**Title**

# Deep learning saliency maps do not accurately highlight diagnostically relevant regions for medical image interpretation

**Authors**

Adriel Saporta BA MBA[1*], Xiaotong Gui BA[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 MS[1†]

**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

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

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

Corresponding author: Pranav Rajpurkar (email: pranavsr@cs.stanford.edu)

Current word count: 3941

## Abstract

Deep learning has enabled automated medical image interpretation at a level often surpassing that of practicing medical experts. However, many clinical practices have cited a lack of model interpretability as reason to delay the use of “black-box” deep neural networks in clinical workflows. Saliency maps, which “explain” a model’s decision 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. In this work, we demonstrate that the most commonly used saliency map generating method, Grad-CAM, results in low performance for 10 pathologies on chest X-rays. We examined under what clinical conditions saliency maps might be more dangerous to use compared to human experts, and found that Grad-CAM performs worse for pathologies that had multiple instances, were smaller in size, and had shapes that were more complex. Moreover, we showed that model confidence was positively correlated with Grad-CAM localization performance, suggesting that saliency maps were safer for clinicians to use as a decision aid when the model had made a positive prediction with high confidence. Our work demonstrates that several important limitations of interpretability techniques for medical imaging must be addressed before use in clinical workflows.

## Introduction

Deep learning has enabled automated medical imaging interpretation at a level shown to surpass that 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

46 “black-box” deep neural networks (DNNs) represents a major barrier to clinical trust and
47 adoption[4,5,6]. In fact, it has been argued that the European Union’s recently adopted
48 General Data Protection Regulation (GDPR) affirms an individual’s right to an explanation
49 in the context of automated decision-making[7]. Although many DNN interpretability
50 techniques have been proposed, rigorous investigation of the accuracy and reliability of
51 these strategies is lacking and necessary before they are integrated into the clinical
52 setting[8].
53
54 One type of DNN interpretation strategy widely used in the context of medical imaging is
55 based on saliency (or pixel-attribution) methods[9–12]. Saliency methods produce heat
56 maps that highlight the areas of the medical image that most influenced the DNN’s
57 prediction. The heat maps help to visualize whether a DNN is concentrating on the same
58 regions of the medical image that a human expert would focus attention on for a given
59 diagnosis, rather than concentrating on a clinically irrelevant part of the medical image or
60 even on confounders in the image[13–15]. However, recent work has shown that saliency
61 methods used to validate model predictions can be misleading in some cases and may
62 lead to increased bias and loss of user trust with concerning implications for clinical
63 translation efforts[16].
64
65 The purpose of this work is to perform a systematic evaluation of the most common
66 saliency method, Grad-CAM[17], on multi-label classification models for medical imaging
67 interpretation from chest X-rays. In order to evaluate how well the saliency method
68 identifies critical areas of an image for diagnosis, we compared the saliency method

segmentations to human expert benchmark and reference annotations. We evaluated the accuracy of the saliency method segmentations and the human expert benchmark segmentations first by calculating their overlap with the human expert reference segmentations, and then by determining whether the segmentations correctly located the pathology of concern, regardless of the exact bounds of the segmentations. We further conducted statistical analyses to better understand how the localization accuracy of saliency methods is affected both by pathological characteristics and also by model confidence.

## Results

### Framework for evaluating a saliency method on multi-label classification models

A model can be trained to perform pixel-level localization in one of two ways: either through supervised learning, in which the model is trained directly on pixel-level segmentations, or through weakly supervised learning, in which the model is trained only on image-level class labels. Because ground truth segmentations for medical imaging can be especially time-consuming and expensive to obtain given the domain expertise required to create them[18], one of the most common saliency methods in medical imaging is a weakly supervised localization technique, Grad-CAM, in which the classification model is never exposed to pixel-level segmentations during training. Instead, Grad-CAM generates a heat map corresponding to each image-level task label that highlights the regions of the input image that, in theory, are most indicative of that task label. This saliency method has 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 in predicting (1) myocardial infarction[19] and hypogycemia[20] from electrocardiograms, (2) visual impairment[21], refractive error[22], and anaemia[23] from retinal photographs (3) long-term mortality[24] and tuberculosis[25] from chest x-ray images, and (4) appendicitis[26], and pulmonary embolism[27] on computed tomography scans.

We use the following framework, which we call CheXplanation, to evaluate Grad-CAM in a multi-label classification setup. We trained and evaluated an ensemble of 30 CNN models on CheXpert, a large publicly available CXR dataset[28]. We then passed each of the 668 CXRs in the dataset's holdout test set into the 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 Grad-CAM to generate 10 heat maps, one for each of the 10 pathologies. We then applied a threshold to the heat maps to produce 10 binary segmentations in order to evaluate their overlap with the human expert reference segmentations (see Fig. 1a). We also extracted the location of the pixel with the largest value from each heat map to determine whether it fell within the bounds of the reference segmentation (see Fig. 1d). In doing so, we could evaluate Grad-CAM's localization performance regardless of the exact bounds of its binary segmentations.

To evaluate how well the saliency method segmentations identified clinically relevant pathology regions of input CXRs ("AI localization performance"), we obtained pixel-level reference segmentations on the holdout test set from two board-certified radiologists who

were asked to segment any of the 10 pathologies that were present in each CXR as determined by the dataset's ground-truth labels (see Fig. 1b). We also established a human benchmark ("expert localization performance") by collecting segmentations from three additional radiologists who were asked to segment the 10 pathologies of interest present in each CXR as determined by the dataset's ground-truth labels (there was no overlap between these three radiologists and the two who drew the reference segmentations) (see Fig. 1c). This second group of radiologists was also asked to locate each pathology present on each CXR using only a single most representative point for that pathology on the CXR to see whether that point fell within the bounds of the reference segmentation (see Fig. 1e). Which point in the CXR is "most representative" varied by pathology, but the point would always lie inside the pathology's segmentation drawn by the radiologist. For example, the most representative point for Pneumothorax would be wherever the pathology is most evident, whereas the most representative point for Cardiomegaly would be the center of the heart. See Supplementary Figs. 1 through 11 for the detailed instructions given to the radiologists.

Our dataset of expert reference segmentations has been made publicly available to encourage further development and evaluation of CXR interpretation models.

