Assessing the (Un)Trustworthiness of Saliency Maps for Localizing Abnormalities in Medical Imaging

Nishanth Arun[†1,2], Nathan Gaw[†3], Praveer Singh[†1], Ken Chang[†1,4], Mehak Aggarwal[1], Bryan Chen[1,4], Katharina Hoebel[1,4], Sharut Gupta[1], Jay Patel[1,4], Mishka Gidwani[1], Julius Adebayo[4], Matthew D. Li[1], and Jayashree Kalpathy-Cramer[*1]

[1] Athinoula A. Martinos Center for Biomedical Imaging, Department of Radiology, Massachusetts General Hospital, Boston, MA, USA
jkalpathy-cramer@mgh.harvard.edu
https://qtim-lab.github.io/
[2] Shiv Nadar University, Greater Noida, India
[3] ASU-Mayo Clinic Center for Innovative Imaging, School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Tempe, AZ, USA
[4] Massachusetts Institute of Technology, Cambridge, MA, USA

Abstract. Saliency maps have become a widely used method to make deep learning models more interpretable by providing post-hoc explanations of classifiers through identification of the most pertinent areas of the input medical image. They are increasingly being used in medical imaging to provide clinically plausible explanations for the decisions the neural network makes. However, the utility and robustness of these visualization maps has not yet been rigorously examined in the context of medical imaging. We posit that trustworthiness in this context requires 1) localization utility, 2) sensitivity to model weight randomization, 3) repeatability, and 4) reproducibility. Using the localization information available in two large public radiology datasets, we quantify the performance of eight commonly used saliency map approaches for the above criteria using area under the precision-recall curves (AUPRC) and structural similarity index (SSIM), comparing their performance to various baseline measures. Using our framework to quantify the trustworthiness of saliency maps, we show that all eight saliency map techniques fail at least one of the criteria and are, in most cases, less trustworthy when compared to the baselines. We suggest that their usage in the high-risk domain of medical imaging warrants additional scrutiny and recommend that detection or segmentation models be used if localization is the desired output of the network.

Keywords: Saliency maps · localization · deep learning.

Deep learning has brought about many promising applications within medical imaging with recent studies showing its potential for key clinical assessments within dermatology, ophthalmology, pathology, oncology, cardiology, and radiology [1,2,3,4,5,6]. One major class of deep neural networks is convolutional neural networks (CNNs), which take raw pixel values as input, and transform them into the output of interest (such as diagnosis of disease [40] or expected clinical outcome [41]) after passing through many layers of non-linear transforms that provide the necessary capacity for higher levels of abstraction. Many of these fully trained CNNs have outperformed conventional methods for various

[†] These authors contributed equally
[*] Corresponding author

medical tasks [42,43,44]. However many clinicians have been hesitant to adopt such methods in their medical workflows largely because of low interpretability of CNNs [7]. In particular, CNNs have been deemed as "black boxes", contributing to a lack of trust in the medical community. Specifically, without explainability, there is risk that the model may perpetuate social biases, have poor generalization across different datasets, or utilize non-biological and spurious correlations [8,9,10]. There has been a well-recognized need for regulations of deep learning in radiology, and recently more focus has been made on creating protocols and standards that establish a safe and informed fusion of deep learning into healthcare [50].

As CNNs have become increasingly popular for classification of medical images, it has become important to find methods that best explain the decisions of these models to establish trust with clinicians. For radiology, a wrong diagnosis can have serious repercussions for a patient and interpretability is crucial to help clinicians understand a prediction made by a CNN [49]. Explanation of an AI model should have enough transparency to confirm model outputs with medical knowledge and provide practical information that can be utilized by the clinicians to guide patient decision-making and management [34]. A likely future application of CNNs in the clinical setting is to help clinicians flag particular images that need further review [52]. Without measures of interpretability, clinicians would resort to guesswork as to why the image was flagged.

Saliency maps (feature attributions) have become a popular approach for post-hoc interpretability of classification CNNs. These maps are designed to highlight the salient components of medical images that are important to the model's prediction. As a result, many deep learning medical imaging studies have used saliency maps to rationalize model prediction and provide localization [11,12,13]. One study developed CheXNeXt, a deep learning algorithm that has the capability to simultaneously detect 14 pathologies from chest radiographs and use class activation mapping (CAM) [47] to localize the areas of the image that are most representative of the particular disease [11]. Another work created MRNet, a deep learning model trained on magnetic resonance images for diagnosing knee injuries and localizing the abnormality using CAM [12]. A study using deep learning to detect anaemia from retinal fundus images also used localization with Grad-CAM [16], Smooth Integrated Gradients [29,27], and Guided-backprop [30] to highlight regions relevant to the model's predictions [13]. All of these studies provided these saliency maps as legitimate localizers of the particular abnormality of interest.

However, the validity of saliency maps has been called into question. A recent study that evaluated a variety of datasets showed that many popular saliency map approaches are not sensitive to model weight or label randomization [14]. Although there is no study that corroborates these findings with medical images [49], there are several works that have demonstrated serious issues with many saliency methods. One study used 3-D CNNs for Alzheimer's Disease classification with magnetic resonance images, and after testing several different models and saliency maps, found that saliency methods varied in robustness in regards to repeated model training [15]. Another study for detection of pneumothorax images using a VGG19 neural network found that only 33% of GradCAM [16] heatmaps for correctly classified pneumothorax images overlapped the correct regions of the pneumothorax with high probability [17]. Young et al. trained an Inception neural network model to differentiate benign skin lesions from melanoma on dermoscopy images and examined the performance of saliency maps generated by GradCAM and Kernel-SHAP [18]. The study found that models of similar accuracy produced different explanations, and GradCAM would even obscure most of the lesion of interest causing any explanation for melanoma classification to be clinically useless. Additionally, only two studies in the medical domain assessed saliency maps' localization capabilities using some ground-truth measure, such as bounding boxes or semantic segmentation [17,45]. However, in Crosby et al. 2020 there was no quantification of the extent of overlap (utility) of GradCAM with the relevant image regions, but

rather a binary measure of whether or not GradCAM's region of highest activation intersected the pneumothorax region. In addition, there was no comprehensive analysis between different types of saliency maps. The other study was preliminary work performed by our group [45].

In this study, we substantially extend the above work by comprehensively evaluating popular saliency map methods for medical imaging classification models trained on the SIIM-ACR Pneumothorax Segmentation and RSNA Pneumonia Detection datasets [19,20] in terms of 4 key criteria for trustworthiness:

1. Utility
2. Sensitivity to weight randomization
3. Repeatability
4. Reproducibility

The combination of these trustworthiness criteria provide a blueprint for us to objectively assess a saliency map's localization capabilities (localization utility), sensitivity to trained model weights (versus randomized weights), and robustness with respect to models trained with the same architectures (repeatability) and different architectures (reproducibility). These criteria are important in order for a clinician to trust the saliency map output for its ability to localize the finding of interest.

Specifically, for localization utility, we assess the performance of these methods in localizing abnormalities in medical imaging by quantifying overlap of the saliency map with pixel-level ground-truth segmentations (for SIIM-ACR Pneumothorax) and bounding boxes (for RSNA Pneumonia). We next assess the effect of model weights on saliency maps to examine how saliency maps change when trained model weights are randomized. In other words, if a given saliency map is not sensitive to model weight randomization, the connection between model training and its interpretability is lacking. Lastly, we quantify repeatability of the saliency maps within the same model architecture and reproducibility across different model architectures. Ideally, if one is to use a saliency map to identify the clinically relevant region of interest in an image, one would want a saliency map that shows similar activations regardless of the instantiation of the model (repeatability) or differing architectures (reproducibility), where all have similar end-training classification performance. Repeatability and reproducibility should be especially emphasized in high-stakes clinical models (as compared to natural image computer vision models) where not only overall model performance is important but also individual predictions and explanations due to the potential downstream effects on individual patient outcomes. Low repeatability or reproducibility would mean that a study that uses saliency maps to substantiate model performance could be undermined by using another model with similar performance that produced a completely different saliency map. Ideally, a robust saliency map would demonstrate high performance in all four trustworthiness criteria. Fig 1 below summarizes the questions that will be addressed in this work. Additionally, to promote reproducibility of our findings, we provide the code we used for all tests performed in this work at this link: `https://github.com/QTIM-Lab/Assessing-Saliency-Maps`.

Results

Model Training and Interpretation

We train our models and generate saliency maps using publicly available chest x-ray (CXR) images from the SIIM-ACR Pneumothorax Segmentation and RSNA Pneumonia Detection datasets [19,20]. We train InceptionV3 [21] and DenseNet121 [22] models using transfer learning [25] on models with ImageNet-pretrained weights employing equal sampling of positive and negative labels for every training batch. We choose InceptionV3 and DenseNet121 network architectures because these are among

