Towards Fairness for the Right Reasons: Using Saliency Maps to Evaluate Bias Removal in Neural Networks

Lukasz Sztukiewicz
research@lukaszsztukiewicz.com
Poznan University of Technology
Poznan, Poland

Ignacy Stępka
Poznan University of Technology
Poznan, Poland

Michal Wiliński
Poznan University of Technology
Poznan, Poland

Jerzy Stefanowski
Poznan University of Technology
Poznan, Poland

Abstract

The growing adoption of machine learning systems has intensified concerns about fairness and bias, highlighting the importance of mitigating harmful biases in AI development. In this paper, we explore the relationship between fairness improvement and the removal of harmful biases in neural networks applied to 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 fairness. This reveals a bidirectional link between ensuring fairness and removing artifacts corresponding to protected attributes, suggesting that models can be made not only fair but **fair for the right reasons** - contributing to the development of more ethical and trustworthy AI systems.

CCS Concepts

• Computing methodologies → Neural networks; **Computer vision**; **• General and reference → Metrics**; **• Human-centered computing → Heat maps**; **Visualization toolkits**; **• Social and professional topics → Men**; **Women**.

Keywords

Deep learning, Fairness, Debiasing, Saliency maps, Artifact Removal

1 Introduction

The widespread adoption of machine learning (ML) systems across a range of domains has not only brought impressive predictive capabilities but also raised critical concerns about fairness and bias. Recent regulations, such as those discussed in [24], underscore the importance of this problem. While some degree of bias is inherent in machine learning models, harmful biases that affect sensitive areas of human life pose ethical issues [7].

Neural networks, especially in computer vision applications, present unique challenges for fairness assessment and bias mitigation [23]. 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 becomes particularly problematic when dealing with protected attributes (such as gender or race), which are high-level features (concepts) that are inaccessible to models operating only on pixel data without explicit featurization. Neural networks are known to develop internal representations that encode not only useful high-level features but also harmful biases [5]. This phenomenon mirrors a well-known problem in tabular data, known as the 'fairness through unawareness' [9], where models can still exhibit biases by inferring protected attributes through proxy or composite features. For example, in the CelebA dataset [15], wearing a necktie is highly correlated with the male gender. It means that the 'wearing necktie feature' 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 [16]. While existing debiasing methods generally improve fairness metrics, they often fail to explicitly address harmful biases encoded in models' internal representations. We examine the relationship between successful fairness improvement and removal of harmful biases from these representations. To address the evaluation gap, 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 based on fine-tuning redirect the model's focus away from protected attributes, thus altering the internal model representations. Building on these insights, we show that techniques originally developed for *artifact removal*, such as the family of ClArC methods [10], can be repurposed for fairness improvement, even though they do not explicitly optimize any fairness metric. Finally, we advocate that fairness should go hand in hand with concept removal to make models **fair for the right reasons**. Our key contributions include:

(1) A set of new XAI-based metrics that quantify shifts in the model's decision-making process based on saliency maps.
(2) Empirical evidence that successful debiasing methods systematically redirect model attention away from protected attributes
(3) Demonstration that artifact removal methods, when applied to localized protected attributes, score excellent on fairness metrics and naturally align with the principle of being *fair for the right reasons*

KDD-UC'25, Toronto, ON, Canada
2025.

2 Related Work

Debiasing methods. are active area of research, usually in the context of tabular data, with a vast landscape of methods applied at various stages of model development [8]. 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[11]. These methods adjust classification thresholds to optimize fairness metrics predefined by the user. The second group focuses on approaches that optimize fairness with adversarial fine-tuning. Zhang's adversarial learning (ZhangAL) [25] formulates debiasing as a minimax optimization game between a predictor and an adversary that attempts to predict protected attributes from model outputs. Savani's adversarial fine-tuning (SavaniAFT) [20] leverages first-order optimization with a critic model that learns to predict bias in minibatches, serving as a differentiable proxy for typically non-differentiable fairness metrics. Finally, the third group focuses on concept-based interventions (artifact removal), exemplified by ClArC variants [2, 10], which operate directly on the model's internal representations through interventions in activation space. These methods use Concept Activation Vectors (CAVs) to identify and manipulate directions in activation space associated with protected attributes. ClArC variants include augmentation during fine-tuning (A-ClArC), or gradient-space regularization (RR-ClArC).

Saliency maps. are explainable AI methods that provide insights into model decision-making process by highlighting regions of input data that influence predictions. Although there are many visualization techniques, they can generally be categorized into gradient-based [17, 22] and relevance-based methods [4]. Integrated Gradients [22] 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) [4] 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 [19].