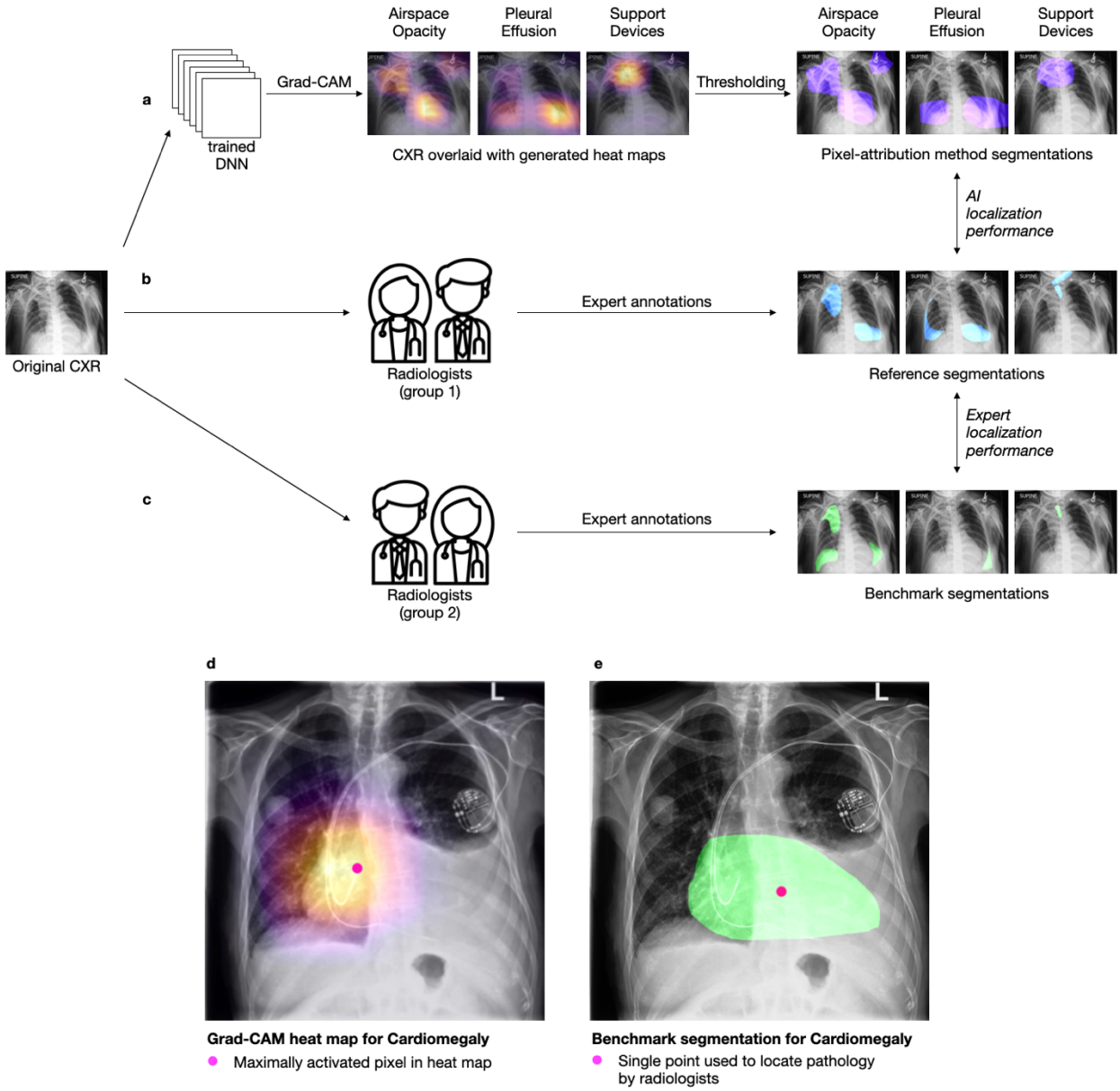



**Fig. 1 | Framework for evaluating a saliency method on multi-label classification models. a**, Left, a CXR image from the holdout test set is passed into an ensemble DNN trained only on CXR images and their corresponding pathology task labels. Grad-CAM is used to generate 10 heat maps for the example CXR, one for each task. Middle, there are three pathologies present in this CXR (Airspace Opacity, Pleural Effusion, and Support Devices). Right, a threshold is applied to the heat maps to produce binary segmentations for each present pathology. **b**, Two board-certified radiologists were asked to segment the pathologies present in the CXR as determined by the dataset's ground-

truth labels. Saliency method segmentations are compared to these reference segmentations to evaluate how well Grad-CAM identifies the clinically relevant areas of the input CXR ("AI localization performance"). **c**, Three radiologists (separate from those in **b**) were asked to segment the pathologies present in the CXR as determined by the dataset's ground-truth labels. These benchmark segmentations are compared to the reference segmentations to determine a human benchmark ("expert localization performance"). **d**, The location of the pixel with the largest value was extracted from each heat map. **e**, In addition to drawing segmentations, the benchmark radiologists were asked to locate each pathology present on each CXR using only a single point on that CXR.

**Evaluating the localization performance of the saliency method**

We used two evaluation schemes to compare the AI localization performance to the expert localization performance. First, we used mean Intersection over Union (mIoU) to measure how much, on average, either the saliency method segmentations or the human benchmark segmentations overlapped with the reference segmentations. Second, we used the pointing game setup[29], in which a "hit" is when the single point used to locate a pathology lies within the reference segmentation and a "miss" is when the single point lies outside the reference segmentation. Localization performance is then calculated as the hit rate across the dataset[30]. See Fig. 2a for mIoU and hit rate for Grad-CAM segmentations on example CXRs.

For nine of the 10 pathologies, the saliency segmentations had a lower overlap with the reference segmentations than did the human benchmark segmentations (see Fig. 2b). Similarly, for nine of the 10 pathologies, the saliency method had a lower hit rate than the human benchmark (see Fig. 2c). Under both the overlap and the pointing game evaluation

schemes, the gap between AI localization performance and expert localization performance was the largest for Lung Lesion: AI mIoU was 93.6% smaller than expert mIoU, and Grad-CAM hit rate was 74.9% smaller than expert hit rate. Lung Lesion is also the pathology for which the saliency method had both the lowest mIoU (0.027) and the lowest hit rate (0.215). Under the hit rate scheme, Support Devices and Pneumothorax displayed the second and third largest gaps, respectively, between AI and expert localization performance (Support Devices: 72.4%; Pneumothorax: 60.4%). Under the overlap scheme, Support Devices and Pneumothorax also displayed the fourth and third largest gaps, respectively, between AI and expert localization performance (Support Devices: 67.4%; Pneumothorax: 68.5%).