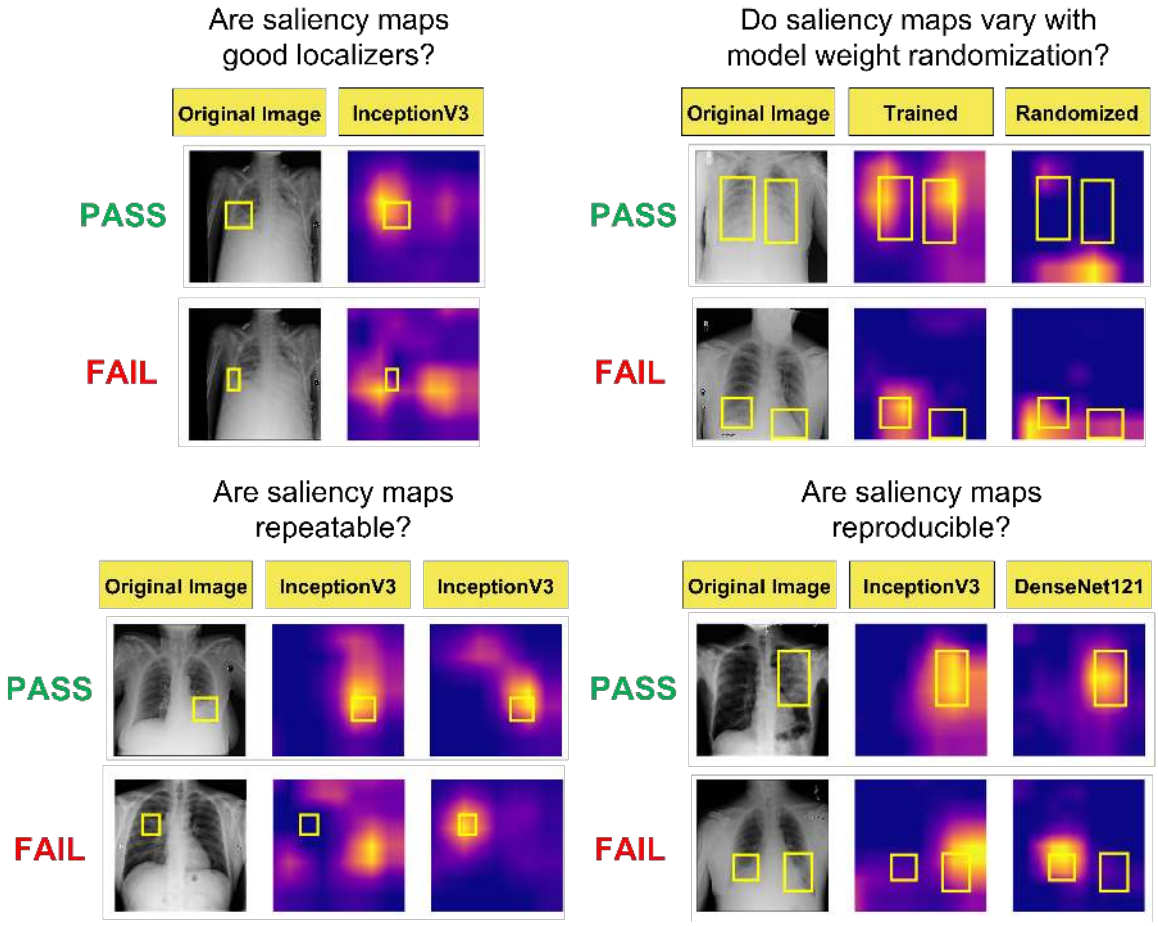


Fig. 1: Visualization of the different questions that will be addressed in this work. Note that the top rows of images and saliency maps demonstrate ideal (and less commonly observed) high performing examples ('PASS'), while the bottom rows of images demonstrate realistic (and more commonly observed) poor performing examples ('FAIL'). First, we examine whether saliency maps are good localizers (in regards to the extent of the maps' overlap with pixel-level segmentations or ground truth bounding boxes). Next, we evaluate whether saliency maps are affected when trained model weights are randomized, indicating how closely the maps reflect model training. Then we generate saliency maps from separately trained InceptionV3 models to assess their repeatability. Finally, we assess the reproducibility by calculating the similarity of saliency maps generated from different models (InceptionV3 and DenseNet121) trained on the same data.

the most commonly used model architectures for classification of chest x-rays and have high performance [23,24]. For the SIIM-ACR Pneumothorax dataset, the InceptionV3 model achieves a test set AUC for classification of 0.889, while the DenseNet121 model achieves a test set AUC for classification of 0.909. For the RSNA Pneumonia dataset, the InceptionV3 model achieves a test set AUC of 0.936

on the test set, while the DenseNet121 model achieves a test set AUC of 0.927. Additional details of model training are provided in the Methods section.

For model interpretation, we evaluate the following saliency maps for their trustworthiness: Gradient Explanation (GRAD), Smoothgrad (SG), Integrated Gradients (IG), Smooth IG (SIG), GradCAM, XRAI, Guided-backprop (GBP), and Guided GradCAM (GGCAM). All methods are summarized and defined in Table 1. We compared the performance of these saliency map techniques against the following baselines: a) in localization utility, a low baseline defined by a single "average" mask obtained by averaging the masks of all images in the training and validation datasets, and a high baseline determined by the AUPRC of segmentation (U-Net) and detection networks (RetinaNet); b) in model weight randomization, the average Structural SIMilarity (SSIM) indices of 50 randomly chosen pairs of saliency maps pertaining to the fully trained model; and c) in repeatability and reproducibility, a low baseline of SSIM = 0.5, and a high baseline determined by the SSIM of U-Net and RetinaNet. Note that a low baseline of SSIM = 0.5 is chosen because SSIM ranges from 0 (for lack of any structural similarity) to 1 (for identical structural similarity), and SSIM = 0.5 marks the midpoint for whether the SSIM is more structurally similar or dissimilar [54,55].

Localization Utility

Segmentation Utility We evaluate the localization utility of each saliency method by quantifying their intersection with ground truth pixel-level segmentations available from the SIIM-ACR Pneumothorax dataset. To capture the intersection between the saliency maps and segmentations, we consider the pixels inside the segmentations to be positive labels and those outside to be negative. Each pixel of the saliency map is thus treated as an output from a binary classifier. Hence, all the pixels in the saliency map can be jointly used to compute the area under the precision-recall curve (AUPRC) utility score [31][1]. More details of how we compute the utility score can be found in the Methods section. An ideal saliency map, from the perspective of utility would have perfect recall (finding all the regions of interest, i.e., pixels with pneumothorax) without labelling any non-pneumothorax pixel as positive (perfect precision). The true negative areas are arguably less important to consider in the metric because they comprise of the vast majority of the image. Additionally, it should be noted that we do not assume that trained classifiers have learned these ground truth segmentations directly. But rather our intention is to highlight the limitations of assuming that saliency maps combined with trained models can be used for localization. We compare the saliency methods with the average of the segmentations across the training and validation sets, as well as a vanilla U-Net [48] trained to learn these segmentations directly.

Saliency maps generated from the InceptionV3 model demonstrate better test set results than those generated from the DenseNet121 model, and are displayed in Fig 2a. For individual saliency maps on InceptionV3, the best performing method is XRAI (AUPRC = 0.224 ± 0.240), while the worst performing method is SIG (AUPRC = 0.024 ± 0.021). It is also interesting to note that using the average of all masks across the Pneumothorax training and validation datasets (AVG) performs as well or better than most of the saliency methods (AUPRC = 0.142 ± 0.149), showing a strong limitation in the saliency maps' utility.

[1] Precision recall curves better serve to be more informative about an algorithm's performance, especially for unbalanced datasets with few positive pixels relative to the number of negative pixels. In the context of findings on medical images, the area under the Receiver Operator Characteristic (ROC) curve can be skewed by the presence of large number of true negatives [37]

Specifically, GBP, GCAM, and GGCAM are not statistically significantly different than the average map, GRAD ($p = 0.0279$), IG ($p< 0.005$), SG ($p < 0.005$), SIG ($p < 0.005$) are all significantly worse and XRAI is significantly better ($p < 0.005$).

Additionally, the U-Net trained on a segmentation task achieves the best performance by far (AUPRC $= 0.404 \pm 0.195$) and the utility of all maps are statistically significantly worse than the U-Net ($p < 0.005$).

The utility of the saliency maps generated using the trained models was higher than the utility of the random models ($p < 0.005$) in all cases except for SG and SIG, where there were no statistical differences.

Table 1: Saliency methods evaluated in this work, along with their corresponding definitions

Saliency Map	Definition
Gradient Explanation (GRAD) [26]	Measures the extent to which a change in a region of the input x affects the prediction $S(x)$ to compute the map $\frac{\partial S}{\partial x}$.
Smoothgrad (SG) [29]	Smoothes the mask obtained using the gradient and integrated gradient saliency methods by stochastically modifying input and performing Gaussian smoothing on the resulting maps.
Integrated Gradients (IG) [27]	Constructs a map by interpolating from a baseline image to the input image and averaging the gradients across these interpolations. We use 25 such interpolations in our experiments to compute the masks.
Smooth IG (SIG) [27,29]	Smoothes an integrated gradients map by stochastically modifying input and performing Gaussian smoothing.
GradCAM (GCAM) [16]	A backpropagation-based method that uses the feature maps of the final convolutional layer to generate heatmaps.
XRAI [28]	Builds on integrated gradients by starting with a baseline image and incrementally adding regions that offer maximal attribution gain.
Guided-backprop (GBP) [30]	Constructs a mask obtained by 'guiding' the conventional backpropagation algorithm to suppress any negative gradients.
Guided GradCAM (GGCAM) [16]	Combines the masks obtained by GradCAM and Guided-backprop in an attempt to minimize the false positives produced by either.

Detection Utility We evaluate detection utility of each saliency method by using the ground truth bounding boxes available from the RSNA Pneumonia Detection dataset. Fig S1, in the Supplementary Material, shows some visualizations of saliency maps generated from InceptionV3 on the RSNA Pneu-