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 [3, 14], established a foundation by quantifying the relevance contained within a bounding box relative to the relevance outside it. This concept has been further developed and implemented as part of the Quantus toolbox [13], which provides a framework for evaluating explanations through various localization metrics. Motzkus et al. [18] 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. Despite these advances, there remains a gap in methods specifically designed to quantify the importance of protected attributes in saliency maps. In this work, we address this limitation by introducing four complementary metrics that evaluate the saliency of protected attributes by measuring the concentration and distribution of relevance within defined regions of interest.

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 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. The proposed metrics can be used with any standard saliency map generation method. We divide them into two groups investigating properties of saliency maps (Sec. 3.1) and the improvement of the debiasing method over Vanilla model (Sec. 3.2).

To establish our framework, we define several key components. The dimensions of an image are given by width N and height M, with p_{ij} representing the intensity (or relevance) of the pixel (i, j). Within this image, we consider an ROI defined by width W and height L, where $W \leq N$ and $L \leq M$. We denote the set of pixels within this ROI as R, such that $|R| = W \cdot L$.

3.1 Saliency Map Evaluation Metric

Here, we introduce metric that evaluate the relationship between ROI and the entire image by analyzing pixel intensity scores in saliency maps.

3.1.1 Rectangle Relevance Fraction (RRF). For ROI is defined as:

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

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. RRF helps us understand the relative contribution of the selected region to the overall decision-making process of the model.

3.2 Improvement over Vanilla Model

To evaluate the effectiveness of debiasing methods, we introduce three metrics that compare the saliency maps of a debiased model with those of the original Vanilla model. These metrics specifically focus on changes within the ROI, helping us quantify how successful the debiasing process has been in reducing the model's reliance on protected attributes.

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

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

where p_{ij}^{v} and p_{ij}^{d} represent pixel intensities in the Vanilla and debiased saliency maps, respectively. A positive ADR value indicates that the Vanilla model 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.

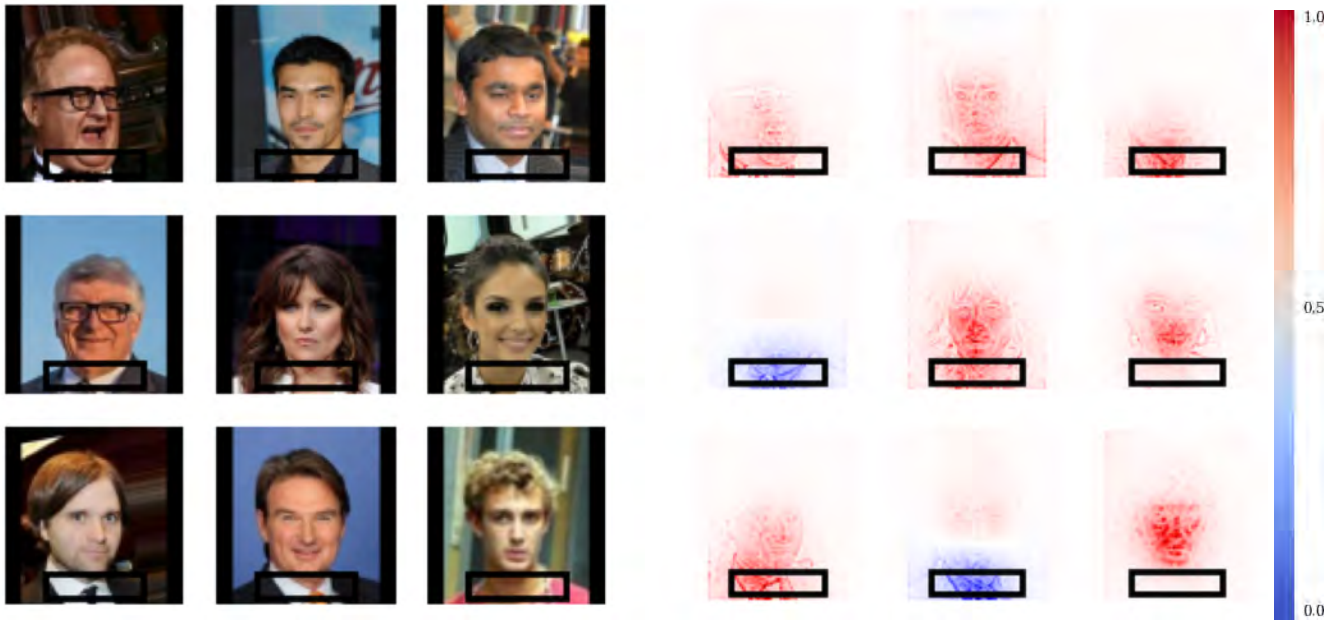

Figure 1: On the left raw images, on the right LRP saliency maps. Red color, above 0.5 indicates positive relevance for the correct ground-truth class, while blue color below 0.5, negative contribution towards that class.

