Title

Benchmarking saliency methods for chest X-ray interpretation

Authors

Adriel Saporta MS MBA[1*], Xiaotong Gui MS[1*], Ashwin Agrawal MS[1*], Anuj Pareek MD PhD[2], Steven QH Truong MBA[3], Chanh DT Nguyen PhD[3,4], Van-Doan Ngo MD[5], Jayne Seekins DO[6], Francis G. Blankenberg MD[6], Andrew Y. Ng PhD[1], Matthew P. Lungren MD MPH[2†], Pranav Rajpurkar PhD [7†]

Affiliations

[1]Stanford University Department of Computer Science, USA

[2]Stanford Center for Artificial Intelligence in Medicine and Imaging, USA

[3]VinBrain, Vietnam

[4]VinUniversity, Vietnam

[5]Vinmec International Hospital, Vietnam

[6]Stanford University School of Medicine, Department of Radiology, USA

[7]Department of Biomedical Informatics, Harvard University, USA

[*]These authors contributed equally: Adriel Saporta, Xiaotong Gui, Ashwin Agrawal

†These authors contributed equally: Matthew P. Lungren, Pranav Rajpurkar

Corresponding author: Pranav Rajpurkar, PhD (pranav_rajpurkar@hms.harvard.edu)

Current word count: 3829

Abstract

Saliency methods, which “explain” deep neural networks by producing heat maps that highlight the areas of the medical image that influence model prediction, are often presented to clinicians as an aid in diagnostic decision-making. Although many saliency methods have been proposed for medical imaging interpretation, rigorous investigation of the accuracy and reliability of these strategies is necessary before they are integrated into the clinical setting. In this work, we quantitatively evaluate three saliency methods (Grad-CAM, Grad-CAM++, and Integrated Gradients) across multiple neural network architectures using two evaluation metrics. We establish the first human benchmark for chest X-ray interpretation in a multilabel classification set up, and examine under what clinical conditions saliency maps might be more prone to failure in localizing important pathologies compared to a human expert benchmark. We find that (i) while Grad-CAM generally localized pathologies better than the two other saliency methods, all three performed significantly worse compared with the human benchmark; (ii) the gap in localization performance between Grad-CAM and the human benchmark was largest for pathologies that had multiple instances, were smaller in size, and had shapes that were more complex; (iii) model confidence was positively correlated with Grad-CAM localization performance. Our work demonstrates that several important limitations of saliency methods must be addressed before we can rely on them for deep learning explainability in medical imaging.

Introduction

Deep learning has enabled automated medical imaging interpretation at the level of practicing experts in some settings[1–3]. While the potential benefits of automated diagnostic models are numerous, lack of model interpretability in the use of "black-box" deep neural networks (DNNs) represents a major barrier to clinical trust and adoption[4–6]. In fact, it has been argued that the European Union's recently adopted General Data Protection Regulation (GDPR) affirms an individual's right to an explanation in the context of automated decision-making[7]. Although the importance of DNN interpretability is widely acknowledged and many techniques have been proposed, little emphasis has been placed on how best to quantitatively evaluate these explainability methods[8].

One type of DNN interpretation strategy widely used in the context of medical imaging is based on saliency (or pixel-attribution) methods[9–12]. Saliency methods produce heat maps highlighting the areas of the medical image that most influenced the DNN's prediction. The heat maps help to visualize whether a DNN is concentrating on the same regions of a medical image that a human expert would focus on, rather than concentrating on a clinically irrelevant part of the medical image or even on confounders in the image[13–15]. Saliency methods have been widely used for a variety of medical imaging tasks and modalities including, but not limited to, visualizing the performance of a convolutional neural network (CNN) in predicting (1) myocardial infarction[16] and hypoglycemia[17] from electrocardiograms, (2) visual impairment[18], refractive error[19], and anaemia[20] from retinal photographs, (3) long-term mortality[21] and tuberculosis[22] from chest X-ray (CXR) images, and (4) appendicitis[23] and pulmonary embolism[24] on computed tomography scans. However, recent work has shown that saliency methods used to validate model

predictions can be misleading in some cases and may lead to increased bias and loss of user trust in high-stakes contexts such as healthcare[25–27]. Therefore, a rigorous investigation of the accuracy and reliability of these strategies is necessary before they are integrated into the clinical setting[28].

In this work, we perform a systematic evaluation of the three most common saliency methods in medical imaging (Grad-CAM[29], Grad-CAM++[30], and Integrated Gradients[31]) using three common CNN architectures (DenseNet121[32], ResNet152[33], Inception-v4[34]). In doing so, we establish the first human benchmark in CXR localization by collecting radiologist segmentations for 10 pathologies using CheXpert, a large publicly available CXR dataset[35]. To compare saliency method segmentations with expert segmentations, we use two metrics to capture localization accuracy: (1) *mean Intersection over Union*, a stricter metric that measures the overlap between the saliency method segmentation and the expert segmentation, and (2) *hit rate*, a less strict metric that does not require the saliency method to locate the full extent of a pathology. We find that (1) while Grad-CAM generally localizes pathologies more accurately than the two other saliency methods, all three perform significantly worse compared with a human radiologist benchmark; (2) the gap in localization performance between Grad-CAM and the human benchmark is largest for pathologies that have multiple instances on the same CXR, are smaller in size, and have shapes that are more complex; (3) model confidence is positively correlated with Grad-CAM localization performance. We publicly release a development dataset of expert segmentations, which we call CheXplanation, to facilitate further research in DNN explainability for medical imaging.

Results

Framework for evaluating saliency methods on multi-label classification models

Three methods were evaluated—Grad-CAM, Grad-CAM++, and Integrated Gradients—in a multi-label classification setup on the CheXpert dataset (Fig. 1a). For each of the three saliency methods, we ran experiments using three CNN architectures previously used on CheXpert: DenseNet121, ResNet152, and Inception-v4. For each combination of saliency method and model architecture, we trained and evaluated an ensemble of 30 CNNs (see Methods for ensembling details). We then passed each of the CXRs in the dataset's holdout test set into the trained ensemble model to obtain image-level predictions for the following 10 pathologies: Atelectasis, Cardiomegaly, Consolidation, Edema, Enlarged Cardiomediastinum, Lung Lesion, Lung Opacity, Pleural Effusion, Pneumothorax, and Support Devices. For each CXR, we used the saliency method to generate heat maps, one for each of the 10 pathologies, and then applied a threshold to each heat map to produce binary segmentations (top row, Fig. 1a). Thresholding is determined per pathology using Otsu's method[36], which iteratively searches for a threshold value that maximizes inter-class pixel intensity variance. We also conducted a sensitivity analysis of localization performance using different thresholds. The result shows that our evaluation of localization performance is robust to different saliency map thresholding values (see Supplementary Fig. 15). Additionally, to calculate the hit rate evaluation metric (described below), we extracted the pixel in the saliency method heat map with the largest value as the single most representative point on the CXR for that pathology.

We obtained two independent sets of pixel-level CXR segmentations on the holdout test set: ground-truth segmentations drawn by two board-certified radiologists (middle row, Fig. 1a) and human benchmark segmentations drawn by a separate group of three board-certified radiologists (bottom row, Fig. 1a). The human benchmark segmentations and the saliency method segmentations were compared with the ground-truth segmentations to establish the human benchmark localization performance and the saliency method localization performance, respectively. Additionally, for the hit rate evaluation metric, the radiologists who drew the benchmark segmentations were also asked to locate a single point on the CXR that was most representative of the pathology at hand (see Supplementary Figs. 1 through 11 for detailed instructions given to the radiologists).

