Reveal to Revise: An Explainable AI Life Cycle for Iterative Bias Correction of Deep Models

Frederik Pahde[1,*] Maximilian Dreyer[1,*] Wojciech Samek[1,2,3,†]
Sebastian Lapuschkin[1,†]

[1]Fraunhofer Heinrich-Hertz-Institute, 10587 Berlin, Germany
[2]Technische Universität Berlin, 10587 Berlin, Germany
[3]BIFOLD – Berlin Institute for the Foundations of Learning and Data, 10587 Berlin, Germany
[†] corresponding authors: {wojciech.samek,sebastian.lapuschkin}@hhi.fraunhofer.de
[*] contributed equally

Abstract

State-of-the-art machine learning models often learn spurious correlations embedded in the training data. This poses risks when deploying these models for high-stake decision-making, such as in medical applications like skin cancer detection. To tackle this problem, we propose Reveal to Revise (R2R), a framework entailing the entire eXplainable Artificial Intelligence (XAI) life cycle, enabling practitioners to iteratively identify, mitigate, and (re-)evaluate spurious model behavior with a minimal amount of human interaction. In the first step (1), R2R *reveals* model weaknesses by finding outliers in attributions or through inspection of latent concepts learned by the model. Secondly (2), the responsible artifacts are *detected* and spatially *localized* in the input data, which is then leveraged to (3) *revise* the model behavior. Concretely, we apply the methods of RRR, CDEP and ClArC for model correction, and (4) (re-)evaluate the model's performance and remaining sensitivity towards the artifact. Using two medical benchmark datasets for Melanoma detection and bone age estimation, we apply our R2R framework to VGG, ResNet and EfficientNet architectures and thereby reveal and correct real dataset-intrinsic artifacts, as well as synthetic variants in a controlled setting. Completing the XAI life cycle, we demonstrate multiple R2R iterations to mitigate different biases. Code is available on https://github.com/maxdreyer/Reveal2Revise.

1 Introduction

Deep Neural Networks (DNNs) have successfully been applied in research and industry for a multitude of complex tasks. This includes various medical applications for which DNNs have even shown to be superior to medical experts, such as with Melanoma detection [4]. However, the reasoning of these highly complex and non-linear models is generally not transparent [16, 17], and as such, their decisions may be biased towards unintended or undesired features [20, 12, 2]. Particularly in high-stake decision processes, such as medical applications, unreliable or poorly understood model behavior may pose severe security risks.

The field of XAI brings light into the black boxes of DNNs and provides a better understanding of their decision processes. As such, local XAI methods reveal (input)

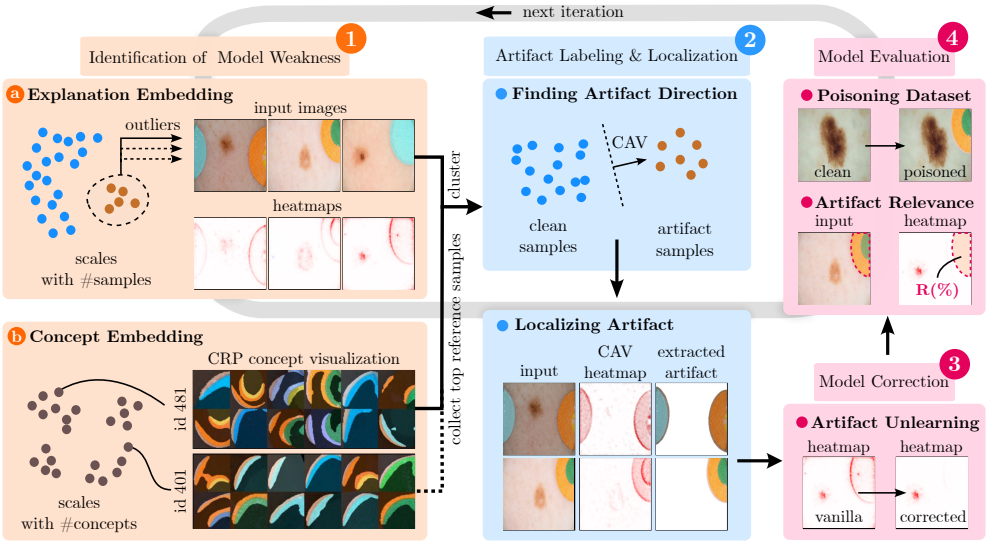

Figure 1: Our R2R life cycle for *revealing* and *revising* spurious behavior of any pre-trained DNN. Firstly, we identify model weaknesses by finding either outliers in explanations using SpRAy (1a) or suspicious concepts using CRP concept visualizations (1b). Secondly (2), SpRAy clusters or collecting the top reference samples allows us to label artifactual samples and to compute an artifact CAV, which we use to model and localize the artifact in latent and input space, respectively. At this point, the artifact localization can be leveraged for (3) model correction, and (4) to evaluate the model's performance on a poisoned test set and measure its remaining attention on the artifact.

features that are most relevant to a model, which, for image data, can be presented as heatmaps. In contrast, global XAI methods (e.g, [10, 12]) reveal general prediction strategies employed or features encoded by a model, which is necessary for the identification and understanding of systematic (mis-)behavior. Acting on the insights from explanations, various methods have been introduced to correct for undesired model behavior [24]. While multiple approaches exist for either *revealing* or *revising* model biases, only few combine both steps, to be applicable as a framework. Such frameworks, however, either rely heavily on human feedback [22, 18], are limited to specific bias types [2], or require labor-intensive annotations for both model evaluation and correction [18, 11].

To that end, we propose Reveal to Revise (R2R), an iterative XAI life cycle requiring low amounts of human interaction that consists of four phases, illustrated in Fig. 1. Specifically, R2R allows to first (1) identify spurious model behavior and secondly, to (2) label and localize artifacts in an automated fashion. The generated annotations are then leveraged to (3) correct and (4) (re-)evaluate the model, followed by a repetition of the entire life cycle if required. For *revealing* model bias, we propose two orthogonal XAI approaches: While Spectral Relevance Analysis (SpRAy) [12] automatically finds outliers in model explanations (potentially caused by the use of spurious features), Concept Relevance Propagation (CRP)[1] precisely communicates the globally learned concepts of a DNN. For model *revision*, we apply and compare the methods of Class Artifact Compensation (ClArC) [2], Contextual Decomposition Explanation Penalization (CDEP) [14] and Right for the Right Reason (RRR) [15], penalizing attention on artifacts via ground truth masks automatically generated in

step (2). The artifact masks are further used for evaluation on a poisoned test set and to measure the remaining attention on the bias. We demonstrate the applicability and high automation of R2R on two medical tasks, including Melanoma detection and bone age estimation, using the VGG-16, ResNet-18 and EfficientNet-B0 DNN architectures. In our experiments, we correct model behavior w.r.t. dataset-intrinsic, as well as synthetic artifacts in a controlled setting. Lastly, we showcase the R2R life cycle through multiple iterations, unveiling and unlearning different biases.