3.2.2 *Decreased Intensity Fraction (DIF).*

DIF quantifies the proportion of pixels within the ROI that show reduced importance after debiasing:

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

This metric calculates the fraction of pixels where the debiased model shows lower saliency values compared to the Vanilla model. The DIF provides insight into how widespread the changes are within the ROI, complementing the ADR's measurement of average change magnitude.

3.2.3 *Rectangle Difference Distribution Testing (RDDT).*

RDDT metric assesses whether the Vanilla model assigns a 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?

The experimental procedure begins by fine-tuning a pre-trained ResNet-50 [12] 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) [6].

In our experiments, we use methods described in Sec. 2, implemented in the DetoxAI library [21]. We compute both quantitative metrics and qualitative summaries for saliency maps generated by LRP and Integrated Gradients, which are also implemented in DetoxAI using the Zennit library [1]. To support reproducibility, we publicly release our code[1], which includes the metrics introduced in this paper as well as scripts used to run the experiments.

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

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

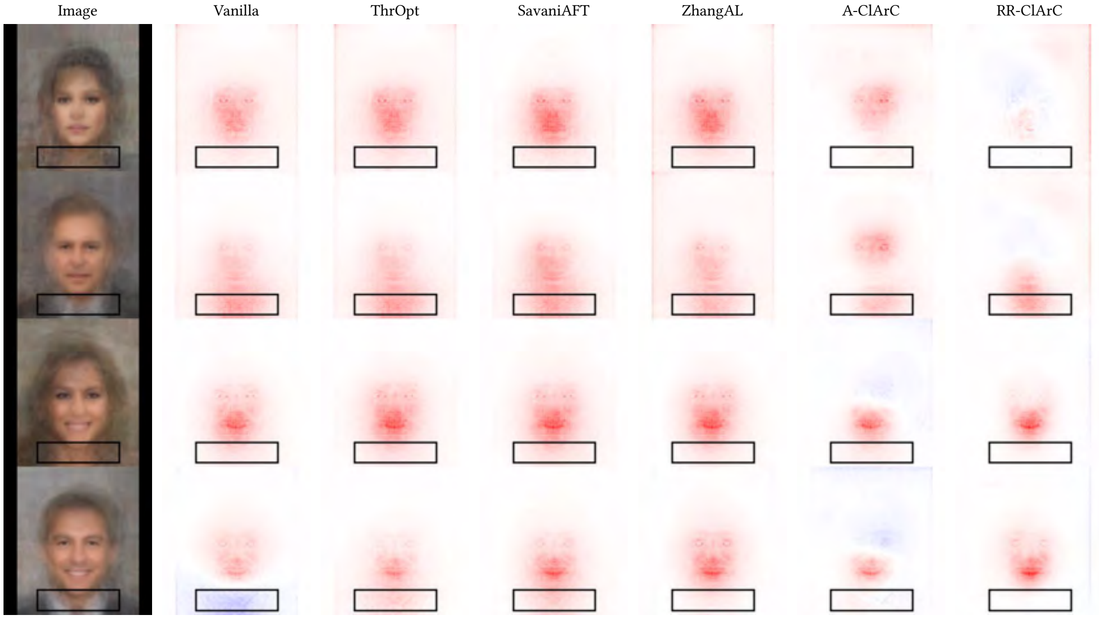

Figure 2: LRP saliency maps for images aggregated by PA-T combinations (starting from top: PA=0 T=0, PA=1 T=0, PA=0 T=1, PA=1 T=1)for a batch size of 128. The protected attribute is *WearingNecktie* and the target is *Smiling*.

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 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.

Metric	Type	Scope	Reported Value	Direction
RRF	Relationship	Per-image	Mean over dataset	min
ADR	Improvement	Per-image	Mean over dataset	max
DIF		Per-image	Mean over dataset	
RDDT		Dataset	Statistical test	

Table 1: Summary of introduced sailency map metrics.

