Revisiting Trustworthiness of Saliency Methods in Radiology AI

Jiajin Zhang, Hanqing Chao, Giridhar Dasegowda, Ge Wang, Mannudeep K. Kalra, Pingkun Yan

From the Department of Biomedical Engineering, Center for Biotechnology and Interdisciplinary Studies, Rensselaer Polytechnic Institute, 110 8th St, Biotech 4231, Troy, NY 12180 (J.Z., H.C., G.W., P.Y.); and Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, Mass (G.D., M.K.K.).

Abstract

Purpose: To determine if saliency maps in radiology artificial intelligence (AI) are vulnerable to subtle perturbations of the input, which could potentially lead to misleading interpretations, using Prediction-Saliency Correlation (PSC) for evaluating the sensitivity and robustness of saliency methods.

Materials and Methods: In this retrospective study, locally trained deep learning models and a research prototype provided by a commercial vender were systematically evaluated on 191,229 chest radiographs from the CheXpert dataset(1,2) and 7,022 MRI images of human brain tumor classification dataset(3). Two radiologists performed a reader study on 270 chest radiographs pairs. A model-agnostic approach for computing the PSC coefficient was used to evaluate the sensitivity and robustness of seven commonly used saliency methods.

Results: Leveraging locally trained model parameters, we revealed the saliency methods' low sensitivity (maximum PSC = 0.25, 95% CI: 0.12, 0.38) and weak robustness (maximum PSC = 0.12, 95% CI: 0.0, 0.25) on the CheXpert dataset. Without model specifics, we also showed that the saliency maps from a commercial prototype could be irrelevant to the model output (area under the receiver operating characteristic curve dropped by 8.6% without affecting the saliency map). The human observer studies confirmed that is difficult for experts to identify the perturbed images, who had less than 44.8% correctness.

Conclusion: Popular saliency methods scored low PSC values on the two datasets of perturbed chest radiographs, indicating weak sensitivity and robustness. The proposed PSC metric provides a valuable quantification tool for validating the trustworthiness of medical AI explainability.

Abbreviations: AI = artificial intelligence, PSC = prediction-saliency correlation, AUC = area under the receiver operating characteristic curve, SSIM = structural similarity index measure.

Summary: Systematic evaluation of saliency methods through subtle perturbations in chest radiographs and brain MRI images demonstrated low sensitivity and robustness of those methods, warranting caution when using saliency methods that may misrepresent changes in AI model prediction.

Key Points:

- A novel evaluation metric, prediction-saliency correlation (PSC), is proposed to systematically quantify the trustworthiness of saliency-based AI explainability.
- The results revealed that the low sensitivity (PSC ≤ 0.25) and weak robustness (PSC ≤ 0.12) of the commonly used saliency methods.
- The findings suggest that the popular saliency maps may misrepresent true model prediction, and thus AI researchers and users should be aware of the vulnerabilities of saliency maps in radiology AI.

Keywords: *Saliency Maps, AI Trustworthiness, Dynamic Consistency, Sensitivity, Robustness*.

Introduction

Explainability is a pillar in supporting applications of artificial intelligence (AI) and machine learning in medicine(4–7). Understanding how and why AI models make particular decisions is critical for building trust in AI-driven applications(4,5,8–10). The predicted value of a disease by an AI model implies the probability of disease presence. However, those values usually do not correlate with the disease probabilities and lack confidence and prediction intervals. Thus, a series of post-hoc explanation approaches have been proposed. Prior research has reported on how AI models work by visualizing the relevant contribution of contribution of each feature to the overall model prediction result(11–17). While researchers and clinicians appreciate the development of explainable AI, it is unclear if the resultant explanations can be trusted. Overlooking the trustworthiness of AI-based saliency methods leaves potential risks to AI-based medical applications.

Saliency maps, also commonly referred to as heat maps, are the most commonly used method for AI explainability(12,16). They are especially important for AI algorithms that target image segmentation, quantification tasks, lesion detection and characterization. When reviewing the AI outputs, radiologists often review these maps to accept or reject AI output findings. Previous empirical works(18,19) demonstrate susceptibility of neural networks to small perturbations in normal inputs, resulting in wrong outputs. Along the same direction, Arun et al.(20) attempted to assess the saliency maps in medical imaging by quantifying their localization capability, variation between randomized networks, repeatability over separately trained models, and reproducibility across different models. However, their methods only demonstrate the generic properties of the saliency approaches but cannot evaluate either the correlation between the saliency maps with the model predictions or the saliency visualization quality for a given specific AI model. Theoretical works(21,22) suggest that such weakness is related to neural networks' lack of local Lipschitz smoothness with respect to the input image space. As a negative result, model outcomes may vary drastically around the original data. There is a lack of comprehensive evidence on how such vulnerabilities in each AI model can substantially limit its practical application. The potential risk associated with those vulnerabilities is underestimated in medical AI, since the model details, such as architecture

and parameters, are usually safeguarded. In addition, previous methods demonstrate only the generic properties of the saliency approaches and cannot evaluate the saliency visualization quality for a specific AI model.

