Domain aware medical image classifier interpretation by counterfactual impact analysis[*]

Dimitrios Lenis, David Major, Maria Wimmer, Astrid Berg, Gert Sluiter, and Katja Bühler

VRVis Zentrum für Virtual Reality und Visualisierung Forschungs-GmbH, Vienna, Austria

Abstract. The success of machine learning methods for computer vision tasks has driven a surge in computer assisted prediction for medicine and biology. Based on a data-driven relationship between input image and pathological classification, these predictors deliver unprecedented accuracy. Yet, the numerous approaches trying to explain the causality of this learned relationship have fallen short: time constraints, coarse, diffuse and at times misleading results, caused by the employment of heuristic techniques like Gaussian noise and blurring, have hindered their clinical adoption.

In this work, we discuss and overcome these obstacles by introducing a neural-network based attribution method, applicable to any trained predictor. Our solution identifies salient regions of an input image in a single forward-pass by measuring the effect of local image-perturbations on a predictor's score. We replace heuristic techniques with a strong neighborhood conditioned inpainting approach, avoiding anatomically implausible, hence adversarial artifacts. We evaluate on public mammography data and compare against existing state-of-the-art methods. Furthermore, we exemplify the approach's generalizability by demonstrating results on chest X-rays. Our solution shows, both quantitatively and qualitatively, a significant reduction of localization ambiguity and clearer conveying results, without sacrificing time efficiency.

Keywords: Explainable AI · XAI · Classifier Decision Visualization · Image Inpainting.

1 Introduction

The last decade's success of machine learning methods for computer-vision tasks has driven a surge in computer assisted prediction for medicine and biology. This has posed a conundrum. Current predictors, predominantly artificial neural networks (ANNs), learn a data-driven relationship between input image and
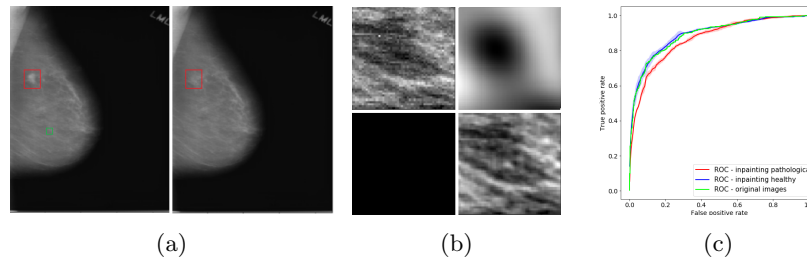

Fig. 1: Overview of marginalization: (a) original with annotated mass (red box) before and after marginalization by our method; (b) local comparisons with popular methods (clockwise): original, blurring [9], inpainting (ours), and averaging [29]; (c) ROC curves of the mammography classifier (green curve) vs. healthy pixel inpainting only in healthy/pathological (blue/red curves) structures.

pathological classification, whose validity, i.e. accuracy and specificity, we can quantitatively test. In contrast, this learned relationship's causality typically remains elusive [1,18,19]. A plethora of approaches have been proposed that aim to fill this gap by explaining causality through identifying and attributing salient image-regions responsible for a predictor's outcome [7,8,9,26,25,28].

Lacking a canonical mapping between an ANN's prediction and its domain, this form of reasoning is predominantly based on *local explanations* (LE), i.e. explicit attribution-maps characterizing image-prediction tuples [18,9]. Typically, these maps are loosely defined as regions with *maximal influence* towards the predictor, implying that any texture change within the attributed area will significantly change the prediction. Besides technical insight, these LE can provide a key benefit for clinical applications: by relating the ANN's algorithmic outcome to the user's a-priori understanding of pathology-causality, they can strengthen confidence in the predictor, thereby increasing its clinical acceptance. To achieve this goal, additional restrictions and clarifications are crucial. Qualitatively, such maps need to be *informative* for its users, i.e. narrow down regions of medical interest, hence coincide with medical knowledge and expectations [21]. Furthermore, the regions' characteristic, i.e. the meaning of *maximal influence*, must be clearly conveyed. Quantitatively, such LE need to be *faithful* to the underpinning predictor, i.e. dependent on architecture, parametrization, and preconditions [2].

