed [6, 7]. However, while existing debiasing methods generally improve fairness metrics, they often fail to explicitly address harmful biases encoded in models' internal representations. To this end, we examine the relationship between successful fairness improvement and removal of harmful biases from these representations. We propose new metrics that quantify the properties of saliency maps given a region of interest, and capture the extent to which biases are removed from the model's decision-making process.

Our findings provide evidence that effective debiasing methods redirect the model's focus away from protected attributes, while explicitly optimizing only the fairness criterion. Furthermore, we observe

Late-breaking work, Demos and Doctoral Consortium, colocated with the 3rd World Conference on eXplainable Artificial Intelligence: July 09–11, 2025, Istanbul, Turkey

[*] Corresponding author.

[†] These authors contributed equally.

jerzy.stefanowski@cs.put.poznan.pl (J. Stefanowski)

0009-0008-7077-0226 (L. Sztukiewicz); 0009-0004-4575-0689 (I. Stępka); 0009-0004-4818-8417 (M. Wiliński); 0000-0002-4949-8271 (J. Stefanowski)

that techniques originally developed for artifact removal, such as the family of ClArC methods [8], also optimize fairness even though their explicit goal is to remove the designated artifact. These findings point to the existence of an inherent relationship between improving fairness and steering the saliency away from the protected attributes.

2. Related Work

Debiasing methods are an active area of research, usually in the context of tabular data, with a vast landscape of methods applied at various stages of model development [7]. The methods employed in our study represent approaches to debiasing in a post-hoc manner, that is, after a model is trained, within a binary classification setup. In our work, we consider three groups of methods. The first group consists of simple threshold optimizers, represented in our experiments by ThrOpt[9]. The second group focuses on approaches that optimize fairness with adversarial fine-tuning and is represented by ZhangAL[10] and SavaniAFT[11]. Finally, the third group focuses on concept-based interventions (artifact removal), exemplified by ClArC variants [12, 8], which operate directly on the model's internal representations utilizing Concept Activation Vectors (CAVs) through interventions in activation space.

Saliency maps are explainable AI methods that provide insights into model decision-making process by highlighting regions of input data that influence predictions. These techniques can generally be categorized into gradient-based [13, 14] and relevance-based methods [15]. Integrated Gradients (IG) [13] attributes predictions to input features by integrating gradients along a path from a baseline to the input, satisfying important axioms, including sensitivity and implementation invariance. Layer-wise Relevance Propagation (LRP) [15] employs a different approach based on a conservation principle, where relevance scores are propagated backward through the network layers while maintaining a constant sum. To improve the faithfulness of our study, we conducted experiments with multiple saliency map methods, each providing a different perspective on model predictions and associated limitations [16].

Quantitative evaluation of saliency maps is crucial for assessing whether models make decisions based on appropriate features rather than biased artifacts or protected attributes. Early approaches, such as the inside-outside ratio [17, 18], established a foundation by quantifying the relevance contained within a bounding box relative to the relevance outside it. This concept has been further developed as part of the Quantus toolbox [19], which provides a framework for evaluating explanations through various localization metrics. Motzkus et al. [20] advanced this approach by adapting the inside-outside metric to compute the ratio of positively attributed relevance within a binary class mask to the overall positive relevance, specifically focusing on the context of individual concepts.

3. Metrics for Saliency Maps

In this section, we present metrics designed to quantify the importance of protected attributes in the model's decision-making process. Our focus is specifically on localized features that can be roughly bounded by rectangular regions of interest (ROIs). These metrics evaluate whether an ROI plays an important role in the model's reasoning by analyzing saliency maps. In principle, they can be used with any standard saliency map generation method that suits the practical needs of an application.

To establish our framework, we define several key components. Image P is a 2D array with p_{ij} representing the intensity (or relevance) of the pixel (i, j). Within this image, we consider a 2D array (ROI) R such that $|R| < |P|$.

Rectangle Relevance Fraction (RRF) provides a direct measure of the ROI's importance in the context of the model's prediction by calculating what percentage of the total relevance falls within the region.

$$\mathbf{RRF} = \frac{\sum_{(i,j)\in R} p_{ij}}{\sum_{(i,j)\in P} p_{ij}} \tag{1}$$

It aids in understanding the relative ROI's contribution to the overall decision-making process of the model.