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. Desired attribute correlations were achieved by manipulating the distributions of binary attributes. We measure correlation using Yule's ϕ:

$$\phi = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{(n_{11}+n_{10})(n_{11}+n_{01})(n_{00}+n_{10})(n_{00}+n_{01})}} \tag{5}$$

where n_{ij} represents the percentage of samples with a particular PA-T combination.

In this experiment, we considered two unique PA-T combinations: *WearingHat–Smiling* and *WearingNecktie–Smiling*, using saliency maps generated with LRP [4] and IntegratedGradients [22]. However, for brevity, we only report the results for *WearingNecktie–Smiling* combination (in Fig. 3 and 4) because the conclusions from all experiment variants are the same. In these plots, we report metrics from Sec. 3, summarized in Tab. 1, along with EqualizedOdds

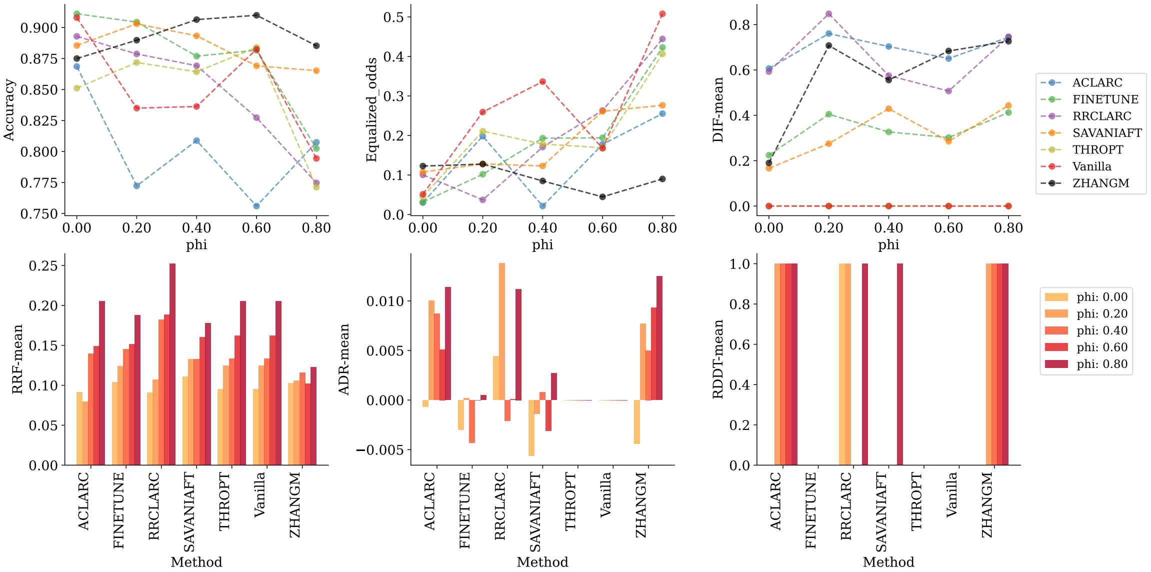

Figure 3: Quantitative metrics for WearingNecktie-Smiling PA-T combination, measured on saliency maps generated with LRP.