2 Related Work

The majority of related works introduce methods to either identify spurious behavior [12, 1], or to align the model behavior with pre-defined priors [15, 14], with only few combining both, such as the eXplanatory Interactive Learning (XIL) framework [22] or the approach introduced by Anders et al. [2]. The former is based on presenting individual local explanations to a human, who, if necessary, provides feedback used for model correction [22, 18]. However, studying individual predictions is slow and labor-extensive, limiting its practicability. In contrast, the authors of [2] use SpRAy [12] for the detection of spurious model behavior and labeling of artifactual samples. In addition to SpRAy, we suggest to study latent features of the model via CRP concept visualizations [1] as a tool for more fine-grained model inspection, catching systematic misbehavior which would not be visible through SpRAy clusters.

Most model correction methods require dense annotations, such as labels for artifactual samples or artifact localization masks, which are either crafted heuristically or by hand [14, 11]. In our R2R framework, we automate the annotation by following [2] for data labeling through SpRAy outlier clusters, or by collecting the most representative samples of bias concepts according to CRP. The spatial artifact localization is further automated by computing artifact heatmaps as outlined in Section 3.1, thereby considerably easing the step from bias identification to correction.

Existing works for model correction measure the performance on the original or clean test set, with corrected models often showing an improved generalization [11, 14]. A more targeted approach for measuring the artifact's influence is the evaluation on poisoned data [18], for which R2R is well suited by using its localization scheme to first extract artifacts and to then poison clean test samples. By precisely localizing artifacts, R2R further allows to measure the model's attention on an artifact through attribution heatmaps.

3 Reveal to Revise Framework

Our *Reveal to Revise* (R2R) framework comprises the entire XAI life cycle, including methods for (1) the identification of model bias, (2) artifact labeling and localization, (3) the correction of detected misbehavior, and (4) the evaluation of the improved model. To that end, we now describe the methods used for R2R.

3.1 Data Artifact Identification and Localization

The identification of spurious data artifacts using CRP concept visualizations or SpRAy clusters is firstly described, followed by our artifact localization approach.

3.1.1 CRP Concept Visualizations

CRP [1] combines global concept visualization techniques with local feature attribution methods. This provides an understanding of the relevance of latent concepts for a prediction and their localization in the input. In this work, we use Layer-wise Relevance Propagation (LRP) [3] for feature attribution under CRP and for heatmaps in general, however, other local XAI methods can be used as well. Jointly with Relevance Maximization [1], CRP is well suited for the identification of spurious concepts by precisely narrowing down the input parts that have been most relevant for model inference, as shown in Fig. 1 (*bottom left*) for band-aid concepts, where irrelevant background is overlaid with black semi-transparent color. The collection of top-ranked reference samples for spurious concepts allows us to label artifactual data.

3.1.2 Explanation Outliers Through SpRAy

Alternatively, SpRAy [12] is a strategy to find outliers in local explanations, which are likely to stem from spurious model behavior, such as the use of a Clever Hans features, *i.e.*, features correlating with a certain class that are unrelated to the actual task. Following [12, 2], we apply SpRAy by clustering latent attributions computed through LRP. The SpRAy clusters then naturally allow us to label data containing the bias.

3.1.3 Artifact Localization

We automate artifact localization by using a Concept Activation Vector (CAV) $\mathbf{h}_l$ to model the artifact in latent space of a layer l, representing the direction from artifactual to non-artifactual samples obtained from a linear classifier. The artifact localization is given by a modified backward pass with LRP for an artifact sample $\mathbf{x}$, where we initialize the relevances $\mathbf{R}_l(\mathbf{x})$ at layer l as

$$\mathbf{R}_l(\mathbf{x}) = \mathbf{a}_l(\mathbf{x}) \circ \mathbf{h}_l \tag{1}$$

with activations $\mathbf{a}_l$ and element-wise multiplication operator $\circ$. This is equivalent to explaining the output from the linear classifier given as $\mathbf{a}_l(\mathbf{x}) \cdot \mathbf{h}_l$. The resulting CAV heatmap can be further processed into a binary mask to crop out the artifact from any corrupted sample, as illustrated in Fig. 1 (*bottom center*).

3.2 Methods for Model Correction

In the following, we present the methods used for mitigating model biases.

3.2.1 ClArC for Latent Space Correction

Methods from the ClArC framework correct model (mis-)behavior w.r.t. an artifact by modeling its direction $\mathbf{h}$ in latent space using CAVs [10]. The framework consists of two methods, namely Augmentive ClArC (a-ClArC) and Projective ClArC (p-ClArC). While a-ClArC adds $\mathbf{h}_l$ to the activations $\mathbf{a}_l$ of layer l for all samples in a fine-tuning phase, hence teaching the model to be invariant towards that direction, p-ClArC suppresses the artifact direction during the test phase and does not require any fine-tuning. More precisely, the perturbed activations $\mathbf{a}'_l$ are given by

$$\mathbf{a}'_l(\mathbf{x}) = \mathbf{a}_l(\mathbf{x}) + \gamma(\mathbf{x})\mathbf{h}_l \tag{2}$$

with perturbation strength $\gamma(\mathbf{x})$ dependent on input $\mathbf{x}$. Parameter $\gamma(\mathbf{x})$ is chosen such that the activation in direction of the CAV is as high as the average value over non-artifactual or artifactual samples for p-ClArC or a-ClArC, respectively.