Average Difference in Region (ADR) provides a direct measure of how the saliency values within the ROI change after debiasing. It is defined as:

$$\mathbf{ADR} = \frac{1}{|R|} \sum_{(i,j) \in R} p^{\mathrm{v}}_{ij} - p^{\mathrm{d}}_{ij} \tag{2}$$

where p^{v}_{ij} and p^{d}_{ij} represent pixel intensities in Vanilla (corresponding to the base model) and debiased saliency maps, respectively. A positive ADR value indicates that Vanilla generally assigned higher importance to pixels within the ROI compared to the debiased model, suggesting a successful reduction in the model's reliance on these features.

Decreased Intensity Fraction (DIF) quantifies the proportion of pixels within the ROI that show reduced importance after debiasing. Specifically, it calculates the fraction of pixels where the debiased model shows lower saliency values compared to the Vanilla model. It is defined as:

$$\mathbf{DIF} = \frac{1}{|R|} \sum_{(i,j) \in R} \mathbb{1}_{\{p^{\mathrm{d}}_{ij} < p^{\mathrm{v}}_{ij}\}} \tag{3}$$

DIF provides insight into how widespread the changes are within the ROI, complementing the ADR's measurement of average change magnitude.

Rectangle Difference Distribution Testing (RDDT) metric assesses whether Vanilla assigns higher importance to pixels within the ROI compared to the debiased model. For each image, we compute the difference between the mean intensities of vanilla and debiased saliency maps within the ROI:

$$d = \mu_{\text{vanilla}} - \mu_{\text{debiased}} \tag{4}$$

where μ_{vanilla} and μ_{debiased} represent the mean pixel intensities within the ROI for the Vanilla and debiased models respectively. We then perform a one-sample t-test on these differences across with $H_0 : \mu_d = 0$ and $H_1 : \mu_d > 0$. The test returns 1 if $p < 0.01$, indicating statistically significant evidence that the Vanilla model assigns a higher importance to the ROI than the debiased model, and 0 otherwise.

4. Experiments

In the experiments below, we aim to explore the following two research questions. **RQ1:** Is there a bidirectional relationship between shifting the importance of pixels in the saliency map out of the ROI and optimizing fairness metrics? **RQ2:** Are debiasing methods capable of decreasing the saliency within ROI w.r.t. a standard end-to-end trained Vanilla model?

For our experiments, we utilize methods detailed in Sec. 2, implemented within the DetoxAI library [21]. We compute metrics and generate visualizations using LRP and Integrated Gradients. To ensure reproducibility, we have open-sourced a GitHub repository containing the relevant implementations [1].

The experimental procedure begins by fine-tuning a pre-trained ResNet-50 [22] on the target task's training set, yielding our Vanilla model. This fine-tuning uses a batch size of 128, the Adam optimizer, and a learning rate of $3 \cdot 10^{-4}$ for a single epoch. Subsequently, we apply the considered debiasing methods using a disjoint hold-out (debias) set. Finally, we evaluate the resulting models on a test set, calculating prediction performance, fairness, and our proposed metrics. Notably, both the training and debias datasets maintain the same *protected attribute-target* (PA-T) correlation, reflecting a common practical scenario where the split strategy is fixed. In contrast, the test set intentionally balances the PA-T correlation to systematically assess predictive performance (Accuracy) and fairness (EqualizedOdds) [23].