We used two evaluation metrics to compare segmentations (Fig. 1b). First, we used *mean Intersection over Union* (mIoU), a stricter metric that measures how much, on average, either the saliency method or benchmark segmentations overlapped with the ground-truth segmentations. Second, we used *hit rate*, a less strict metric that does not require the saliency method or benchmark annotators to locate the full extent of a pathology. Hit rate is based on the pointing game setup[37], in which credit is given if the most representative point identified by the saliency method or the benchmark annotators lies within the ground-truth segmentation. A “hit” indicates that the correct region of the CXR was located regardless of the exact bounds of the binary segmentations. Localization performance is then calculated as the hit rate across the dataset[38].

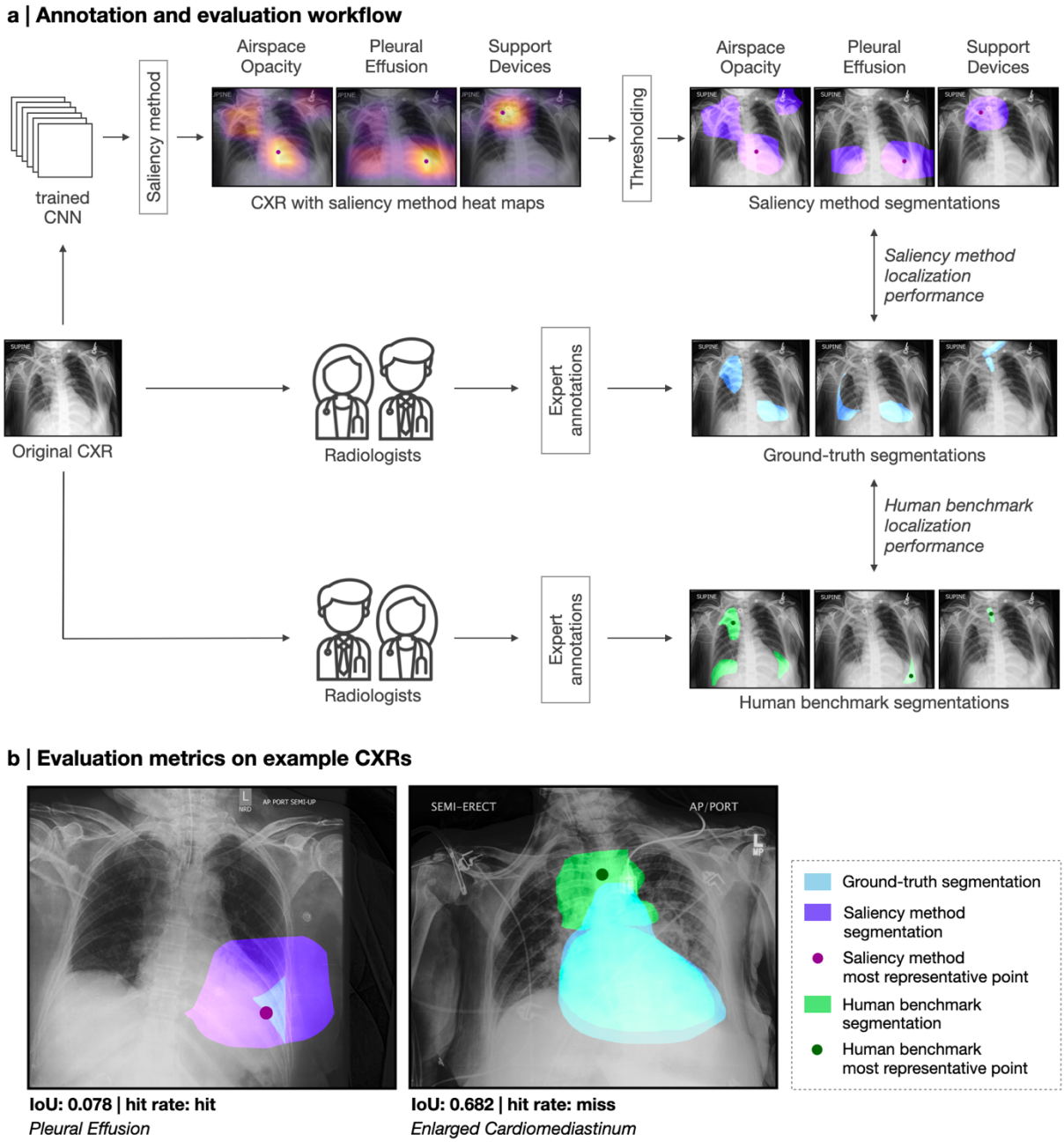


Fig. 1 | Framework for evaluating saliency methods on multi-label classification models. a, Top row left: a CXR image from the holdout test set is passed into an ensemble CNN trained only on CXR images and their corresponding pathology task labels. Saliency method is used to generate 10 heat maps for the example CXR, one for each task. The pixel in the heat map with the largest value is determined to be the single most representative point on the CXR for that pathology. Top row middle: there are three pathologies present in this CXR (Airspace Opacity, Pleural Effusion, and Support Devices). Top row right: a threshold is applied to the heat maps to produce binary segmentations for each present pathology. Middle row: Two board-certified radiologists were asked to segment the pathologies that were present in the CXR as determined by the dataset's ground-truth labels. Saliency method annotations are compared to these

ground-truth annotations to evaluate how well saliency method identifies clinically-relevant areas of the input CXR ("saliency method localization performance"). Bottom row: Two board-certified radiologists (separate from those in middle row) were also asked to segment the pathologies that were present in the CXR as determined by the dataset's ground-truth labels. In addition, these radiologists were asked to locate the single point on the CXR that was most representative of each present pathology. These benchmark annotations are compared to the ground-truth annotations to determine a human benchmark ("human benchmark localization performance"). **b**, Left: CXR with ground-truth and saliency method annotations for Pleural Effusion. The segmentations have a low overlap (IoU is 0.078), but pointing game is a "hit" since the saliency method's most representative point is inside of the ground-truth segmentation. Right, CXR with ground-truth and human benchmark annotations for Enlarged Cardiomediastinum. The segmentations have a high overlap (IoU is 0.682), but pointing game is a "miss" since saliency method's most representative point is outside of the ground-truth segmentation.

Evaluating localization performance of the saliency methods against the human benchmark

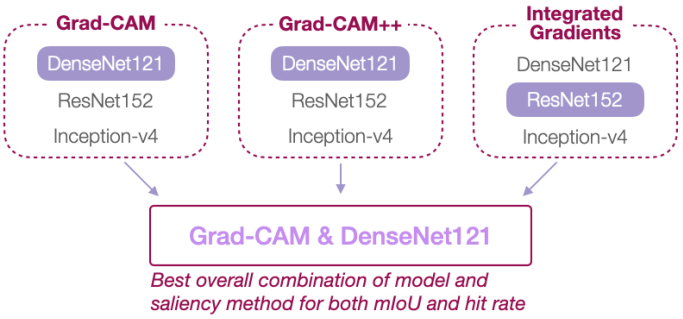
In order to compare the localization performance of the saliency methods with the human benchmark, we ran eighteen experiments, one for each combination of saliency method (Grad-CAM, Grad-CAM++, or Integrated Gradients) and CNN architecture (DenseNet121, ResNet152, or Inception-v4) using one of the two evaluation metrics (mIoU or hit rate). For each evaluation metric, we chose the combination of saliency method and architecture that demonstrated the best localization performance (Fig. 2a). We found that Grad-CAM with DenseNet121 had the highest mIoU performance and the highest hit rate performance. Accordingly, we compared Grad-CAM with DenseNet121 ("saliency method pipeline") with the human benchmark using both mIoU and hit rate. The localization performance for each pathology is reported on the true positive slice of the dataset (CXRs that contain both saliency method and human benchmark segmentations when the ground-truth label of the pathology is positive). Localization performance was calculated this way so that saliency methods were not penalized by DNN classification error: while the benchmark radiologists were provided with ground-truth labels when

annotating the dataset, saliency method segmentations were created based on labels predicted by the model. (See Supplementary Fig. 16 for localization performance results on the full dataset.)