3.2.2 RRR and CDEP for Correction through Prior Knowledge

Model correction using RRR [15] or CDEP [14] is based on an additional λ-weighted loss term (besides the cross-entropy loss $\mathcal{L}_{\text{CE}}$) for neural network training that aligns the use of features by the model f_θ, described by an explanation $\exp_\theta$, to a defined prior explanation $\exp_{\text{prior}}$. The authors of RRR propose to penalize the model's attention on unfavorable artifacts using the input gradient w.r.t. the cross-entropy loss, leading to

$$\mathcal{L}_{\text{RRR}}\left(\exp_\theta(\mathbf{x}), \exp_{\text{prior}}(\mathbf{x})\right) = \left\|\nabla_{\mathbf{x}}\mathcal{L}_{\text{CE}}\left(f_\theta(\mathbf{x}), y_{\text{true}}\right) \circ \mathbf{M}_{\text{prior}}(\mathbf{x})\right\|_2{}^2 \tag{3}$$

with a binary mask $\mathbf{M}_{\text{prior}}(\mathbf{x})$ localizing an artifact and class label y_{true}. We further adapt the RRR loss to increase stability in regard to the high variance of DNN gradients by using the cosine similarity (instead of L2-norm) and compute the gradient w.r.t. the predicted logit p, leading to

$$\mathcal{L}_{\text{RRR}}\left(\exp_\theta(\mathbf{x}), \exp_{\text{prior}}(\mathbf{x})\right) = \frac{|\nabla_{\mathbf{x}} f_\theta(\mathbf{x})_p| \cdot \mathbf{M}_{\text{prior}}(\mathbf{x})}{\|\nabla_{\mathbf{x}} f_\theta(\mathbf{x})_p\|_2 \|\mathbf{M}_{\text{prior}}(\mathbf{x})\|_2}. \tag{4}$$

Alternatively, CDEP [14] proposes to use CD [13] importance scores $\boldsymbol{\beta}(\mathbf{x}_s)$ for a feature subset $\mathbf{x}_s$ based on the forward pass instead of gradient to align the model's attention. Penalizing artifact features via masked input $\mathbf{x}_M$ results in

$$\mathcal{L}_{\text{CDEP}}\left(\exp_\theta(\mathbf{x}), \exp_{\text{prior}}(\mathbf{x})\right) = \left\| \frac{e^{\boldsymbol{\beta}(\mathbf{x}_M)}}{e^{\boldsymbol{\beta}(\mathbf{x}_M)} + e^{\boldsymbol{\beta}(\mathbf{x}-\mathbf{x}_M)}} \right\|_1. \tag{5}$$

4 Experiments

The experimental section is divided into the two parts of (1) identification, mitigation and evaluation of spurious model behavior with various correction methods and (2) showcasing the whole R2R framework in an iterative fashion.

4.1 Experimental Setup

We train VGG-16 [19], ResNet-18 [9] and EfficientNet-B0 [21] models on the ISIC 2019 dataset [23, 6, 7] for skin lesion classification and Pediatric Bone Age dataset [8] for bone age estimation based on hand radiographs. Besides evaluating our methodology on data-intrinsic artifacts occurring in these datasets, we artificially insert an artifact into data samples in a controlled setting. Specifically, we insert a "Clever Hans" text (shown in Fig. 2) into a subset of training samples of one specific class. See Appendix A.2 for additional experiment details.

4.2 Revealing and Revising Spurious Model Behavior

Revealing Bias: In the first step of the R2R life cycle, we can reveal the use of several artifacts by the examined models, including the well-known band-aid, ruler and skin marker [5] and our synthetic Clever Hans for the ISIC dataset, as shown in Fig. 2 for VGG-16. Here, we show concept visualizations and cropped out artifacts based on our automatic artifact localization scheme described in Section 3.1. The "band-aid"

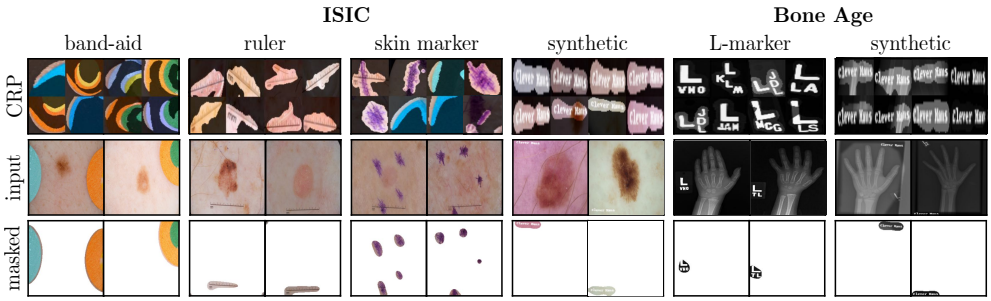

Figure 2: Overview of artifacts with CRP visualization of corresponding concepts (*top*), input samples (*middle*), and cropped out artifacts (*bottom*) using our artifact localization method. Shown are band-aid, ruler, skin marker, and synthetic artifacts for the ISIC dataset, as well as "L"-marker and synthetic artifacts for the Bone Age dataset.

use can be further identified via SpRAy, as illustrated in Fig. 3 (*right*). Examplary artifact CAV heatmaps for all data-intrinsic artifacts are given in Appendix A.3.1.

Besides the synthetic Clever Hans for bone age classification, we encountered the use of "L" markings, resulting from physical lead markers placed by radiologist to specify the anatomical side. Interestingly, the "L" markings are larger for hands of younger children, as all hands are scaled to similar size [8], offering the model to learn a shortcut by estimating the bone age based on the relative size of the "L" markings, instead of valid features. While we revealed the "L" marking bias using CRP, we did not find corresponding SpRAy clusters, underlining the importance of both approaches for model investigation.

Revising Model Behavior: Having revealed spurious behavior, we now revise the models, beginning with model correction. Specifically, we correct for the band-aid, "L" markings as well as synthetic artifacts. The skin marker and ruler artifacts are corrected for in iterative fashion in Section 4.3. For all methods (RRR, CDEP[1] and ClArC), including a *Vanilla* model without correction, we fine-tune the models' last dense layers for 10 epochs. Note that both RRR and CDEP require artifact masks to unlearn the undesired behavior. As part of R2R, we propose measures to automate this step by using the artifact localization strategy described in Section 3.1. Further note, that once generated, artifact localizations can be used for *all* investigated models. See Appendix A.2 for additional fine-tuning details.

