



# Right for the Wrong Reason: Can Interpretable ML Techniques Detect Spurious Correlations?

Susu Sun<sup>1(✉)</sup>, Lisa M. Koch<sup>2,3</sup>, and Christian F. Baumgartner<sup>1</sup>

<sup>1</sup> Cluster of Excellence – ML for Science,  
University of Tübingen, Tübingen, Germany  
{susu.sun,christian.baumgartner}@uni-tuebingen.de

<sup>2</sup> Hertie Institute for AI in Brain Health,  
University of Tübingen, Tübingen, Germany

<sup>3</sup> Institute of Ophthalmic Research, University of Tübingen, Tübingen, Germany  
lisa.koch@uni-tuebingen.de

**Abstract.** While deep neural network models offer unmatched classification performance, they are prone to learning spurious correlations in the data. Such dependencies on confounding information can be difficult to detect using performance metrics if the test data comes from the same distribution as the training data. Interpretable ML methods such as post-hoc explanations or inherently interpretable classifiers promise to identify faulty model reasoning. However, there is mixed evidence whether many of these techniques are actually able to do so. In this paper, we propose a rigorous evaluation strategy to assess an explanation technique’s ability to correctly identify spurious correlations. Using this strategy, we evaluate five post-hoc explanation techniques and one inherently interpretable method for their ability to detect three types of artificially added confounders in a chest x-ray diagnosis task. We find that the post-hoc technique SHAP, as well as the inherently interpretable Attri-Net provide the best performance and can be used to reliably identify faulty model behavior.

**Keywords:** Interpretable machine learning · Confounder detection

## 1 Introduction

Black-box neural network classifiers offer enormous potential for computer-aided diagnosis and prediction in medical imaging applications but, unfortunately, they also have a strong tendency to learn spurious correlations in the data [11]. For the development and safe deployment of machine learning (ML) systems it is

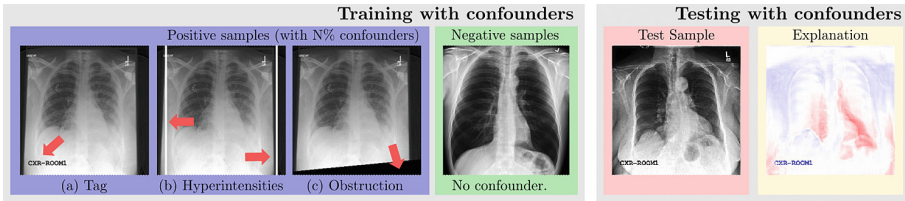
---

**Supplementary Information** The online version contains supplementary material available at [https://doi.org/10.1007/978-3-031-43895-0\\_40](https://doi.org/10.1007/978-3-031-43895-0_40).

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023  
H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14221, pp. 425–434, 2023.  
[https://doi.org/10.1007/978-3-031-43895-0\\_40](https://doi.org/10.1007/978-3-031-43895-0_40)

essential to understand what information the classifiers are basing their decisions on, such that reliance on spurious correlations may be identified.

Spurious correlations arise when the training data are confounded by additional variables that are unrelated to the diagnostic information we want to predict. For instance, older patients in our training data may be more likely to present with a disease than younger patients. A classifier trained on this data may inadvertently learn to base its decision on image features related to age rather than pathology. Crucially, such faulty behavior cannot be identified using classification performance metrics such as area under the ROC curve (AUC) if the testing data contains the same confounding information as the training data, since the classifier predicts the *right* thing, but for the *wrong* reason. If undetected, however, such spurious correlations may lead to serious safety implications after deployment.



**Fig. 1. Overview.** We train classifiers on datasets with three types of artificially added confounders highlighted by arrows. We then evaluate the ability of explanation techniques to correctly identify reliance on these confounders (shown Attri-Net [24]).

Interpretable ML approaches may be used as a powerful tool to detect spurious correlations during development or after deployment of an ML system. Currently, the most widely used explanation modality are *visual* explanations, which highlight the pixels in the input image that are responsible for a particular decision. Common strategies include methods which leverage the gradient of the prediction with respect to the input image [7, 19, 22, 23, 25], explain the predictions by counterfactually generating an image of the opposite class [9, 18, 20, 24], interpret the feature map of the last layer before the classification [8, 10, 27], or methods that build a local approximation of the decision function such as LIME [16], or SHAP [15].