A trustworthy saliency approach should meet two general conditions: **1) Sensitivity.** A saliency map should change accordingly when the model's prediction for that image substantially alters due to the input change; **2) Robustness.** A saliency map should stay consistent when the model's prediction remains unchanged after an input image is randomly transformed without impacting the image content. In other words, a trustworthy saliency map should be consistent with the model prediction, not just for a specific example at a given state, but dynamically consistent when the model prediction changes.

In this study, we sought to determine if saliency maps in radiology AI are vulnerable to subtle perturbations in their input that can lead to misleading results. We present a novel systematic approach for quantifying the trustworthiness of saliency explanations of given medical AI models. Specifically, we propose a model-agnostic and generalizable measurement to quantitatively analyze saliency method robustness called prediction-saliency correlation (PSC), which depicts the correlation between changes in model predictions and changes in the corresponding saliency maps, to quantitatively analyze both the robustness and sensitivity. We then illustrate the uses of this approach on commonly employed AI models and saliency map methods.

Materials and Methods

Study Design

This is a retrospective study for quantifying the trustworthiness of the most popular explanation methods in radiological AI. This study was exempt from institutional review board approval

and Health Insurance Portability and Accountability Act-compliant because fully de-identified public datasets were used. Seven of the most commonly used saliency methods in medical AI applications were selected for the investigation, including Vanilla Back Propagation (Vanilla BP)(23), Vanilla BP × Image(12), Gradient-weighted Class Activation Mapping (GradCAM)(14), Guided-GradCAM(14), Integrated Gradients (IG)(15), Smoothed Gradients (SG)(16), and eXplanation with Ranked Area Integrals (XRAI)(17)(23). A representative list of medical AI papers using the above saliency methods are presented in **Supplementary Table S0**. We trained two widely used networks (Resnet-152 and DenseNet-121) as our baseline models tasked to identify atelectasis, cardiomegaly, consolidation, edema, and pleural effusion on chest radiographs. We quantitatively verified the sensitivity and robustness on the two most commonly used convolutional neural networks (CNNs) in medical image classifications, DenseNet-121, and ResNet-152, trained on the CheXpert dataset. In addition, we processed the chest radiographs with and without perturbation with an AI-based chest radiograph research prototype, where we had no access to the model architecture and parameters. This is to test whether the proposed method can be generalized to such 'black-box' situations. To demonstrate the efficacy of our proposed evaluation framework on other imaging modalities, analysis tasks and deep learning model architectures, additional experiments were performed with a ResNet-50 model trained on a brain tumor multi-class classification MRI dataset(3). A preliminary version of this work was presented at MICCAI 2022(24). Compared with our previous MICCAI presentation, this manuscript under consideration includes significant extensions, such as additional technical innovations, numerical experiments, human observer studies, and mathematical analysis.

Dataset Preparation

We demonstrate the discovered issues of saliency maps on a multi-label classification task using a chest radiograph dataset, CheXpert[1]. The original dataset consists of 223,648 publicly available chest radiographs (including both frontal and lateral projections) from 64,740 patients (40.6% female, mean age 59.6± [SD] 16.8 years; 59.4% male, mean age 58.6±16.3 years).