Under the overlap evaluation scheme, the only pathology for which the saliency method (mIoU 0.145 95% CI [0.126, 0.163]) outperformed the human benchmark (mIoU 0.122, 95% CI [0.108, 0.136]) was Atelectasis. Under the hit rate evaluation scheme, the only pathology for which the saliency method (hit rate 0.624 95% CI [0.462, 0.771]) outperformed the human benchmark (hit rate 0.513 95% CI [0.355, 0.674]) was for Consolidation. However, the differences were statistically insignificant. Both the saliency method and the human benchmark achieved their highest mIoUs for Cardiomegaly (AI: 0.275 95% CI [0.242, 0.30]); expert: 0.712 95% CI [0.692, 0.731]). AI hit rate was largest for Cardiomegaly (0.903 95% CI [0.857, 0.945]) and Enlarged Cardiomediastinum (0.761 95% CI [0.712, 0.81]). Expert hit rate was above 0.95 for Pneumothorax (1.0 95% CI [1.0, 1.0]), Cardiomegaly (0.971 95% CI [0.944, 0.994]), and Enlarged Cardiomediastinum (0.953 95% CI [0.927, 0.975]).

We also evaluated whether and how the gap between AI and expert localization performance changed from when overlap was used as an evaluation metric to when hit rate was used as an evaluation metric. Percentage decrease from expert to AI localization performance fell the most from mIoU to hit rate for Consolidation (41.2% - (-21.6%) = 62.8%), Cardiomegaly (61.4% - 7.1% = 54.3%), and Enlarged Cardiomediastinum (70.0% - 20.1% = 49.9%). Percentage decrease from expert to AI localization performance increased by far the most from mIoU to hit rate for Atelectasis (-18.9% - 59.0% = -77.9%).

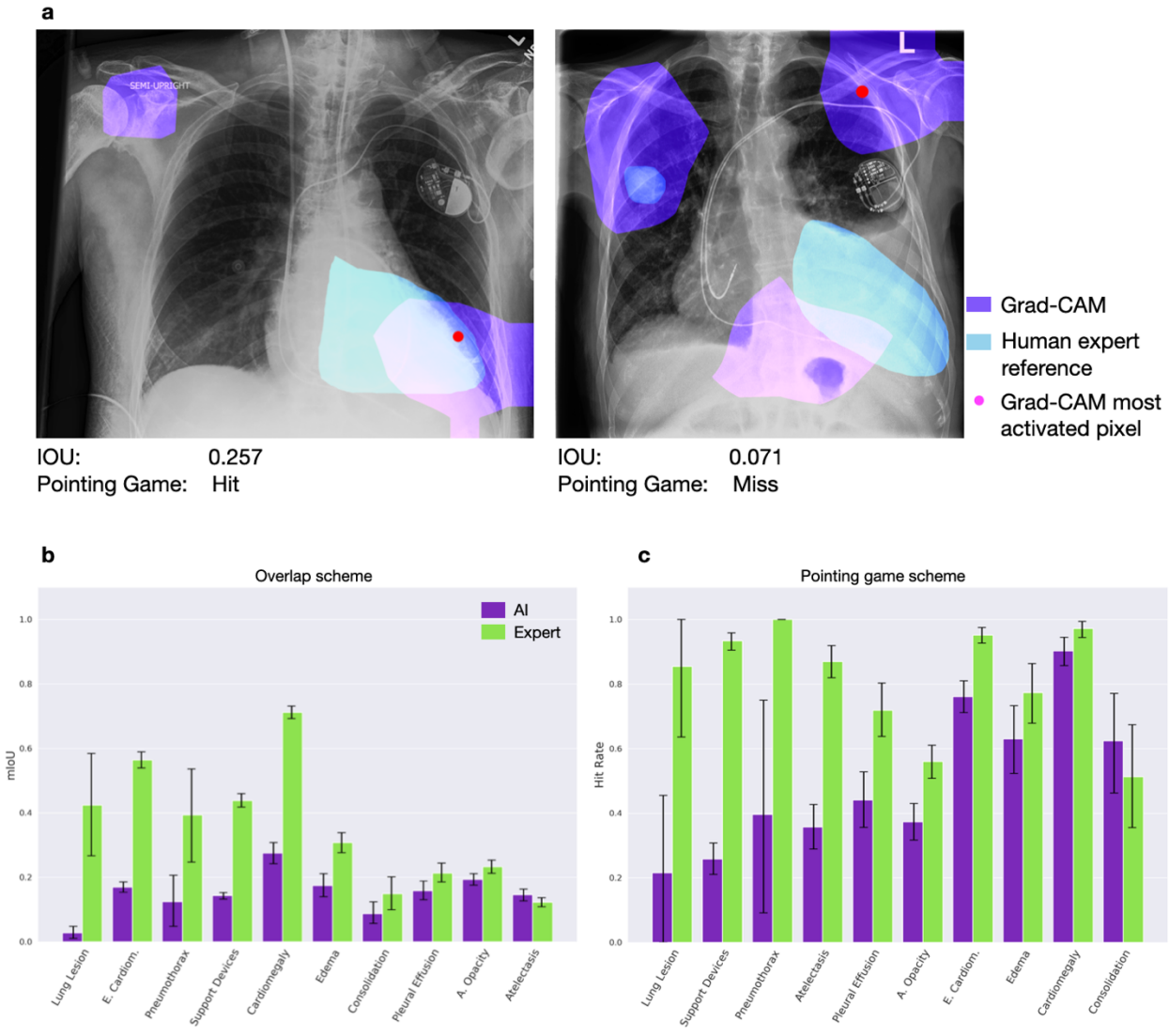


**Fig. 2 | Evaluating the localization performance of the attribution method. a**, Grad-CAM and human expert reference segmentations for two CXRs with Airspace Opacity. Left, IoU score is 0.257, and pointing game is a "hit" since Grad-CAM's most activated pixel is inside of the reference segmentation. Right, IoU score is 0.071, and pointing game is a "miss" since Grad-CAM's most activated pixel is outside of the reference segmentation. **b**, Comparing AI and expert localization performances under the overlap evaluation scheme. Pathologies are sorted on the x-axis in descending order of percentage decrease from expert mIoU to AI mIoU. **c**, Comparing AI and expert localization performances under the hit rate evaluation scheme. Pathologies are sorted on the x-axis in descending order of percentage decrease from expert hit rate to AI hit rate for each pathology. The black error bars indicate 95% bootstrap confidence interval.