The majority of visual explanation methods are *post-hoc* techniques, meaning a heuristic is applied to any trained model (e.g. a ResNet [13]) to approximately understand the decision mechanism for a given data point. However, post-hoc techniques are by definition only approximations and many techniques have been found to suffer from serious limitations [4, 12, 26]. *Inherently interpretable* techniques on the other hand build custom architectures that are designed to directly reveal the reasoning of the classifier to the user without the need for approximations. This class of methods does not suffer from the same limitations as post-hoc methods, and it has been argued that inherently interpretable approaches should

be preferred in high-stakes applications such as medical image analysis [17]. For instance, if a classifier bases its decision on a spurious signal, an inherently interpretable classifier should by definition reveal this relationship.

Inherently interpretable visual explanation approaches are much less widely explored than post-hoc techniques, but there has recently been an increased interest in the topic. Two recently proposed methods in this category are the attribution network (Attri-Net) [24], and convolutional dynamic alignment networks (CoDA-Nets) [6]. Attri-Net first produces human-interpretable feature attribution maps for each disease category using a GAN-based counterfactual generator [24]. Then makes the final prediction with simple logistic regression classifiers based on those feature attribution maps. CoDA-Nets express neural networks as input dependent linear transformation [6]. Both approaches produce explanations on the pixel level of the input images.

**Related Work on Comparing Explanation Techniques.** A number of works have studied the quality of post-hoc explanation techniques. The vast majority of work focuses exclusively on gradient-based approaches (e.g. [5, 21]). In their landmark study, Adebayo et al. [1] find that commonly used gradient-based explanation techniques do not pass some basic sanity checks. Arun et al. [5] extends this work to weakly supervised localisation in one of the few papers in this domain focusing on medical data. Both papers, however, do not consider other types of commonly used approaches such as counterfactual methods, or local function approximations such as LIME or SHAP.

A small number of works specifically investigate explanations’ sensitivity to spurious correlations. In closely related work to ours, Adebayo et al. [3] explore a large library of post-hoc explanation techniques including LIME and SHAP, for detecting spurious image backgrounds in a bird versus dog classification task and find that many techniques are in fact able to detect the spurious background. In subsequent work, the same authors explore the usefulness of four post-hoc gradient-based explanation methods for identifying spurious correlations in hand and knee radiographs [2] and come to the conclusion that the examined methods are ineffective at identifying spurious correlations. We note that prior work is inconclusive on the usefulness of explanation techniques for identifying spurious correlations. In particular, in the medical context it is still unclear if commonly used explanation techniques are suitable for the detection of spurious correlations. Moreover, there is, to our knowledge, no evidence for the supposition that inherently interpretable techniques are better suited for this task.

**Contributions.** We present a rigorous evaluation of post-hoc explanations and inherently interpretable techniques for the identification of spurious correlations in a medical imaging task. Specifically, we focus on the task of diagnosing cardiomegaly from chest x-ray data with three types of synthetically generated spurious correlations (see Fig. 1). To identify whether an explanation correctly

identifies a model’s reliance on spurious correlations, we propose two quantitative metrics which are highly reflective of our qualitative findings. In contrast to the majority of prior work we focus on a wide range of different explanation approaches including counterfactual techniques and local function approximations, as well as post-hoc techniques and an inherently interpretable approach. Our analysis yields actionable insights which will be useful for a wide audience of ML practitioners.

## 2 A Framework for Evaluating Explanation Techniques

In the following, we introduce our evaluation strategy and proposed evaluation metrics, the studied confounders, as well as the evaluated explanation techniques. The strategy and evaluation metrics are generic and can also be applied to different problems. The confounders are engineered to correspond to realistic image artifacts that can appear in chest x-ray imaging<sup>1</sup>.

### 2.1 Evaluation Strategy

We assume a setting in which the development data for a binary neural network based classifier contains an unknown spurious correlation with the target label. To quantitatively study this setting, we create training data with artificial spurious correlations by adding a confounding effect (e.g. a hospital tag) in a percentage of the cases with a positive label, where we vary the percentage  $p \in \{0, 20, 50, 80, 100\}$ . E.g., for  $p = 100\%$  all of the positive images in the training set will have an artificial confounder, and for  $p = 0\%$  there is no spurious signal. With increasing  $p$  the reliance on a spurious signal becomes more likely. The images with a negative label remain untouched.

In the evaluation, we consider a scenario in which the test data contain the same confounder type with the same proportion  $p$  used in the respective trainings. In this case, we can not tell if a classifier relies on the confounded features from classification performance. Our aim, therefore, is to investigate whether explanation techniques can identify that the classifier predicts the *right thing for the wrong reason*.