The dominant class of methods follow a *direct approach*. Utilizing an ANN's assumed analytic nature and its layered architecture, they typically employ a modified backpropagation approach to backtrack the ANN's activation to the input image [26,30]. While efficiently applicable, the resulting maps lack a clear a-priori interpretation, are potentially incomplete, coarse, and may deliver misleading information [2,8,9,31]. Thereby they are potentially neither *informative* nor *faithful*, thus pose an inherent risk in medical environments.

In contrast, *reference based* LE approaches directly manipulate the input image and analyze the resulting prediction's differences [9]. They aim to as-

sess an image-region's influence on prediction by counterfactual reasoning: how would the prediction score vary, if the region's image-information would be missing, i.e. its contribution marginalized? The prevailing heuristic approaches, e.g. Gaussian noise and blurring or replacement by a predefined colour [29,8,9], have been advanced to local neighborhood [31] and stronger conditional generative models [7,28]. Reference based LEs have the advantage of an a-priori clear and intuitively conveyable meaning of their result, hence address *informativeness* for end-users. However, their applicability for medical imaging hinges on the utilized marginalization technique, i.e. the mapping between potentially pathological tissue representations and their healthy equivalent. Resulting *prediction-neutral* regions need to depict healthy tissue per definition. Contradictory, the presented approaches introduce noise and thereby possibly pathological indications or anatomically implausible tissue (cf. Fig. 1). Hence, they violate the needed *faithfulness* [9].

While dedicated generative adversarial networks (GANs) for medical images deliver significantly improved results, applications are hindered by possible resolutions and limited control over the globally acting models [3,4,5,6]. In [22], the locally acting, but globally conditioned, per-pixel reconstruction of partial convolution inpainting (PCI) [20] is favoured over GANs, thereby enforcing anatomically sound, image specific replacements. While overcoming out-of-domain issues, this gradient descent based optimization method works iteratively, hence cannot be used in time restrictive environments.

Contribution: We introduce a *resource efficient* reference based *faithful* and *informative* attribution method for real time pathology classifier interpretation. Utilizing a specialized ANN and exploiting PCI's local per-pixel reconstruction, conditioned on a global healthy tissue representation, we are able to enforce anatomically sound, image specific marginalization, without sacrificing computational efficiency. We formulate the ANN's objective function as a quantitative prediction problem under strict area constraints, thereby clarifying the resulting attribution map's a-priori meaning. We evaluate the approach on public mammography data and compare against two existing state-of-the-art methods. Furthermore, we exemplify the method's generalizability by demonstrating results on a second unrelated task, namely chest X-ray data. Our solution shows, both quantitatively and qualitatively, a significant reduction of localization ambiguity and clearer conveying results without sacrificing time efficiency.

2   Methods

Given a pathology classifier's prediction for an input image, we want to estimate its cause by attributing the specific pixel-regions that substantially influenced the predictor's outcome. Informally, we search for the image-area that, if changed, results in a *sufficiently healthy* image able to *fool the classifier*. The resulting attribution-map needs to be *informative* for the user and *faithful* to its underpinning classifier. While we can quantitatively test for the latter, the former is an ill-posed problem. We therefore formalize as follows:
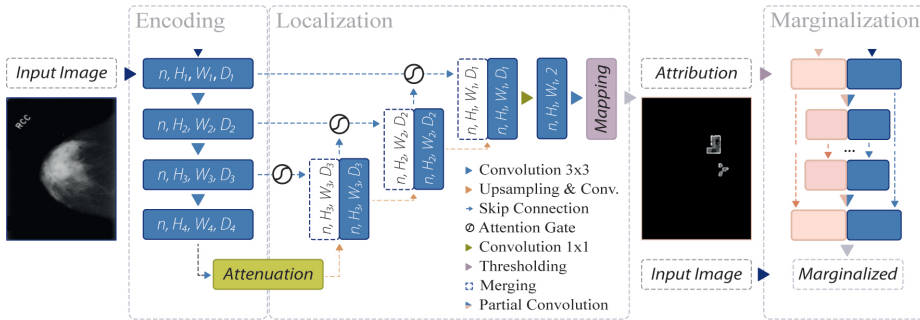
Fig. 2: Attribution framework: The input image is encoded using the classifier's features (left) and attenuated to enclose pathological regions (middle). During training, counterfactual images are produced by the marginalization-net (right), fed by thresholded attribution (pink blocks) and input image (blue blocks).