**Table 1 | Percentage decrease from expert localization to AI localization for each pathology**

| Pathology | Expert mIoU | AI mIoU | % decrease mIoU | Expert hit rate | AI hit rate | % decrease hit rate |
|---|---|---|---|---|---|---|
| Lung Lesion | 0.424 | 0.027 | 93.6 | 0.855 | 0.215 | 74.9 |
| E. Cardiom. | 0.564 | 0.169 | 70.0 | 0.953 | 0.761 | 20.1 |
| Pneumothorax | 0.393 | 0.124 | 68.4 | 1.000 | 0.396 | 60.4 |
| Support Devices | 0.438 | 0.143 | 67.4 | 0.934 | 0.258 | 72.4 |
| Cardiomegaly | 0.712 | 0.275 | 61.4 | 0.971 | 0.903 | 7.1 |
| Edema | 0.307 | 0.174 | 43.3 | 0.773 | 0.630 | 18.5 |
| Consolidation | 0.148 | 0.087 | 41.2 | 0.513 | 0.624 | -21.6 |
| Pleural Effusion | 0.213 | 0.158 | 25.8 | 0.719 | 0.441 | 38.7 |
| Airspace Opacity | 0.232 | 0.193 | 16.8 | 0.560 | 0.317 | 33.4 |
| Atelectasis | 0.122 | 0.145 | -18.9 | 0.869 | 0.289 | 59.0 |

## Characterizing the gaps between AI localization performance and expert localization performance

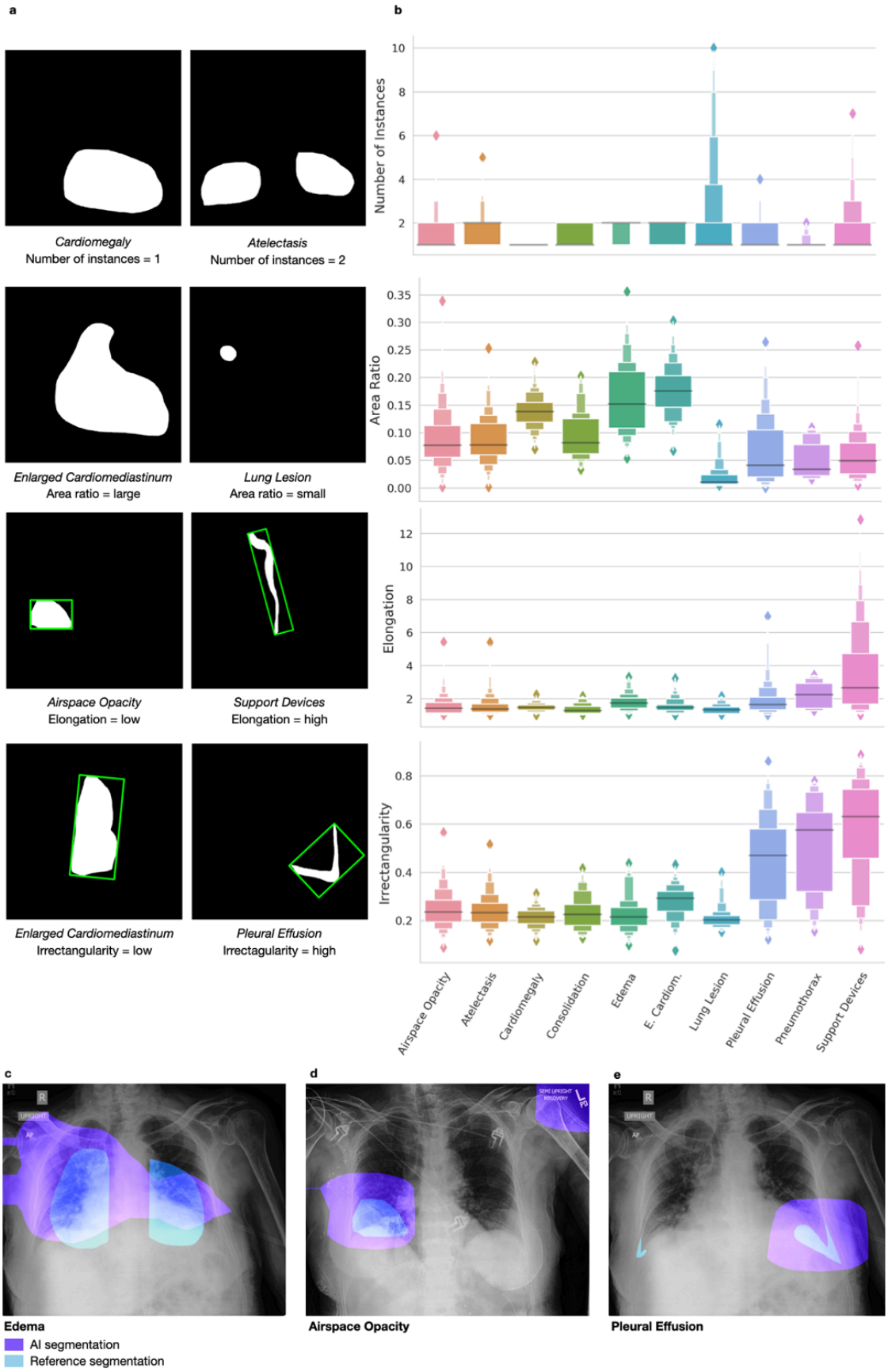
In order to better understand under what circumstances the AI localization performance was closer to, or further from, the expert localization performance, we first conducted a

qualitative analysis by visually inspecting both the saliency method segmentations and the benchmark radiologist segmentations with a radiologist. Then, to inform our qualitative interpretation, we conducted a statistical analysis to quantify how both the saliency method segmentations and the benchmark radiologist segmentations performed in the presence of four pathological characteristics[31]: (1) number of instances for a given pathology (for example, bilateral Pleural Effusion would have two instances, whereas there is usually only one instance for Cardiomegaly), (2) area ratio (pathology area with respect to the area of the whole CXR), (3) elongation, and (4) irrectangularity (the last two features were meant to measure the complexity of the pathology's shape). See Fig. 3a for example segmentations with the above 4 characteristics. See Fig. 3b for the distribution of the four pathological characteristics across all 10 pathologies. Lung Lesion had the largest number of instances on average, and the lowest mean area ratio among all the pathologies. Support Devices and Pneumothorax had the two highest mean elongation and irrectangularity values, respectively.

For each evaluation scheme (overlap and hit rate), we ran 12 simple linear regressions: four with the AI evaluation metric (IoU or hit rate) as the response variable, four with the expert evaluation metric as the response variable, and four with the difference between the expert and AI evaluation metrics as the response variable. Each group of four regressions used the above four pathological characteristics as the regression's single attribute, respectively, and only CXRs with a positive label were included in each regression (n=1534). Each regression coefficient can be interpreted as the effect of that

pathological characteristic on the evaluation metric at hand. See Table 2 for coefficients from the regressions under the overlap and hit rate evaluation schemes.