We perform all experiments on chest x-ray images from the widely used CheXpert dataset [14], where we focus on the binary classification task on disease cardiomegaly. We divided our dataset into a training (80%), validation (10%) test (10%) set.

### 2.2 Studied Confounders

We study three types of confounders inspired by real-world artefacts. Firstly, we investigate a hospital tag placed in the lower left corner of the image (see

---

<sup>1</sup> Our code can be found under <https://github.com/ss-sun/right-for-the-wrong-reason>.

Fig. 1a). Secondly, we add vertical lines of hyperintense signal that can be caused by foreign materials on the light path assembly (see Fig. 1b). Lastly, we consider an oblique occlusion of the image in the lower part of the image, which is an artefact that we observed for many images in the CheXpert dataset (see Fig. 1c).

### 2.3 Evaluation Metrics for Measuring Confounder Detection

We propose two novel metrics which reflect an explanation’s ability to correctly identify spurious correlations.

**Confounder Sensitivity (CS).** Firstly, the explanations should be able to correctly attribute the confounder if classifier bases its decision on it. We assess this property by summing the number of true positive attributions divided by the total number of confounded pixels for each test image. We consider a pixel a true positive if it is part of the pixels affected by the confounder *and* in the top 10% attributed pixels according to a visual explanation. Thus the maximum sensitivity of 1 is obtained if all confounded pixels are in the top 10% of the attributions. Note that we do not penalise attributions outside of the confounding label as those can still also be correct. To guarantee that we only evaluate on samples for which the prediction is actually influenced by the confounder, we only include images for which the prediction with and without the confounding label is of the opposite class. To reduce computation times we use a maximum of 100 samples for each evaluation. An optimal explanation methods should obtain a CS score of 0 if the data contains  $p = 0\%$  confounded data points, since in that case the spurious signal should not be attributed. For increasing  $p$  the confounder sensitivity should increase, i.e. the explanation should reflect the classifiers increasing reliance on the confounder.

**Sensitivity to Prediction Changes via Explanation NCC.** Secondly, the explanations should not be invariant to changes in classifier prediction. That is, if the classifier’s prediction for a specific image changes when adding or removing a confounder, then the explanations should also be different. We measure this property using the average normalised cross correlation (NCC) between explanations of test images when confounders were either present or absent. Again, we only evaluate on images for which the prediction changes when adding the confounder as in these cases, we know the classifier is relying on confounders, and we evaluate a maximum of 100 samples. An optimal explanation method should obtain a high NCC score if the training data contains  $p = 0\%$  confounded data points, since in that case the explanation with and without the confounder should be similar. For increasing  $p$  the NCC score should decrease to reflect the classifiers increasing reliance on the confounder.

### 2.4 Evaluated Explanation Methods

We evaluated five post-hoc techniques with representative examples from the approaches mentioned in the introduction: Guided Backpropagation [23] and Grad-CAM [19] (gradient-based), Gifsplanation (counterfactual), and LIME [16]

and SHAP partition explainer [15] (local linear approximations). All post-hoc techniques were applied to a standard black-box ResNet50 model. We further investigated the interpretable visual explanation method Attri-Net [24]. We used the default parameters for all methods. We found CoDA-Nets [6] required lengthy hyperparameter tuning for each type of experiment, and decided to exclude it in this paper.