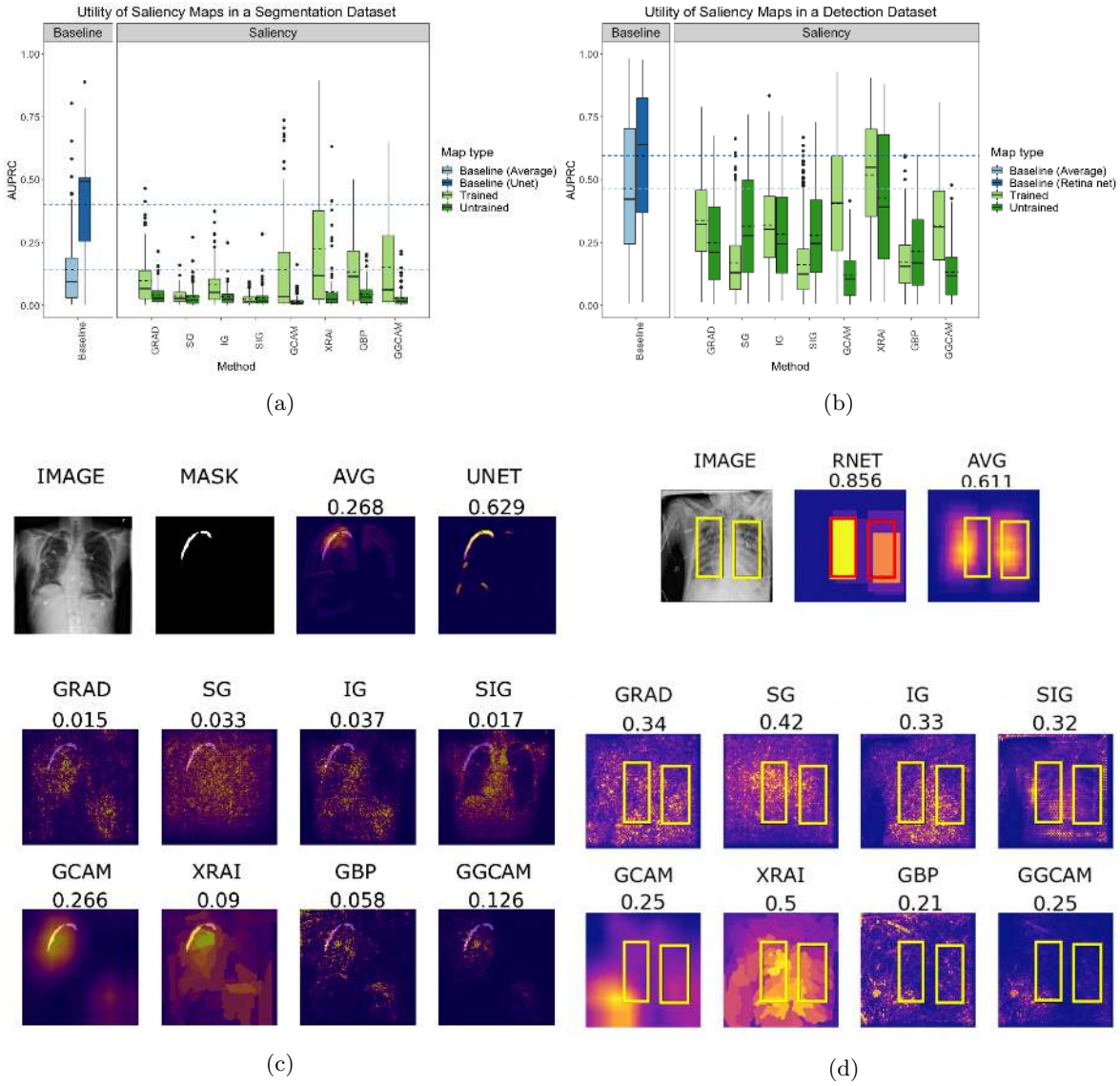


Fig. 2: (a) test set segmentation utility scores for SIIM-ACR Pneumothorax segmentation dataset; (b) test set bounding box detection utility scores for RSNA Pneumonia detection dataset; each box plot represents the distribution of scores across the test data sets for each saliency map, with a solid line denoting the median and dashed line denoting the mean; results are compared to a low baseline using the average segmentation/bounding box of the training and validation sets (light blue) and high baseline using U-Net/RetinaNet (dark blue); (c) example saliency maps on SIIM-ACR Pneumothorax dataset with corresponding utility scores; (d) example saliency maps on RSNA Pneumonia dataset with corresponding utility scores; "Average Mask"/"AVG" refers to using the average of all ground-truth masks (for pneumothorax) or bounding boxes (for pneumonia) across the training and validation datasets; "U-Net"/"UNET" refers to using the vanilla U-Net trained on a segmentation task for localization of pneumothorax; "RetinaNet"/"RNET" refers to using RetinaNet to generate bounding boxes for localizing pneumonia with bounding boxes.

monia Detection dataset. In the same fashion as the Segmentation Utility subsection, we calculate the AUPRC utility score considering pixels inside the bounding boxes to be positive labels and those outside to be negative. We compare the saliency methods with the average of the bounding boxes across the training and validation sets, as well as a RetinaNet [51] trained to learn these bounding boxes directly.

Results for the test set are shown in Fig 2b. The best performing saliency method is XRAI (AUPRC $= 0.519 \pm 0.220$), while the worst performing method is SIG (AUPRC $= 0.160 \pm 0.128$). It is interesting to note that using the average of all bounding boxes across the Pneumonia training and validation datasets performs better than all of the methods (AUPRC $= 0.465 \pm 0.268$) except for XRAI, which was significantly better ($p < 0.005$). RetinaNet trained to generate bounding boxes achieves better performance than all the saliency methods by far (AUPRC $= 0.596 \pm 0.260$, $p < 0.005$).

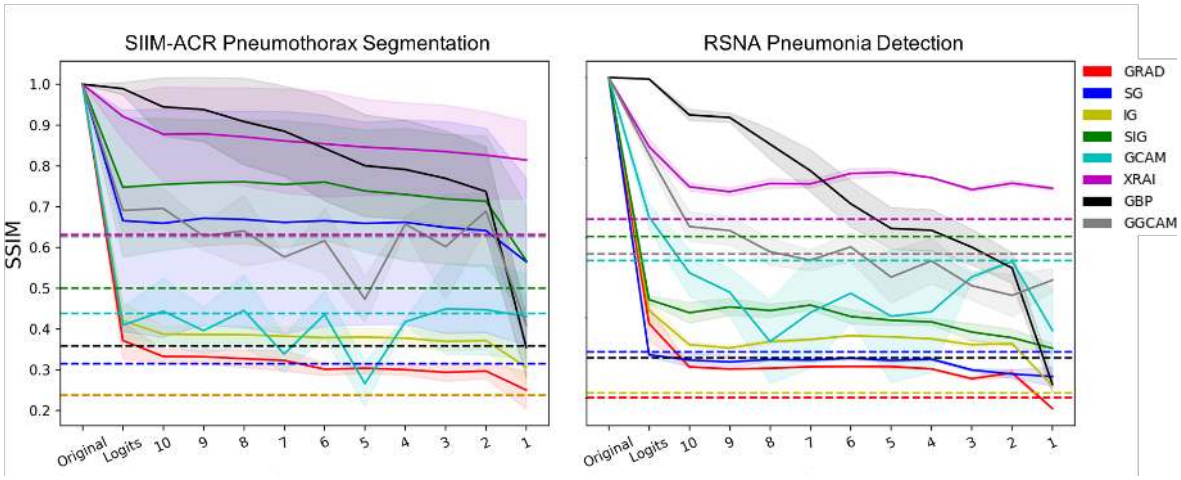
The utility of the saliency maps generated using the trained models was higher than the utility of the random models for GRAD, GCAM, XRAI and GGCAM ($p < 0.005$). The random model had higher utility for SG and SIG ($p < 0.005$) and there were no statistical differences for IG ($p = 0.83$) and GBP ($p = 0.6$).

Sensitivity to Trained vs Random Model Weights

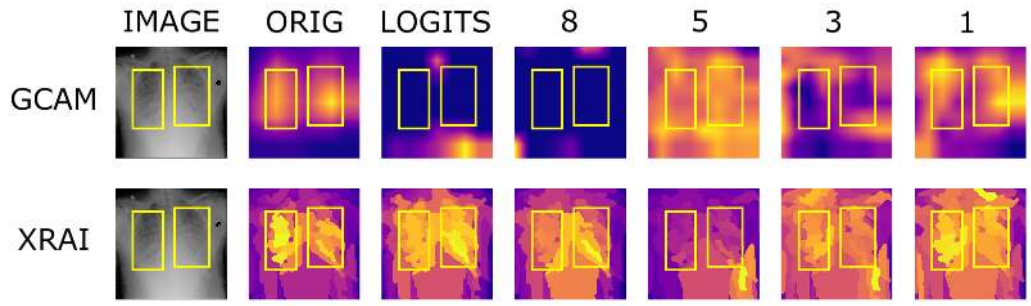
To investigate the sensitivity of saliency methods under changes to model parameters and identify potential correlation of particular layers to changes in the maps, we employ cascading randomization [14]. In cascading randomization, we successively randomize the weights of the trained model beginning from the top layer to the bottom one, which results in erasing the learned weights in a gradual fashion. We use the Structural SIMilarity (SSIM) index of the original saliency map with the saliency maps generated from the model after each randomization step to assess the change of the corresponding saliency maps [32][2]. Adebayo et al. [14] established that SSIM does not necessarily decrease with progressive randomization by running experiments on ImageNet with an InceptionV3 model [21]. We extend those findings to the medical imaging domain by using cascading randomization on the InceptionV3 model described above. This is highlighted by the visualizations in Fig 3a, which monitor the saliency maps (via SSIM index) with the model subject to cascading randomization. Fig 3b shows an example image of saliency map degradation from cascading randomization. Fig S2, in the Supplementary Material, shows some additional examples. Table S1, in the Supplementary Material, shows the mean SSIM index scores of saliency maps for fully randomized models (after cascading randomization has reached the bottom-most layer), as well as the corresponding degradation thresholds, defined below.

We define that a saliency map has reached degradation when the SSIM for the saliency map goes below the degradation threshold. In this work, we define the degradation threshold as the average SSIMs of 50 randomly chosen pairs of saliency maps pertaining to the fully trained model. For both pneumothorax and pneumonia datasets, we observe that the saliency maps that fall below this degradation threshold when cascading randomization has reached the bottom-most layer (i.e., fully randomized) include GradCAM, GBP, and GGCAM, showing a dependency on the trained model weights. We also note that although XRAI performed the best in the utility tests, for both pneumothorax and pneumonia datsets, XRAI does not reach the degradation threshold when fully randomized, showing invariance to the trained model parameters.

[2] Note that we did not take the absolute values of the maps before computing the SSIM and instead used the raw saliency maps to get a clearer picture of map stability.



(a)



(b)

Fig. 3: (a) Structural SIMilarity (SSIM) index under cascading randomization of modules on InceptionV3 for SIIM-ACR Pneumothorax segmentation dataset and RSNA Pneumonia detection dataset; note that the colored dotted lines correspond to the degradation threshold for each saliency map; they are generated by the average SSIMs of 50 randomly chosen pairs of saliency maps pertaining to the fully trained model; a saliency model successfully reaches degradation if it goes below its corresponding degradation threshold; (b) example image from RSNA Pneumonia detection dataset to visualize saliency map degradation from cascading randomization; 'Logits' refers to the Logit Layer (final layer) of the InceptionV3 model and layer blocks 1 through 10 refer to blocks Mixed1 through Mixed10 in the original InceptionV3 architecture.

Repeatability and Reproducibility

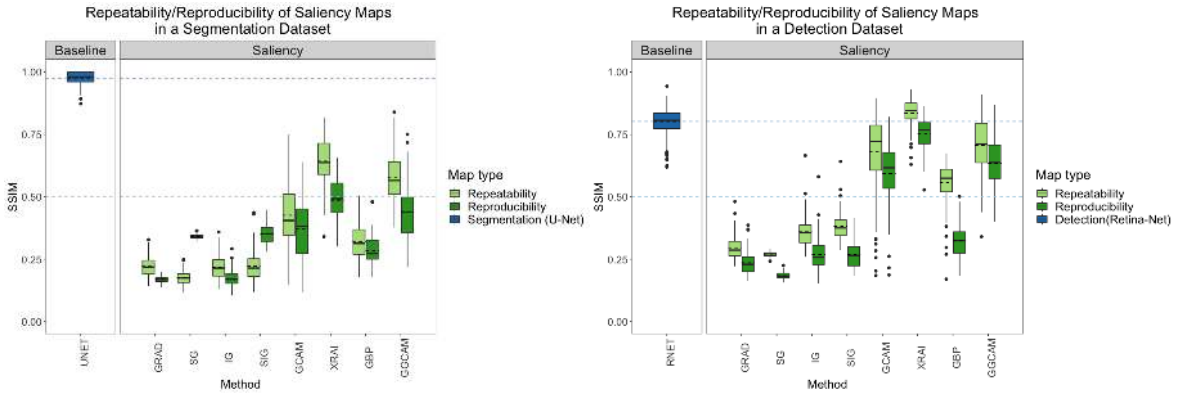
We conduct repeatability tests on the saliency methods by comparing maps from a) different randomly initialized instances of models with the same architecture trained to convergence (intra-architecture repeatability) and b) models with different architectures each trained to convergence (inter-architecture reproducibility) using SSIM between saliency maps produced from each model. These experiments are designed to test if the saliency methods produce similar maps with a different set of trained weights and whether they are architecture agnostic (assuming that models with different trained weights or architectures have similar classification performance). Although there is no constraint that indicates that interpretations should be the same across models, an ideal trait of a saliency map would be to have some degree of robustness across models with different trained weights or architectures. For our baselines, have a low baseline of SSIM = 0.5 (since SSIM = 0.5 marks the midpoint for whether the SSIM is more structurally similar or dissimilar), and a high baseline of repeatability/reproducibility of separately trained U-Nets (for the pneumothorax dataset) and RetinaNets (for the pneumonia dataset).