Our qualitative analysis uncovered three patterns in the saliency method segmentations that were associated with lower localization performance. First, we observed that when multiple instances of a single pathology are present in a CXR, instead of highlighting each distinct instance of the pathology separately, the saliency method segmentation often highlights one large confluent area that encompasses all of the instances (see Fig. 3c). Second, we found that saliency method segmentations tend to be significantly larger than either the human benchmark or reference segmentations, and often fail to respect clear anatomical boundaries (see Fig. 3d). Correspondingly, the AI overlap coefficient for area ratio was 0.556 (95% CI [0.510, 0.601]), suggesting that as a pathology's area ratio decreases, AI localization performance worsens. Furthermore, under the overlap evaluation scheme, the gap between AI localization performance and expert localization performance increases as a pathology's area ratio decreases: the area ratio coefficient was -0.151 (95% CI [–0.233, -0.07]) when the difference between expert and AI overlap was the response variable. Third, our qualitative analysis showed that, in segmenting complex and elongated pathologies, when the AI segmentations include the pathology, they also frequently enclose significant portions of the CXR where the pathology is not present (see Fig. 3e). Similarly, our statistical analysis demonstrated that AI localization performance worsens both as a pathology's elongation increases (overlap coefficient = -0.375 95% CI [-0.453, -0.297]), and as a pathology's irrectangularity increases (overlap coefficient = -0.205 95% CI [–0.245, -0.165]).


a
b
Cardiomegaly
Number of instances = 1
Atelectasis
Number of instances = 2
Enlarged Cardiomediastinum
Area ratio = large
Lung Lesion
Area ratio = small
Airspace Opacity
Elongation = low
Support Devices
Elongation = high
Enlarged Cardiomediastinum
Irrectangularity = low
Pleural Effusion
Irrectagularity = high
Number of Instances
Area Ratio
Elongation
Irrectangularity
Airspace Opacity
Atelectasis
Cardiomegaly
Consolidation
Edema
E. Cardiom.
Lung Lesion
Pleural Effusion
Pneumothorax
Support Devices
c
d
e
Edema
Airspace Opacity
Pleural Effusion
AI segmentation
Reference segmentation

263

**Fig. 3 | Characterizing the gaps between AI localization performance and expert localization performance.** **a**, Example segmentations with four pathological characteristics: (1) number of instances (top row), (2) area ratio (second row), (3) elongation (third row), and (4) irrectangularity (fourth row). Elongation and irrectangularity were calculated by fitting a rectangle of minimum area enclosing the binary mask. Elongation = maxLength/minLength. Irrectangularity = 1 - (area of segmentation/area of enclosing rectangle). **b**, Distribution of the four pathological characteristics across all 10 pathologies in letter value plot style. The black horizontal line in each box indicates the median for that pathology, and from the middle each box represents the increasing (middle to up) and decreasing (middle to bottom) quantile. **c**, Multiple instances of Edema in this CXR are shown by the reference segmentations in blue. Instead of highlighting each distinct instance of the pathology separately, the AI segmentation in purple highlights one large confluent area that tries to encompass both instances. **d**, Airspace Opacity in this CXR is shown by the reference segmentation in blue. Not only is the AI segmentation larger than the Airspace Opacity, but it also fails to respect clear anatomical boundaries by highlighting an area outside of the chest cavity. **e**, Bilateral Pleural Effusion in this CXR is shown by the reference segmentations in blue. On the right, instead of highlighting the distinct V-shape of the pathology, the AI segmentation highlights a large portion of the CXR where the pathology is not present in trying to enclose the whole pathology.

Our statistical analysis showed that the first and second trends listed above held not only for the AI segmentations, but also for the expert segmentations: as the number of instances of a pathology increases, expert localization performance worsens (overlap coefficient = -0.178 95% CI [–0.334, -0.021]), and as the area ratio of a pathology increases, expert localization performance improves (overlap coefficient = 0.404 95% CI [0.334, 0.475]). However, our statistical analysis showed no evidence that the third trend holds for the expert segmentations: expert localization performance does not worsen as

pathology complexity increases (irrectangularity overlap coefficient = 0.073 [0.016, 0.13]; elongation overlap coefficient was not statistically significant).

The results of the above experiments using hit/miss as an evaluation metric were consistent with the results when using overlap as an evaluation metric. In both experiments, we reported the 95% confidence interval and the Bonferroni corrected p-values.

**Table 2 | Coefficients from regressions on pathological characteristics**

| Pathological characteristic (independent variable) | Coefficient using AI localization as response | Coefficient using expert localization as response | Coefficient using localization difference (expert - AI) |
|---|---|---|---|
| IoU | | | |
| Number of instances | 0.003 (-0.111, 0.116) | -0.178 (-0.334, -0.021) * | -0.18 (-0.355, -0.006) * |
| Area ratio | 0.556 (0.510, 0.601) *** | 0.404 (0.334, 0.475) *** | -0.151 (-0.233, -0.07) *** |
| Elongation | -0.375 (-0.453, -0.297) *** | 0.067 (-0.043, 0.178) | 0.442 (0.321, 0.563) *** |
| Irrectangularity | -0.205 (-0.245, -0.165) *** | 0.073 (0.016, 0.13) ** | 0.278 (0.216, 0.341) *** |
| Hit/Miss | | | |
| Number of instances | -0.101 (-1.307, -0.714) *** | 0.403 (0.175, 0.632) ** | 1.413 (1.059, 1.768) *** |
| Area ratio | 1.111 (0.982, 1.24) *** | 0.313 (0.207, 0.42) *** | -0.78 (-0.962, -0.633) *** |
| Elongation | -0.842 (-1.05, -0.634) *** | 0.267 (0.105, 0.428) ** | 1.108 (0.86, 1.357) *** |
| Irrectangularity | -0.536 (-0.642, -0.429) *** | 0.169 (0.0852, 0.252) *** | 0.704 (0.577, 0.832) *** |

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

**Effect of model confidence on AI localization performance**

Since the saliency method is highly dependent on the DNN's architecture, we conducted statistical analyses to determine whether there was any correlation between the model's confidence in its prediction and AI localization performance. We first ran a simple

305 regression for each pathology using the model's probability output for the pathology as
306 the single independent variable and using the IoU of the AI segmentation with reference
307 segmentation as the response variable. We then performed a simple regression that uses
308 the same approach as above, but that includes all 10 pathologies. For each of the 11
309 regressions, we excluded true negative cases in order to calculate the IoU score for the
310 expert segmentations. In addition to the linear regression coefficients, we also computed
311 the Spearman correlation coefficients to capture any potential non-linear associations
312 (see Table 3).
313
314 We found that for all pathologies, model confidence was positively correlated with AI
315 localization performance. The p-values for all of the coefficients were below 0.001 except
316 for the coefficients for Pneumothorax (n=11) and Lung Lesion (n=50), the two pathologies
317 for which we had the fewest positive examples. Of all the pathologies, model confidence
318 for positive predictions of Enlarged Cardiomediastinum had the largest linear regression
319 coefficient with AI localization performance (1.974, p-value = 2.523e-24). Model
320 confidence for positive predictions of Pneumothorax had the largest Spearman correlation
321 coefficient with AI localization performance (0.734, p-value = 0.01), although the
322 coefficient was not as statistically significant as the Spearman correlation coefficient for
323 Pleural Effusion (0.69, p-value=8.08e-24), the second largest of all the pathologies.
324 Combining all of the pathologies (n=2365), the linear regression coefficient was 0.109
325 (95% CI [0.083, 0.135]), and the Spearman correlation coefficient was 0.285 (95%CI
326 [0.239, 0.331]). We also performed analogous experiments using hit rate as the response
327 variable and found comparable results (see Supplementary Table 1).