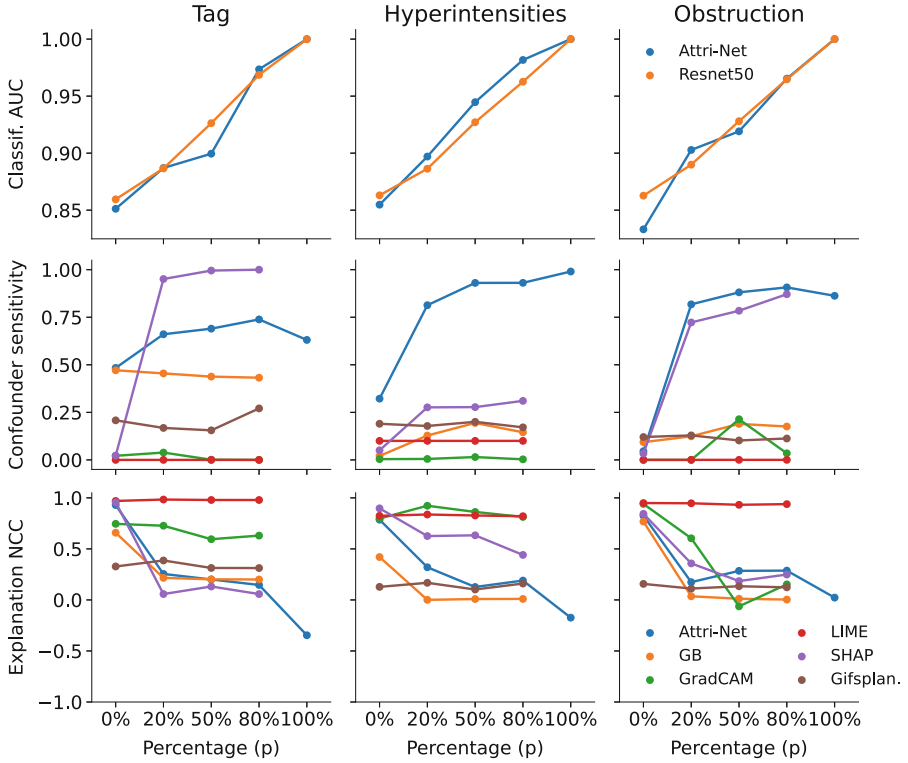
### 3 Results

We first established the classifiers' performance in the presence of confounders, then compared all techniques in their ability to identify such confounders.

**Classification Performance.** Both investigated classifiers, the ResNet50 and the inherently interpretable Attri-Net, performed similarly in terms of classification AUC (first row of Fig. 2). For all three confounders, classification AUC consistently increased with increasing contamination  $p$  of the training dataset. This indicated that the classifiers increasingly relied on the spurious signal. For  $p = 100\%$  contamination, where the confounder was present on all positive training examples, both classifiers reached almost a perfect classification AUC of 1.

**Explanations.** We analysed the explanations' ability to identify confounders by reporting confounder sensitivity (CS, middle row in Fig. 2) and explanation NCC (bottom row in Fig. 2). Out of the investigated methods Attri-Net and SHAP were closest to the ideal behaviour of high confounder sensitivity and low explanation NCC for  $p > 0\%$ . We found that SHAP performed extremely well in detecting tag confounders, but struggled with hyperintensities confounders. This can be explained by the fact that the tag confounder is relatively small and thus is more likely to be completely covered by the superpixels in SHAP. Overall, the inherently interpretable Attri-Net technique achieved the best balance. In agreement with related literature we found that gradient-based explanation methods performed poorly. In particular, Guided Backpropagation displayed similar CS-scores no matter if the classifier relies on a spurious signal ( $p > 0\%$ ) or not ( $p = 0\%$ ). Note that some results for  $p = 100\%$  were missing because no data points fulfilled the criterion of the prediction being flipped with and without the confounders.

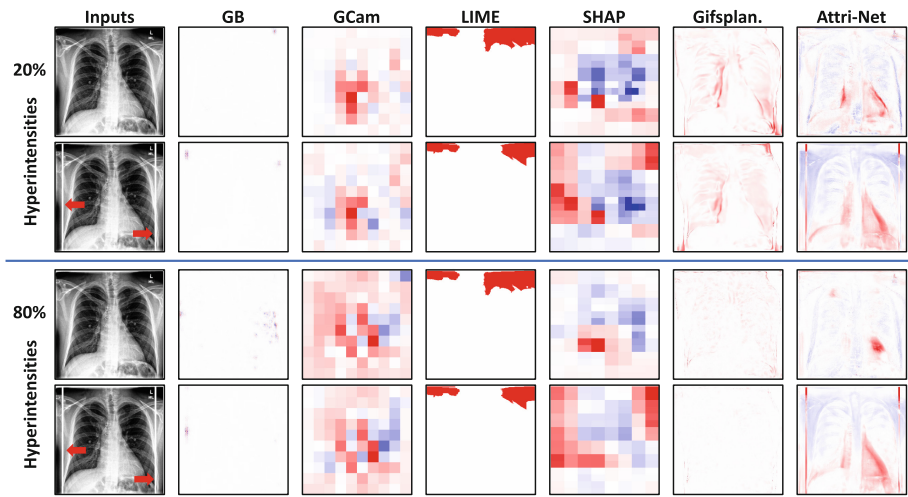
Figures 3, 4 & 5 contain examples explanations for the hyperintensity, tag, and edge confounder, respectively. Our qualitative analysis of the results confirms the quantitative findings, with SHAP and Attri-Net providing the most intuitive explanations. In particular, in the challenging hyperintensities scenario (see Fig. 3) AttriNet was the only method able to highlight the confounders in a human-interpretable fashion. We note that in all examples when a confounder was present, SHAP tended to highlight only the confounder, while Attri-Net also highlighted features related to Cardiomegaly. This may reflect the different decision mechanisms of the ResNet50 and the Attri-Net.