Only 191,229 frontal chest radiographs (191,027 from the original training set and 202 from the original validation set) were included in our study. Since the test set of the CheXpert dataset is not publicly available, we further randomly split the original training set including 191,027 frontal images into training and validation sets with a ratio of 6:1. The 202 frontal chest radiographs of the original validation set were used as our test set. In addition, we also include a human brain tumor MRI classification dataset(3) which consists of 7,022 images. More detailed information for the brain tumor MRI dataset is provided in **Supplementary Sec. V–1. Dataset and Model Preparation.**

Quantitative Analysis of Trustworthiness

As shown in **Fig. 1**, we evaluate the dynamic consistency from two aspects, i.e., the *sensitivity* and the *robustness*. For each image $\boldsymbol{x}_i$ in a set of test images with size of N, we first obtain its prediction $\boldsymbol{p}_i$ and the saliency map $\mathbf{m}_i$ with an AI model. We then alter each image $\boldsymbol{x}_i$ to produce a new image $\boldsymbol{x}_i'$. The specific alteration depends on which property is being examined, as detailed below in this section. The model will calculate a new prediction $\boldsymbol{p}_i'$ and generate a corresponding saliency map $\mathbf{m}_i'$ for the new image.

To evaluate sensitivity of each saliency method, we observed whether changes in model predictions due to alteration of input images resulted in corresponding changes to the saliency maps. Specifically, as shown in the left of **Fig. 1**, by adopting the optimization techniques from adversarial attacks(18,25–29), we identified the slightly perturbed radiographs that caused the AI model to predict a different result but had saliency maps close to those of the original input. The word "perturb" means making very small changes to the pixel values of an image, *i.e.*, an original radiograph in our study. Such small changes are usually imperceptible to human but can lead to output change of an AI model. For each of the five observations, every radiograph is perturbed such that the model prediction will be flipped, *i.e.*, from 'observation exists' to 'no observation' and vice versa. In the meanwhile, the optimization algorithm keeps the saliency map unchanged. The technical details of the image perturbation algorithm are presented in **Supplementary Sec. II. Sensitivity Examination**.

The robustness of dynamic consistency demands the saliency maps to remain consistent when the predictions do not change. Robustness evaluates if a saliency map can stay consistent with the model's prediction output when randomly perturbing an image without impacting the model prediction. As shown in the right part of **Fig. 1**, we use similar optimization techniques as in the above sensitivity experiments to investigate if it is possible to pull a saliency map towards an arbitrary pattern while keeping the model predictions unaffected. The target pattern was designed as a square at the top right corner of the saliency map. More specifically, for each of the five observations, we perturbed the input radiograph to distract the saliency map from the pre-defined target square pattern. Meanwhile, the model prediction of the radiograph remains unchanged by optimizing the perturbation. To quantify the findings, we evaluated changes in model performance and the similarities of the saliency maps generated on the perturbed images to the original saliency maps and the target saliency map, respectively. More detailed mathematical derivations are presented in **Supplementary Sec. II. Robustness Examination**.

Statistical Analysis

The sensitivity and robustness of saliency methods are uniformly quantified by the prediction-saliency correlation (PSC) coefficient proposed in this study. This coefficient is defined by the Pearson correlation between variations in model predictions and changes in their corresponding saliency maps, both of which are gauged using the Jensen–Shannon divergence. The PSC coefficient ranges from -1 to +1: a -1 value signifies a perfectly negative correlation, 0 suggests no correlation, and +1 denotes a perfectly positive correlation. A PSC value above 0.5 is considered of having a high degree of correlation. We used the changes of the area under the receiver operating characteristic curve (AUC) to evaluate the changes of the model predictions. The structural similarity index measure (SSIM) was used to quantify the changes of saliency maps. The average PSC of the five findings was used to quantify the overall performance of each saliency method. Mathematical derivations of PSC are presented in **Supplementary Sec. II. Prediction-Saliency Correlation**. The significance tests for AUC comparison were performed using the z-test as detailed by Zhou et al(30). while the confidence intervals for the

AUC values were computed based on the methodology proposed by Hanley(31). The significance of each individual finding is evaluated, and the p values ($p < 0.05$ indicated a statistically significant difference) are reported in the **Supplementary Section III**. We reported the averaged performance over all findings in the main paper, which has no associated p values.

Human Observer Study

Two radiologists (M.K.K. with 15 years of experience in thoracic imaging and G.D. with 2 years of post-doctoral experience in thoracic imaging) were presented with 270 pairs of perturbed and original chest radiographs from the CheXpert dataset. These pairs consisted of 120 pairs from the sensitivity experiments in Table 1 and 150 pairs from the robustness experiments in Table 2. Neither abnormality nor pathology was inserted or overlaid on the altered radiographs. The order of altered and original radiographs was randomly assigned for each pair. The radiologists were asked to identify the perturbated image from the pair. The two radiologists were first provided another 150 pairs of images for training, where the images were clearly labelled. Both radiologists separately and independently assessed the radiographs.