**Table 3 | Overlap: 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) *** | 0.69 (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 saliency methods, which are widely used in clinical practice for DNN prediction explainability. We demonstrate that saliency maps are consistently worse than expert radiologists at localizing a variety of pathologies on CXRs. We use qualitative and quantitative analyses to establish that AI localization performance is furthest from expert localization performance in the face of pathologies that have multiple instances, are smaller in size, and have 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 show that model assurance is positively correlated with AI localization performance,

which could indicate that the saliency methods are safer to use as a decision aid to clinicians when the model has made a positive prediction with high confidence. Finally, since IoU computes the overlap of two segmentations but pointing game hit rate better captures diagnostic attention, we suggest using both metrics to evaluate both AI and expert localization performance.

Our work has several potential implications for patient care. Heat maps generated using saliency methods are advocated as clinical decision support in the hope that the heat maps not only improve clinical decision-making, but also encourage clinicians to trust model predictions[32–34]. However, we found that AI localization performance, on balance, performed worse than expert localization across multiple analyses. This is consistent with recent work focused on localizing a single pathology, Pneumothorax, in CXRs[35]. Our work expanded this exploration and found that common saliency methods may underperform for many important pathologies on CXRs. 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[25].

Determining when saliency methods are more likely to succeed or fail in localizing pathologies could have further implications for patient care. That knowledge could inform not only under what clinical conditions saliency methods might be safer to use, but also how we might improve saliency methods in the future. We found that AI localization performance worsens in the presence of pathologies that have multiple instances. We

also found that AI localization performance worsens in the presence of pathologies that are smaller in size compared with the CXR image. This result explains why, under both the overlap and hit rate schemes, the gap between AI and expert localization performance was the largest for Lung Lesion, whose mean area ratio was the smallest of the 10 pathologies we explored. Moreover, AI localization performance under both evaluation schemes was best for Enlarged Cardiomediastinum and Cardiomegaly, which had the largest and third largest area ratios, respectively, of the 10 pathologies, suggesting that saliency methods might be safer to use in the context of these two pathologies, or pathologies with similar characteristics. Grad-CAM segmentations often fail to respect clear anatomical boundaries, and we hypothesize that this is an algorithmic artifact of Grad-CAM, whose feature map sized (14 x 14) heatmap is interpolated to the original image dimension (usually 2000 x 2000), resulting in coarse resolution. We also found that AI localization performance worsens in the presence of pathologies whose shapes are more complex. AI localization for Pneumothorax and Support Devices, both of which were more elongated and complex than any of the other conditions, underperformed compared to expert localization performance; however, this performance gap must also be considered in the context of the model training data prevalence, and future work may explore the impact of training data prevalence on the localization performance of saliency methods.

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 expert segmentations had relatively low overlap with the reference

386 segmentations (the highest expert mIoU was 0.712 for Cardiomegaly). One potential
387 explanation for this consistent underperformance is that pathologies can be hard to
388 distinguish, especially without clinical context. Furthermore, whereas many people might
389 agree on how to segment, say, a cat or a stop sign in traditional computer vision tasks,
390 radiologists use a certain amount of clinical discretion when defining the boundaries of a
391 pathology on a CXR. There can also be institutional and geographic differences in how
392 radiologists are taught to recognize pathologies, and studies have shown that there can
393 be high interobserver variability in the interpretation of CXRs[36–38]. We sought to address
394 this with the hit rate evaluation metric, which highlights when two radiologists share the
395 same diagnostic intention, even if it is less exact than IoU in comparing segmentations
396 directly. Expert performance using hit rate was above 0.95 for four pathologies
397 (Pneumothorax, Cardiomegaly, Support Devices, and Enlarged Cardiomediastinum);
398 these are pathologies for which there is often little disagreement between radiologists
399 about where the pathologies are located, even if the expert segmentations are noisy. The
400 only pathology for which AI localization performance was better than expert localization
401 performance under the hit rate scheme was Consolidation. However, because the hit rate
402 scheme required the benchmark radiologists to select only one point on the CXR, even if
403 there were multiple instances of the pathology present (as is often the case with
404 Consolidation), it is likely that the hit rate setup unfairly penalized expert performance in
405 this case and that it is not the best evaluation metric to use for Consolidation. Further
406 work is needed to validate this hypothesis and to demonstrate which segmentation
407 evaluation metrics, even beyond overlap and hit rate, are more appropriate for which
408 pathologies when evaluating saliency methods for the clinical setting.

Our work builds upon several studies investigating the validity of saliency maps in localization[39,40] and upon some early work on trustworthiness of saliency methods to explain DNNs in medical imaging[41]. We substantially extend the body of literature by doing a comprehensive analysis on a multi-label classification task using the most popular saliency method in the medical context, Grad-CAM. Our work analyzes 10 different commonly occurring clinical pathologies as opposed to only Pneumothorax and Pneumonia. In addition to demonstrating that Grad-CAM is not yet ready for clinical use, we also highlight the strengths and weaknesses of Grad-CAM by doing quantitative statistical analysis, thus opening future avenues of research to improve Grad-CAM in particular and saliency methods in general. Finally, we establish new ground truth and benchmark segmentations on 10 different CXR observations facilitating future research on attribution methods.

There are several limitations of our work. We chose to evaluate Grad-CAM, since it has become one of the most popular explainability methods for CXRs, but future work may also evaluate other saliency methods, including Integrated Gradients[42], WILDCAT[43], and SmoothGrad[44]. While we used DenseNet121 as the underlying model architecture for the saliency method, since it was shown to produce the best classification results on the CheXpert dataset, future work should explore and compare different model architectures for CXR explainability. Our dataset had only 11 CXRs with Pneumothorax and 50 CXRs with Lung Lesion. Future work should investigate the impact of pathology prevalence in the training data on AI localization performance. Some pathologies, such as effusions

and cardiomegaly, are always in the same place in a frontal view CXR; others, such as lesions and opacities, can occur in different locations on a CXR. Future work could investigate how a pathology's location on a CXR, and the consistency of that location, affect AI localization performance. Finally, we compared Grad-CAM-generated pixel-level segmentations to human expert pixel-level segmentations, but future work might explore how AI localization performance changes when comparing bounding-box annotations, instead of pixel-level segmentations.