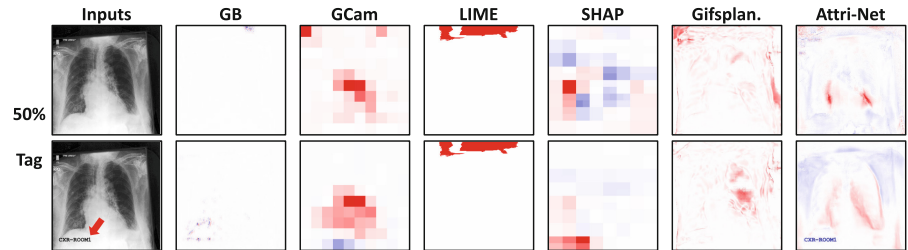
**Fig. 2. (top row)** Classification AUC of Attri-Net and Resnet50 on images containing hospital tags (left), hyperintensities (middle) or obstruction confounders (right column). The classifiers were trained with a varying proportion of confounders present in the positive examples in the training set (shown on the x-axes). **(bottom rows)** The explanation techniques’ ability to identify confounders in terms of confounder sensitivity (middle row) and explanation NCC (bottom row, lower is better).

## 4 Discussion

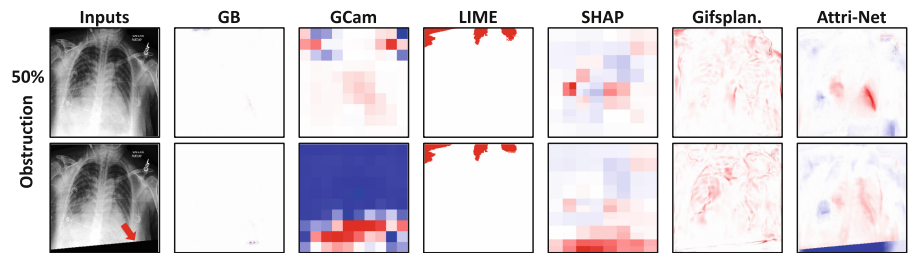
In this paper, we proposed an evaluation strategy to assess the ability of visual explanations to correctly identify a classifier’s reliance on a spurious signal. We specifically focused on the scenario where the classifier is predicting the right thing, but for the wrong reason, which is highly significant for the safe development of ML-based diagnosis and prediction systems. Using this strategy, we assessed the performance of five post-hoc explanation techniques and one inherently interpretable technique with three realistic confounding signals. We found that the inherently interpretable Attri-Net technique, as well as the post-hoc SHAP technique performed the best, with Attri-Net yielding the most balanced performance. Both techniques are suitable for finding false reliance on a spurious signals. We also observed that the variation in the explanations’ sparsity makes them perform differently in detecting spurious signals of different sizes and



**Fig. 3.** Explanations for one example image with and without hyperintensities confounders. We show results for models trained on 20% (**top rows**) and 80% (**bottom rows**) confounded data points, respectively.



**Fig. 4.** Explanations for one example image with (**top**) and without (**bottom**) a tag confounder for models trained on 50% confounded data points.



**Fig. 5.** Explanation for one example image with (**top**) and without (**bottom**) an obstruction confounder for models trained on 50% confounded data points.



shapes. In agreement with prior work, we found that gradient based techniques performed less robustly in our experiments.

From our experiments we draw two main conclusions. Firstly, practitioners looking to check for spurious correlations in a trained black-box model such as a ResNet should give preference to SHAP which provided the best performance out of the post-hoc techniques in our experiments. Secondly, an inherently interpretable technique, namely Attri-Net, performed the best in our experiments providing evidence to the supposition by Rudin et al. [17] that inherently interpretable techniques may provide a fruitful avenue for future work.

A major limitation of our study is the limited number of techniques we examined. Thus a primary focus of future work will be to scale our experiments to a wider range of techniques. Future work will also focus on human-in-the-loop experiments, as we believe, this will be the ultimate assessment of the usefulness of different explanation techniques.