Let I denote an image of a domain $\mathcal{I}$ with pixels on a discrete grid $m_1 \times m_2$, c a fixed pathology-class, and f a classifier capable of estimating $p(c|I)$, the probability of c for I. Also, let M denote the attribution-map for image I and class c, hence $M \in M^{m_1 \times m_2}(\{0,1\})$. Furthermore, assume a function $\pi(M)$ proficient in marginalizing all pixel regions attributed by M in I such that the result of the operation is still within the domain of f. Hence, $\pi(M)$ yields a new image similar to I, but where we know all regions attributed by M to be healthy per definition. Therefore, assuming I depicts a pathological case and M attributes only pathology pixel representations, $\pi(M)$ is a healthy counterfactual image to I. In any case $p(c|\pi(M))$ is well defined. Using this notation, we can formalize what an *informative* map $\hat{M}$ means, hence give it an a-priori, testable semantic meaning. We define it as

$$\hat{M} := \underset{M \in \hat{\mathcal{M}}}{\operatorname{argmin}} \, d(M) \quad \text{where} \quad \hat{\mathcal{M}} := \{p(c|\pi(M)) \leq \theta, d(M) \leq \delta, M \in \mathcal{S}\},$$

where θ is the classification-threshold, d a metric measuring the attributed area, δ a constant limiting the attributed area, and $\mathcal{S}$ the set of compact and connected masks. Any map of $M^{m_1 \times m_2}(\{0,1\})$ can be (differentiably) mapped into $\mathcal{S}$ by taking the smoothed maximum of a convolution with a Gaussian kernel [16,9]. In this form, $\hat{M}$ is clearly defined, and can be intuitively understood by end-users.

Solving for $\hat{M}$ requires choosing (i) an appropriate measure d (e.g. the map area in pixels), (ii) an appropriate size-limit δ (e.g. n times average mass-size for mammography), and (iii) a fitting marginalization technique $\pi(\cdot)$. In the following we describe how we solve for $\hat{M}$ through an ANN, and overcome the out-of-domain obstacles by partial convolution [20] for marginalization.

2.1   Architecture

Iteratively finding solutions for $\hat{M}$ is typically time-consuming [9,22]. Therefore, we develop a dedicated ANN, capable of finding the desired attribution

in a single forward pass. To this end, the network learns on multiple resolutions, to combine relevant classifier-extracted features (cf. Fig. 2). Inspired by [8], we build on a U-Net architecture, where the down-sampling, encoding branch consists of the trained classifier without its classification layers. These features, $x_{i,j,l}$, are subsequentially passed through a feature-filter, performing $x_{i,j,l} \cdot \sigma((W_m \rho(W_l^\mathsf{T} x_{i,j,l} + b_l) + b_m))$ where ρ is an element-wise nonlinearity (namely a rectified linear unit), σ a normalization function (sigmoid function) and W. resp. b. linear transformation parameters. This is similar to additive attention, which, compared to multiplicative attention, has shown better performance on high dimensional input-features [24]. The upsampling branch consists of four consecutive blocks of: upsampling by a factor of two, followed by convolution and merging with attention-gate weighted features from the classifier of the corresponding resolution scale. After final upsampling back to input-resolution, we apply 1×1 conv. of depth two, resulting in two channels $c_{1,2}$. The final attribution-map $\hat{M}$ is derived through thresholding $\frac{|c_1|}{|c_1|+|c_2|}$. Intuitively, the network attenuates the classifier's final features, generating an initial localization. This coarse map is subsequently refined by additional weighting and information from higher resolution features (cf. Fig. 2). We train the network, by minimizing

$$\mathcal{L}(M) = \phi(M) + \psi(M) + \lambda \cdot \mathcal{R}(M), \text{ s.t. } d(M) \leq \delta$$

where $\phi(M) := -1 \cdot \log(p(c|\pi(M)))$, $\psi(M) := \log(\text{odds}(I)) - \log(\text{odds}(\pi(M)))$, and $\text{odds}(I) = \frac{p(c|I)}{1-p(c|I)}$, hence weigh the probability of the marginalized image, enforcing $p(c|\pi(M)) \leq \theta$. We introduced an additional regularization-term: a weighted version of total variation [23], which experimentally greatly improved convergence. All terms where normalized through a generalized logistic function. The inequality constraint was enforced by the method proposed in [15]. Note that after mapping into $\mathcal{S}$, any solution to $\mathcal{L}$ will also estimate $\hat{M}$, thereby yielding our desired attribution-map. The parametrization is task/classifier-dependent and will be described in the following sections.