In conclusion, we demonstrate that not only should we not yet rely on saliency methods for deep learning explainability in CXRs, but saliency methods are particularly brittle in the presence of pathologies that have multiple instances, are smaller in size, and have shapes that are more complex. Although our findings suggest that care should be taken when deploying saliency methods into clinical practice, we demonstrate scenarios in which heat maps generated by saliency methods are most consistent with human expert annotations. Our work serves as the foundation for future work that rigorously evaluates a range of saliency methods using a variety of evaluation metrics before deep learning explainability techniques are integrated into the medical workflow.

## Methods

### Ethical and information governance approvals.

This study does not involve human subject participants.

454 **Dataset and clinical taxonomy.** *Dataset description.* The localization experiments were
455 performed using CheXpert, a large public dataset for chest X-ray interpretation. The
456 CheXpert dataset contains 224,316 chest X-rays for 65,240 patients labeled for the
457 presence of 14 observations (13 pathologies and an observation of "No Finding") as
458 positive, negative, or uncertain. The CheXpert validation set consists of 234 chest X-rays
459 from 200 patients randomly sampled from the full dataset and was labeled according to
460 the consensus of three board-certified radiologists. The test set consists of 668 chest X-
461 rays from 500 patients not included in the training or validation sets and was labeled
462 according to the consensus of five board-certified radiologists. See Supplementary Table
463 2 for dataset summary statistics.
464
465 *Reference segmentation.* The chest X-rays in our validation set and test set were
466 manually segmented by two board-certified radiologists with 18 and 27 years of
467 experience, using the annotation software tool MD.ai[45] (see Supplementary Figs. 12
468 through 14). The radiologists were asked to contour the region of interest for all
469 observations in the chest X-rays for which there was a positive ground truth label in the
470 CheXpert dataset. For a pathology with multiple instances, all the instances were
471 contoured. For Support Devices, radiologists were asked to contour any implanted or
472 invasive devices including pacemakers, PICC/central catheters, chest tubes,
473 endotracheal tubes, feeding tubes and stents and ignore ECG lead wires or external
474 stickers visible in the chest X-ray. Finally, of the 14 observations labeled in the CheXpert
475 dataset, Fracture, Pleural Other, and Pneumonia were not segmented because they
476 either had low prevalence and/or ill-defined boundaries unfit for segmentation.

*Evaluating the expert performance using benchmark segmentation.* To evaluate the expert performance on the test set using the IoU evaluation method, three radiologists, certified in Vietnam with 9, 10, and 18 years of experience, were asked to segment the regions of interest for all observations in the chest X-rays for which there was a positive ground truth label in the CheXpert dataset. These radiologists were also provided the same instructions for contouring as were provided to the radiologists drawing the reference segmentations. To extract the "maximally activated" point from the benchmark segmentations, we asked the same radiologists to locate each pathology present on each CXR using only a single most representative point for that pathology on the CXR (see Supplementary Figs. 1 through 11 for the detailed instructions given to the radiologists). There was no overlap between these three radiologists and the two who drew the reference segmentations.

**Classification network architecture and training protocol.** *Multi-label classification model.* The model takes as input a single-view chest X-ray and outputs the probability for each of the 14 observations. In case of availability of more than one view, the models output the maximum probability of the observations across the views. Each chest X-ray was resized to 320×320 pixels and normalized before it was fed into the network. The DenseNet121 model architecture[46] was used. Cross-entropy loss was used to train the model. The Adam optimizer[47] was used with default β-parameters of $\beta 1 = 0.9$ and $\beta 2 = 0.999$, and the learning rate was fixed at $1 \times 10{-}4$ for the duration of the training. Batches were sampled using a fixed batch size of 16 images.

*Ensembling.* An ensemble of 30 DenseNet121 checkpoints was created to improve the performance of the model. The 30 checkpoints were generated by training the model for 3 epochs and selecting the 10 checkpoints from each epoch 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. See Supplementary Table 3 for the performance of the model on each of the pathologies.

**DNN interpretation strategy.** *Saliency method*. Grad-CAM was used to visualize the decision made by the classification network. Grad-CAM uses the gradients of the target flowing into the final convolutional layer to produce a saliency map that highlights the regions on which the model focuses while making the decision. The saliency map outputted by Grad-CAM was resized to the original image dimension. It was then normalized using max-min normalization and then converted into a binary segmentation using binary thresholding (Otsu's method[48]). To further ensure that the final binary segmentation is consistent with model probability output, another layer of thresholding was applied 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 (Intersection over Union) IoU is

the ratio between the area of overlap and the area of union between the ground truth and the predicted areas, ranging from 0–1 with 0 signifying no overlap and 1 signifying perfectly overlapping segmentation. We then compared the mean Intersection over Union (mIoU) of Grad-CAM and radiologist benchmark on each pathology. The mIoU is the average IoU of all the images in the test dataset. True negatives where both segmentations were labeled as all 0s are excluded in the mean calculation. 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.

**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. Area ratio area is the area of the pathology divided by the total image area. 3.4. 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 - the 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 three groups of simple linear regressions, with expert and AI evaluation metrics and their differences as the response variables. Each group has four regressions using the above four pathological characteristics as the regression's single attribute, respectively, and only CXRs with a positive label were included in each regression (n=1534). All features are normalized using min-max normalization so that they are comparable on scales of magnitudes. We report the 95% confidence interval and p-value of the regression coefficients.

**Data Availability**

CheXpert data is available at https://stanfordmlgroup.github.io/competitions/chexpert/. The validation set and corresponding benchmark radiologist annotations will be available online for the purpose of extending the study.

**Code Availability**

All code used to produce the results of the paper will be in a public repository for the purpose of reproducing the study. The link to the code will be added to the text of the paper for the camera-ready version.

## Competing Interests

There are no competing interests.

## Author Contributions

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

## References


1. Rajpurkar, P. *et al.* CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. *ArXiv171105225 Cs Stat* (2017).
2. Baselli, G., Codari, M. & Sardanelli, F. Opening the black box of machine learning in radiology: can the proximity of annotated cases be a way? *Eur. Radiol. Exp.* **4**, 30 (2020).
3. Litjens, G. *et al.* A survey on deep learning in medical image analysis. *Med. Image Anal.* **42**, 60–88 (2017).
4. Wang, F., Kaushal, R. & Khullar, D. Should Health Care Demand Interpretable Artificial Intelligence or Accept “Black Box” Medicine? *Ann. Intern. Med.* **172**, 59–60 (2019).

5. 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.