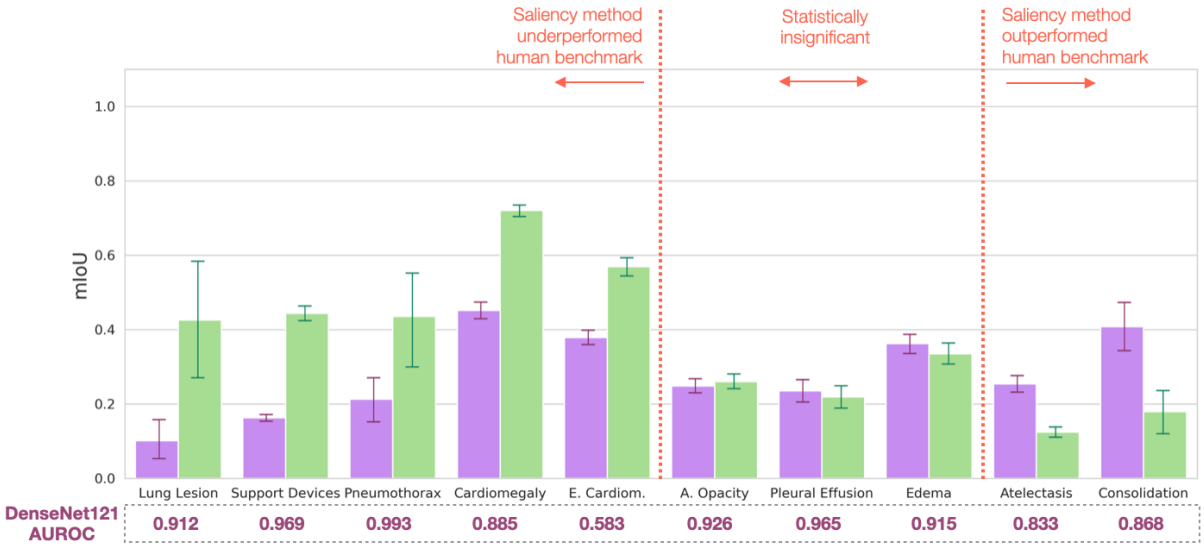
We found that the saliency method pipeline demonstrated significantly worse localization performance when compared with the human benchmark using both mIoU (Fig. 2b) and hit rate (Fig. 2c) as an evaluation metric, regardless of model classification AUROC. For each metric, we report the 95% confidence intervals using the bootstrap method with 1,000 bootstrap samples[39]. For five of the 10 pathologies, the saliency method pipeline had a significantly lower mIoU than the human benchmark. For example, the saliency method pipeline had one of the highest AUROC scores of the 10 pathologies for Support Devices (0.969), but had among the worst localization performance for Support Devices when using both mIoU (0.163 [95% CI 0.154, 0.172]) and hit rate (0.357 [95% CI 0.303, 0.408]) as evaluation metrics. On two pathologies (Atelectasis and Consolidation) the saliency method pipeline significantly outperformed the human benchmark. On average, across all 10 pathologies, mIoU saliency method pipeline performance was 26.6% [95% CI 18.1%, 35.0%] worse than the human benchmark, with Lung Lesion displaying the largest gap in performance (76.2% [95% CI 59.1%, 87.5%] worse than the human benchmark) (Supplementary Table 4). Consolidation was the pathology on which the mIoU saliency method pipeline performance exceeded the human benchmark the most, by 56.1% [95% CI 42.7%, 69.4%]. For seven of the 10 pathologies, the saliency method pipeline had a significantly lower hit rate than the human benchmark. On average, hit rate saliency method pipeline performance was 29.4% [95% CI 15.0%, 43.2%] worse than the

human benchmark (Supplementary Table 5), with Lung Lesion again displaying the largest gap in performance (65.9% [95% CI 35.3%, 91.7%] worse than the human benchmark). The hit rate saliency method pipeline did not significantly outperform the human benchmark on any of the 10 pathologies; for the remaining three of the 10 pathologies, the hit rate performance differences between the saliency method pipeline and the human benchmark were not statistically significant. Therefore, while the saliency method pipeline significantly underperformed the human benchmark regardless of evaluation metric used, the average performance gap was larger when using hit rate as an evaluation metric than when using mIoU as an evaluation metric.

a | Selection strategy for saliency method pipeline



b | mIoU localization performances



c | Hit rate localization performances

210

Fig. 2 | Evaluating the localization performance of the saliency methods against the human benchmark. a, The selection strategy for the mIoU saliency method pipeline and the hit rate saliency method pipeline. For each saliency method, the best model architecture is selected (highlighted in purple). Then, the best saliency method + model architecture pair, of the three, is selected. The selection strategy was performed twice, once using mIoU as the evaluation metric and once using hit rate as the evaluation metric. The best saliency method + model architecture pair was the same for both mIoU and hit rate: Grad-CAM + DenseNet121. **b**, Comparing saliency method and human benchmark localization performances under the overlap evaluation scheme (mIoU). Pathologies, along with their DenseNet121 AUROCs, are sorted on the x-axis in descending order of percentage decrease from human benchmark mIoU to saliency method pipeline mIoU for each pathology. **c**, Comparing saliency method and human benchmark localization performances under the hit rate evaluation scheme. Pathologies, along with their DenseNet121 AUROCs, are sorted on the x-axis in descending order of percentage decrease from human benchmark hit rate to saliency method pipeline hit rate for each pathology.

Characterizing the underperformance of the saliency method pipeline localization

In order to better understand the underperformance of the saliency method pipeline localization, we first conducted a qualitative analysis with a radiologist by visually inspecting both the segmentations produced by the saliency method pipeline (Grad-CAM with DenseNet121) and the human benchmark segmentations. We found that, in general, saliency method segmentations fail to capture the geometric nuances of a given pathology, and instead produce coarse, low-resolution heat maps. Specifically, our qualitative analysis found that the performance of the saliency method depended on three pathological characteristics (Fig. 3a): (1) *number of instances*: when a pathology had multiple instances on a CXR, the saliency method segmentation often highlighted one large confluent area, instead of highlighting each distinct instance of the pathology separately; (2) *size*: saliency method segmentations tended to be significantly larger than human expert segmentations, often failing to respect clear anatomical boundaries; (3) *shape complexity*: the saliency method segmentations for pathologies with complex shapes frequently included significant portions of the CXR where the pathology is not present.

Informed by our qualitative analysis and previous work in histology[40], we defined four geometric features for our quantitative analysis (Fig. 3b): (1) *number of instances* (for example, bilateral Pleural Effusion would have two instances, whereas there is only one instance for Cardiomegaly), (2) *size* (pathology area with respect to the area of the whole CXR), (3) *elongation* and (4) *irrectangularity* (the last two features measure the complexity of the pathology shape and were calculated by fitting a rectangle of minimum area enclosing the binary mask). See Supplementary Fig. 17 for the distribution of the four pathological characteristics across all 10 pathologies.