2.2   Marginalization

As we need to derive $p(c|\pi(M))$, our goal is to marginalize arbitrary image regions marked by our network during its training process. Therefore, we aim for an image inpainting method to replace pathological tissue by healthy appearance. The result should resemble valid global anatomical appearance with high quality local texture. To address the these criteria we apply the U-Net like architecture with partial convolution blocks of [20] which gets an image and a hole mask as input (cf. Fig. 2). Partial convolution considers only unmasked inputs in a current sliding window to compute its output. Where it succeeded, hole mask positions are eliminated. This mechanism helps conditioning on local texture. The loss function ($\mathcal{L}_{PCI}$) balances local per-pixel reconstruction quality of masked/unmasked regions ($\mathcal{L}_{hole}/\mathcal{L}_{valid}$), against globally sound anatomical appearance ($\mathcal{L}_{perc}, \mathcal{L}_{style}$). An additional total variation term ($\mathcal{L}_{tv}$) ensures a smooth transition between hole and present image regions in the final result.

This yields $\mathcal{L}_{PCI} = \mathcal{L}_{valid} + 6 \cdot \mathcal{L}_{hole} + 0.05 \cdot \mathcal{L}_{perc} + 120 \cdot \mathcal{L}_{style} + 0.1 \cdot \mathcal{L}_{tv}$ where parametrization follows [20]. The architecture's contraction path consists of 8 partial convolution blocks with a stride of 2. The kernels of depth 64, 128, 256, 512, ..., 512 have sizes 7, 5, 5, 3, ..., 3. The expansion path, a mirrored version of the contraction path, contains upsampling layers with a factor of 2, kernel size of 3 at every layer, and a final filterdepth of 3. Each block contains batch normalization (BN) and ReLU/LeakyReLU (alpha=0.2) activations in the contraction/expansion paths which are connected by skip connections. Zero padding of the input was applied to control resolution shrinkage and keep aspect ratio.

3   Experimental Setup

Datasets: We evaluated our framework on two different datasets, on mammography scans and on chest X-ray images. For mammography, we complemented the 1565 annotated, pathological CBIS-DDSM scans containing masses [17] with 2778 healthy DDSM images [10] and downsampled them to 576x448 pixels. Data was split into 1231/2000 mass/healthy samples for training, and into 334/778 scans for testing. There was no patient-wise overlap between the training/test data. We demonstrate generalization on a private collection of healthy and tuberculotic (TBC) frontal chest X-ray images, at a downsampled resolution of 256x256. We split healthy images into sets of 1700/135 for training respectively validation set, and TBC cases into 700/70. The test set contains 52 healthy and 52 TBC samples. No pixel-wise GT information was provided for this data.

Classifiers: The backbone of our mammography attribution network is a MobileNet [11] classifier for distinguishing between healthy samples and scans with masses. The network was trained using the Adam optimizer with batchsize of 4 and learning rate of 1e-5 for 250 epochs with early stopping. The network was pretrained with 50k 224x224 pixel patches from the training data for the same task. The TBC attribution utilized a DenseNet-121 [12] classifier for the binary classification task of healthy or TBC cases. It was trained using the SGD momentum optimizer with a batchsize of 32 and learning rate of 1e-5 for 2000 epochs. This network was pretrained on the CheXpert dataset [13].

Marginalization: The chest X-ray images have one magnitude smaller resolution than the mammography scans, thus we removed the bottom-most blocks from the contraction and expansion paths. Both inpainter networks were trained on healthy training samples with a batch size of 1 for mammography and 5 for chest X-ray. Training was done in two phases, the first phase with BN after each partial convolution layer and the second with BN only in the expansion path. The network for the mass classification task was trained with learning rates of 1e-5/1e-6 and for the TBC classification task of 2e-4/1e-5 for the two phases. For each image irregular masks were generated which mimic possible configurations during the attribution network training [20].

Attribution: We used the last four resolution-scales of each classifier, and in all cases the features immediately after the activation function, following the convolution. The weights of the pre-trained ANNs were kept fixed during the

complete process. Filterdepths of the upsampling convolution blocks correspond to the equivalent down-sampling filters, filter-size is fixed to 1×1. Upsampling itself is done via neighborhood upsampling. We used standard gradient descent, and a cyclic learning rate [27], varying between 1e-6 and 1e-4, and trained for up to 5000 epochs with early stopping. We thresholded the masks at 0.55, and used a Gaussian RBF with $\sigma = $ 5e-2, and a smoothing parameter of 30. All trainable weights where random-normal initialized.