Data availability

The chest radiograph datasets used in this study are available in the Stanford CheXpert database under accession code https://stanfordmlgroup.github.io/competitions/chexpert. All data needed to evaluate the findings in the paper are presented in the paper and/or the supplementary material. Additional data related to this paper, such as the detailed reader test data, may be requested from the authors.

Results

Sensitivity Examination

Fig. 2 shows a chest radiograph with atelectasis as an example to demonstrate how the predictions and the saliency maps may diverge from each other. Although the perturbed images look identical to the original image, the probabilities of atelectasis predicted by the model (DenseNet-121) dropped from 67.1% to 2%. The highlighted regions of saliency maps for the perturbed images were similar to those for the original images (SSIM > 0.76), suggesting that the saliency maps failed to reflect the changes of the model predictions, i.e., the sensitivity may be low. More results of each individual class are included in **Supplementary III. 1**.

Table 1 shows the quantification results. All the reported numbers are the averages over the five classes. The overall averaged PSC was no greater than 0.26. Given that the value of PSC was in the range of [-1,1], this is considered as a weak association. We further examined the details. On the perturbed images, the AUCs of both models degraded drastically from 0.88 to 0.01. However, the corresponding saliency maps for the perturbed images were similar to the saliency maps for the original images, with a mean SSIM $\geq$ 0.76. Such inconsistency between the prediction variation and the saliency preservation is reflected by the small PSC $\leq$ 0.25. We also performed similar sensitivity evaluation on the MRI classification dataset and reached similar conclusions. These experimental results are reported in **Supplementary Sec. V – Sensitivity Examination**.

Further, we examined the sensitivity of a research prototype model provided by a commercial vendor, henceforth referred to as the "commercial prototype" for brevity. Since we do not have access to the architecture and parameters of the model, we generate altered images based on three in-house models that we trained and fed these images to the model. The saliency maps were generated by the commercial prototype itself. To avoid conflict of interest, we

visualized the saliency maps using our colormap. The corresponding results are presented in **Fig. 3**. Four cases of the four shared classes (atelectasis, cardiomegaly, consolidation, and pleural effusion) between our pretrained local proxy model trained on CheXpert and the commercial prototype are shown. The results indicate that the generated perturbed images do not substantially change the model saliency maps. The quantitative results also support this observation, as the similarities (SSIM) between the saliency maps on the perturbed images and the original ones are greater than 0.88 on all four classes. However, the perturbed images caused the average AUC to drop by 8.6% ($p<0.01$) on the four classes.

Robustness Examination

An example case is presented in **Fig. 4** (Results of each individual class are included in **Supplementary III. 3**). The DenseNet-121 predictions for probability of atelectasis, using perturbed images were consistent with the original prediction (58.1%). However, the perturbed images successfully misled the saliency maps. For all the saliency methods, the perturbed images shifted the saliency areas toward the targeted top right corner of the image.

Table 2 indicates that the deep neural networks performed consistently on the original and perturbed images, with the AUC remaining the same at 0.88. However, the saliency maps on the perturbed images were dramatically different from the original saliency maps (mean $SSIM_{org} \leq 0.51$). Such inconsistency between the prediction variation and the saliency preservation is reflected by the small $PSC \leq 0.12$. At the same time, the saliency maps from the perturbed images share strong similarities with the target saliency map, with mean $SSIM_{tgt} = 0.65$ on GradCAM and mean $SSIM_{tgt} \geq 0.82$ on all other saliency methods. We also performed similar robustness evaluation on the MRI classification dataset and reached similar conclusions. These experimental results are reported in **Supplementary Sec. V Robustness Examination.**

Human Observer Study

At the time of testing, the two physicians correctly pointed out 63/270 (23.3% for GD [2 years

of experience]) and 121/270 (44.8% for MKK [15 years of experience]) altered radiographs, respectively, indicating that the alterations of the image alterations are difficult to spot by human experts, even for a thoracic radiologist with 15 years of subspecialty experience.

