Towards Fixing Clever-Hans Predictors with Counterfactual Knowledge Distillation

Sidney Bender [1] Christopher Anders [1,5] Pattarawat Chormai [1,2] Heike Marxfeld[4]
Jan Herrmann[4] Grégoire Montavon [3,5]

[1]Technische Universität Berlin [2]Max Planck School of Cognition [3]FU Berlin [4]BASF [5]BIFOLD

Summary

Deciding from pathological images whether tissues of animals treated with different substances contain any abnormal cells, is a difficult and repetitive task. Therefore, it is desirable to support the human decision process with predictive models, which can then be worked into automatized solutions. However, an automated solution begs the following questions:

1. How does the actual decision boundary of the model look?
2. How can we influence the decision boundary post hoc according to our intuition?

Yolo[3] Follicle Detection and LRP[1] explanations

Detection: Is there any of those follicles in this image?

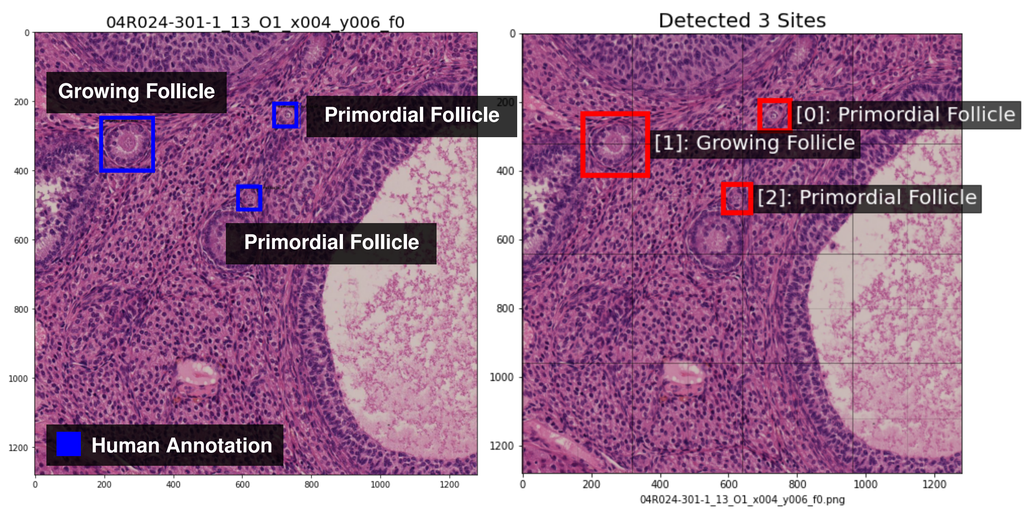


Figure 1. Annotated and detected follicles in a validation image.

Explanation: What features did the model consider for its decision?

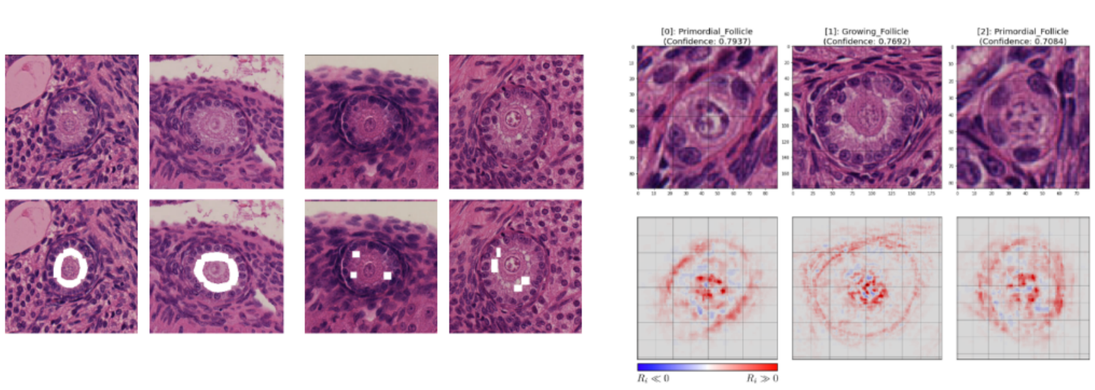


Figure 2. On the left side are the features the model should look at and on the right side are the ones it looks at according to LRP.

Successes? The marked features look very similar to how humans would also detect follicles, namely a round shape and the nucleus in the middle of the structure.