For each evaluation metric, we ran 8 simple linear regressions: four with the evaluation metric (IoU or hit rate) of the saliency method pipeline (Grad-CAM with DenseNet121) as the dependent variable (to understand the relationship between the geometric features of a pathology and saliency method localization performance), and four with the difference between the evaluation metrics of the saliency method pipeline and the human benchmark as the dependent variable (to understand the relationship between the geometric features of a pathology and the gap in localization performance between the saliency method pipeline and the human benchmark). Each regression used one of the four geometric features as a single independent variable, and only the true positive slice was included in each regression. Each feature was normalized using z-score normalization and the regression coefficient can be interpreted as the effect of that geometric feature on the evaluation metric at hand. See Table 1 for coefficients from the regressions using both evaluation metrics, where we also report the 95% confidence

interval and the Bonferroni corrected p-values. For confidence intervals and p-values, we used the standard calculation for linear models.

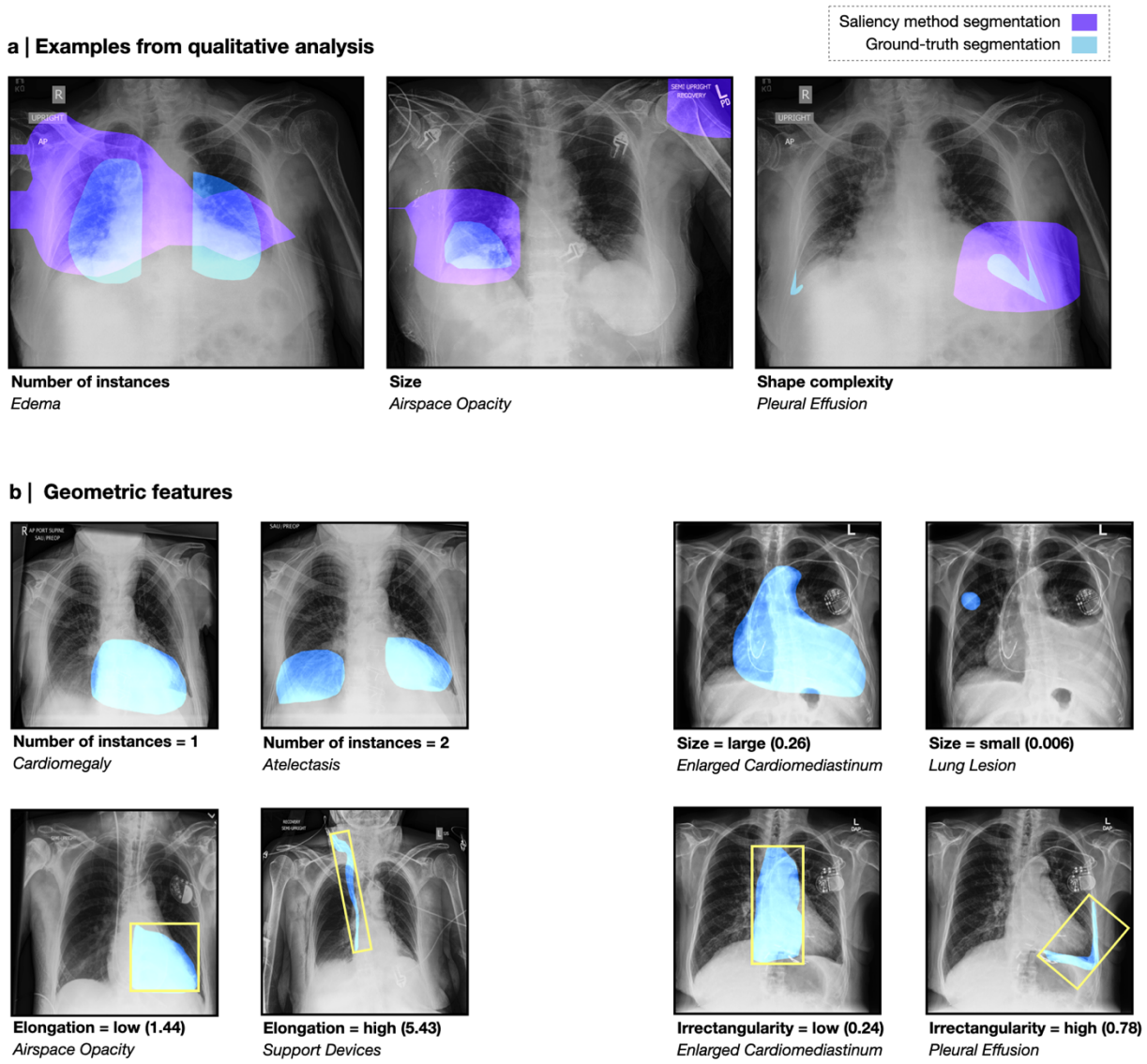


Fig. 3 | Characterizing the underperformance of saliency method localization. a, Example CXRs that highlight the three pathological characteristics identified by our qualitative analysis: (1) Left, number of instances; (2) Middle, size; and (3) Right, shape complexity. **b**, Example CXRs with the four geometric features used in our quantitative analysis: (1) Top row left, number of instances; (2) Top row right, size = area of segmentation/area of CXR; (3) Bottom row left, elongation; and (4) Bottom row right, irrectangularity. Elongation and irrectangularity were calculated by fitting a rectangle of minimum area enclosing the binary mask (as indicated by the yellow rectangles). Elongation = maxAxis/minAxis. Irrectangularity = 1 - (area of segmentation/area of enclosing rectangle).

Our statistical analysis showed that as the area ratio of a pathology increased, mIoU saliency method localization performance improved (0.566 [95% CI 0.526, 0.606]). We also found that as elongation and irrectangularity increased, mIoU saliency method localization performance worsened (elongation: -0.425 [95% CI -0.497, -0.354], irrectangularity: -0.256 [95% CI -0.292, -0.219]). We observed that the effects of these three geometric features were similar for hit rate saliency method localization performance in terms of levels of statistical significance and direction of the effects. However, there was no evidence that the number of instances of a pathology had a significant effect on either mIoU (-0.115 [95% CI -0.220, -0.009]) or hit rate (-0.051 [95% CI -0.364, 0.244]) saliency method localization. Therefore, regardless of evaluation metric, saliency method localization performance suffered in the presence of pathologies that were small in size and complex in shape.

We found that these same three pathological characteristics—larger size, and higher elongation and irrectangularity—characterized the *gap* in mIoU localization performance between saliency method and human benchmark. We observed that the *gap* in hit rate localization performance was significantly characterized by all four geometric features (number of instances, size, elongation, and irrectangularity). As the number of instances increased, despite no significant change in hit rate localization performance itself, the *gap* in hit rate localization performance between saliency method and the human benchmark increased (0.470 [95% CI 0.114, 0.825]). This suggests that the saliency method performs especially poorly in the face of a multi-instance diagnosis.