We examine the repeatability of saliency methods from two separately trained InceptionV3 models, and the reproducibility of saliency methods from fully trained InceptionV3 and DenseNet121 models. Figs 4a and 4b summarize the results for the SIIM-ACR Pneumothorax and RSNA Pneumonia datasets, respectively. In the SIIM-ACR Pneumothorax dataset, the baseline U-Net achieves a value of SSIM $= 0.976 \pm 0.024$, which is much greater performance than any of the saliency maps ($p < 0.005$). The lower SSIM = 0.5 baseline is greater than the performance of all saliency maps except for the repeatability of XRAI and GGCAM. Among the saliency maps, XRAI has the best repeatability (SSIM $= 0.643 \pm 0.092$), while SG has the worst repeatability (SSIM $= 0.176 \pm 0.027$). For reproducibility, XRAI performs the best (SSIM $= 0.487 \pm 0.084$), while GRAD performs the worst (SSIM $= 0.169 \pm 0.013$). Repeatability was higher than reproducibility for GRAD, IG, XRAI, GBP, and GGCAM. Surprisingly, reproducibility was higher than repeatability for SG and SIG.

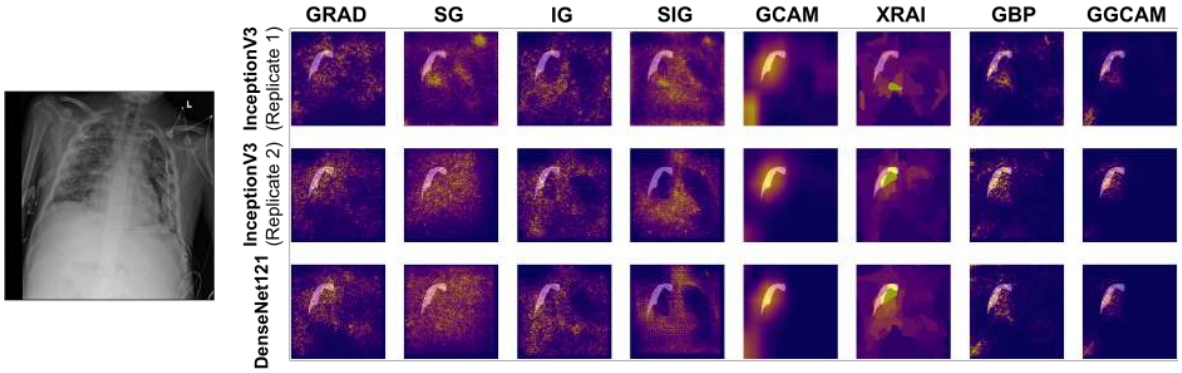
In the RSNA Pneumonia dataset, the baseline RetinaNet achieves a value of SSIM $= 0.802 \pm 0.048$, which is only exceeded by XRAI's repeatability score ($p < 0.005$). The lower SSIM = 0.5 baseline is greater than the performance of all saliency maps except the repeatability and reproducibility of GCAM, GGCAM, and XRAI, and the repeatability of GBP. Among the saliency maps, XRAI has the best repeatability (SSIM $= 0.836 \pm 0.055$), while SG has the worst (SSIM $= 0.270 \pm 0.009$). For reproducibility, XRAI performs the best (SSIM $= 0.754 \pm 0.062$), while SG performs the worst (SSIM $= 0.184 \pm 0.014$). Repeatability was higher than reproducibility for all methods ($p \ll 0.005$). Overall, XRAI has the highest repeatability and reproducibility across different datasets. In Figs 4c and 4d, we observe the repeatability and reproducibility of a two example images from the SIIM-ACR Pneumothorax dataset and RSNA Pneumonia dataset, respectively. The first two rows are saliency maps generated from two separately trained InceptionV3 models (Replicates 1 and 2) to demonstrate repeatability, and the last row are saliency maps generated from a DenseNet121 model to demonstrate reproducibility. Fig S3, in the Supplementary Material, shows a couple of additional examples for repeatability and reproducibility.

Discussion and Conclusion

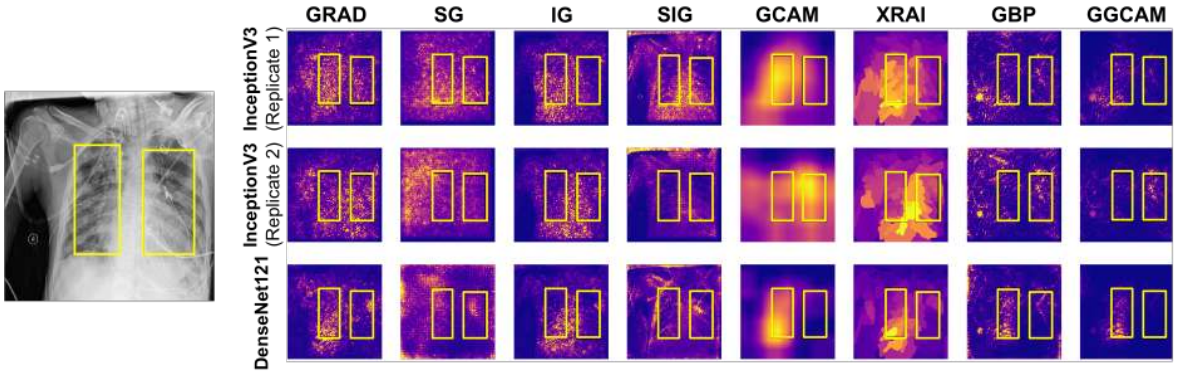
In this study, we evaluated the performance and robustness of several popular saliency map techniques on medical images. By considering utility with respect to localization, sensitivity to model weight randomization, repeatability and reproducibility, we demonstrate that none of the saliency maps meet



(a) (b)



(c)



(d)

Fig. 4: Comparison of repeatability and reproducibility scores for all saliency methods for (a) SIIM-ACR Pneumothorax segmentation dataset and (b) RSNA Pneumonia detection dataset; each box plot represents the distribution of scores across the test data sets for each saliency map, with a solid line denoting the median and dashed line denoting the mean; results are compared to a low baseline of SSIM = 0.5 (light blue dashed line) and high baseline using U-Net/RetinaNet (dark blue box plot and dashed line); two examples of repeatability (InceptionV3 Replicates 1 and 2) and reproducibility (InceptionV3 and DenseNet121) for the (c) SIIM Pneumothorax dataset with transparent segmentations and (d) RSNA Pneumonia dataset with yellow bounding boxes

all tested criteria and their credibility should be critically evaluated prior to integration into medical imaging pipelines. This is especially important because many recent deep learning based clinical studies rely on saliency maps for interpretability of deep learning models without noting and critically evaluating their inherent limitations. Table 2 demonstrates the overall results for each saliency map across all tests. It is clear that none of the maps demonstrate a superior performance in all four defined trustworthiness criteria, and in fact most of them are inferior to their corresponding baselines. For their high baseline methods, the utility, repeatability, and reproducibility tasks utilize networks that train specifically as localizers (i.e., U-Net and RetinaNet). With the exception of XRAI on repeatability in the pneumonia dataset, all of the saliency maps perform worse than U-Net and RetinaNet. This highlights a severe limitation in saliency maps as a whole, and shows that using models trained directly on localization tasks (such as U-Net and RetinaNet) greatly improves the results. Additional insights for the results of each task are provided in the Supplementary Material. Depending on the desired outcome of interpretability, there are alternative techniques besides saliency methods that can be employed. One approach would be to train CNNs that output traditional handcrafted features (such as shape and texture) as intermediates [46]. This approach would provide some explainability but is limited by the utility and reliability of the handcrafted features. Another approach may also be to use interpretable models in the first place. Rudin argues that instead of creating methods to explain black box models trained for high-stakes decision making, we should instead put our focus on designing models that are inherently interpretable [35]. Rudin further argues that there is not necessarily a tradeoff between accuracy and interpretability, especially if the input data is well structured (i.e., features are meaningful). Thus moving forward from the results in this work, it would be wise to consider multiple avenues for improving model interpretation.

There are a few limitations to our study. First, we only evaluated saliency maps for two medical datasets, both consisting of chest radiographs. Future studies will examine more medical imaging datasets, including different image modalities and diseases. Additionally, we only performed tests on two neural network architectures, though these are commonly used networks in the literature for chest radiograph analysis [23,24]. As a next step, we can examine the effect of other neural network architectures to determine if they result in saliency maps that are more repeatable and reproducible. Third, we focus only on the ability of saliency maps to localize pathology and thus the utility metrics were calculated using the regions-of-interest specifically (bounding boxes for pneumonia and segmentation maps for pneumothorax). These regions-of-interest may not include other image features that can contribute to classification algorithm performance, known as hidden stratification. For example, a chest tube in an image would imply the presence of a pneumothorax, but much of the chest tube may not be in the region-of-interest [53]. More global features could also contribute to classification. For example, low lung volumes and portable radiograph technique may suggest that the patient is hospitalized, which could be associated with likelihood of pneumonia. These features would also not be covered in the regions-of-interest. Future work can evaluate the utility of saliency maps to localize these other features. We could also investigate incorporating saliency maps as a part of neural network training and evaluate if this type of approach results in maps that have higher utility than maps that are generated post-model training [36].