**Acknowledgements.** Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany’s Excellence Strategy - EXC number 2064/1 - Project number 390727645. The authors acknowledge support of the Carl Zeiss Foundation in the project “Certification and Foundations of Safe Machine Learning Systems in Healthcare” and the Hertie Foundation. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Susu Sun, Lisa M. Koch, and Christian F. Baumgartner.

## References

1. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. In: *Advances in Neural Information Processing Systems* 31 (2018)
2. Adebayo, J., Muelly, M., Abelson, H., Kim, B.: Post hoc explanations may be ineffective for detecting unknown spurious correlation. In: *International Conference on Learning Representations* (2022)
3. Adebayo, J., Muelly, M., Llicardi, I., Kim, B.: Debugging tests for model explanations. arXiv preprint [arXiv:2011.05429](https://arxiv.org/abs/2011.05429) (2020)
4. Alvarez-Melis, D., Jaakkola, T.S.: On the robustness of interpretability methods. arXiv preprint [arXiv:1806.08049](https://arxiv.org/abs/1806.08049) (2018)
5. Arun, N., et al.: Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiol.: Artif. Intell.* **3**(6), e200267 (2021)
6. Bohle, M., Fritz, M., Schiele, B.: Convolutional dynamic alignment networks for interpretable classifications. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10029–10038 (2021)
7. Boreiko, V., et al.: Visual explanations for the detection of diabetic retinopathy from retinal fundus images. In: *Medical Image Computing and Computer Assisted Intervention* (2022)
8. Brendel, W., Bethge, M.: Approximating CNNs with bag-of-local-features models works surprisingly well on ImageNet. arXiv preprint [arXiv:1904.00760](https://arxiv.org/abs/1904.00760) (2019)
9. Cohen, J.P., et al.: Gifsplanation via Latent Shift: a simple autoencoder approach to counterfactual generation for chest X-rays. In: *Medical Imaging with Deep Learning*, pp. 74–104. PMLR (2021)

10. Djoumessi, K.R., et al.: Sparse activations for interpretable disease grading. arXiv preprint [arXiv:TODO](#) (2023)
11. Geirhos, R., et al.: Shortcut learning in deep neural networks. *Nat. Mach. Intell.* **2**(11), 665–673 (2020)
12. Han, T., Srinivas, S., Lakkaraju, H.: Which explanation should i choose? a function approximation perspective to characterizing post hoc explanations. arXiv preprint [arXiv:2206.01254](#) (2022)
13. 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)
14. Irvin, J., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: *Proceedings of the AAAI conference on artificial intelligence*. vol. 33, pp. 590–597 (2019)
15. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: *Advances in Neural Information Processing Systems* 30 (2017)
16. Ribeiro, M.T., Singh, S., Guestrin, C.: why should i trust you? explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144 (2016)
17. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**(5), 206–215 (2019)
18. Samangouei, P., Saeedi, A., Nakagawa, L., Silberman, N.: ExplainGAN: model explanation via decision boundary crossing transformations. In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 666–681 (2018)
19. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: visual explanations from deep networks via gradient-based localization. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626 (2017)
20. Singla, S., Pollack, B., Chen, J., Batmanghelich, K.: Explanation by progressive exaggeration. arXiv preprint [arXiv:1911.00483](#) (2019)
21. Sixt, L., Granz, M., Landgraf, T.: When explanations lie: why many modified BP attributions fail. In: *International Conference on Machine Learning*, pp. 9046–9057. PMLR (2020)
22. Smilkov, D., Thorat, N., Kim, B., Viégas, F., Wattenberg, M.: SmoothGrad: removing noise by adding noise. arXiv preprint [arXiv:1706.03825](#) (2017)
23. Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for simplicity: the all convolutional net. arXiv preprint [arXiv:1412.6806](#) (2014)
24. Sun, S., Woerner, S., Maier, A., Koch, L.M., Baumgartner, C.F.: Inherently interpretable multi-label classification using class-specific counterfactuals. arXiv preprint [arXiv:2303.00500](#) (2023)
25. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: *International Conference on Machine Learning*, pp. 3319–3328. PMLR (2017)
26. White, A., Garcez, A.D.: Measurable counterfactual local explanations for any classifier. arXiv preprint [arXiv:1908.03020](#) (2019)
27. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: *Proceedings of the IEEE Conference on Computer Vision and Precognition*, pp. 2921–2929 (2016)