Table 1 | Coefficients from regressions on geometric features of pathologies

Geometric feature (independent variable)	Coefficient using saliency method localization (dependent variable)	Coefficient using localization difference (human benchmark - saliency method) (dependent variable)
IoU		
Number of instances	-0.115 (-0.220, -0.009)	-0.072 (-0.237, -0.094)
Size	0.566 (0.526, 0.606) ***	-0.154 (-0.231, -0.076) ***
Elongation	-0.425 (-0.497, -0.354) ***	0.476 (0.362, 0.589) ***
Irrectangularity	-0.256 (-0.292 -0.219) ***	0.307 (0.249, 0.366) ***
Hit/Miss		
Number of instances	-0.051 (-0.346, 0.244)	0.470 (0.114, 0.825) *
Size	1.269 (1.146, 1.391) ***	-0.944 (-1.104, -0.785) ***
Elongation	-0.849 (-1.053, -0.646) ***	1.110 (0.865, 1.354) ***
Irrectangularity	-0.519 (-0.624, -0.415) ***	0.689 (0.564, 0.815) ***

* p-value < 0.05, ** p-value < 0.01, *** p-value < 0.001

Effect of model confidence on saliency method localization performance

We also conducted statistical analyses to determine whether there was any correlation between the model's confidence in its prediction and saliency method pipeline performance (Table 2). We first ran a simple regression for each pathology using the model's probability output as the single independent variable and using the saliency method IoU as the dependent variable. We then performed a simple regression that uses the same approach as above, but that includes all 10 pathologies. For each of the 11 regressions, we used the full dataset since the analysis of false positives and false negatives was also of interest. In addition to the linear regression coefficients, we also computed the Spearman correlation coefficients to capture any potential non-linear associations.

We found that for all pathologies, model confidence was positively correlated with mIoU saliency method pipeline performance. The p-values for all coefficients were below 0.001 except for the coefficients for Pneumothorax (n=11) and Lung Lesion (n=50), the two pathologies for which we had the fewest positive examples. Of all the pathologies, model confidence for positive predictions of Enlarged Cardiomediastinum had the largest linear regression coefficient with mIoU saliency method pipeline performance (1.974, p-value<0.001). Model confidence for positive predictions of Pneumothorax had the largest Spearman correlation coefficient with mIoU saliency method pipeline performance (0.734, p-value<0.01), followed by Pleural Effusion (0.690, p-value<0.001). Combining all pathologies (n=2365), the linear regression coefficient was 0.109 (95% CI [0.083, 0.135]), and the Spearman correlation coefficient was 0.285 (95% CI [0.239, 0.331]). We also performed analogous experiments using hit rate as the dependent variable and found comparable results (Supplementary Table 1).

Table 2 | mIoU: Coefficients from regressions on model assurance

Pathology	CXRs (*n*)	Linear regression coefficient	Spearman correlation coefficient
Airspace Opacity	381	0.714 (0.601, 0.826) ***	0.577 (0.542, 0.61) ***
Atelectasis	296	0.489 (0.333, 0.645) ***	0.348 (0.303, 0.391) ***
Cardiomegaly	229	0.679 (0.535, 0.823) ***	0.592 (0.559, 0.624) ***
Consolidation	120	1.155 (0.674, 1.635) ***	0.384 (0.341, 0.426) ***
Edema	124	0.642 (0.459, 0.826)***	0.548 (0.512, 0.582) ***
Enlarged Cardiomediastinum	668	1.974 (1.608, 2.340) ***	0.428 (0.386, 0.468) ***
Lung Lesion	50	0.218 (0.087, 0.349) **	0.509 (0.47, 0.545) ***
Pleural Effusion	159	0.632 (0.489, 0.776) ***	0.690 (0.663, 0.715) ***
Pneumothorax	11	0.446 (0.108, 0.783) *	0.734 (0.710, 0.756) **
Support Devices	327	0.211 (0.172, 0.25) ***	0.468 (0.428, 0.506) ***
All pathologies	2365	0.109 (0.083, 0.135) ***	0.285 (0.239, 0.331) ***

* p-value < 0.05, ** p-value < 0.01, *** p-value < 0.001

Discussion

The purpose of this work was to evaluate the performance of some of the most used saliency methods (Grad-CAM, Grad-CAM++, Integrated Gradients) for deep learning explainability using a variety of model architectures. We establish the first human benchmark for CXR localization in a multilabel classification setup and demonstrate that saliency maps are consistently worse than expert radiologists regardless of model classification AUROC. We use qualitative and quantitative analyses to establish that saliency method localization performance is most inferior to expert localization performance when a pathology has multiple instances, is smaller in size, or has shapes that are more complex, suggesting that deep learning explainability as a clinical interface may be less reliable and less useful when used for pathologies with those characteristics. We also show that model assurance is positively correlated with saliency method localization performance, which could indicate that saliency methods are safer to use as a decision aid to clinicians when the model has made a positive prediction with high confidence.

While there are several public CXR datasets with image-level labels annotated by expert radiologists, including MIMIC-CXR[41] and ChestX-ray8[42], and several datasets containing segmentations for a single pathology, including SIIM-ACR Pneumothorax Segmentation[43] and RSNA Pneumonia Detection[44], to our knowledge there are no other publicly available CXR datasets with multilabel expert segmentations. By publicly releasing a development dataset, CheXplanation, of 234 images with 885 expert segmentations, and a competition

with a test set of 668 images, we hope to encourage the further development of saliency methods and other explainability techniques for medical imaging.

Our work has several potential implications for human-AI collaboration in the context of medical decision-making. Heat maps generated using saliency methods are advocated as clinical decision support in the hope that they not only improve clinical decision-making, but also encourage clinicians to trust model predictions[45–47]. Many of the large CXR vendors[48–50] use localization methods to provide pathology visualization in their computer-aided detection (CAD) products. In addition to being used for clinical interpretation, saliency method heat maps are also used for the evaluation of CXR interpretation models, for quality improvement (QI) and quality assurance (QA) in clinical practices, and for dataset annotation[51]. However, we found that saliency method localization performance, on balance, performed worse than expert localization across multiple analyses and across many important pathologies (our findings are consistent with recent work focused on localizing a single pathology, Pneumothorax, in CXRs[52]). If used in clinical practice, heat maps that incorrectly highlight medical images may exacerbate well documented biases (chiefly, automation bias) and erode trust in model predictions (even when model output is correct), limiting clinical translation[22].

Since IoU computes the overlap of two segmentations but pointing game hit rate better captures diagnostic attention, we suggest using both metrics when evaluating localization performance in the context of medical imaging. While IoU is a commonly used metric for evaluating semantic segmentation outputs, there are inherent limitations to the metric in

the pathological context. This is indicated by our finding that even the human benchmark segmentations had low overlap with the ground truth segmentations (the highest expert mIoU was 0.720 for Cardiomegaly). One potential explanation for this consistent underperformance is that pathologies can be hard to distinguish, especially without clinical context. Furthermore, whereas many people might agree on how to segment, say, a cat or a stop sign in traditional computer vision tasks, radiologists use a certain amount of clinical discretion when defining the boundaries of a pathology on a CXR. There can also be institutional and geographic differences in how radiologists are taught to recognize pathologies, and studies have shown that there can be high interobserver variability in the interpretation of CXRs[53–55]. We sought to address this with the hit rate evaluation metric, which highlights when two radiologists share the same diagnostic intention, even if it is less exact than IoU in comparing segmentations directly. The human benchmark localization using hit rate was above 0.9 for four pathologies (Pneumothorax, Cardiomegaly, Support Devices, and Enlarged Cardiomediastinum); these are pathologies for which there is often little disagreement between radiologists about where the pathologies are located, even if the expert segmentations are noisy. Further work is needed to demonstrate which segmentation evaluation metrics, even beyond overlap and hit rate, are more appropriate for which pathologies when evaluating saliency methods for the clinical setting.