[1] https://github.com/DetoxAI/saliency-fairness-metrics

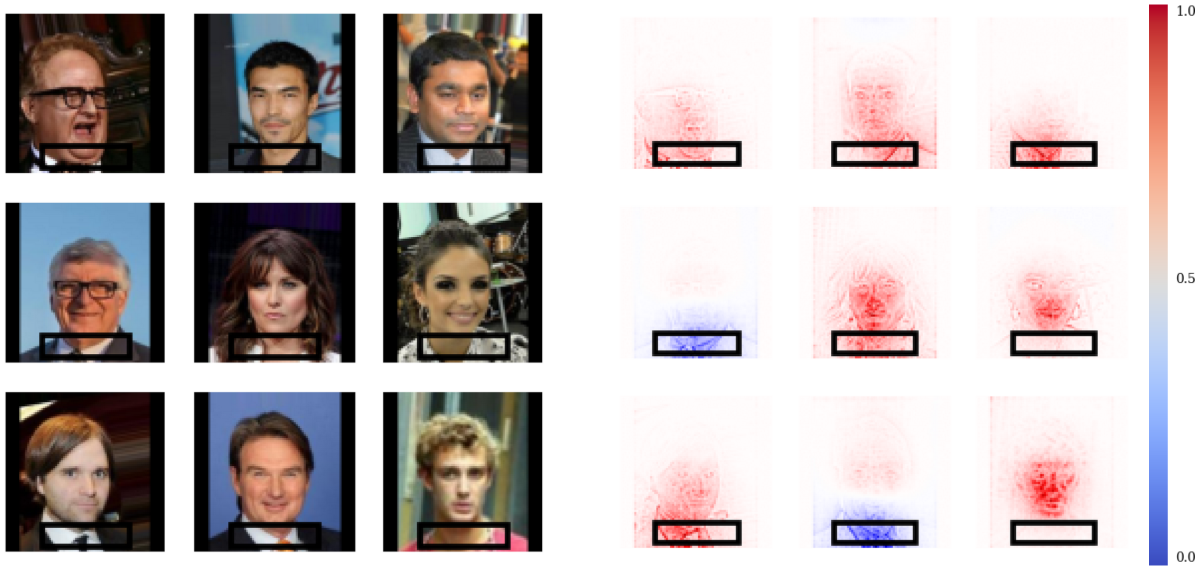

Figure 1: The left panel shows raw images, and the right panel, corresponding LRP saliency maps. In the saliency maps, red hues indicate positive relevance the true class, while blue hues indicate negative contributions.

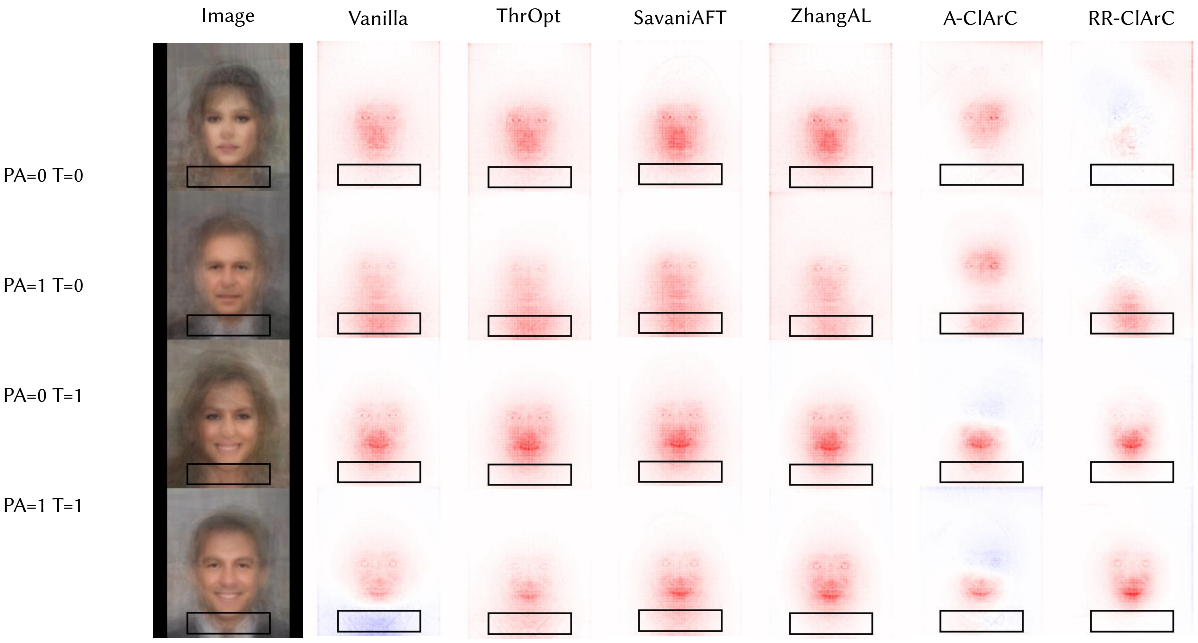

Figure 2: LRP saliency maps, averaged over a batch of 128 images and grouped by protected attribute (WearingNecktie) and target (Smiling) combinations. PA=1 indicates WearingNecktie, T=1 indicates Smiling.