4   Results and Conclusion

Marginalization: To evaluate the inpainter network we assessed how much the classification score of an image changes, when pathological tissue is replaced.

Thus, we computed ROC curves using the classifier on all test samples (i) without any inpainting as reference, and for comparison, randomly sampled inpainting (ii) only in healthy respective (iii) pathological scans over 10 runs (Fig. 1). The clear distance between the ROC curves of the mammography image classifiers without any inpainting, yielding an AUC of 0.89, and with inpainting in pathological regions, resulting in an AUC of 0.86, shows that the classifier is sensitive to changes around pathological regions of the image. Moreover, it is visible that the ROC curves of inpainting in healthy tissues with an AUC of 0.89 follow closely the unaffected classifier's ROC curve (Fig. 1). The AUC scores for the TBC classifier without and with inpainting in healthy tissue are 0.89 and 0.88 which proves the above mentioned observations. Pathological tissue inpainting was ommitted in this case due to the lack of pixel-wise annotations.

Attribution: We compared our attribution network against the gradient explanation *saliency map* [26] (SAL), and the network/gradient-derived *Grad-CAM* [25] visualizations. We limited our comparisons to these direct approaches, as they are widely used within medical imaging [13], and inherently valid [2]. Popular *reference based* approaches either utilize blurring, noise or some other heuristic [9,8,31], or were not available [7], therefore could not be considered. Quantitatively, we relate (i) the result-maps $\hat{M}$ to both organ, and ground truth (GT) annotations, and (ii) to each other. Particularly for (i) we studied the Hausdorff distances H between GT and $\hat{M}$ indicating location proximity. Lower values demonstrate better localization in respect to the pathology. Further, we performed a weak localization experiment [8,9]: per image, we derived bounding boxes (BB) for each connected component of GT and $\hat{M}$ attributions. A GT BB counts as found, if any $\hat{M}$ BB has an $IOU \leq 0.125$. We chose this threshold, as a proficient classifier presumably focuses on the masses' boundaries and neighborhoods, thereby limiting possible BB-overlap. We report average localization L. For (ii) we derived the area ratio A between $\hat{M}$ and organ-mask (breast-area) or whole image (chest X-ray). Again, lower values indicate a smaller thereby clearer map. Due to missing GT we could only derive (ii) for TBC. All measurements were performed on binary masks, hence GradCAM and SAL had to be thresholded. We chose the $50, 75, 90$ percentiles, i.e. compared $50, 25, 10$ percent of the map-points. Where multiple pathologies, or mapping results occurred we
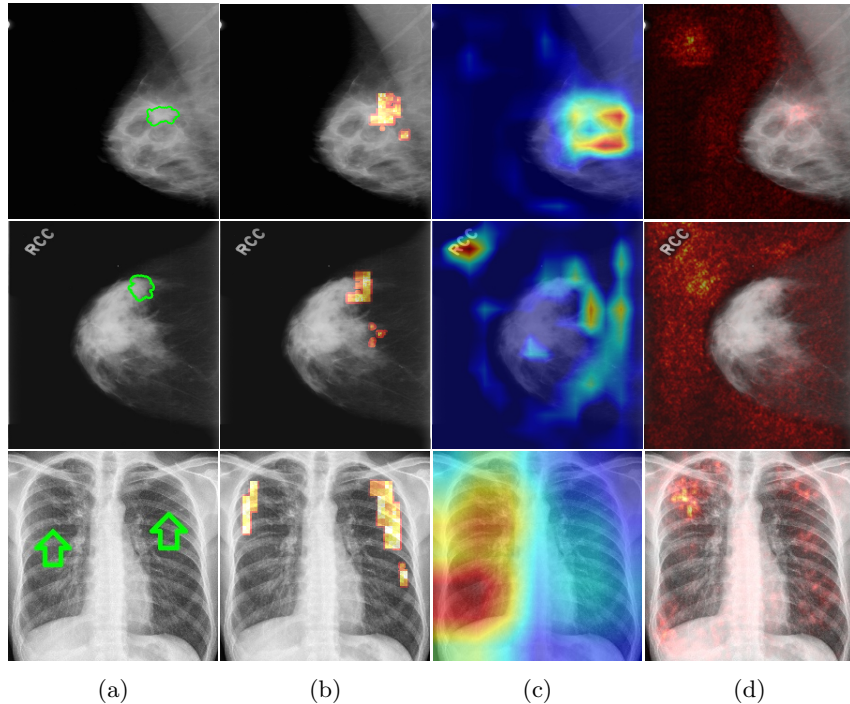
| (a) | (b) | (c) | (d) |