Discussion

In this study, we introduce a novel assessment metric, namely the PSC coefficient, to provide an intuitive and quantitative evaluation of the trustworthiness of widely used saliency maps. The PSC coefficient can serve as an evaluator to quantify the sensitivity and robustness of explanation methods. The quantitative and qualitative results in the **Results** section show that commonly adopted saliency methods in medical AI applications can produce misleading interpretations. The saliency methods demonstrated low sensitivity (PSC<0.25) and robustness (PSC<0.12) on multiple radiographs. All the findings suggest that either the predictions or the saliency maps of the models have undergone tremendous changes. In addition, the saliency maps generated by the commercial AI software may be neither relevant nor robust to perturbation added to the images without knowing the model specifics. The human observer studies verified that the perturbed images are difficult to identify even by one human expert with extensive experience in chest radiography. These results indicate that for deep learning models, the sensitivity and robustness of all the seven saliency methods was weak, i.e., *the generated saliency maps may not be relevant to the model predictions*. Notably, the radiologist with 15 years of clinical experience demonstrated a substantially stronger ability to correctly identify altered radiographs compared to the other physician with only 2 years of experience. This observation suggests a potential correlation between the clinical expertise and the capacity to discern subtle perturbations.

To this end, we proposed a model-agnostic method for saliency map trustworthiness evaluation, which is generalizable to the commonly available saliency methods. Our method of induced perturbation can help establish the trustworthiness of the explainability of AI outputs and assess individual or multiple AI models for susceptibility to similar or different types of perturbation. This is a major clinical implication of our work. Our findings in **Supplementary III. 2 and III. 4** indicate that even if multiple saliency methods were applied

and we obtained consistent results, it is still possible that none of the saliency maps is faithful to the model predictions. These evaluation results on the commercial prototype suggest that the concerns about the trustworthiness are valid even for commercial AI systems trained with tremendous data. Our future work will also focus on refining the extent, distribution, and patterns of perturbations to simulate variations in patients, diseases, and imaging parameters. Such work can help reduce the cost and time needed for thorough validation of AI models and uncover the implications of deploying non-generalizable or non-explainable AI models.

Many clinical end-users of AI might not be aware of the explainability aspects of AI; therefore, these aspects are often not realized and/or applied in clinical practices. Few AI models have safety valves where certain AI outputs are not generated in the presence of issues related to AI explainability. For example, AI models should exercise caution or not describe cardiothoracic ratio or presence of enlarged cardiac silhouette on portable, supine radiographs as opposed to an upright posterior-anterior radiographs. On CT, variations in reconstructed section thickness can profoundly influence AI-based estimation of nodule size, growth, and attenuation characteristics over serial CT examinations. A lack of explanation on if and how the AI model accounts for such variations in acquisition technique and measured findings can mislead the clinical end-users or cause them to discard all AI outputs. The addition and awareness of explainability aspects to the models can thus help improve adoption and proper use of AI algorithm outputs. Such explainability would be especially helpful given the profound variations in patient factors (supine versus upright radiographs or radiographs with low lung volumes), acquisition factors (radiation dose and image quality, including artifacts) as well as image reconstruction techniques (differences in section thickness and kernels).

Our work has limitations. The evaluation presented in this work mainly focused on the most popular attribution-based AI explanation methods(32,33). However, there exist other explanation techniques, such as counterfactual explanations(33,34). Our future work will extend the gradient-based evaluation to the counterfactual-based explanation methods.

In conclusion, we proposed a model-agnostic method to dynamically evaluate the trustworthiness of saliency maps used for explaining the results of AI models. Our findings suggest that the commonly used saliency methods in medical AI can produce interpretations inconsistent with the model predictions. Thus, it is important to establish the trustworthiness

of the saliency methods in clinical adoption of AI models. Furthermore, our future work will extend the evaluation from gradient-based methods to the counterfactual explanation methods to determine their trustworthiness.

References

1. Irvin J, et al. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. Proc AAAI Conf Artif Intell. 2019;33:590-597. **doi**: 10.1609/aaai.v33i01.3301590. Published July 17, 2019. Accessed July 17, 2019.

2. Garbin C, Rajpurkar P, Irvin J, Lungren MP, Marques O. Structured dataset documentation: a datasheet for CheXpert. **doi**: arXive:2105.03020, Published May 5, 2021. Accessed May 5, 2021.

3. Msoud Nickparvar. Brain Tumor MRI Dataset. **doi**:10.34740/KAGGLE/DSV/2645886. Published May 5, 2021. Accessed May 5, 2021.

4. Shen Y, et al. Artificial intelligence system reduces false-positive findings in the interpretation of breast ultrasound exams. Nat Commun. 2021;12:5645. **doi**:10.1038/s41467-021-26023-2, Published September 14, 2021. Accessed September 24, 2021.