4.1. Qualitative assessment

We perform a qualitative assessment of the debiasing by inspecting the relevancy maps before and after applying different debiasing methods. Fig. 2 presents LRP saliency maps for images aggregated by PA-T combinations, where the protected attribute is *WearingNecktie* and the target attribute is *Smiling*. The black rectangles highlight the ROI roughly corresponding to the necktie area (see Fig. 1).

Several key observations can be made from these visualizations. The Vanilla model (second column) shows considerable attention to the necktie region, particularly for the (PA=1, T=0) combination, indicating that the model has learned to associate the necktie area with its predictions. Interestingly, for the (PA=1, T=1) combination (bottom row), the necktie area shows strong negative relevance (blue), suggesting the model uses this feature to make negative predictions about smiling.

Simple threshold optimization (ThrOpt) does not substantially alter the saliency patterns compared

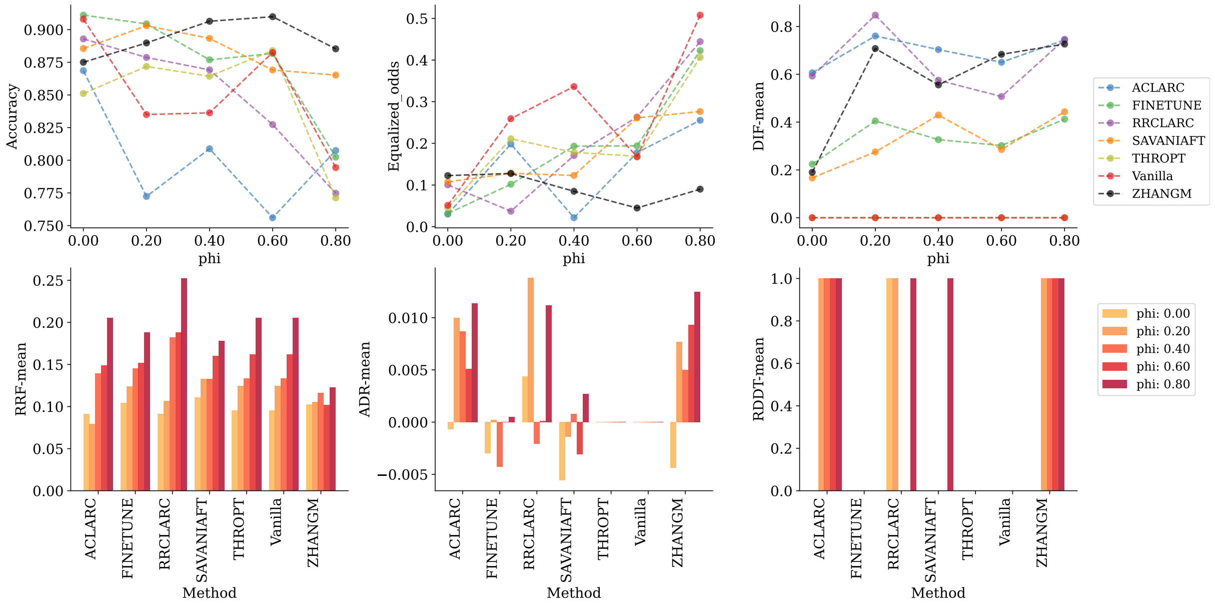

Figure 3: Quantitative metrics for WearingNecktie-Smiling PA-T classification task, measured on saliency maps generated with LRP. Metrics in the upper row are supposed to be minimized, while in the lower row, maximized.

to Vanilla, maintaining similar attention to the necktie area. This suggests that merely adjusting classification thresholds does not change the underlying reasoning of the model. Adversarial fine-tuning methods (SavaniAFT and ZhangAL) show modest reductions in the attention to the ROI but largely preserve the overall saliency patterns of the Vanilla model. The ClArC-based methods show the most noticeable shifts. A-ClArC reduces the saliency in the necktie region across all PA-T combinations, redirecting attention to facial features, relevant to the *Smiling* attribute. RR-ClArC shows the most visible improvements, excluding the second row, almost completely eliminating the relevance from ROI. These observations suggest that, while all debiasing methods may improve fairness metrics, they differ in how they alter the model's underlying decision-making process. Methods from the ClArC family most effectively redirect the model's attention away from the protected attribute region.