We evaluate the effectiveness of model corrections based on two metrics: the attributed fraction of relevance to artifacts and prediction performance on both the original and a poisoned test set (in terms of F1-score and accuracy). Whereas in the synthetic case, we simply insert the artifact into all samples to poison the test set, data-intrinsic artifacts are cropped from random artifactual samples using our artifact localization strategy. Note that artifacts might overlap clinically informative features in poisoned samples, limiting the comparability of *poisoned* and *original* test performance. As shown in Tab. 1 (ISIC 2019) and Appendix A.3 (Bone Age), we are generally able to improve model behavior with all methods. The only exception is the synthetic artifact for VGG-16, where only RRR mitigates the bias to a certain extent, indicating that the artifact signal is too strong for the model. Here, fine-tuning only the last layer is not sufficient to learn alternative prediction strategies.

[1]CDEP is not applied to EfficientNets, as existing implementations are incompatible.

Table 1: Model correction results for two ISIC dataset artifacts (band-aid | synthetic). Arrows indicate whether low (↓) or high (↑) scores are better with best scores bold.

architecture	method	↓ artifact relevance (%)	↑ F1 (%) *poisoned*	↑ F1 (%) *original*	↑ accuracy (%) *poisoned*	↑ accuracy (%) *original*
VGG-16	*Vanilla*	45.5 | 76.3	59.7 | 7.7	73.9 | 79.0	71.5 | 19.1	80.1 | 86.9
	RRR	**14.3** | **12.0**	**64.2** | **39.2**	71.8 | 77.7	**74.4** | **32.4**	78.0 | 85.4
	CDEP	23.7 | 78.4	62.8 | 7.2	73.9 | **79.0**	72.3 | 18.9	80.2 | 86.9
	p-ClArC	41.9 | 76.1	61.8 | 7.6	**74.0** | 78.1	73.0 | 19.1	**80.3** | 85.4
	a-ClArC	42.8 | 75.5	62.4 | 12.5	70.3 | 76.5	73.1 | 21.0	78.4 | **88.9**
ResNet-18	*Vanilla*	33.1 | 37.6	68.2 | 39.0	79.1 | 82.1	76.8 | 35.6	83.3 | 89.5
	RRR	30.3 | **16.9**	70.4 | **70.4**	**79.7** | 79.1	77.1 | **75.7**	**83.4** | 84.8
	CDEP	**25.4** | 22.2	**71.5** | 60.9	75.9 | 81.6	**77.5** | 59.4	81.5 | 87.9
	p-ClArC	32.0 | 33.6	69.2 | 38.9	78.3 | **81.8**	75.9 | 34.4	82.5 | **89.1**
	a-ClArC	32.9 | 38.4	70.1 | 52.9	78.3 | 80.5	76.2 | 45.3	81.1 | 88.9
Efficient-Net-B0	*Vanilla*	45.6 | 63.9	72.2 | 38.8	81.8 | 84.7	80.1 | 30.2	85.4 | 90.8
	RRR	**34.5** | **24.6**	**74.0** | 65.8	81.3 | 83.3	80.1 | 65.9	84.6 | 89.8
	p-ClArC	41.3 | 62.5	73.1 | 38.7	**82.0** | **84.4**	**80.4** | 29.8	**85.5** | **90.5**
	a-ClArC	45.7 | 65.6	72.7 | **72.4**	81.8 | 81.4	80.1 | **79.4**	84.9 | 87.3

Interestingly, despite successfully decreasing the models' output sensitivity towards artifacts, applying a-ClArC barely decreases the relevance attributed to artifacts in input space. This might result from ClArC methods not directly penalizing the use of artifacts, but instead encouraging the model to develop alternative prediction strategies. Overall, RRR yields the most consistent results, constantly reducing the artifact relevance while increasing the model performance on poisoned test sets. Both observations are underlined by heatmaps for revised models in Fig. 8 (Appendix A.3), where RRR and CDEP visibly reduce the model attention on the artifacts.

4.3 Iterative Model Correction with R2R

Showcasing the full R2R life cycle (as shown in Fig. 1), we now perform multiple R2R iterations, revealing and revising undesired model behavior step by step. Specifically, we successively correct the VGG-16 model w.r.t. the skin marker, band-aid, and ruler artifacts discovered in Section 4.2 using RRR. In order to prevent the model from re-learning previously unlearned artifacts, we keep the previous artifact-specific RRR losses intact. Thus, we are able to correct for all artifacts, with evaluation results given in Tab. 2, applying the same metrics as in Section 4.2. In Fig. 3, we show exemplary attribution heatmaps for all artifacts after each iteration. While there are large amounts of relevance on all artifacts initially, it can successfully be reduced in the according iterations to correct the model behavior w.r.t. skin marker (SM), band-aids (BA), and rulers (R). It is to note, that correcting for the skin marker also (slightly) improved the model w.r.t. other artifacts, which might result from corresponding latent features that are not independent, as shown by CRP visualizations in Fig. 2 for skin marker. Moreover, we show the SpRAy embedding of training samples after the first iteration in Fig. 3 (*right*), revealing an isolated cluster with samples containing the band-aid artifact, which dissipates after the correction step.

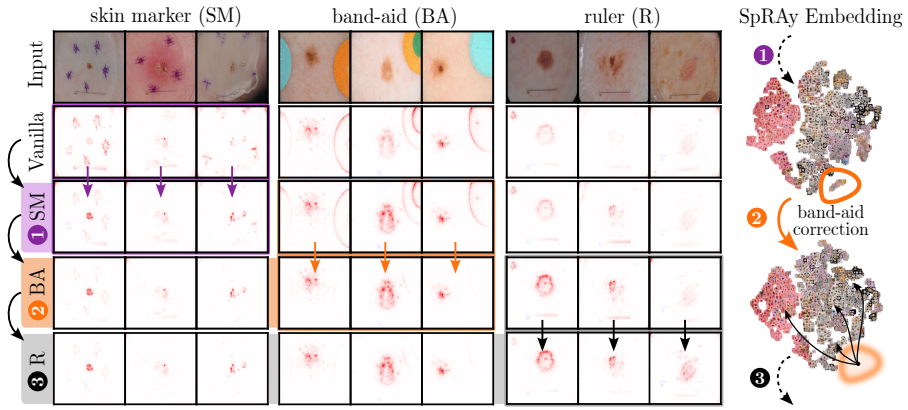

Figure 3: The effect of iterative model correction on relevances attributed to artifacts for each iteration (*left*) and the band-aid artifact cluster from SpRAy, which dissipates after its correction step (*right*).