Our work builds upon several studies investigating the validity of saliency maps for localization[56,57] and upon some early work on the trustworthiness of saliency methods to explain DNNs in medical imaging[58]. However, as recent work has shown[31], evaluating

saliency methods is inherently difficult given that they are post-hoc techniques. To illustrate this, consider the following models and saliency methods as described by some oracle: (1) a model *M_bad* that has perfect AUROC for a given image classification task, but that we know does *not* localize well (i.e. because the model picks up on confounders in the image); (2) a model *M_good* that also has perfect AUROC, but that we know *does* localize well (i.e. is looking at relevant regions of the image); (3) a saliency method *S_bad* that does *not* properly reflect the model's attention; and (4) a saliency method *S_good* that *does* properly reflect the model's attention. Let us say that we are evaluating the following pipeline: we first classify an image and we then apply a saliency method post hoc. Imagine that our evaluation reveals poor localization performance as measured by mIoU or hit rate (as was the case in our findings). There are three possible pipelines (combinations of model and saliency method) that would lead to this scenario: (1) *M_bad* + *S_good*; (2) *M_good* + *S_bad*; and (3) *M_bad* + *S_bad*. The first scenario (*M_bad* + *S_good*) is the one for which saliency methods were originally intended: we have a working saliency method that properly alerts us to models picking up on confounders. The second scenario (*M_good* + *S_bad*) is our nightmare scenario: we have a working model whose attention is appropriately directed, but we reject it based on a poorly localizing saliency method. Because all three scenarios result in poor localization performance, it is difficult—if not impossible—to know whether poor localization performance is attributable to the model or to the saliency method (or to both). While we cannot say whether models or saliency methods are failing in the context of medical imaging, we can say that we should not rely on saliency methods to evaluate model localization. Future work should explore potential techniques for localization performance attribution.

There are several limitations of our work. First, we did not investigate the impact of pathology prevalence in the training data on saliency method localization performance. Second, some pathologies, such as effusions and cardiomegaly, are in similar locations across frontal view CXRs, while others, such as lesions and opacities, can vary in locations across CXRs. Future work could investigate how the location of pathologies on a CXR in the training/test data distribution, and the consistency of those locations, affect saliency method localization performance. Third, while we compared saliency method-generated pixel-level segmentations to human expert pixel-level segmentations, future work might explore how saliency method localization performance changes when comparing bounding-box annotations, instead of pixel-level segmentations. Finally, the impact of saliency methods on the trust and efficacy of users is underexplored.

In conclusion, we present a rigorous evaluation of a range of saliency methods and a human benchmark dataset, which can serve as a foundation for future work exploring deep learning explainability techniques. This work is a reminder that care should be taken when leveraging common saliency methods for deep learning-based workflows for medical imaging.

Methods

Ethical and information governance approvals.

This study does not involve human subject participants.

443 **Dataset and clinical taxonomy.** *Dataset description.* The localization experiments were
444 performed using CheXpert, a large public dataset for chest X-ray interpretation. The
445 CheXpert dataset contains 224,316 chest X-rays for 65,240 patients labeled for the
446 presence of 14 observations (13 pathologies and an observation of "No Finding") as
447 positive, negative, or uncertain. The CheXpert validation set consists of 234 chest X-rays
448 from 200 patients randomly sampled from the full dataset and was labeled according to
449 the consensus of three board-certified radiologists. The test set consists of 668 chest X-
450 rays from 500 patients not included in the training or validation sets and was labeled
451 according to the consensus of five board-certified radiologists. See Supplementary Table
452 2 for dataset summary statistics.
453
454 *Ground-truth segmentation.* The chest X-rays in our validation set and test set were
455 manually segmented by two board-certified radiologists with 18 and 27 years of
456 experience, using the annotation software tool MD.ai[59] (see Supplementary Figs. 12
457 through 14). The radiologists were asked to contour the region of interest for all
458 observations in the chest X-rays for which there was a positive ground truth label in the
459 CheXpert dataset. For a pathology with multiple instances, all the instances were
460 contoured. For Support Devices, radiologists were asked to contour any implanted or
461 invasive devices including pacemakers, PICC/central catheters, chest tubes,
462 endotracheal tubes, feeding tubes and stents and ignore ECG lead wires or external
463 stickers visible in the chest X-ray. Finally, of the 14 observations labeled in the CheXpert
464 dataset, Fracture, Pleural Other, Pneumonia, and No Finding were not segmented

465 because they either had low prevalence and/or ill-defined boundaries unfit for
466 segmentation.
467
468 *Evaluating the expert performance using benchmark segmentation.* To evaluate the
469 expert performance on the test set using the IoU evaluation method, three radiologists,
470 certified in Vietnam with 9, 10, and 18 years of experience, were asked to segment the
471 regions of interest for all observations in the chest X-rays for which there was a positive
472 ground truth label in the CheXpert dataset. These radiologists were also provided the
473 same instructions for contouring as were provided to the radiologists drawing the
474 reference segmentations. To extract the "maximally activated" point from the benchmark
475 segmentations, we asked the same radiologists to locate each pathology present on each
476 CXR using only a single most representative point for that pathology on the CXR (see
477 Supplementary Figs. 1 through 11 for the detailed instructions given to the radiologists).
478 There was no overlap between these three radiologists and the two who drew the
479 reference segmentations.
480
481 **Classification network architecture and training protocol.** *Multi-label classification*
482 *model.* The model takes as input a single-view chest X-ray and outputs the probability for
483 each of the 14 observations. In case of availability of more than one view, the models
484 output the maximum probability of the observations across the views. Each chest X-ray
485 was resized to 320×320 pixels and normalized before it was fed into the network. The
486 model architectures (DenseNet121, ResNet152, and Inception-v4) were used. Cross-
487 entropy loss was used to train the model. The Adam optimizer[60] was used with default β-

parameters of $\beta 1 = 0.9$ and $\beta 2 = 0.999$. The learning rate was hyperparameter tuned for the different model architectures. The best learning rate for each architecture was: 1×10^{-4} for DenseNet121, 1×10^{-5} for ResNet152, 1×10^{-5} for Inceptionv4. Batches were sampled using a fixed batch size of 16 images.

Ensembling. We use an ensemble of checkpoints to create both predictions and saliency maps to maximize model performance. In order to capture uncertainties inherent in radiograph interpretation, we train our models using four uncertainty handling strategies outlined in CheXpert: Ignoring, Zeroes, Ones, and 3-Class Classification. For each of the four uncertainty handling strategies, we train our model three separate times, each time saving the 10 checkpoints across the three epochs with the highest average AUC across 5 observations selected for their clinical importance and prevalence in the validation set: Atelectasis, Cardiomegaly, Consolidation, Edema, and Pleural Effusion. In total, after training, we have saved 4 x 30 = 120 checkpoints for a given model. Then, from the 120 saved checkpoints for that model, we select the top 10 performing checkpoints for each pathology. For each CXR and each task, we compute the predictions and saliency maps using the relevant checkpoints. We then take the mean both of the predictions and of the saliency maps to create the final set of predictions and saliency maps for the ensemble model. See Supplementary Table 3 for the performance of the model on each of the pathologies.

CNN interpretation strategy. Saliency methods (Grad-CAM, Grad-CAM++, and Integrated Gradients) were used to visualize the decision made by the classification

network. The saliency map was resized to the original image dimension using bilinear interpolation. It was then normalized using max-min normalization and then converted into a binary segmentation using binary thresholding (Otsu's method). We also reported mIoU localization performance using different saliency map thresholding values. We first applied max-min normalizations to the saliency maps so that each value gets transformed into a decimal between 0 and 1. We then passed in a range of threshold values from 0.2 to 0.8 to create binary segmentations and calculated the mIoU score per pathology under each threshold on the validation set. Then for the analysis with the full dataset (see Supplementary Figure 16), we further ensure that the final binary segmentation is consistent with model probability output by applying another layer of thresholding such that the segmentation mask produced all zeros if the predicted probability was below a chosen level. The probability threshold is searched on the interval of [0,0.8] with steps of 0.1. The exact value is determined per pathology by maximizing the mIoU on validation set.