4.2. Quantitative experiments

While the CelebA dataset exhibits inherent attribute correlations, we artificially enforced specific PA-T correlations in our experimental framework to amplify the biases. This was done by rebalancing the dataset by undersampling attribute combinations to control their correlation with the target, as captured by Yule's correlation coefficient ϕ.

In this experiment, we considered two PA-T combinations: *WearingHat–Smiling* and *WearingNecktie–Smiling*, using saliency maps generated with LRP [15] and IntegratedGradients [13]. However, in the following, we only report the results for LRP and *WearingNecktie–Smiling* combination (in Fig. 3), while we move the rest to the Appendix, because the conclusions from all experiment variants are the same. In these plots, we report metrics from Sec. 3 along with EqualizedOdds calculated as: $EqualizedOdds = \max\left(|TPR_{PA=1} - TPR_{PA=0}|, |FPR_{PA=1} - FPR_{PA=0}|\right)$, where TPR and FPR stand for true and false positive rates respectively, and $PA = 0$, $PA = 1$ protected attribute value assignments.

First, it is clear that as ϕ increases, all methods achieve a higher EqualizedOdds value, which indicates more bias in their predictions. The best performing method for this metric is ZhangAL, which optimizes it directly internally. However, most methods decrease the EqualizedOdds score w.r.t. Vanilla's, confirming that they are effective.

ThrOpt, a post-hoc classification threshold optimization method, does not shift the relevancy in or out of the ROI. Its bars are empty for ADR and RDDT and equal to Vanilla on DIF and RRF, indicating that no change in the saliency maps was recorded. This is expected since ThrOptdoes not intervene into the reasoning process. This method decreases in accuracy as the correlation grows larger.

SavaniAFT and ZhangAL both perform well across most metrics. ZhangAL scores remarkably well in saliency map-based metrics. It lowers all but one metric value in the first row of the plot, showing that it moves the saliency out of ROI. As correlation grows, accuracy of the model also grows. In addition, it also scores visibly well on the metrics in the lower row, which measure the improvement over the Vanilla model within the ROI. This provides evidence that optimizing with a fairness-oriented objective as a fine-tuning step can significantly shift the model's reasoning process.

RR-ClArC and A-ClArC do not optimize any fairness objective. Yet, they effectively debias the model (as captured by EqualizedOdds) and significantly shift model relevancy within the ROI. Both score high at DIF and ADR, and often appear on RDDT (the more bars the better). Regarding attention outside the ROI, they tend to lower the RRF with respect to Vanilla, which suggests that more attention is given to features outside the ROI, - the desired outcome. Both methods cause decrease in accuracy.

5. Conclusion

Experiments show that effective debiasing methods decrease saliency within the ROI compared to the Vanilla model, which positively answers RQ2. Both qualitative and quantitative analyses reveal that while threshold optimization (ThrOpt) produces no changes in saliency maps, fine-tuning-based approaches yield significant improvements. Notably, ZhangAL and SavaniAFT and ClArC-based methods (A-ClArC and RR-ClArC) redirect the attention away from protected features towards task-relevant features such as facial expressions for smile detection. For the latter, the saliency redirection is stronger while achieving competitive EqualizedOdds, despite not directly optimizing any fairness objective.

These findings provide evidence for a bidirectional relationship between shifting pixel importance in saliency maps away from regions of interest and optimizing fairness metrics, validating the premise of RQ1. They confirm that methods that effectively redirect model attention away from protected attributes tend to score better on EqualizedOdds, and vice versa.

We believe that this research provides useful evidence for further work on fairness methods, which could adapt concept removal methods directly in the field of fair machine learning.