Fig. 3: Result attribution heatmaps for mammography [17] and chest X-ray [14]: (a) original image overlayed with annotation contours (and arrows for missing GT), (b) our attribution framework. (c) GradCAM [25] (d) Saliency [26].

used the median for a robust estimation per image. Statistically significant difference between all resulting findings was formalized using Wilcoxon signed-rank tests, for $\alpha < 0.05$. Additionally we followed [2], and tested our network with randomised parametrization (labels have no effect in our case).

As seen in Table 1, our framework achieves significantly lower H, than either GradCAM or SAL at all threshold levels. Moreover, we report significantly better weak localization (L) which underlines the higher accuracy of our approach. Qualitatively our attribution-maps are tighter focused (c.f. Fig. 3(b)) and enclose the masses. The former is also expressed by the lower overlap values A. All p-values where significantly below 1e-2, hardening our results. Randomization of the ANN's weights yields pure noise maps, hence we pass [2]'s checks.

Timing: We estimated the time needed for a single attribution map, one forward pass, by averaging over ten times repeated map derivations for all images of the resp. test sets. These were compared with the analogous timings of GRAD and SAL. Additionally, as a reference for iterative methods, we compared with [22] that, using same marginalization technique, yields equivalent maps.

Our model is capable of deriving 75 mammography maps per second (mps) utilizing a GPU (NVIDIA Titan RTX). This compares favourably to both GRAD

P	H_{ours}	H_{grad}	H_{sal}	L_{ours}	L_{grad}	L_{sal}
50	**188.12**±68.3	296.29±54.4	240.83±36.2	**0.45**	0.06	0.27
75	**188.12**±68.3	274.86±40.0	257.85±38.6	**0.45**	0.23	0.30
90	**188.12**±68.3	243.80±59.6	259.57±43.7	**0.45**	0.28	0.25

P	A_{ours}^{mammo}	A_{grad}^{mammo}	A_{sal}^{mammo}	A_{ours}^{tbc}	A_{grad}^{tbc}	A_{sal}^{tbc}
50	**0.07**±0.04	1.10±0.10	1.10±.14	**0.06**±0.0	0.50±0.0	0.50±0.0
75	**0.07**±0.04	0.55±0.21	0.55±0.2	**0.06**±0.0	0.25±0.0	0.25±0.0
90	**0.07**±0.04	0.22±0.40	0.22±0.43	**0.06**±0.0	0.10±0.0	0.10±0.0

Table 1: Top: Hausdorff distances H and weak localization results L, relating maps $\hat{M}$ to GT ; Bottom: relating maps $\hat{M}$ to the organ resp. image-size

and SAL, 50 resp. 31 mps, and significantly outperforms the iterative method (27 seconds per map). Considering the smaller X-ray images, these throughputs increase up to a factor of three, sufficient even for real time environments.

Conclusion: In this work, we proposed a novel neural network based attribution method for real time interpretation of pathology classifiers. Our reference based approach enforces domain aware marginalization, without sacrificing computational efficiency. Overcoming these common obstacles, our approach can provide further confidence, and thereby increase critical user acceptance. We compared our method with state-of-the-art techniques on two different tasks, and show favorable results throughout. This underlines the suitability of our approach as an interpretation tool in radiology workflows.

References

1. Adadi, A., Berrada, M.: Peeking inside the black-box: A survey on explainable artificial intelligence (xai). IEEE Access **6**, 52138–52160 (2018)
2. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. In: Proceedings of NIPS. pp. 9505–9515 (2018)
3. Andermatt, S., Horváth, A., Pezold, S., Cattin, P.: Pathology segmentation using distributional differences to images of healthy origin. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. LNCS, vol. 11383, pp. 228–238. Springer (2019)
4. Baumgartner, C., Koch, L., Tezcan, K., Ang, J., Konukoglu, E.: Visual feature attribution using Wasserstein GANs. In: Proceedings of CVPR. pp. 8309–8319 (2017)
5. Becker, A., Jendele, L., Skopek, O., Berger, N., Ghafoor, S., Marcon, M., Konukoglu, E.: Injecting and removing suspicious features in breast imaging with CycleGAN: A pilot study of automated adversarial attacks using neural networks on small images. European Journal of Radiology **120**, 108649 (2019)
6. Bermudez, C., Plassard, A., Davis, L., Newton, A., Resnick, S., Landman, B.: Learning implicit brain MRI manifolds with deep learning. In: SPIE Medical Imaging. vol. 10574, pp. 408–414 (2018)
7. Chang, C.H., Creager, E., Goldenberg, A., Duvenaud, D.: Explaining image classifiers by counterfactual generation. In: Proceedings of ICLR (2019)