Problems? The real decision boundary of follicle types is whether there is a second layer of supporting cells, but this second layer is not highlighted at all.

Solution? Apply an explanation method more specialized to explaining class boundaries and actively adapt class boundaries if they still do not follow human intuition.

Counterfactual Explanations

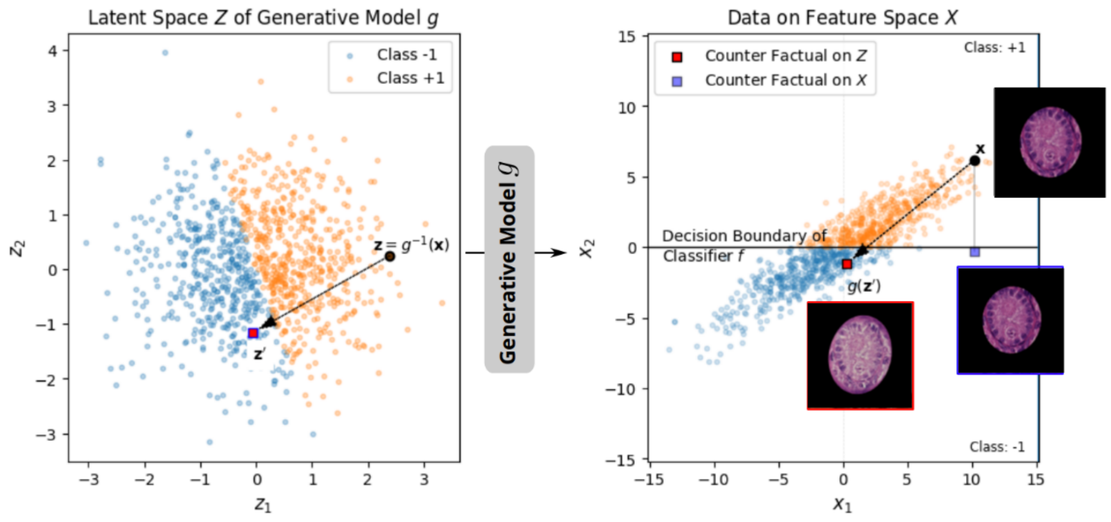


Figure 3. Diffeomorphic Counterfactual Explanations[2]. As can be seen one has to use a generative model as a regularizer instead of directly using gradient ascend in input space to avoid creating adversarial examples.

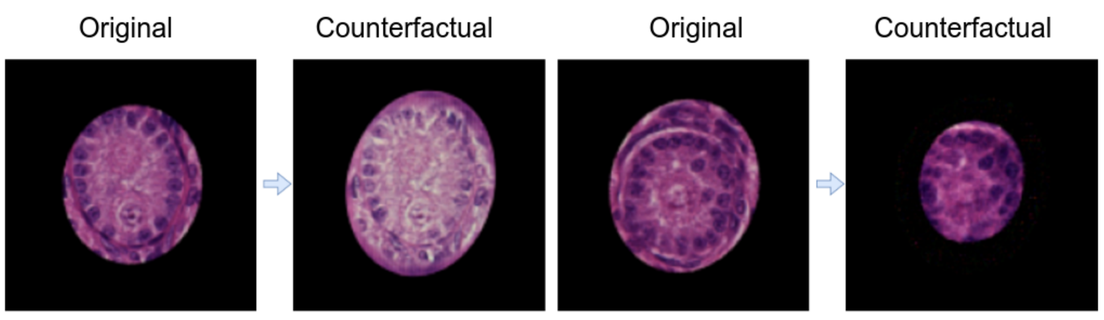
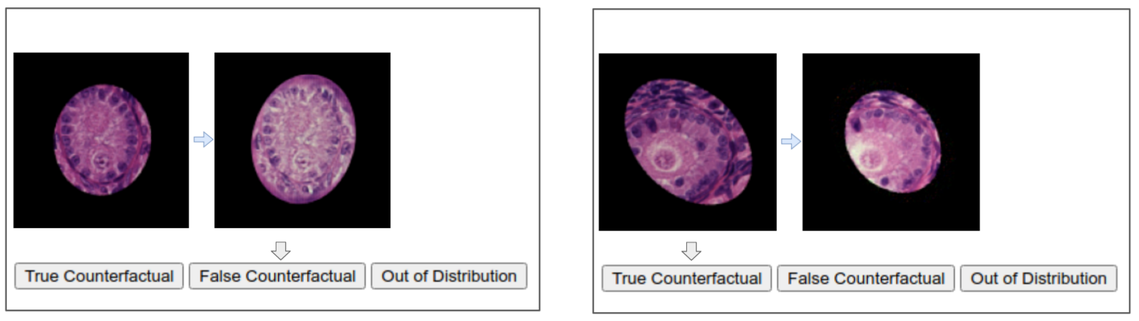


Figure 4. Counterfactual Explanations. Shows that the classifier relies on the confounder size.