5. DeGrave AJ, Janizek JD, Lee S-I. AI for radiographic COVID-19 detection selects shortcuts over signal. Nat Mach Intell. 2021;3:610-619. **doi**:10.1038/s42256-021-00338-7, Published May 31, 2021. Accessed July 11, 2021.

6. Arnaout R, et al. An ensemble of neural networks provides expert-level prenatal detection of complex congenital heart disease. Nat Med. 2021;27:882-891. **doi**: 10.1038/s41591-021-01342-5, Published April 8, 2021. Accessed May 14, 2021.

7. Abitbol JL, Karsai M. Interpretable socioeconomic status inference from aerial imagery through urban patterns. Nat Mach Intell. 2020;2:684-692. **doi**: 10.1038/s42256-020-00243-5, Published September 25, 2020. Accessed October 26, 2020.

8. Gonzalez-Gonzalo C, Liefers B, van Ginneken B, Sanchez CI. Iterative Augmentation of Visual Evidence for Weakly-Supervised Lesion Localization in Deep Interpretability Frameworks: Application to Color Fundus Images. IEEE Trans Med Imaging. 2020;39:3499-3511. **doi**: 10.1109/TMI.2020.2994463, Published May 28, 2020. Accessed November 6, 2020.

9. Mitani A, et al. Detection of anaemia from retinal fundus images via deep learning. Nat Biomed Eng. 2020;4:18-27. **doi**: 10.1038/s41551-019-0487-z, Published December 23, 2019. Accessed January 1, 2020.

10. Sayres R, et al. Using a Deep Learning Algorithm and Integrated Gradients Explanation to Assist Grading for Diabetic Retinopathy. Ophthalmology. 2019;126:552-564. **doi**: https://doi.org/10.1016/j.ophtha.2018.11.016, Published November 14, 2018. Accessed December 13, 2018.

11. Ribeiro MT, Singh S, Guestrin C. 'Why Should I Trust You?': Explaining the Predictions

of Any Classifier. In: Proceedings of the 22nd ACM SIGKDD Int Conf on Knowl discovery and data mining, 2016;22:1135-1144, **doi**: 10.18653/v1/N16-3020, Published June, 2016. Accessed June, 2016.

12. Shrikumar A, Greenside P, Kundaje A. Learning Important Features Through Propagating Activation Differences. In: Proceedings of the Int Conf on Mach Learn (ICML), 2019;2145-3153, **doi**: 10.5555/3305890.3306006, Published August 6, 2017. Accessed August 6, 2017.

13. Shrikumar A, Greenside P, Shcherbina A, Kundaje A. Not Just a Black Box: Learning Important Features Through Propagating Activation Differences. **doi**: arXiv:1605.01713, Published August 6, 2017. Accessed August 6, 2017.

14. Selvaraju RR, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE Int Conf on Computer Vision (ICCV), 2017;618-626, **doi**: 10.1109/ICCV.2017.74, Published October 22, 2017. Accessed December 25, 2017.

15. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. In: Proceedings of the 34th International Conference on Machine Learning (ICML'17), 2017;70:3319-3328, **doi**: 10.5555/3305890.3306024, Published August 6, 2017. Accessed August 6, 2017.

16. Smilkov D, Thorat N, Kim B, Viégas F, Wattenberg M. SmoothGrad: removing noise by adding noise. **doi**: arXiv:1706.03825, Published June, 2017. Accessed June, 2017.

17. Kapishnikov A, Bolukbasi T, Viégas F, Terry M. XRAI: Better Attributions Through Regions. In: Proceedings of the IEEE Int Conf on Computer Vision (ICCV), 2019;4948-4957, **doi**: 10.1109/ICCV.2019.00505, Published August 6, 2019. Accessed August 6, 2019.

18. Goodfellow IJ, Shlens J, Szegedy C. Explaining and Harnessing Adversarial Examples. **doi**: arXiv:1412.6572, Published March 20, 2015. Accessed March 20, 2015.

19. Kurakin A, Goodfellow IJ, Bengio S. Adversarial Examples in the Physical World. In: Artificial Intelligence Safety and Security. 1st ed. Chapman and Hall/CRC; 2018; 14.

20. Arun N, Gaw N, Singh P, et al. Assessing the Trustworthiness of Saliency Maps for Localizing Abnormalities in Medical Imaging. Radiol Artif Intell. 2021;3(6):e200267. **doi**: 10.1148/ryai.2021200267.