Table 2: Iterative R2R results for ISIC artifacts (skin marker: SM | band-aid: BA | ruler: R). Arrows show whether low (↓)/high (↑) scores are better, best in bold.

R2R iteration	corrected artifacts	↓ artifact relevance (%)			↑ F1 (%) *poisoned*		↑ F1 (%) *original*	↑ accuracy (%) *poisoned*		↑ accuracy (%) *original*
0	-	18.4	45.5	24.2	61.3	59.7	60.5	73.9	71.8	71.5	68.7	**80.1**
1	SM	13.1	35.0	21.3	61.6	61.0	60.7	73.8	72.2	72.6	68.4	80.0
2	SM, BA	**12.8**	16.8	16.8	61.5	**63.6**	61.1	73.9	72.3	**74.6**	68.6	79.7
3	SM, BA, R	14.6	**15.7**	**8.5**	**62.0**	63.4	**64.0**	**74.0**	**72.4**	74.5	**71.8**	79.9

5 Conclusion

We present R2R, an XAI life cycle to reveal and revise spurious model behavior requiring minimal human interaction via high automation. To *reveal* model bias, R2R relies on CRP and SpRAy. Whereas SpRAy automatically points out Clever Hans behavior by analyzing large sets of attribution data, CRP allows for a fine-grained investigation of spurious concepts learned by a model. Moreover, CRP is ideal for large datasets, as the concept space dimension remains constant. By automatically localizing artifacts, we successfully perform model *revision*, thereby reducing attention on the artifact and leading to improved performance on corrupted data. When applying R2R iteratively, we did not find the emergence of new biases, which, however, might happen if larger parts of the model are fine-tuned or retrained to correct strong biases. Future research directions include the application to non-localizable artifacts, and addressing fairness issues in DNNs.

Acknowledgements

This work was supported by the Federal Ministry of Education and Research (BMBF) as grants [SyReal (01IS21069B), BIFOLD (01IS18025A, 01IS18037I)]; the European Union's Horizon 2020 research and innovation programme (EU Horizon 2020) as grant [iToBoS (965221)]; the state of Berlin within the innovation support program ProFIT (IBB) as grant [BerDiBa (10174498)]; and the German Research Foundation [DFG KI-FOR 5363].

References

[1] Achtibat, R., Dreyer, M., Eisenbraun, I., Bosse, S., Wiegand, T., Samek, W., Lapuschkin, S.: From "where" to "what": Towards human-understandable explanations through concept relevance propagation. arXiv preprint arXiv:2206.03208 (2022)

[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., 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), e0130140 (2015)

[4] Brinker, T.J., Hekler, A., Enk, A.H., Klode, J., Hauschild, A., Berking, C., Schilling, B., Haferkamp, S., Schadendorf, D., Holland-Letz, T., et al.: Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. European Journal of Cancer **113**, 47–54 (2019)

[5] Cassidy, B., Kendrick, C., Brodzicki, A., Jaworek-Korjakowska, J., Yap, M.H.: Analysis of the isic image datasets: Usage, benchmarks and recommendations. Medical image analysis **75**, 102305 (2022)

[6] Codella, N.C., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., et al.: Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018). pp. 168–172. IEEE (2018)

[7] Combalia, M., Codella, N.C., Rotemberg, V., Helba, B., Vilaplana, V., Reiter, O., Carrera, C., Barreiro, A., Halpern, A.C., Puig, S., et al.: Bcn20000: Dermoscopic lesions in the wild. arXiv preprint arXiv:1908.02288 (2019)

[8] Halabi, S.S., Prevedello, L.M., Kalpathy-Cramer, J., Mamonov, A.B., Bilbily, A., Cicero, M., Pan, I., Pereira, L.A., Sousa, R.T., Abdala, N., et al.: The rsna pediatric bone age machine learning challenge. Radiology **290**(2), 498–503 (2019)

[9] 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)

[10] Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al.: Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In: International conference on machine learning. pp. 2668–2677. PMLR (2018)

[11] Kim, B., Kim, H., Kim, K., Kim, S., Kim, J.: Learning not to learn: Training deep neural networks with biased data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9012–9020 (2019)

[12] Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., Müller, K.R.: Unmasking clever hans predictors and assessing what machines really learn. Nature communications **10**(1), 1096 (2019)

[13] Murdoch, W.J., Liu, P.J., Yu, B.: Beyond word importance: Contextual decomposition to extract interactions from lstms. arXiv preprint arXiv:1801.05453 (2018)

[14] Rieger, L., Singh, C., Murdoch, W., Yu, B.: Interpretations are useful: penalizing explanations to align neural networks with prior knowledge. In: International conference on machine learning. pp. 8116–8126. PMLR (2020)

[15] Ross, A.S., Hughes, M.C., Doshi-Velez, F.: Right for the right reasons: Training differentiable models by constraining their explanations. arXiv preprint arXiv:1703.03717 (2017)

[16] 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)

[17] Samek, W., Montavon, G., Lapuschkin, S., Anders, C.J., Müller, K.R.: Explaining deep neural networks and beyond: A review of methods and applications. Proceedings of the IEEE **109**(3), 247–278 (2021)

[18] Schramowski, P., Stammer, W., Teso, S., Brugger, A., Herbert, F., Shao, X., Luigs, H.G., Mahlein, A.K., Kersting, K.: Making deep neural networks right for the right scientific reasons by interacting with their explanations. Nature Machine Intelligence **2**(8), 476–486 (2020)

[19] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)

[20] Stock, P., Cisse, M.: Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 498–512 (2018)

[21] Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. pp. 6105–6114. PMLR (2019)

[22] Teso, S., Kersting, K.: Explanatory interactive machine learning. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. pp. 239–245 (2019)

[23] Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data **5**(1), 1–9 (2018)

[24] Weber, L., Lapuschkin, S., Binder, A., Samek, W.: Beyond explaining: Opportunities and challenges of xai-based model improvement. Information Fusion (2022)

A Appendix

A.1 *Reveal to Revise* Algorithm

Input: Biased Model f, Training Data X, y
Output: Corrected Model $\hat{f}$
$\hat{f} \leftarrow f$