Segmentation evaluation metrics. Localization performance of each segmentation was evaluated using Intersection over Union (IoU) score. The IoU is the ratio between the area of overlap and the area of union between the ground truth and the predicted areas, ranging from 0 to 1 with 0 signifying no overlap and 1 signifying perfectly overlapping segmentation. Confidence intervals are calculated using bootstrapping with 1000 bootstrap samples. The variance in the width of CI across pathologies can be explained by difference in sample sizes. For the percentage decrease from expert mIoU to AI mIoU, we bootstrapped the difference between human benchmark and saliency method

localization and created the 95% confidence intervals. The confidence intervals for hit rates were calculated in the same fashion.

Statistical analysis.

Pathology Characteristics. We used four features to characterize the pathologies. (1) Number of instances is defined as the number of disjoint components in the segmentation. (2) Size is the area of the pathology divided by the total image area. (3) and (4) Elongation and irrectangularity are geometric features that measure shape complexities. They were designed to quantify what radiologists qualitatively described as focal or diffused. To calculate the metrics, a rectangle of minimum area enclosing the contour is fitted to each pathology. Elongation is defined as the ratio of the rectangle's longer side to short side. Irrectangularity = 1 - (area of segmentation/area of enclosing rectangle), with values ranging from 0 to 1 with 1 being very irrectangular. When there are multiple instances within one pathology, we used the characteristics of the dominant instance (largest in perimeter).

Model Confidence. We used the probability output of the DNN architecture for model confidence. The probabilities were normalized using max-min normalization per pathology before aggregation.

Linear Regression. For each evaluation scheme (overlap and hit rate), we ran two groups of simple linear regressions, with AI evaluation metrics and their differences as the response variables. Each group has four regressions using the above four pathological

characteristics as the regressions' single attribute, respectively, and only the true positive slice was included in each regression. All features are normalized using min-max normalization so that they are comparable on scales of magnitudes. We report the 95% confidence interval and Bonferroni adjusted p-value of the regression coefficients.

Acknowledgements

We would like to acknowledge MD.ai for generously providing us access to their annotation platform. We would like to acknowledge Weights & Biases for generously providing us access to their experiment tracking tools.

Data Availability

CheXpert data is available at https://stanfordmlgroup.github.io/competitions/chexpert/. The validation set and corresponding benchmark radiologist annotations can be downloaded from https://stanfordmlgroup.github.io/competitions/chexplanation/.

Code Availability

All code used to produce the results of the paper is available in the following public repository for the purpose of reproducing the study: https://github.com/stanfordmlgroup/cheXplanation.

Competing Interests

M.L. is an advisor for and/or has research funded by GE, Philips, Carestream, Nines Radiology, Segmed, Centaur Labs, Microsoft, and BunkerHill.

580

581 **Author Contributions**

582 (1) Conceptualization: P.R. and A.P., (2) Design: P.R., A.P., A.S., X.G. and A.A., (3) Data
583 analysis and interpretation: A.S., X.G., A.A., P.R., A.P., S.T., C.N., V.N., J.S., and F.B.,
584 (4) Drafting of the manuscript: A.S., X.G., A.A., and P.R., (5) Critical revision of the
585 manuscript for important intellectual content: A.P, S.T., C.N., V.N., J.S., F.B, A.N., and
586 M.L., (6) Supervision: A.N., M.L, and P.R.

587

588 **References**

589 1. Rajpurkar, P. *et al.* Deep learning for chest radiograph diagnosis: A retrospective
590 comparison of the CheXNeXt algorithm to practicing radiologists. *PLOS Med.* **15**,
591 e1002686 (2018).
592 2. Rajpurkar, P. *et al.* CheXNet: Radiologist-Level Pneumonia Detection on Chest X-
593 Rays with Deep Learning. *ArXiv171105225 Cs Stat* (2017).
594 3. Bien, N. *et al.* Deep-learning-assisted diagnosis for knee magnetic resonance
595 imaging: Development and retrospective validation of MRNet. *PLOS Med.* **15**,
596 e1002699 (2018).
597 4. Baselli, G., Codari, M. & Sardanelli, F. Opening the black box of machine learning in
598 radiology: can the proximity of annotated cases be a way? *Eur. Radiol. Exp.* **4**, 30
599 (2020).
600 5. Litjens, G. *et al.* A survey on deep learning in medical image analysis. *Med. Image*
601 *Anal.* **42**, 60–88 (2017).

602 6. Wang, F., Kaushal, R. & Khullar, D. Should Health Care Demand Interpretable
603 Artificial Intelligence or Accept “Black Box” Medicine? *Ann. Intern. Med.* **172**, 59–60
604 (2019).
605 7. Goodman, B. & Flaxman, S. European Union regulations on algorithmic decision-
606 making and a ‘right to explanation’. *AI Mag.* **38**, 50–57 (2017).
607 8. Venugopal, V., Takhar, R., Gupta, S., Saboo, A. & Mahajan, V. *Clinical Explainability*
608 *Failure (CEF) & Explainability Failure Ratio (EFR) – changing the way we*
609 *validate classification algorithms?* 2020.08.12.20169607
610 https://www.medrxiv.org/content/10.1101/2020.08.12.20169607v1 (2020)
611 doi:10.1101/2020.08.12.20169607.
612 9. Pasa, F., Golkov, V., Pfeiffer, F., Cremers, D. & Pfeiffer, D. Efficient Deep Network
613 Architectures for Fast Chest X-Ray Tuberculosis Screening and Visualization. *Sci.*
614 *Rep.* **9**, 6268 (2019).
615 10. Simonyan, K., Vedaldi, A. & Zisserman, A. Deep Inside Convolutional Networks:
616 Visualising Image Classification Models and Saliency Maps. *ArXiv13126034 Cs*
617 (2014).
618 11. Aggarwal, M. *et al.* Towards Trainable Saliency Maps in Medical Imaging.
619 *ArXiv201107482 Cs Eess* (2020).
620 12. Tjoa, E. & Guan, C. Quantifying Explainability of Saliency Methods in Deep
621 Neural Networks. *ArXiv200902899 Cs* (2020).
622 13. Badgeley, M. A. *et al.* Deep learning predicts hip fracture using confounding
623 patient and healthcare variables. *Npj Digit. Med.* **2**, 1–10 (2019).