Method	Utility		Randomization	Repeatability		Reproducibility	
	AVG	UNET		LOW	UNET	LOW	UNET
GRAD	FAIL	FAIL	FAIL	FAIL	FAIL	FAIL	FAIL
SG	FAIL	FAIL	FAIL	FAIL	FAIL	FAIL	FAIL
IG	FAIL	FAIL	FAIL	FAIL	FAIL	FAIL	FAIL
SIG	FAIL	FAIL	FAIL	FAIL	FAIL	FAIL	FAIL
GCAM	FAIL	FAIL	PASS	FAIL	FAIL	FAIL	FAIL
XRAI	PASS	FAIL	FAIL	PASS	FAIL	FAIL	FAIL
GBP	FAIL	FAIL	PASS	FAIL	FAIL	FAIL	FAIL
GGCAM	FAIL	FAIL	PASS	PASS	FAIL	FAIL	FAIL

(a) SIIM-ACR Pneumothorax Segmentation

Method	Utility		Randomization	Repeatability		Reproducibility	
	AVG	RNET		LOW	RNET	LOW	RNET
GRAD	FAIL	FAIL	PASS	FAIL	FAIL	FAIL	FAIL
SG	FAIL	FAIL	PASS	FAIL	FAIL	FAIL	FAIL
IG	FAIL	FAIL	FAIL	FAIL	FAIL	FAIL	FAIL
SIG	FAIL	FAIL	PASS	FAIL	FAIL	FAIL	FAIL
GCAM	FAIL	FAIL	PASS	PASS	FAIL	PASS	FAIL
XRAI	PASS	FAIL	FAIL	PASS	PASS	PASS	FAIL
GBP	FAIL	FAIL	PASS	PASS	FAIL	FAIL	FAIL
GGCAM	FAIL	FAIL	PASS	PASS	FAIL	PASS	FAIL

(b) RSNA Pneumonia Detection

Table 2: Summary of all the results for experiments on the (a) SIIM-ACR Pneumothorax segmentation dataset and (b) RSNA Pneumonia detection dataset. The Utility column compares maps to the average of all maps across the training and validation datasets (AVG) and a U-Net (UNET) trained for pneumothorax segmentation, or a RetinaNet (RNET) trained for pneumonia detection. The Randomization column compares the SSIM scores of the saliency maps when cascading randomization is completed to the bottom-most layer with the degradation threshold defined in the Results section. The Repeatability and Reproducibility columns compare SSIM scores of saliency maps with the low baseline of SSIM = 0.5 (LOW) and two independently trained U-Nets for pneumothorax segmentation or two independently trained RetinaNets for pneumonia detection.

Methods

First, we utilize the ground truth semantic segmentations from the SIIM-ACR Pneumothorax dataset to evaluate the localization utility of saliency maps to give meaningful localization for a pixel-level segmentation task. Then, we leverage the ground-truth bounding box coordinates provided in the RSNA Pneumonia dataset to establish the utility of saliency maps for detection task. Next, we examine changes in saliency maps caused by weight randomization in model layers (assessed using the Structural SIMilarity (SSIM) index). Finally, we use SSIM to examine the reproducibility and repeatibility of each saliency map method. Details for the tests are provided in the Results section and the subsections below.

Data Preparation

The SIIM-ACR Pneumothorax dataset consists of a total of 10675 images, split in a $90:10$ ratio for training + validation and test sets respectively. The combined training + validation set is in turn spit in a $90:10$ ratio to create separate training and a validation sets. The final training set is comprised of 8646 images with 1931 positive cases; the validation set contains 961 images with 202 positive cases and the test set contains 1068 images with 246 positive cases.

In the RSNA Pneumonia dataset, only those cases that exhibit the presence of pneumonia or those that are classified as normal are considered (removing the cases that are classified as 'not pneumonia, not normal' to focus the localization task only on pneumonia as an abnormality). This results in a curated dataset comprising 14863 samples. This curated dataset is split in a similar fashion as described above for the SIIM-ACR Pneumothorax dataset. The final training set is comprised of 12039 images with 4870 positive cases; the validation set contains 1338 images with 541 positive cases and the test set contains 1486 images with 601 positive cases.

Model Training

We train the InceptionV3 models used for our analysis via transfer learning using ImageNet-pretrained model weights [38] by replacing the final fully connected layer (`https://github.com/keras-team/keras-applications/blob/master/keras_applications/inception_v3.py`). The learning rate is set at $1e-4$ after experimenting with different values, and the models are trained for a maximum of 20 epochs. Early stopping [39] is used while monitoring the validation AUC with a patience of 4 epochs.

In a similar fashion, we use an ImageNet-pretrained DenseNet121 model (`https://github.com/keras-team/keras-applications/blob/master/keras_applications/densenet.py`) for reproducibility analysis. The learning rate is set at $7e-5$ and the models are trained for a maximum of 30 epochs. Early stopping [39] is used while monitoring the validation AUC with a patience of 5 epochs. Both the InceptionV3 and the DenseNet121 models are trained using Binary Cross Entropy loss as given below

$$H_p(q) = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log p(y_i) + (1-y_i)\log(1-p(y_i)))$$

Cascading randomization is performed on the InceptionV3 model to reset the parameter values sampled from a truncated normal distribution, in accordance with Adebayo et al. [14]. The models trained on the SIIM-ACR Pneumothorax dataset report test set AUCs of 0.889 (InceptionV3) and 0.909

(DenseNet121), while the fully trained models on the curated RSNA Pneumonia dataset report test set AUCs of 0.936 (InceptionV3) and 0.927 (DenseNet121). The corresponding completely random models both record AUCs of 0.5 as expected. The trained InceptionV3 replicate used for the repeatability experiment on the RSNA Pneumonia dataset reported a test set AUC of 0.940 and the correspoding replicate for the SIIM-ACR Pneumothorax dataset reported a test set AUC of 0.907.

For the segmentation baseline model, we use the base U-Net architecture [48] with no pretraining. During training, the Adam optimizer is used with default beta values, a mix of focal and dice loss, batch size of 4, and learning rate of 1e-04. The learning rate is decreased by a factor of 0.1 when the validation loss fails to decrease for more than 3 epochs. The best model (based on the validation loss) is trained for up to 75 epochs, using early stopping if the loss has not decreased for 15 epochs. Additionally, in order to equally sample the positive and negative pneumothorax cases, balanced sampling is used on the images. For the final output, the sigmoid function is applied to each cell (pixel). The mean non-zero cell value in this output is taken (if it exists, otherwise we output 0) and used as the model's classification output for calculating the AUC of the model. We obtain a test AUC score of 0.893. To generate the final image, the sigmoided U-Net output is multiplied by 255 and rounded to scale the values into a final output image.

For the object detection baseline model, we use RetinaNet with ResNet101 base architecture pretrained on ImageNet (`https://github.com/yhenon/pytorch-retinanet.git`). During training, the Adam opimizer is used with hyper-parameters set at 1e-4 learning rate, 4 batch-size and, 0.9 and 0.999 beta values. To get a pixel-wise output for evaluation, a continuous mask is constructed with all detection boxes having greater than 0.01 probability. Similar to how the non-maximum suppression algorithm works for appropriate box selection, each pixel is assigned a probability value corresponding to the highest scoring box covering that pixel. The result is a heatmap where each pixel value is directly associated with probability of presence of abnormality in that pixel region. We obtain a test AUROC of 0.949 and test AUPRC of 0.932.

Utility Metric

We choose the area under the precision recall curve (AUPRC) as the metric to capture localizability of saliency maps as the relatively small size of the ground truth segmentation masks and bounding boxes necessitated an approach that would account for the class imbalance.

Precision is defined as the ratio of True Positives (TP) to Predicted Positives (TP+FP) while Recall is defined as the ratio of True Positives to Ground Truth Positives (TP+FN). Since neither of these account for the number of True Negatives, they make ideal candidates for our analysis. A PR curve shows the trade-off between precision and recall across different decision thresholds by varying the threshold to plot corresponding precision and recall values. An example curve for a saliency map is as shown below in Fig S4 in the Supplementary Material.

Statistical analysis

Statistical analyses were performed in RStudio version 1.2.5033 using R 3.6 and the *lmer*, *lmerTest*, *ggplot2* and *multComp* packages. A linear mixed effects model was used to study the following in order to establish the trustworthiness of the eight saliency map methods using the following statistical methods. All tests were two-sided with alpha level set at 0.05 for statistical significance. The Tukey's HSD test was used for post-hoc analysis, and the following questions were examined:

1. Is there a difference between the utility of each of the saliency map methods derived from the trained classification models compared to the localization baseline models (detection or segmentation) as measured using AUPRC?
2. Is there a difference between the trained models and the "average" mask in terms of the utility of the map?
3. Is there a difference in the utility of trained models compared to their associated random model?
4. Is there a difference between the repeatability/reproducibility of each of the saliency map methods compared to the localization baselines, measured using SSIM?

References

1. Esteva, Andre and Kuprel, Brett and Novoa, Roberto A and Ko, Justin and Swetter, Susan M and Blau, Helen M and Thrun, Sebastian,"Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, Vol. 542, No. 7639, 115–118,(2017). Nature Publishing Group.
2. Chang, Ken and Beers, Andrew L and Brink, Laura and Patel, Jay B and Singh, Praveer and Arun, Nishanth T and Hoebel, Katharina V and Gaw, Nathan and Shah, Meesam and Pisano, Etta D and Tilkin, Mike and Coombs, Laura P and Dreyer, Keith J and Allen, Bibb and Agarwal, Sheela and Kalpathy-Cramer, Jayashree,"Multi-Institutional Assessment and Crowdsourcing Evaluation of Deep Learning for Automated Classification of Breast Density," *JACR*, (2020). American College of Radiology.
3. Campanella, Gabriele and Hanna, Matthew G and Geneslaw, Luke and Miraflor, Allen and Silva, Vitor Werneck Krauss and Busam, Klaus J and Brogi, Edi and Reuter, Victor E and Klimstra, David S and Fuchs, Thomas J,"Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nature medicine*, Vol. 25, No. 8, 1301–1309,(2019). Nature Publishing Group.
4. Chang, Ken and Beers, Andrew L and Bai, Harrison X and Brown, James M and Ly, K Ina and Li, Xuejun and Senders, Joeky T and Kavouridis, Vasileios K and Boaro, Alessandro and Su, Chang and others,"Automatic assessment of glioma burden: a deep learning algorithm for fully automated volumetric and bidimensional measurement," *Neuro-oncology*, Vol. 21, No. 11, 1412–1422,(2019). Oxford University Press US.
5. Ghorbani, Amirata and Ouyang, David and Abid, Abubakar and He, Bryan and Chen, Jonathan H and Harrington, Robert A and Liang, David H and Ashley, Euan A and Zou, James Y,"Deep Learning Interpretation of Echocardiograms," *bioRxiv*, 681676,(2019). Cold Spring Harbor Laboratory.
6. Li, Matthew D and Chang, Ken and Bearce, Ben and Chang, Connie Y and Huang, Ambrose J and Campbell, J Peter and Brown, James M and Singh, Praveer and Hoebel, Katharina V and Erdo, "Siamese neural networks for continuous disease severity evaluation and change detection in medical imaging," *NPJ digital medicine*, No. 3, 1,(1–9). 2020.
7. Wang, Fei and Kaushal, Rainu and Khullar, Dhruv,"Should Health Care Demand Interpretable Artificial Intelligence or Accept Black Box Medicine?," *Annals of Internal Medicine*,(2020).
8. Zou, James and Schiebinger, Londa,"AI can be sexist and racist–it's time to make it fair," *Nature Publishing Group*,(2018).
9. Ilyas, Andrew and Santurkar, Shibani and Tsipras, Dimitris and Engstrom, Logan and Tran, Brandon and Madry, Aleksander,"Adversarial examples are not bugs, they are features," *Advances in Neural Information Processing Systems*, 125–136,(2019).
10. Winkler, Julia K and Fink, Christine and Toberer, Ferdinand and Enk, Alexander and Deinlein, Teresa and Hofmann-Wellenhof, Rainer and Thomas, Luc and Lallas, Aimilios and Blum, Andreas and Stolz, Wilhelm and others,"Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition," *JAMA dermatology*, Vol. 155, No. 10, 1135–1141,(2019). American Medical Association.
11. Rajpurkar, Pranav and Irvin, Jeremy and Zhu, Kaylie and Yang, Brandon and Mehta, Hershel and Duan, Tony and Ding, Daisy and Bagul, Aarti and Langlotz, Curtis and Shpanskaya, Katie and others,"Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," *arXiv preprint arXiv:1711.05225*,(2017).
12. Bien, Nicholas and Rajpurkar, Pranav and Ball, Robyn L and Irvin, Jeremy and Park, Allison and Jones, Erik and Bereket, Michael and Patel, Bhavik N and Yeom, Kristen W and Shpanskaya, Katie and others,"Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet," *PLoS medicine*, Vol. 15, No. 11, e1002699,(2018). Public Library of Science.
13. Mitani, Akinori and Huang, Abigail and Venugopalan, Subhashini and Corrado, Greg S and Peng, Lily and Webster, Dale R and Hammel, Naama and Liu, Yun and Varadarajan, Avinash V,"Detection of anaemia from retinal fundus images via deep learning," *Nature Biomedical Engineering*, 1–10,(2019). Nature Publishing Group.