21. Cohen J. M., Rosenfeld E. & Kolter J. Z. Certified Adversarial Robustness via Randomized Smoothing. in Int Conf on Mach Learn, PMLR 97:1310-1320, 2019; **doi**: 10.48550/arXiv.1902.02918, Published June, 2019. Accessed June, 2019.

22. Qin C et al., Adversarial robustness through local linearization. In: Proc of the 33rd Inter Conf on Neural Info Proc Sys (NeurIPS), 2019; 1240:13842–13853, **doi**:

10.5555/3454287.3455527, Published December, 2019. Accessed December, 2019.

23. Simonyan K, Vedaldi A, Zisserman A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. arXiv:13126034 [cs]. 2014; **doi**: http://arxiv.org/abs/1312.6034. Accessed December 19, 2021.

24. Zhang J, Chao H, Dasegowda G, Wang G, Kalra MK, Yan P. Overlooked Trustworthiness of Saliency Maps. In: Wang L, Dou Q, Fletcher PT, Speidel S, Li S, editors. Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. Cham: Springer Nature Switzerland; 2022. p. 451–461. **doi**: 10.1007/978-3-031-16437-8_43. Accessed October 9, 2022.

25. Bortsova G, González-Gonzalo C, Wetstein SC, et al. Adversarial Attack Vulnerability of Medical Image Analysis Systems: Unexplored Factors. Medical Image Analysis. 2021;73:102141. **doi**: 10.1016/j.media.2021.102141. Published June 10, 2021. Accessed June 17, 2021.

26. Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS. Adversarial attacks on medical machine learning. Science. 2019;363(6433):1287–1289. **doi**: 10.1126/science.aaw4399. Accessed March 22 2019.

27. Xu M, Zhang T, Li Z, Liu M, Zhang D. Towards Evaluating the Robustness of Deep Diagnostic Models by Adversarial Attack. Medical Image Analysis. 2021;69:101977. **doi**: 10.1016/j.media.2021.101977. Published January 18, 2021. Accessed January 22, 2021.

28. Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks. arXiv:13126199 [cs]. 2014; **doi**: http://arxiv.org/abs/1312.6199. Accessed May 27, 2021.

29. Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A. Towards Deep Learning Models Resistant to Adversarial Attacks. arXiv:170606083 [cs, stat]. 2019; **doi**: http://arxiv.org/abs/1706.06083. Accessed May 27, 2021.

30. Zhou X-H, Obuchowski NA, McClish DK. Statistical Methods in Diagnostic Medicine. 2nd ed. March 2011;592 Pages. ISBN: 978-0-470-18314-4.

31. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143(1):29–36. **doi**: 10.1148/radiology.143.1.7063747. Accessed April, 1982.

32. Singh A, Sengupta S, Lakshminarayanan V. Explainable Deep Learning Models in Medical Image Analysis. Journal of Imaging. Multidisciplinary Digital Publishing Institute; 2020;6(6):52. **doi**: 10.3390/jimaging6060052. Published June 17, 2020. Accessed June 20, 2020.

33. Goyal Y, Wu Z, Ernst J, Batra D, Parikh D, Lee S. Counterfactual Visual Explanations. arXiv; 2019. **doi**: http://arxiv.org/abs/1904.07451. Accessed February 3, 2023.

34. Atad M, Dmytrenko V, Li Y, et al. CheXplaining in Style: Counterfactual Explanations for Chest X-rays using StyleGAN. arXiv; 2022. **doi**: http://arxiv.org/abs/2207.07553. Accessed February 3, 2023.

Tables and Figures

Table 1. Quantification Results of Saliency Sensitivity on ResNet-152 and DenseNet-121

Model	Evaluation metrics	Vanilla BP	Vanilla BP × Image	GradCAM	Guided GradCAM	IG	SG	XRAI
ResNet-152	**AUC on the perturbed radiographs** (AUC_{origin} = 0.88)	0.03	0.05	0.05	0.05	0.02	0.02	0.02
	Saliency map similarity (SSIM)	0.78	0.79	0.86	0.88	0.84	0.87	0.87
	PSC	0.21	0.22	0.25	0.23	0.04	0.13	0.12
DenseNet-121	**AUC on the perturbed radiographs** (AUC_{origin} = 0.88)	0.02	0.04	0.04	0.04	0.04	0.01	0.04
	Saliency map similarity (SSIM)	0.77	0.78	0.76	0.78	0.88	0.83	0.78
	PSC	0.25	0.26	0.25	0.21	0.25	0.02	0.26

Note.—All values in the table are averages over the five classes, including atelectasis, cardiomegaly, consolidation, edema, and pleural effusion. AUC = area under the receiver operating characteristic curve, PSC = prediction-saliency correlation, SSIM = structural similarity index measure, Vanilla BP = vanilla back propagation, Vanilla BP × Image = vanilla back propagation times image, GradCAM = gradient-weighted class activation mapping, IG = integrated gradients, SG = smoothed gradients, XRAI = eXplanation with Ranked Area Integrals.