`/* (1) Identification of model weaknesses via SpRAy (1a) or CRP concpt visualizations (1b) */`
$w \leftarrow \text{identifyWeaknesses}(\hat{f}, X)$

while $w \neq \{\}$ **do**
  `/* (2) Artifact labeling and localization with CAV */`
  $X_{art} \leftarrow \text{labelArtifactualSamples}(\hat{f}, X, w)$
  $h_{art} \leftarrow \text{fitCAV}(\hat{f}, X_{art}, X)$
  $M_{art} \leftarrow \text{localizeArtifacts}(\hat{f}, h_{art}, X_{art})$

  `/* (3) Model correction via RRR, CDEP or ClArC */`
  if *Correction Method from ClArC-Framework* **then**
    `/* Correct with artifact direction in latent space */`
    $\hat{f} \leftarrow \text{correctModel}(\hat{f}, X, h_{art})$
  else
    `/* Correct with (automatically detected) artifact masks */`
    $\hat{f} \leftarrow \text{correctModel}(\hat{f}, X, M_{art})$
  end

  `/* (4) (Re-)evaluate Model */`
  $\text{evaluateModel}(\hat{f}, X, M_{art}, y)$

  `/* Repeat from (1) */`
  $w \leftarrow \text{identifyWeaknesses}(\hat{f}, X)$
end

Algorithm 1: R2R Algorithm as outlined in Fig. 1: We identify model weaknesses by finding either outliers in explanations using SpRAy (1a) or suspicious concepts using CRP concept visualizations (1b). Secondly (2), SpRAy clusters or collecting the top reference samples allows us to label artifactual samples and to compute an artifact CAV, which we use to model and localize the artifact in latent and input space, respectively. At this point, the artifact localization can be leveraged for (3) model correction, and (4) to evaluate the model's performance on a poisoned test set and measure its remaining attention on the artifact. This algorithm can be repeated until no more model weaknesses can be detected.

A.2 Experimental Details

Table 3: Details for ISIC2019 and Pediatric Bone Age datasets. All images are resized to the same input size and normalized to mean μ and standard deviation σ. In the controlled setting, we insert a "Clever Hans"-text with random size, position, and rotation into a chosen class with given probability p. Each dataset is split into training data to train the models, as well as to detect and unlearn undesired behavior, validation data to pick optimal λ values for correction methods, and test data for evaluation.

dataset	number samples	input size	norm. (μ/σ)	classes	split size train/val/test	Clever Hans class (p)
ISIC2019	25331	224×224	0.5/0.5	MEL, NV, BCC, AK, BKL, DF, VASC, SCC	0.8/0.1/0.1	MEL (10%)
Bone Age	12611	224×224	0.5/0.5	0-46, 47-91, 92-137, 138-182, 183-228 (months)	0.8/0.1/0.1	0-46 (50%)

Table 4: Training details for examined architectures. Weights are pre-trained from the PyTorch model zoo. The learning rate is divided by 10 after 50 and 80 epochs.

architecture	optimizer	loss	epochs (ISIC/Bone Age)	initial learning rate
VGG-16	SGD	Cross Entropy	150/100	0.005
Resnet-18	SGD	Cross Entropy	150/100	0.005
EfficientNet-B0	Adam	Cross Entropy	150/100	0.001

Table 5: Details for model correction performed for 10 epochs using SGD optimizer with learning rate of 10^{-4}. We test λ-values in range $\{1, 5, 10, \ldots, 10^4\}$ (CDEP/RRR) or different layers (ClArC) and pick the best ones on the validation performance. Best hyperparameters are given for (ISIC band-aid | synthetic | Bone Age "L"-marker | synthetic). The RRR loss is adapted to handle high variance of DNN gradients by using cosine similarity (instead of L2-norm) and absolute gradient w.r.t. the predicted logit.

method	best λ/layers (VGG-16)	best λ/layer (ResNet-18)	best λ/layer (EfficientNet-B0)
RRR	500 | 200 | 500 | 500	500 | 500 | 5000 | 1000	500 | 100 | 500 | 1000
CDEP	10 | 1 | 100 | 5	50 | 50 | 100 | 50	-
p-ClArC	`features` 28 | 21 | 26 | 26	`layer` 3 | 2 | 4 | 2	`features` 6 | 3 | 6 | 7
a-ClArC	`features` 14 | 21 | 28 | 19	`layer` 4 | 3 | 3 | 2	`features` 8 | 6 | 2 | 6