14. Adebayo, Julius and Gilmer, Justin and Muelly, Michael and Goodfellow, Ian and Hardt, Moritz and Kim, Been,"Sanity checks for saliency maps," *Advances in Neural Information Processing Systems*, 9505–9515,(2018).
15. Eitel, Fabian and Ritter, Kerstin and Alzheimer's Disease Neuroimaging Initiative (ADNI and others), *Testing the Robustness of Attribution Methods for Convolutional Neural Networks in MRI-Based Alzheimer's Disease Classification*, Springer, 2019.
16. Selvaraju, Ramprasaath R and Cogswell, Michael and Das, Abhishek and Vedantam, Ramakrishna and Parikh, Devi and Batra, Dhruv,"Grad-cam: Visual explanations from deep networks via gradient-based localization," *Proceedings of the IEEE international conference on computer vision*, 618–626, (2017).
17. Crosby, Jennie and Chen, Sophia and Li, Feng and MacMahon, Heber and Giger, Maryellen, "Network output visualization to uncover limitations of deep learning detection of pneumothorax," *Medical Imaging 2020: Image Perception, Observer Performance, and Technology Assessment*, Vol. 11316, 113160O, (2020). International Society for Optics and Photonics.
18. Young, Kyle and Booth, Gareth and Simpson, Becks and Dutton, Reuben and Shrapnel, Sally, *Deep neural network or dermatologist?*, Springer, 2019.
19. SIIM-ACR Pneumothorax Segmentation, Kaggle, (2019). `https://www.kaggle.com/c/siim-acr-pneumothorax-segmentation`.
20. RSNA Pneumonia Detection, Kaggle, (2018). `https://www.kaggle.com/c/rsna-pneumonia-detection-challenge`.
21. Szegedy, Christian and Vanhoucke, Vincent and Ioffe, Sergey and Shlens, Jon and Wojna, Zbigniew,"Rethinking the inception architecture for computer vision," *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826,(2016).
22. Huang, Gao and Liu, Zhuang and Van Der Maaten, Laurens and Weinberger, Kilian Q,"Densely connected convolutional networks," *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708,(2017).
23. Narin, Ali and Kaya, Ceren and Pamuk, Ziynet,"Automatic detection of coronavirus disease (COVID-19) using X-ray images and deep convolutional neural networks," *arXiv preprint arXiv:2003.10849*,(2020).
24. Rajpurkar, Pranav and Irvin, Jeremy and Ball, Robyn L and Zhu, Kaylie and Yang, Brandon and Mehta, Hershel and Duan, Tony and Ding, Daisy and Bagul, Aarti and Langlotz, Curtis P and others,"Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists," *PLoS medicine*, Vol. 15, No. 11, e1002686,(2018). Public Library of Science.
25. Bengio, Yoshua,"Deep learning of representations for unsupervised and transfer learning," *Proceedings of ICML workshop on unsupervised and transfer learning*, 17–36,(2012).
26. Simonyan, Karen and Vedaldi, Andrea and Zisserman, Andrew,"Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*,(2013).
27. Sundararajan, Mukund and Taly, Ankur and Yan, Qiqi,"Axiomatic attribution for deep networks," *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 3319–3328,(2017). JMLR. org.
28. Kapishnikov, Andrei and Bolukbasi, Tolga and Vi,"XRAI Better Attributions Through Regions," *Proceedings of the IEEE International Conference on Computer Vision*, 4948–4957,(2019).
29. Smilkov, Daniel and Thorat, Nikhil and Kim, Been and Vi,"Smoothgrad: removing noise by adding noise," *arXiv preprint arXiv:1706.03825*,(2017).
30. Springenberg, Jost Tobias and Dosovitskiy, Alexey and Brox, Thomas and Riedmiller, Martin,"Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806*,(2014).
31. Boyd, Kendrick and Eng, Kevin H and Page, C David,"Area under the precision-recall curve: point estimates and confidence intervals," *Joint European conference on machine learning and knowledge discovery in databases*, 451–466,(2013). Springer.
32. Zar, Jerrold H,"Spearman rank correlation," *Encyclopedia of Biostatistics*, Vol. 7,(2005). Wiley Online Library.
33. Nie, Weili and Zhang, Yang and Patel, Ankit,"A theoretical explanation for perplexing behaviors of backpropagation-based visualizations," *arXiv preprint arXiv:1805.07039*,(2018).

34. Tonekaboni, Sana and Joshi, Shalmali and McCradden, Melissa D and Goldenberg, Anna,"What clinicians want: contextualizing explainable machine learning for clinical end use," *arXiv preprint arXiv:1905.05134*,(2019).
35. Rudin, Cynthia,"Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, Vol. 1, No. 5, 206–215,(2019). Nature Publishing Group.
36. Alzantot, Moustafa and Widdicombe, Amy and Julier, Simon and Srivastava, Mani,"NeuroMask: Explaining Predictions of Deep Neural Networks through Mask Learning," *2019 IEEE International Conference on Smart Computing (SMARTCOMP)*, 81–86,(2019). IEEE.
37. Ozenne, Brice and Subtil, Fabien and Maucort-Boulch, Delphine,"The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases," *Journal of clinical epidemiology*, Vol. 68, No. 8, 855–859,(2015). Elsevier.
38. Krizhevsky, Alex and Sutskever, Ilya and Hinton, Geoffrey E,"Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, 1097–1105,(2012).
39. Zhang, Zhanpeng and Luo, Ping and Loy, Chen Change and Tang, Xiaoou,"Facial landmark detection by deep multi-task learning," *European conference on computer vision*, 94–108,(2014).
40. Farooq, Ammarah and Anwar, SyedMuhammad and Awais, Muhammad and Rehman, Saad,"A deep CNN based multi-class classification of Alzheimer's disease using MRI," *2017 IEEE International Conference on Imaging systems and techniques (IST)*, 1–6,(2017). IEEE.
41. Nie, Dong and Zhang, Han and Adeli, Ehsan and Liu, Luyan and Shen, Dinggang,"3D deep learning for multi-modal imaging-guided survival time prediction of brain tumor patients," *International conference on medical image computing and computer-assisted intervention*, 212–220,(2016). Springer.
42. Shin, Younghak and Balasingham, Ilangko,"Comparison of hand-craft feature based SVM and CNN based deep learning framework for automatic polyp classification," *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 3277–3280,(2017). IEEE.
43. Saba, Tanzila and Khan, Muhammad Attique and Rehman, Amjad and Marie-Sainte, Souad Larabi,"Region extraction and classification of skin cancer: A heterogeneous framework of deep CNN features fusion and reduction," *Journal of medical systems*, Vol. 43, No. 9, 289,(2019). Springer.
44. Jeyaraj, Pandia Rajan and Nadar, Edward Rajan Samuel,"Computer-assisted medical image classification for early diagnosis of oral cancer employing deep learning algorithm," *Journal of cancer research and clinical oncology*, Vol. 145, No. 4, 829–837,(2019). Springer.
45. Arun, Nishanth Thumbavanam and Gaw, Nathan and Singh, Praveer and Chang, Ken and Hoebel, Katharina Viktoria and Patel, Jay and Gidwani, Mishka and Kalpathy-Cramer, Jayashree,"Assessing the validity of saliency maps for abnormality localization in medical imaging," *Medical Imaging with Deep Learning*,(2020).
46. Lou, Bin and Doken, Semihcan and Zhuang, Tingliang and Wingerter, Danielle and Gidwani, Mishka and Mistry, Nilesh and Ladic, Lance and Kamen, Ali and Abazeed, Mohamed E,"An image-based deep learning framework for individualising radiotherapy dose: a retrospective analysis of outcome prediction," *The Lancet Digital Health*, Vol. 1, No. 3, e136–e147,(2019). Elsevier.
47. Zhou, Bolei and Khosla, Aditya and Lapedriza, Agata and Oliva, Aude and Torralba, Antonio,"Learning deep features for discriminative localization," *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2921–2929,(2016).
48. Ronneberger, Olaf and Fischer, Philipp and Brox, Thomas,"U-net: Convolutional networks for biomedical image segmentation," *International Conference on Medical image computing and computer-assisted intervention*, 234–241,(2015). Springer.
49. Reyes, Mauricio and Meier, Raphael and Pereira, Sergio and Silva, Carlos A and Dahlweid, Fried-Michael and von Tengg-Kobligk, Hendrik and Summers, Ronald M and Wiest, Roland, "On the Interpretability of Artificial Intelligence in Radiology: Challenges and Opportunities," *Radiology: Artificial Intelligence*, Vol. 2, No. 3,(2020). RSNA.