Table 2. Quantification Results of Saliency Robustness on ResNet-152 and DenseNet-121.

Models			Vanilla BP	Vanilla BP × Image	GradCAM	Guided GradCAM	IG	SG	XRAI
ResNet-152	**AUC on the perturbed radiographs** (AUC_{origin} = 0.88)		0.88	0.88	0.88	0.88	0.88	0.88	0.88
	Saliency map similarity	$SSIM_{org}$	0.46	0.54	0.32	0.38	0.47	0.51	0.47
		$SSIM_{tgt}$	0.94	0.93	0.65	0.74	0.93	0.82	0.85
	PSC		0.01	0.01	0.04	0.06	0.03	0.06	0.04
DenseNet-121	**AUC on the perturbed radiographs** (AUC_{origin} = 0.88)		0.88	0.88	0.88	0.88	0.88	0.88	0.88
	Saliency map similarity	$SSIM_{org}$	0.47	0.53	0.32	0.40	0.47	0.51	0.48
		$SSIM_{tgt}$	0.88	0.93	0.65	0.89	0.93	0.88	0.83
	PSC		0.01	0.01	0.13	0.12	0.07	0.12	0.09

Note.—All values in the table are averages over the five classes: atelectasis, cardiomegaly, consolidation, edema, and pleural effusion. $SSIM_{org}$ and $SSIM_{tgt}$ denote the SSIM between the new and the original saliency maps and the SSIM between the new and the target saliency maps, respectively. AUC = area under the receiver operating characteristic curve, PSC = prediction-saliency correlation, SSIM = structural similarity index measure, Vanilla BP = vanilla back propagation, Vanilla BP × Image = vanilla back propagation times image, GradCAM = gradient-weighted class activation mapping, IG = integrated gradients, SG = smoothed gradients, XRAI = eXplanation with Ranked Area Integrals.

Fig. 1 Overview of the proposed methods. The trustworthiness of saliency maps can be examined from two aspects: sensitivity and robustness. An adversarial image $\boldsymbol{x_i^p}$ generated by prediction attack examines the sensitivity between saliency map and model output. Another adversarial image $\boldsymbol{x_i^s}$ generated by saliency attack evaluates if saliency maps are resistant to the model output change. In both cases, the adversarial images look no different from the original image. AI = artificial intelligence.

Fig. 2. Saliency maps lacking sensitivity to predictions of DenseNet-121. The color bar indicates the intensity of the saliency maps. Probabilities of atelectasis are shown at the bottom of saliency maps for original and perturbed images. The images show that highly similar saliency maps of frontal chest radiographs may be associated with very different model predictions. Vanilla BP = vanilla back propagation, Vanilla BP × Image = vanilla back propagation times image, GradCAM = gradient-weighted class activation mapping, IG = integrated gradients, SG = smoothed gradients, XRAI = eXplanation with Ranked Area Integrals.

Fig. 3. Saliency sensitivity evaluation of a commercially available artificial intelligence (AI) software. The color bar indicates the intensity of the saliency maps. By perturbating the original chest radiographs (top row), perturbed images (bottom row) are generated via attacking a proxy model. The perturbed images were then fed to a commercially available medical AI model. Note the large variations of the predicted probabilities (at the bottom of each image) from the original to the perturbed images on different findings, despite only minor changes to the saliency maps.

Fig. 4. Example saliency maps lacking robustness to saliency tampering of ResNet-152. The color bar indicates the intensity of the saliency maps. The target region of saliency maps on chest radiographs have been manipulated (third row), but the predicted probability (at the bottom of each image) of atelectasis remains similar to the original prediction. Vanilla BP = vanilla back propagation, Vanilla BP × Image = vanilla back propagation times image, GradCAM = gradient-weighted class activation mapping, IG = integrated gradients, SG = smoothed gradients, XRAI = eXplanation with Ranked Area Integrals.

Competing interests

The authors declare no competing interests.