Acknowledgments

Authors Ignacy Stępka, Lukasz Sztukiewicz, Michal Wiliński were supported by a Ministry of Science and Higher Education grant No. MNiSW/2025/DPI/56 under the FERS program, co-financed by the European Union. Jerzy Stefanowski was supported by a National Science Centre grant (No. 2022/47/D/ST6/01770).

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT, LeChat, and Grammarly for: Grammar and spelling checking, paraphrasing and rewording. After using these tools/services, the authors reviewed and edited the content as needed and assume full responsibility for the content of the publication.

References

[1] M. Buyl, T. De Bie, Inherent limitations of ai fairness, Commun. ACM 67 (2024) 48–55.

[2] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: automated decisions and the gdpr, Harvard Journal of Law and Technology 31 (2018).

[3] H. Tian, T. Zhu, W. Liu, W. Zhou, Image fairness in deep learning: Problems, models, and challenges, Neural Computing and Applications 34 (2022) 12875–12893.

[4] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, A. T. Kalai, Man is to computer programmer as woman is to homemaker? debiasing word embeddings, Advances in Neural Information Processing Systems 29 (2016).

[5] Z. Liu, P. Luo, X. Wang, X. Tang, Deep learning face attributes in the wild, in: Proceedings of International Conference on Computer Vision (ICCV), 2015, pp. 3730–3738.

[6] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM Computing Surveys 54 (2021).

[7] S. Caton, C. Haas, Fairness in machine learning: A survey, ACM Comput. Surv. 56 (2024).

[8] M. Dreyer, F. Pahde, C. J. Anders, W. Samek, S. Lapuschkin, From hope to safety: Unlearning biases of deep models via gradient penalization in latent space, Proceedings of the AAAI Conference on Artificial Intelligence 38 (2024) 21046–21054.

[9] M. Hardt, E. Price, N. Srebro, Equality of opportunity in supervised learning, in: Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16, Curran Associates Inc., Red Hook, NY, USA, 2016, p. 3323–3331.

[10] B. H. Zhang, B. Lemoine, M. Mitchell, Mitigating unwanted biases with adversarial learning, in: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 335–340.

[11] Y. Savani, C. White, N. S. Govindarajulu, Intra-processing methods for debiasing neural networks, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 2798–2810.

[12] C. J. Anders, L. Weber, D. Neumann, W. Samek, K.-R. Müller, S. Lapuschkin, Finding and removing clever hans: Using explanation methods to debug and improve deep models, Information Fusion 77 (2022) 261–295.

[13] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: International Conference on Machine Learning, PMLR, 2017, pp. 3319–3328.

[14] C. Molnar, Interpretable Machine Learning, 2 ed., 2022.

[15] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PLOS ONE 10 (2015) 1–46.

[16] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence 1 (2019) 206–215.

[17] M. Kohlbrenner, A. Bauer, S. Nakajima, A. Binder, W. Samek, S. Lapuschkin, Towards best practice in explaining neural network decisions with lrp, in: 2020 International Joint Conference on Neural Networks (IJCNN), IEEE, 2020, pp. 1–7.

[18] S. Bach, A. Binder, G. Montavon, K.-R. Müller, W. Samek, Analyzing classifiers: Fisher vectors and deep neural networks, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015) 2912–2920.

[19] A. Hedström, L. Weber, D. Krakowczyk, D. Bareeva, F. Motzkus, W. Samek, S. Lapuschkin, M. M.-C. Höhne, Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond, Journal of Machine Learning Research 24 (2023) 1–11.

[20] F. Motzkus, G. Mikriukov, C. Hellert, U. Schmid, Locally testing model detections for semantic global concepts, in: World Conference on Explainable Artificial Intelligence, Springer, 2024, pp. 137–159.

[21] I. Stępka, L. Sztukiewicz, M. Wiliński, J. Stefanowski, DetoxAI: a Python toolkit for debiasing deep learning models in computer vision, 2025. URL: https://arxiv.org/abs/2505.05492. arXiv:2505.05492.

[22] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[23] D. Brzezinski, J. Stachowiak, J. Stefanowski, I. Szczech, R. Susmaga, S. Aksenyuk, U. Ivashka, O. Yasinskyi, Properties of fairness measures in the context of varying class imbalance and protected group ratios, ACM Transactions on Knowledge Discovery from Data 18 (2024) 1–18.