A.3 Additional Results

A.3.1 Artifact Localization

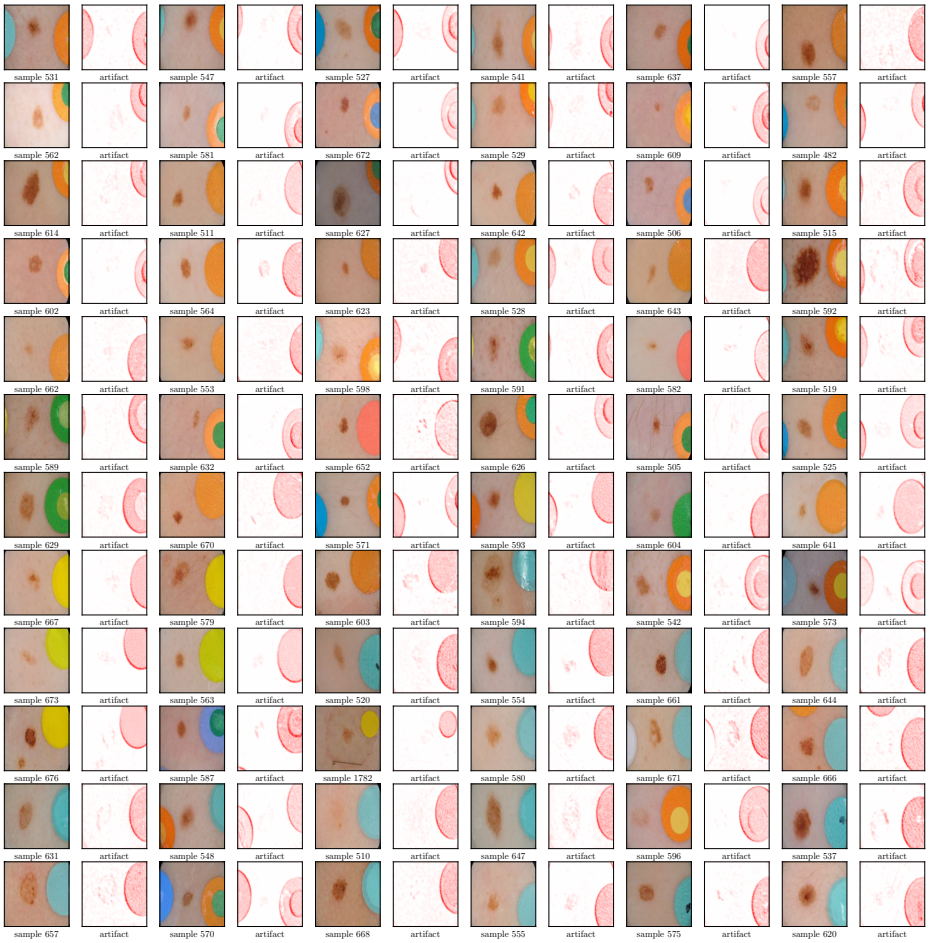

Figure 4: Examples for the R2R automatic concept localization scheme using artifact CAVs for the band-aid artifact of the ISIC dataset. Shown are 36 artifact samples with corresponding CAV heatmaps. Artifacts have been localized in layer `features.7` of the VGG-16 model.

Figure 5: Examples for the R2R automatic concept localization scheme using artifact CAVs for the skin marker artifact of the ISIC dataset. Shown are 36 artifact samples with corresponding CAV heatmaps. Artifacts have been localized in layer `features.12` of the VGG-16 model.

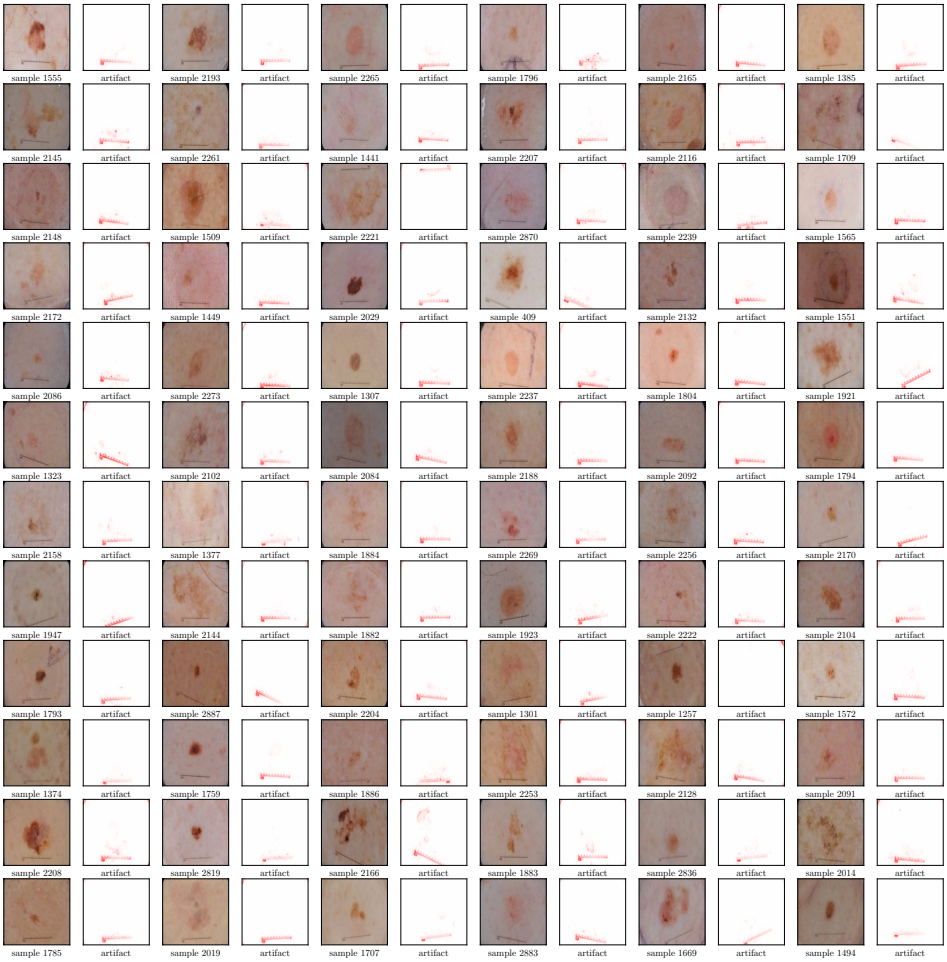

Figure 6: Examples for the R2R automatic concept localization scheme using artifact CAVs for the ruler artifact of the ISIC dataset. Shown are 36 artifact samples with corresponding CAV heatmaps. Artifacts have been localized in layer `features.28` of the VGG-16 model.

Figure 7: Examples for the R2R automatic concept localization scheme using artifact CAVs for the "L"-marker artifact of the Bone Age Estimation dataset. Shown are 36 artifact samples with corresponding CAV heatmaps. Artifacts have been localized in layer `features.28` of the VGG-16 model.

A.3.2 Model Correction

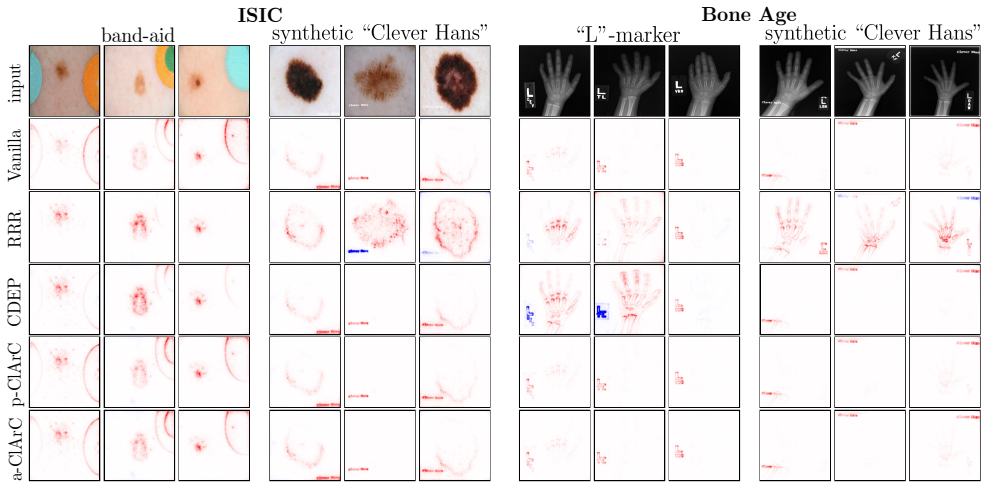

Figure 8: Explanation heatmaps for corrected VGG-16 models using different methods for ISIC2019 (band-aid | synthetic) and Bone Age ("L"-marker | synthetic) artifacts.