8. Dabkowski, P., Gal, Y.: Real time image saliency for black box classifiers. In: Proceedings of NIPS. pp. 6967–6976 (2017)

9. Fong, R., Patrick, M., Vedaldi, A.: Understanding deep networks via extremal perturbations and smooth masks. In: Proceedings of ICCV (2019)

10. Heath, M., Bowyer, K., Kopans, D., Moore, R., Kegelmeyer, W.: The digital database for screening mammography. In: Proceedings of IWDM. pp. 212–218 (2000)

11. Howard, A., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)

12. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of CVPR (2017)

13. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R.L., Shpanskaya, K.S., Seekins, J., Mong, D.A., Halabi, S.S., Sandberg, J.K., Jones, R., Larson, D.B., Langlotz, C.P., Patel, B.N., Lungren, M.P., Ng, A.Y.: CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. AAAI (2019)

14. Jaeger, S., Candemir, S., Antani, s., Wáng, Y., Lu, P., Thoma, G.: Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. Quantitative Imaging in Medicine and Surgery **4**(6) (2014)

15. Kervadec, H., Dolz, J., Tang, M., Granger, E., Boykov, Y., Ayed, I.: Constrained-CNN losses for weakly supervised segmentation. Medical Image Analysis **54**, 88–99 (2019)

16. Lange, M., Zühlke, D., Holz, O., Villmann, T.: Applications of lp-norms and their smooth approximations for gradient based learning vector quantization. In: Proceedings of ESANN (2014)

17. Lee, R., Gimenez, F., Hoogi, A., Rubin, D.: Curated breast imaging subset of DDSM. The Cancer Imaging Archive **8** (2016)

18. Lipton, Z.: The mythos of model interpretability. ACM Queue **16**(3), 30:31–30:57 (2018)

19. Litjens, G., Kooi, T., Bejnordi, B., Setio, A., Ciompi, F., Ghafoorian, M., van der Laak, J., van Ginneken, B., Sánchez, C.: A survey on deep learning in medical image analysis. Medical Image Analysis **42**, 60–88 (2017)

20. Liu, G., Reda, F., Shih, K., Wang, T.C., Tao, A., Catanzaro, B.: Image inpainting for irregular holes using partial convolutions. In: Proceedings of ECCV. pp. 85–100 (2018)

21. Lombrozo, T.: The structure and function of explanations. Trends in Cognitive Sciences **10**(10), 464–70 (2006)

22. Major, D., Lenis, D., Wimmer, M., Sluiter, G., Berg, A., Bühler, K.: Interpreting medical image classifiers by optimization based counterfactual impact analysis. In: Proceedings of ISBI (2020)

23. Peng, J., Kervadec, H., Dolz, J., Ayed, I.B., Pedersoli, M., Desrosiers, C.: Discretely-constrained deep network for weakly supervised segmentation (2019)

24. Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., Rueckert, D.: Attention gated networks: Learning to leverage salient regions in medical images. Medical Image Analysis **53**, 197 – 207 (2019)

25. 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 ICCV (2017)

26. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)
27. Smith, L.N.: No more pesky learning rate guessing games. CoRR (2015)
28. Uzunova, H., Ehrhardt, J., Kepp, T., Handels, H.: Interpretable explanations of black box classifiers applied on medical images by meaningful perturbations using variational autoencoders. In: SPIE Medical Imaging. vol. 10949, pp. 264–271 (2019)
29. Zeiler, M., Fergus, R.: Visualizing and understanding convolutional networks. In: Proceedings of ECCV. LNCS, vol. 8689, pp. 818–833. Springer (2014)
30. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of CVPR. pp. 2921–2929 (2016)
31. Zintgraf, L., Cohen, T., Adel, T., Welling, M.: Visualizing deep neural network decisions: Prediction difference analysis. In: Proceedings of ICLR (2017)