A. Extra visualizations

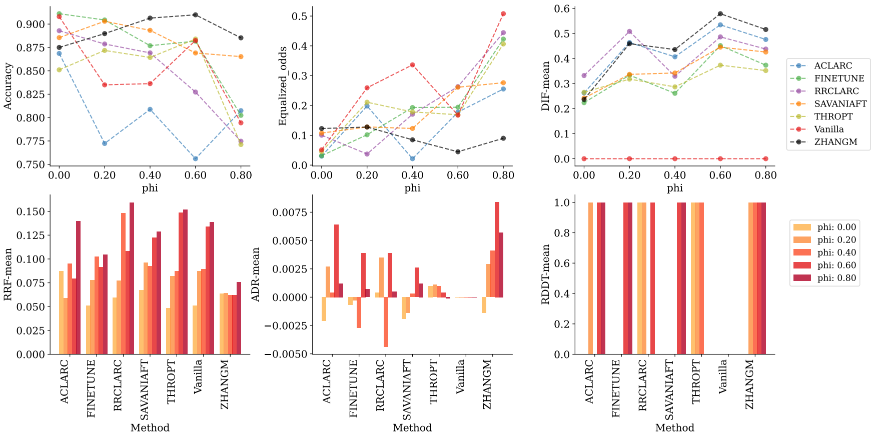

Figure 4: Metric values for the IG attributions and *WearingNecktie* protected attribute.

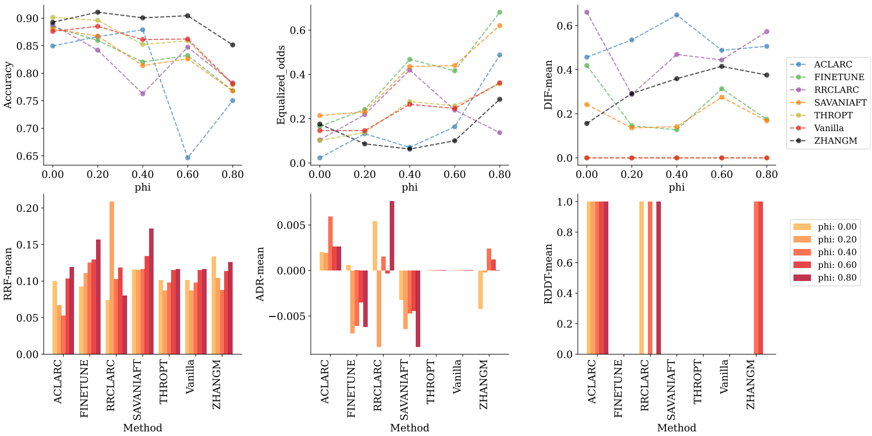

Figure 5: Metric values for the LRP attributions and *WearingHat* protected attribute

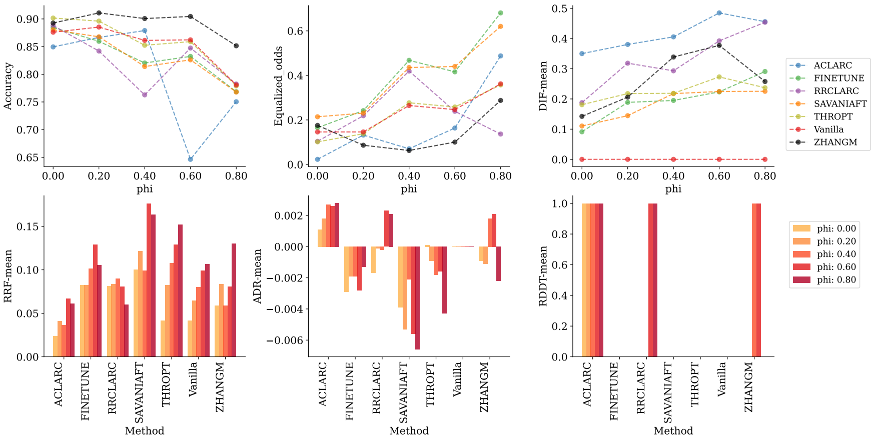

Figure 6: Metric values for the IG attributions and *WearingHat* protected attribute