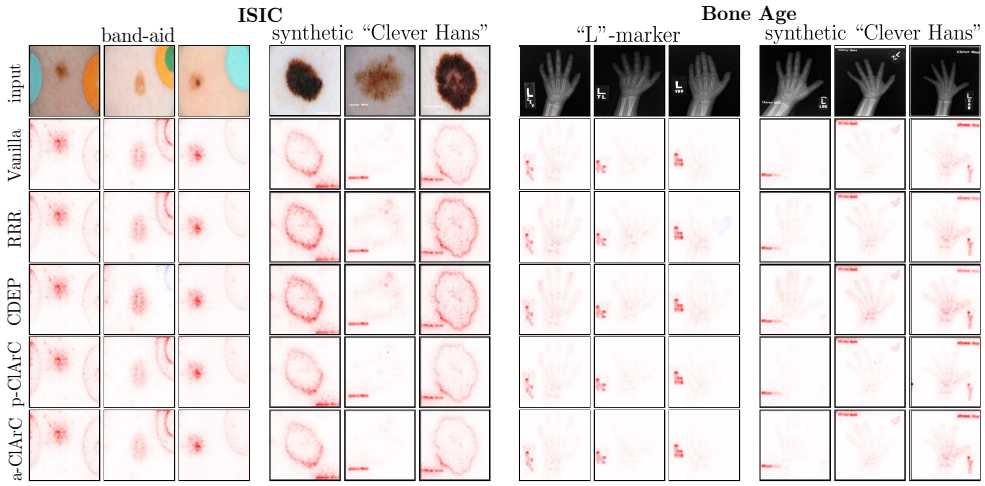

Figure 9: Explanation heatmaps for corrected ResNet-18 models using different methods for ISIC2019 (band-aid | synthetic) and Bone Age ("L"-marker | synthetic) artifacts.

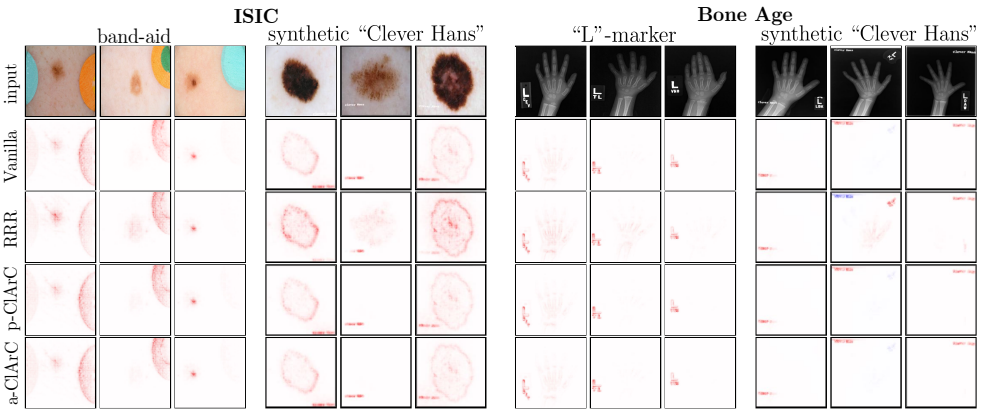

Figure 10: Explanation heatmaps for corrected EfficientNet-B0 models using different methods for ISIC2019 (band-aid | synthetic) and Bone Age ("L"-marker | synthetic) artifacts.

Table 6: Results for the Bone Age Estimation dataset artifacts ("L"-marker | synthetic). Arrows show whether low (↓)/high (↑) scores are better with best in bold.

architecture	method	↓ artifact relevance (%)	↑ F1 (%) *poisoned*	↑ F1 (%) *original*	↑ accuracy (%) *poisoned*	↑ accuracy (%) *original*
VGG-16	*Vanilla*	29.6 | 57.5	61.4 | 16.2	76.6 | 82.3	62.2 | 13.6	78.6 | 80.3
	RRR	**9.6** | **9.0**	**66.1** | 44.6	**77.4** | 82.2	**65.7** | 58.5	78.5 | 80.7
	CDEP	21.4 | 21.4	62.8 | 26.9	76.3 | 82.7	63.9 | 23.4	**78.7** | 80.8
	p-ClArC	29.0 | 48.5	62.2 | 23.5	76.7 | 57.2	62.6 | 25.0	**78.7** | 58.8
	a-ClArC	29.9 | 47.2	63.2 | **55.1**	76.1 | **84.3**	64.2 | **64.7**	78.1 | **81.8**
ResNet-18	*Vanilla*	23.3 | 38.4	65.5 | 51.9	72.5 | 80.1	66.6 | 64.9	75.9 | 79.2
	RRR	**16.4** | 27.7	67.4 | 59.1	71.0 | 79.0	68.9 | **68.0**	75.6 | 78.5
	CDEP	19.6 | **17.9**	65.7 | 56.3	71.9 | 77.5	66.6 | 64.0	74.9 | 76.4
	p-ClArC	21.4 | 29.2	66.5 | 53.7	**72.6** | **79.9**	67.7 | 66.5	**76.1** | **79.1**
	a-ClArC	21.4 | 29.5	**70.0** | **66.2**	70.3 | 77.1	**74.4** | 67.6	**76.1** | 76.1
EfficientNet-B0	*Vanilla*	35.1 | 69.1	69.5 | 49.4	75.0 | 81.6	73.6 | 60.3	78.1 | 81.4
	RRR	**24.0** | 51.7	**71.2** | 58.5	74.9 | **82.0**	**74.3** | **65.5**	**78.0** | **81.6**
	p-ClArC	31.1 | 66.2	70.4 | 50.4	74.9 | 81.7	74.0 | 61.1	77.9 | 81.3
	a-ClArC	35.3 | **24.7**	70.5 | **61.3**	**75.4** | 78.7	72.9 | 64.9	77.7 | 77.0