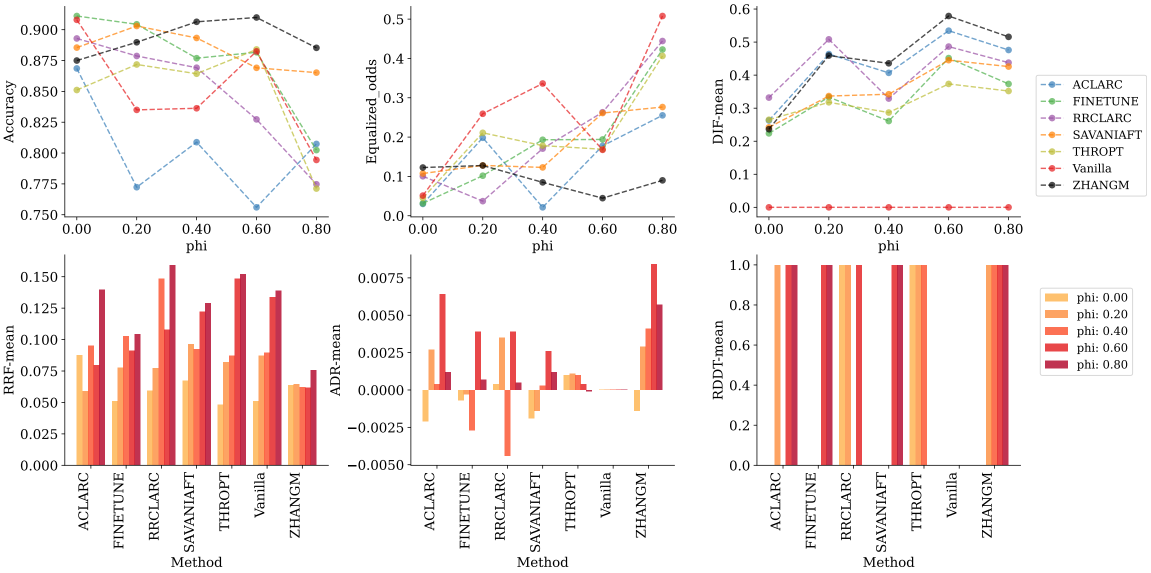

Figure 4: Quantitative metrics for WearingNecktie-Smiling PA-T combination, measured on saliency maps generated with IG.

metric calculated with a following formula:

$$EqualizedOdds = \max\left(|TPR_{PA=1} - TPR_{PA=0}|, |FPR_{PA=1} - FPR_{PA=0}|\right) \quad (6)$$

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 on this metric is ZhangAL, which directly optimizes it internally. However, most methods decrease the EqualizedOdds score w.r.t. Vanilla's, which shows 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 no changes are made to the reasoning process with this method.

SavaniAFT and ZhangAL both perform well across most metrics. ZhangAL scores remarkably well across saliency map-based metrics. It consistently lowers the metric values in the first row of the plot, showing that it moves the saliency out of ROI. Moreover, 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.

In particular, RR-ClArC and A-ClArC methods 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 RRF with respect to Vanilla, which suggests that more attention is given to features outside the ROI, what is the desired outcome.

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. This confirms our hypothesis that successful concept unlearning should shift attention from protected attributes and opens an avenue to employ such methods for fairness, even though they do not explicitly optimize the fairness objective.

We believe this work encourages the development of fairness methods that go beyond metric optimization to ensure that models are not only fair but **fair for the right reasons** — i.e., their decisions are not driven by biased or protected attributes encoded in internal representations.

Acknowledgement. The work was supported by a Ministry of Science and Higher Education grant (No. MNiSW/2025/DPI/56) as part of the FERS programme, co-financed by the European Union.

References

[1] Anders, C.J., Neumann, D., Samek, W., Müller, K.R., Lapuschkin, S.: Software for dataset-wide xai: From local explanations to global insights with Zennit, CoRelAy, and ViRelAy (2021)

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

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

[4] Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLOS ONE **10**(7), 1–46 (07 2015)

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

[6] Brzezinski, D., Stachowiak, J., Stefanowski, J., Szczech, I., Susmaga, R., Aksenyuk, S., Ivashka, U., Yasinskyi, O.: Properties of fairness measures in the context of varying class imbalance and protected group ratios. ACM Transactions on Knowledge Discovery from Data **18**(7), 1–18 (2024)

[7] Buyl, M., De Bie, T.: Inherent limitations of ai fairness. Commun. ACM **67**(2), 48–55 (Jan 2024)

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

[9] Datta, A., Tschantz, M., Datta, A.: Automated experiments on ad privacy settings. Proceedings on Privacy Enhancing Technologies **1** (04 2015)

[10] Dreyer, M., Pahde, F., Anders, C.J., Samek, W., Lapuschkin, S.: From hope to safety: Unlearning biases of deep models via gradient penalization in latent space. Proceedings of the AAAI Conference on Artificial Intelligence **38**(19), 21046–21054 (Mar 2024)

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

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

[13] Hedström, A., Weber, L., Krakowczyk, D., Bareeva, D., Motzkus, F., Samek, W., Lapuschkin, S., Höhne, M.M.C.: Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond. Journal of Machine Learning Research **24**(34), 1–11 (2023)

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

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

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

[17] Molnar, C.: Interpretable Machine Learning. 2 edn. (2022)

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

[19] Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence **1**(5), 206–215 (2019)

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

[21] Stępka, I., Sztukiewicz, L., Wiliński, M., Stefanowski, J.: DetoxAI: a Python package for debiasing neural networks (2025), https://github.com/DetoxAI/detoxai

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

[23] Tian, H., Zhu, T., Liu, W., Zhou, W.: Image fairness in deep learning: Problems, models, and challenges. Neural Computing and Applications **34**(15), 12875–12893 (Aug 2022)

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

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