50. Parikh, Ravi B and Obermeyer, Ziad and Navathe, Amol S, "Regulation of predictive analytics in medicine," *Science*, Vol. 363, No. 6429, 810–812,(2019). American Association for the Advancement of Science.
51. Lin, Tsung-Yi and Goyal, Priya and Girshick, Ross and He, Kaiming and Dollár, Piotr, "Focal loss for dense object detection," *Proceedings of the IEEE international conference on computer vision*, 2980–2988,(2017). IEEE.
52. Titano, Joseph J and Badgeley, Marcus and Schefflein, Javin and Pain, Margaret and Su, Andres and Cai, Michael and Swinburne, Nathaniel and Zech, John and Kim, Jun and Bederson, Joshua and others,"Automated deep-neural-network surveillance of cranial images for acute neurologic events," *Nature medicine*, Vol. 24, No. 9, 1337–1341,(2018). Nature Publishing Group.
53. Oakden-Rayner, Luke and Dunnmon, Jared and Carneiro, Gustavo and R, "Hidden stratification causes clinically meaningful failures in machine learning for medical imaging," *Proceedings of the ACM Conference on Health, Inference, and Learning*, 2020.
54. Renieblas, Gabriel Prieto and Nogués, Agustín Turrero and González, Alberto Muñoz and León, Nieves Gómez and Del Castillo, Eduardo Guibelalde, "Structural similarity index family for image quality assessment in radiological images," *Journal of medical imaging*, Vol. 4, No. 3, (2017). International Society for Optics and Photonics.
55. Brooks, Alan C and Zhao, Xiaonan and Pappas, Thrasyvoulos N,"Structural similarity quality metrics in a coding context: exploring the space of realistic distortions," *IEEE Transactions on image processing*, Vol. 17, No. 8, 1261–1273,(2008). IEEE.

Acknowledgements

Research reported in this publication was supported by a training grant from the National Institute of Biomedical Imaging and Bioengineering (NIBIB) of the National Institutes of Health under award number 5T32EB1680 to K. Chang and J. B. Patel and by the National Cancer Institute (NCI) of the National Institutes of Health under Award Number F30CA239407 to K. Chang. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

This publication was supported from the Martinos Scholars fund to K. Hoebel. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the Martinos Scholars fund.

This study was supported by National Institutes of Health (NIH) grants U01CA154601, U24CA180927, U24CA180918, and U01CA242879, and National Science Foundation (NSF) grant NSF1622542 to J. Kalpathy-Cramer.

This research was carried out in whole or in part at the Athinoula A. Martinos Center for Biomedical Imaging at the Massachusetts General Hospital, using resources provided by the Center for Functional Neuroimaging Technologies, P41EB015896, a P41 Biotechnology Resource Grant supported by the National Institute of Biomedical Imaging and Bioengineering (NIBIB), National Institutes of Health.

Supplementary Material

Supplementary Discussion

In regards to evaluating localization utility for segmentation of pneumothorax images, maximum utility is achieved by XRAI at AUPRC $= 0.224 \pm 0.240$. Additionally, in case of evaluating localization utility for detection of pneumonia images, maximum utility is achieved by XRAI at AUPRC $= 0.519 \pm 0.220$. Although this suggests a substantial improvement from the pneumothorax results, it is important to observe that ground truth bounding boxes for detection tasks do not have complex shapes and cover a superset of pixels from a corresponding mask for segmentation tasks. This leads to less difficult localization task.

In general, the particular localization tasks presented in this paper can be incredibly difficult due to the overlapping structures present in 2D chest radiographs, as well as subtle changes in texture that can be challenging to detect [17]. The challenges of the pneumothorax and pneumonia chest radiograph datasets serve to demonstrate the limitations on localization abilities of saliency maps. These findings highlight the necessity to build saliency maps that better consider complex shapes of the object of interest (segmentation utility), as well as have improved localization capabilities in general (detection utility). To inform future saliency map development, we can consider some aspects of the better performing ones. Among the saliency maps that were tested, XRAI performed the best for both types of utility tasks. For each test image in the dataset, XRAI segments the image into small regions, iteratively evaluates the relevance of each region to the model prediction, and aggregates the smaller regions into a larger region based on the relevance scores [28]. This iterative evaluation of small patches within the image likely gives XRAI an advantage over other methods, since it results in maps with better fine-grained localization catering to adjacent spatial neighborhoods, thus achieving a higher recall and precision than the other methods. In the detection utility task, GradCAM performed the second best. GradCAM [16] generates maps with smooth activations (since activations are derived from gradients that are up-sampled from a lower-dimensional convolutional layer). This smoothness likely plays a role in having an improved AUPRC since the map is less granular compared to other saliency maps. GradCAM's smoothness property provides a strong advantage with high recall since more activations within the area covered by the segmentations or bounding boxes likely have higher values. However, with segmentations being more granular than the up-sampled GradCAM map, there would be a more severe penalty when GradCAM shows high activations in areas outside the segmented area. Therefore, GradCAM would be safer to use on applications with less granular localization, such as a detection task (with bounding boxes). When considering how to build a saliency map with improved pixel-level localization, some of these aspects of XRAI and GradCAM may be of value to consider.

Additionally, for both datasets, under the application of cascading randomization across the different layers of the InceptionV3 model, GradCAM and GGCAM demonstrate a degraded SSIM that goes below the previously defined degradation threshold. This demonstrates GradCAM's sensitivity to trained model weights versus randomly initialized weights, which should be an important consideration when choosing a saliency map to use for prediction interpretation (if a saliency map does not change considerably after trained model weights have been randomized, this implies that the saliency map is not model weight-sensitive). GradCAM forward propagates test images through the model to obtain a prediction, then backpropagates the gradient of the predicted class to the the desired convolutional feature map [16]. As a result, there is a high sensitivity to the value of the weights in the model. In contrast, other saliency methods that do not show as much sensitivity to the value of model weights are more affected by details in the image input (e.g., contrast, sharpness, etc.) than by

the trained model weights. Although GBP demonstrated degradation when cascading randomization reached the bottom-most layer, the SSIM remained above the degradation threshold across all other layers indicating a limitation of GBP's sensitivity to trained weights. There is a study that found that GBP performs a partial image recovery that is unrelated to the trained neural network's decisions [33]. The study found that backward ReLU and local connections in the trained CNNs are two main contributors to this effect. In regards to local connections, with the exception of neurons in the first layer of a CNN, each neuron is only associated with a small, local group of pixels from the input image (i.e., all the input that goes into a neuron is spatially close to each other). When backward ReLU performed, theoretical analysis has shown that local connections cause the resulting visualizations to be more human-interpretable but less class-sensitive [33]. Properties like these should be accounted for when developing a saliency technique.

Finally, for both datasets, XRAI demonstrates the highest repeatability score between two separately trained models with the same architecture and also the highest reproducibility between two separately trained models with different architectures. XRAI's aggregation of smaller regions into larger regions likely reduces the influence of variability across trained models with similar architectures. Thus, the overall model weight distribution should remain the same in a specific area for a particular image even if the models are separately trained. The aggregation of smaller regions to larger regions likely also has a stabilizing effect even on models of different architectures. However, these properties have not yet been extensively studied since XRAI is a fairly new saliency method. It is also notable that the spatial resolution of the generated heatmaps are not the same across methods. In particular, GradCAM's resolution depends on the dimensions of the final convolutional layer of the network, which is up-sampled to fit the dimensions of the model input images (in our case the final dimensions of the DenseNet121 and InceptionV3 models is 10×10) [16]. A single activation change in the 10×10 feature map might result in a high degree of variability in the corresponding enlarged GradCAM generated map, hence influencing the repeatability and reproducibility scores.

Supplementary Figures

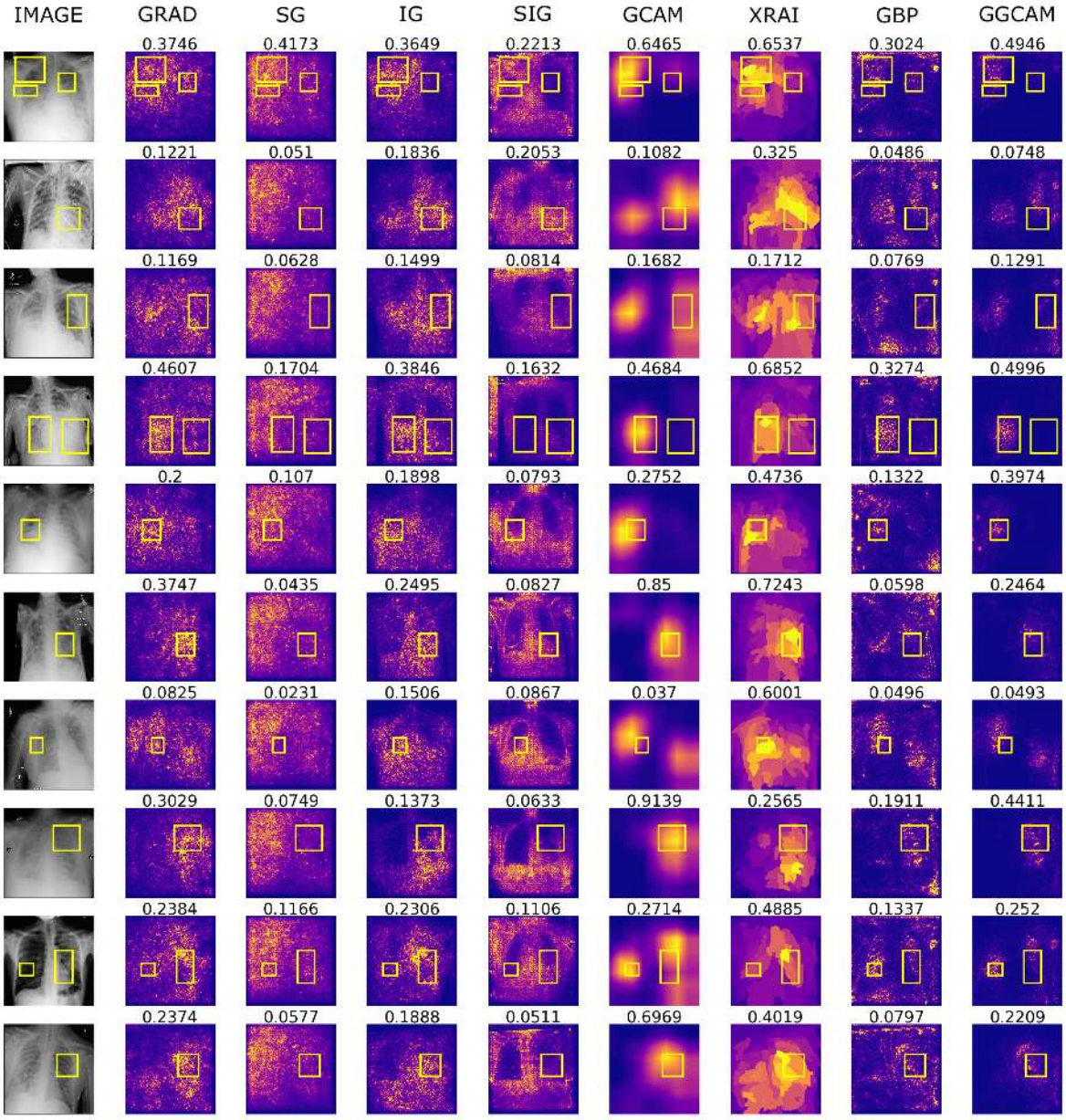


Fig. S1: Additional visualizations of saliency maps generated using an InceptionV3 on examples from the RSNA Pneumonia dataset (with yellow bounding boxes). Numbers above each image denote the SSIM score.