6. Pasa, F., Golkov, V., Pfeiffer, F., Cremers, D. & Pfeiffer, D. Efficient Deep Network Architectures for Fast Chest X-Ray Tuberculosis Screening and Visualization. *Sci. Rep.* **9**, 6268 (2019).

7. Simonyan, K., Vedaldi, A. & Zisserman, A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *ArXiv13126034 Cs* (2014).

8. Aggarwal, M. *et al.* Towards Trainable Saliency Maps in Medical Imaging. *ArXiv201107482 Cs Eess* (2020).

9. Tjoa, E. & Guan, C. Quantifying Explainability of Saliency Methods in Deep Neural Networks. *ArXiv200902899 Cs* (2020).

10. Badgeley, M. A. *et al.* Deep learning predicts hip fracture using confounding patient and healthcare variables. *Npj Digit. Med.* **2**, 1–10 (2019).

11. Zech, J. R. *et al.* Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Med.* **15**, e1002683 (2018).

12. DeGrave, A. J., Janizek, J. D. & Lee, S.-I. AI for radiographic COVID-19 detection selects shortcuts over signal. *medRxiv* (2020) doi:10.1101/2020.09.13.20193565.

13. 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).

612 14. Selvaraju, R. R. *et al.* Grad-CAM: Visual Explanations from Deep Networks via
613 Gradient-based Localization. *Int. J. Comput. Vis.* **128**, 336–359 (2020).
614 15. Rizwan I Haque, I. & Neubert, J. Deep learning approaches to biomedical image
615 segmentation. *Inform. Med. Unlocked* **18**, 100297 (2020).
616 16. Makimoto, H. *et al.* Performance of a convolutional neural network derived from
617 an ECG database in recognizing myocardial infarction. *Sci. Rep.* **10**, 8445 (2020).
618 17. Raghunath, S. *et al.* Prediction of mortality from 12-lead electrocardiogram
619 voltage data using a deep neural network. *Nat. Med.* **26**, 886–891 (2020).
620 18. Porumb, M., Stranges, S., Pescapè, A. & Pecchia, L. Precision Medicine and
621 Artificial Intelligence: A Pilot Study on Deep Learning for Hypoglycemic Events
622 Detection based on ECG. *Sci. Rep.* **10**, 1–16 (2020).
623 19. Tham, Y.-C. *et al.* Referral for disease-related visual impairment using retinal
624 photograph-based deep learning: a proof-of-concept, model development study.
625 *Lancet Digit. Health* **3**, e29–e40 (2021).
626 20. Varadarajan, A. V. *et al.* Deep Learning for Predicting Refractive Error From
627 Retinal Fundus Images. *Invest. Ophthalmol. Vis. Sci.* **59**, 2861–2868 (2018).
628 21. Mitani, A. *et al.* Detection of anaemia from retinal fundus images via deep
629 learning. *Nat. Biomed. Eng.* **4**, 18–27 (2020).
630 22. Deep Learning to Assess Long-term Mortality From Chest Radiographs |
631 Pulmonary Medicine | JAMA Network Open | JAMA Network. https://jamanetwork-
632 com.stanford.idm.oclc.org/journals/jamanetworkopen/fullarticle/2738349.
633 23. Rajpurkar, P. *et al.* CheXaid: deep learning assistance for physician diagnosis of
634 tuberculosis using chest x-rays in patients with HIV. *Npj Digit. Med.* **3**, 1–8 (2020).

24. Tschandl, P. *et al.* Human–computer collaboration for skin cancer recognition. *Nat. Med.* **26**, 1229–1234 (2020).

25. 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).

26. Bien, N. *et al.* Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. *PLOS Med.* **15**, e1002699 (2018).

27. 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).

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

29. 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).

30. Vrabac, D. *et al.* DLBCL-Morph: Morphological features computed using deep learning for an annotated digital DLBCL image set. *ArXiv200908123 Cs* (2020).

31. Steiner, D. F. *et al.* Impact of Deep Learning Assistance on the Histopathologic Review of Lymph Nodes for Metastatic Breast Cancer. *Am. J. Surg. Pathol.* **42**, 1636–1646 (2018).

32. Uyumazturk, B. *et al.* Deep Learning for the Digital Pathologic Diagnosis of Cholangiocarcinoma and Hepatocellular Carcinoma: Evaluating the Impact of a Web-based Diagnostic Assistant. *ArXiv191107372 Eess* (2019).

33. Park, A. *et al.* Deep Learning–Assisted Diagnosis of Cerebral Aneurysms Using the HeadXNet Model. *JAMA Netw. Open* **2**, e195600 (2019).

34. Crosby, J., Chen, S., Li, F., MacMahon, H. & Giger, M. Network output visualization to uncover limitations of deep learning detection of pneumothorax. in *Medical Imaging 2020: Image Perception, Observer Performance, and Technology Assessment* vol. 11316 113160O (International Society for Optics and Photonics, 2020).

35. Melbye, H. & Dale, K. Interobserver Variability in the Radiographic Diagnosis of Adult Outpatient Pneumonia. *Acta Radiol.* **33**, 79–81 (1992).

36. Herman, P. G. *et al.* Disagreements in Chest Roentgen Interpretation. *CHEST* **68**, 278–282 (1975).

37. Albaum, M. N. *et al.* Interobserver Reliability of the Chest Radiograph in Community-Acquired Pneumonia. *CHEST* **110**, 343–350 (1996).

38. Arun, N. T. *et al.* Assessing the validity of saliency maps for abnormality localization in medical imaging. *ArXiv200600063 Cs* (2020).

39. Graziani, M., Lompech, T., Müller, H. & Andrearczyk, V. *Evaluation and Comparison of CNN Visual Explanations for Histopathology*. (2020).

40. Arun, N. *et al.* Assessing the (Un)Trustworthiness of Saliency Maps for Localizing Abnormalities in Medical Imaging. *ArXiv200802766 Cs* (2020).

41. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic Attribution for Deep Networks. *ArXiv170301365 Cs* (2017).

42. Durand, T., Mordan, T., Thome, N. & Cord, M. WILDCAT: Weakly Supervised Learning of Deep ConvNets for Image Classification, Pointwise Localization and

Segmentation. in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 5957–5966 (IEEE, 2017). doi:10.1109/CVPR.2017.631.

43. Smilkov, D., Thorat, N., Kim, B., Viégas, F. & Wattenberg, M. SmoothGrad: removing noise by adding noise. *ArXiv170603825 Cs Stat* (2017).

44. MD.ai. https://www.md.ai/.

45. Huang, G., Liu, Z., van der Maaten, L. & Weinberger, K. Q. Densely Connected Convolutional Networks. *ArXiv160806993 Cs* (2018).

46. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *ArXiv14126980 Cs* (2017).

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