624 14. Zech, J. R. *et al.* Variable generalization performance of a deep learning model
625 to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Med.* **15**,
626 e1002683 (2018).
627 15. DeGrave, A. J., Janizek, J. D. & Lee, S.-I. AI for radiographic COVID-19
628 detection selects shortcuts over signal. *medRxiv* (2020)
629 doi:10.1101/2020.09.13.20193565.
630 16. Makimoto, H. *et al.* Performance of a convolutional neural network derived from
631 an ECG database in recognizing myocardial infarction. *Sci. Rep.* **10**, 8445 (2020).
632 17. Porumb, M., Stranges, S., Pescapè, A. & Pecchia, L. Precision Medicine and
633 Artificial Intelligence: A Pilot Study on Deep Learning for Hypoglycemic Events
634 Detection based on ECG. *Sci. Rep.* **10**, 1–16 (2020).
635 18. Tham, Y.-C. *et al.* Referral for disease-related visual impairment using retinal
636 photograph-based deep learning: a proof-of-concept, model development study.
637 *Lancet Digit. Health* **3**, e29–e40 (2021).
638 19. Varadarajan, A. V. *et al.* Deep Learning for Predicting Refractive Error From
639 Retinal Fundus Images. *Invest. Ophthalmol. Vis. Sci.* **59**, 2861–2868 (2018).
640 20. Mitani, A. *et al.* Detection of anaemia from retinal fundus images via deep
641 learning. *Nat. Biomed. Eng.* **4**, 18–27 (2020).
642 21. Deep Learning to Assess Long-term Mortality From Chest Radiographs |
643 Pulmonary Medicine | JAMA Network Open | JAMA Network. https://jamanetwork-
644 com.stanford.idm.oclc.org/journals/jamanetworkopen/fullarticle/2738349.
645 22. Rajpurkar, P. *et al.* CheXaid: deep learning assistance for physician diagnosis of
646 tuberculosis using chest x-rays in patients with HIV. *Npj Digit. Med.* **3**, 1–8 (2020).

23. Rajpurkar, P. *et al.* AppendiXNet: Deep Learning for Diagnosis of Appendicitis from A Small Dataset of CT Exams Using Video Pretraining. *Sci. Rep.* **10**, 3958 (2020).

24. Huang, S.-C. *et al.* PENet—a scalable deep-learning model for automated diagnosis of pulmonary embolism using volumetric CT imaging. *Npj Digit. Med.* **3**, 1–9 (2020).

25. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215 (2019).

26. Eitel, F. & Ritter, K. Testing the robustness of attribution methods for convolutional neural networks in MRI-based Alzheimer's disease classification. *ArXiv190908856 Cs Eess* (2019).

27. Young, K., Booth, G., Simpson, B., Dutton, R. & Shrapnel, S. Deep neural network or dermatologist? *ArXiv190806612 Cs Eess Stat* **11797**, 48–55 (2019).

28. Reyes, M., Meier, R., Pereira, S., Silva, C. A. & Dahlweid, F.-M. On the Interpretability of Artificial Intelligence in Radiology: Challenges and Opportunities. **2**, 12.

29. Selvaraju, R. R. *et al.* Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *Int. J. Comput. Vis.* **128**, 336–359 (2020).

30. Chattopadhay, A., Sarkar, A., Howlader, P. & Balasubramanian, V. N. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)* 839–847 (2018). doi:10.1109/WACV.2018.00097.

31. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic attribution for deep networks. in *Proceedings of the 34th International Conference on Machine Learning - Volume 70* 3319–3328 (JMLR.org, 2017).

32. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely Connected Convolutional Networks. in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2261–2269 (IEEE, 2017). doi:10.1109/CVPR.2017.243.

33. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778 (IEEE, 2016). doi:10.1109/CVPR.2016.90.

34. Szegedy, C. *et al.* Going deeper with convolutions. in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 1–9 (2015). doi:10.1109/CVPR.2015.7298594.

35. Irvin, J. *et al.* CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. *Proc. AAAI Conf. Artif. Intell.* **33**, 590–597 (2019).

36. Otsu, N. A Threshold Selection Method from Gray-Level Histograms. *IEEE Trans Syst. Man Cybern.* 62–66 (1979).

37. Zhang, J., Lin, Z., Brandt, J., Shen, X. & Sclaroff, S. Top-down Neural Attention by Excitation Backprop. *ArXiv160800507 Cs* (2016).

38. Kim, H.-E. *et al.* Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. *Lancet Digit. Health* **2**, e138–e148 (2020).

39. Efron, B. & Tibshirani, R. J. *An Introduction to the Bootstrap*. (CRC Press, 1994).

691 40. Vrabac, D. *et al.* DLBCL-Morph: Morphological features computed using deep
692 learning for an annotated digital DLBCL image set. *ArXiv200908123 Cs* (2020).
693 41. Johnson, A. E. W. *et al.* MIMIC-CXR, a de-identified publicly available database
694 of chest radiographs with free-text reports. *Sci. Data* **6**, 317 (2019).
695 42. Wang, X. *et al.* ChestX-ray8: Hospital-Scale Chest X-Ray Database and
696 Benchmarks on Weakly-Supervised Classification and Localization of Common
697 Thorax Diseases. in 2097–2106 (2017).
698 43. SIIM-ACR Pneumothorax Segmentation. https://kaggle.com/c/siim-acr-
699 pneumothorax-segmentation.
700 44. RSNA Pneumonia Detection Challenge. https://kaggle.com/c/rsna-pneumonia-
701 detection-challenge.
702 45. Steiner, D. F. *et al.* Impact of Deep Learning Assistance on the Histopathologic
703 Review of Lymph Nodes for Metastatic Breast Cancer. *Am. J. Surg. Pathol.* **42**,
704 1636–1646 (2018).
705 46. Uyumazturk, B. *et al.* Deep Learning for the Digital Pathologic Diagnosis of
706 Cholangiocarcinoma and Hepatocellular Carcinoma: Evaluating the Impact of a Web-
707 based Diagnostic Assistant. *ArXiv191107372 Eess* (2019).
708 47. Park, A. *et al.* Deep Learning–Assisted Diagnosis of Cerebral Aneurysms Using
709 the HeadXNet Model. *JAMA Netw. Open* **2**, e195600 (2019).
710 48. Annalise.ai - Medical imaging AI, by clinicians for clinicians. *Annalise.ai*
711 https://annalise.ai/.
712 49. Lunit Inc. https://www.lunit.io/en.
713 50. Qure.ai | Artificial Intelligence for Radiology. https://qure.ai/.

714 51. Gadgil, S., Endo, M., Wen, E., Ng, A. Y. & Rajpurkar, P. CheXseg: Combining
715 Expert Annotations with DNN-generated Saliency Maps for X-ray Segmentation.
716 *ArXiv210210484 Cs* (2021).
717 52. Crosby, J., Chen, S., Li, F., MacMahon, H. & Giger, M. Network output
718 visualization to uncover limitations of deep learning detection of pneumothorax. in
719 *Medical Imaging 2020: Image Perception, Observer Performance, and Technology*
720 *Assessment* vol. 11316 113160O (International Society for Optics and Photonics,
721 2020).
722 53. Melbye, H. & Dale, K. Interobserver Variability in the Radiographic Diagnosis of
723 Adult Outpatient Pneumonia. *Acta Radiol.* **33**, 79–81 (1992).
724 54. Herman, P. G. *et al.* Disagreements in Chest Roentgen Interpretation. *CHEST*
725 **68**, 278–282 (1975).
726 55. Albaum, M. N. *et al.* Interobserver Reliability of the Chest Radiograph in
727 Community-Acquired Pneumonia. *CHEST* **110**, 343–350 (1996).
728 56. Arun, N. T. *et al.* Assessing the validity of saliency maps for abnormality
729 localization in medical imaging. *ArXiv200600063 Cs* (2020).
730 57. Graziani, M., Lompech, T., Müller, H. & Andrearczyk, V. *Evaluation and*
731 *Comparison of CNN Visual Explanations for Histopathology*. (2020).
732 58. Arun, N. *et al.* Assessing the (Un)Trustworthiness of Saliency Maps for Localizing
733 Abnormalities in Medical Imaging. *ArXiv200802766 Cs* (2020).
734 59. MD.ai. https://www.md.ai/.
735 60. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization.
736 *ArXiv14126980 Cs* (2017).

737