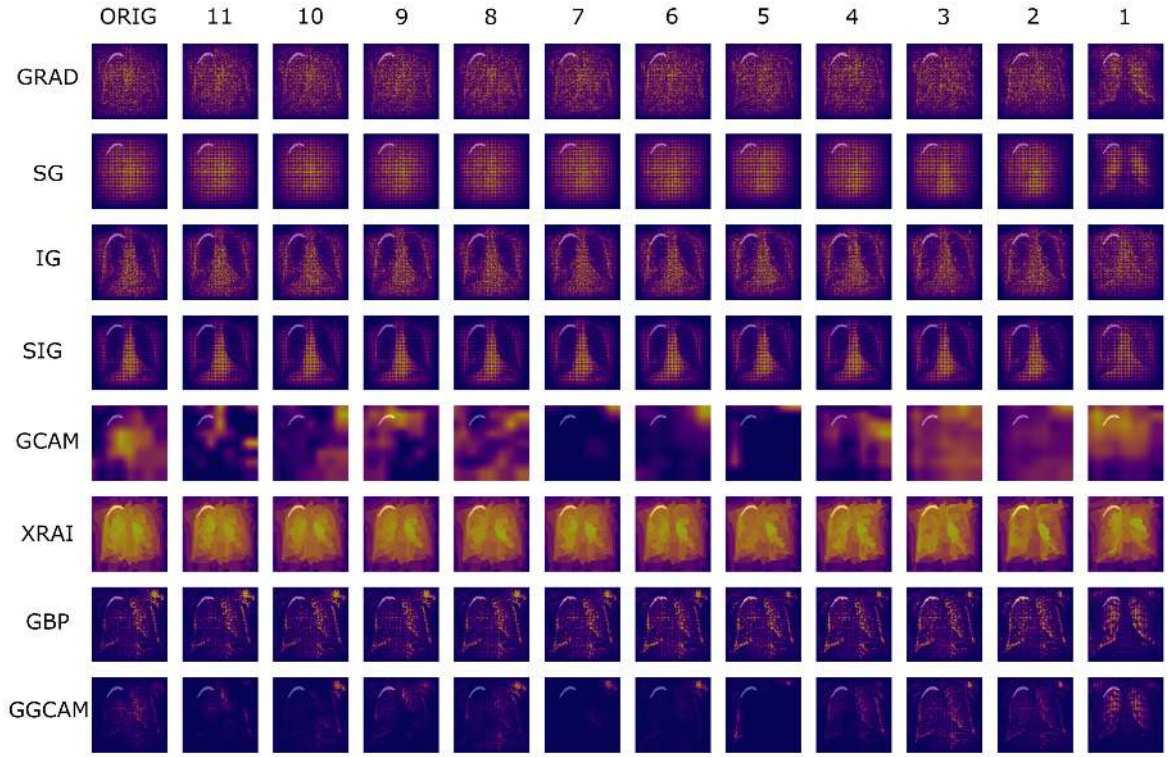
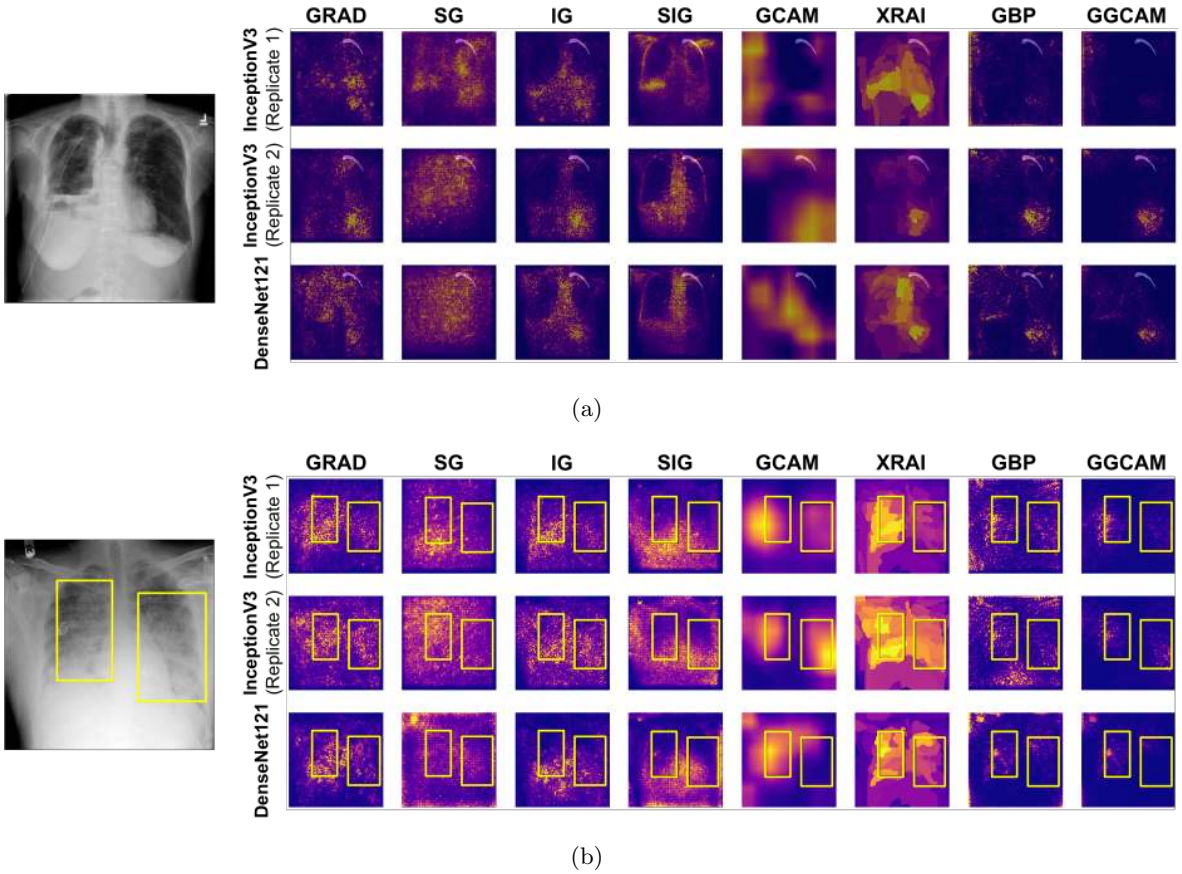


Fig. S2: Additional visualizations of cascading randomization using InceptionV3 on examples from the SIIM Pneumothorax dataset (with transparent segmentations).

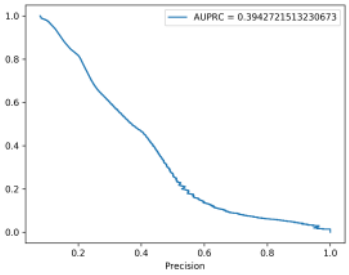


(a)

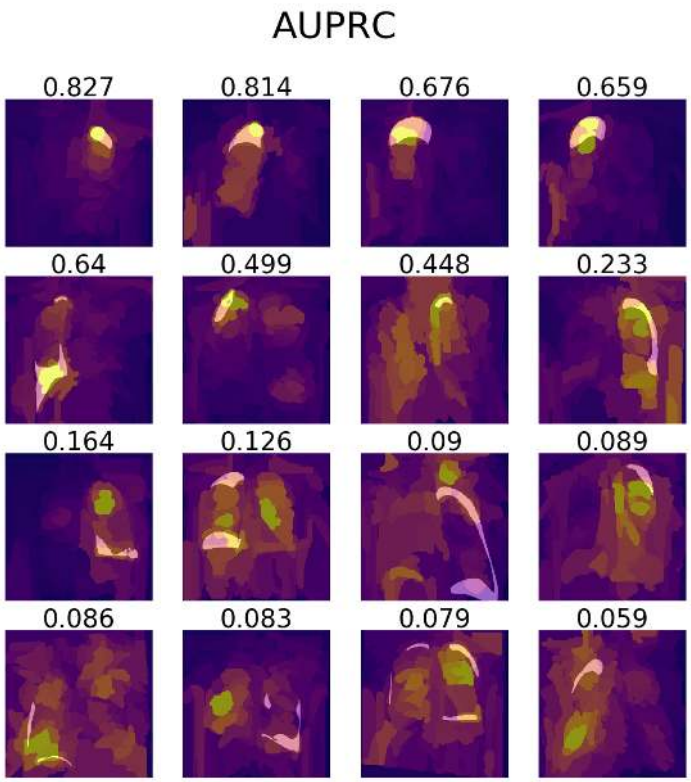
(b)

Fig. S3: Two additional examples of repeatability (InceptionV3 Replicates 1 and 2) and reproducibility (InceptionV3 and DenseNet121) for the (a) SIIM-ACR Pneumothorax dataset with transparent segmentations and (b) RSNA Pneumonia dataset with yellow bounding boxes

(a)



(b)



(c)

Fig. S4: (a) An example of a transparent pneumothorax segmentation compared to its corresponding GradCAM saliency map; (b) the corresponding precision-recall curve for the saliency map's overlap with the ground truth segmentation; (c) examples of XRAI saliency map overlap with ground truth pneumothorax segmentation with corresponding value of AUPRC.

Supplementary Tables

Table S1: Mean Structural SIMilarity (SSIM) index scores of saliency maps for fully randomized models (i.e., after cascading randomization has reached the bottom-most layer) and degradation thresholds (defined in the Results section) for the SIIM-ACR Pneumothorax segmentation and RSNA Pneumonia detection datasets. GRAD-Gradient Explanation; SG-Smooth Gradients; IG-Integrated Gradients; SIG-Smooth Integrated Gradients; GCAM-GradCAM; GBP-Guided-backprop; GGCAM-Guided GradCAM

	SSIM of Fully Randomized Model		**Degradation Threshold**	
	Segmentation	**Detection**	**Segmentation**	**Detection**
GRAD	0.250 ± 0.046	0.173 ± 0.001	0.239 ± 0.046	0.200 ± 0.029
SG	0.565 ± 0.204	0.253 ± 0.024	0.314 ± 0.204	0.315 ± 0.010
IG	0.306 ± 0.043	0.230 ± 0.001	0.239 ± 0.043	0.212 ± 0.036
SIG	0.567 ± 0.134	0.324 ± 0.013	0.501 ± 0.134	0.600 ± 0.114
GCAM	0.430 ± 0.106	0.369 ± 0.103	0.435 ± 0.106	0.542 ± 0.137
XRAI	0.814 ± 0.095	0.723 ± 0.005	0.631 ± 0.095	0.646 ± 0.053
GBP	0.355 ± 0.076	0.234 ± 0.022	0.359 ± 0.076	0.299 ± 0.079
GGCAM	0.409 ± 0.062	0.494 ± 0.029	0.627 ± 0.062	0.564 ± 0.092