Human-in-the-loop Teacher



(a) **Case A**: False counterfactual, that is contrary to the intuitive decision boundary of human experts.

(b) **Case B**: True counterfactual, that agrees with intuitive decision boundary of human experts.

Figure 5. Comparison of two counterfactual cases. We embedded the feedback mechanism with the buttons in a web interface, that is started when executing CFKD.

Feedback Accuracy The ratio of the expert saying a created counterfactual is a true counterfactual we call feedback accuracy. Intuitively it is 0 if the expert and the model never agree and 1 if they always agree. It thereby measures how similar the model decides to the expert.

Counterfactual Knowledge Distillation (CFKD)

How can counterfactual explanations be used to actively influence the decision boundary?

1. Create counterfactual image.
2. Let human-in-the-loop decide whether it is Case A or Case B.
3. If Case A: Select original label.
4. If Case B: Select flipped label.
5. Add counterfactual image and selected label to the training data.
6. Retrain the model.
7. Go back to the first step until feedback accuracy is as wished.

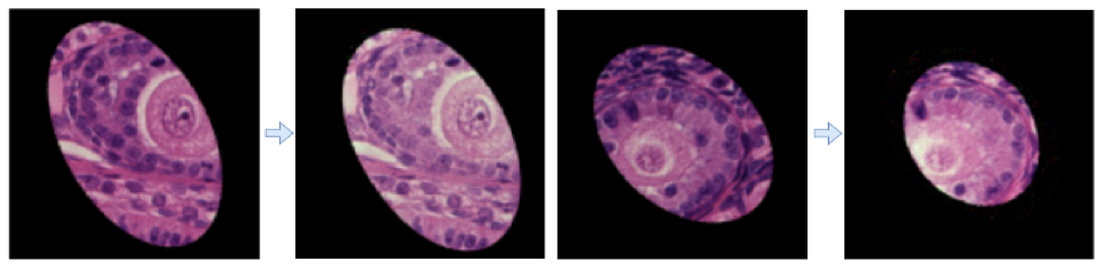


Figure 6. Counterfactual Explanations after applying CFKD. Shows that the classifier does not rely on the confounder size anymore.

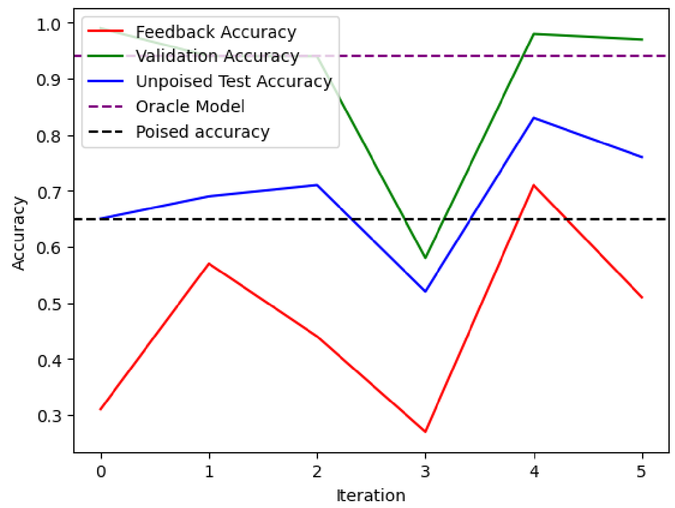


Figure 7. Quantitative Results on colorectal cancer dataset. Here the confounder strength was controlled by subsampling the dataset based on the staining of the image and creating a correlation between cancer and staining.

References

[1] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*, 10(7):e0130140, 07 2015.

[2] Ann-Kathrin Dombrowski, Jan E Gerken, Klaus-Robert Müller, and Pan Kessel. Diffeomorphic counterfactuals with generative models. *arXiv preprint arXiv:2206.05075*, 2022.

[3] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 779–788. IEEE Computer Society, 2016.