

1 **Title**

2 Benchmarking saliency methods for chest X-ray interpretation

3

4 **Authors**

5 Adriel Saporta MS MBA^{1*}, Xiaotong Gui MS^{1*}, Ashwin Agrawal MS^{1*}, Anuj Pareek MD

6 PhD², Steven QH Truong MBA³, Chanh DT Nguyen PhD^{3,4}, Van-Doan Ngo MD⁵, Jayne

7 Seekins DO⁶, Francis G. Blankenberg MD⁶, Andrew Y. Ng PhD¹, Matthew P. Lungren MD

8 MPH^{2†}, Pranav Rajpurkar PhD ^{7†}

9

10 **Affiliations**

11 ¹Stanford University Department of Computer Science, USA

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

13 ³VinBrain, Vietnam

14 ⁴VinUniversity, Vietnam

15 ⁵Vinmec International Hospital, Vietnam

16 ⁶Stanford University School of Medicine, Department of Radiology, USA

17 ⁷Department of Biomedical Informatics, Harvard University, USA

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

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

20

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

22

23 Current word count: 3829

24 **Abstract**

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

44

45 **Introduction**

46 Deep learning has enabled automated medical imaging interpretation at the level of
47 practicing experts in some settings¹⁻³. While the potential benefits of automated
48 diagnostic models are numerous, lack of model interpretability in the use of “black-box”
49 deep neural networks (DNNs) represents a major barrier to clinical trust and adoption⁴⁻⁶.
50 In fact, it has been argued that the European Union’s recently adopted General Data
51 Protection Regulation (GDPR) affirms an individual’s right to an explanation in the context
52 of automated decision-making⁷. Although the importance of DNN interpretability is widely
53 acknowledged and many techniques have been proposed, little emphasis has been
54 placed on how best to quantitatively evaluate these explainability methods⁸.

55
56 One type of DNN interpretation strategy widely used in the context of medical imaging is
57 based on saliency (or pixel-attribution) methods⁹⁻¹². Saliency methods produce heat
58 maps highlighting the areas of the medical image that most influenced the DNN’s
59 prediction. The heat maps help to visualize whether a DNN is concentrating on the same
60 regions of a medical image that a human expert would focus on, rather than concentrating
61 on a clinically irrelevant part of the medical image or even on confounders in the image¹³⁻
62 ¹⁵. Saliency methods have been widely used for a variety of medical imaging tasks and
63 modalities including, but not limited to, visualizing the performance of a convolutional
64 neural network (CNN) in predicting (1) myocardial infarction¹⁶ and hypoglycemia¹⁷ from
65 electrocardiograms, (2) visual impairment¹⁸, refractive error¹⁹, and anaemia²⁰ from retinal
66 photographs, (3) long-term mortality²¹ and tuberculosis²² from chest X-ray (CXR) images,
67 and (4) appendicitis²³ and pulmonary embolism²⁴ on computed tomography scans.
68 However, recent work has shown that saliency methods used to validate model

69 predictions can be misleading in some cases and may lead to increased bias and loss of
70 user trust in high-stakes contexts such as healthcare^{25–27}. Therefore, a rigorous
71 investigation of the accuracy and reliability of these strategies is necessary before they
72 are integrated into the clinical setting²⁸.

73

74 In this work, we perform a systematic evaluation of the three most common saliency
75 methods in medical imaging (Grad-CAM²⁹, Grad-CAM++³⁰, and Integrated Gradients³¹)
76 using three common CNN architectures (DenseNet121³², ResNet152³³, Inception-v4³⁴).

77 In doing so, we establish the first human benchmark in CXR localization by collecting
78 radiologist segmentations for 10 pathologies using CheXpert, a large publicly available
79 CXR dataset³⁵. To compare saliency method segmentations with expert segmentations,
80 we use two metrics to capture localization accuracy: (1) *mean Intersection over Union*, a
81 stricter metric that measures the overlap between the saliency method segmentation and
82 the expert segmentation, and (2) *hit rate*, a less strict metric that does not require the
83 saliency method to locate the full extent of a pathology. We find that (1) while Grad-CAM
84 generally localizes pathologies more accurately than the two other saliency methods, all
85 three perform significantly worse compared with a human radiologist benchmark; (2) the
86 gap in localization performance between Grad-CAM and the human benchmark is largest
87 for pathologies that have multiple instances on the same CXR, are smaller in size, and
88 have shapes that are more complex; (3) model confidence is positively correlated with
89 Grad-CAM localization performance. We publicly release a development dataset of expert
90 segmentations, which we call CheXplanation, to facilitate further research in DNN
91 explainability for medical imaging.

92

93 **Results**

94 **Framework for evaluating saliency methods on multi-label classification models**

95 Three methods were evaluated—Grad-CAM, Grad-CAM++, and Integrated Gradients—
96 in a multi-label classification setup on the CheXpert dataset (Fig. 1a). For each of the
97 three saliency methods, we ran experiments using three CNN architectures previously
98 used on CheXpert: DenseNet121, ResNet152, and Inception-v4. For each combination
99 of saliency method and model architecture, we trained and evaluated an ensemble of 30
100 CNNs (see Methods for ensembling details). We then passed each of the CXRs in the
101 dataset's holdout test set into the trained ensemble model to obtain image-level
102 predictions for the following 10 pathologies: Atelectasis, Cardiomegaly, Consolidation,
103 Edema, Enlarged Cardiomeastinum, Lung Lesion, Lung Opacity, Pleural Effusion,
104 Pneumothorax, and Support Devices. For each CXR, we used the saliency method to
105 generate heat maps, one for each of the 10 pathologies, and then applied a threshold to
106 each heat map to produce binary segmentations (top row, Fig. 1a). Thresholding is
107 determined per pathology using Otsu's method³⁶, which iteratively searches for a
108 threshold value that maximizes inter-class pixel intensity variance. We also conducted a
109 sensitivity analysis of localization performance using different thresholds. The result
110 shows that our evaluation of localization performance is robust to different saliency map
111 thresholding values (see Supplementary Fig. 15). Additionally, to calculate the hit rate
112 evaluation metric (described below), we extracted the pixel in the saliency method heat
113 map with the largest value as the single most representative point on the CXR for that
114 pathology.

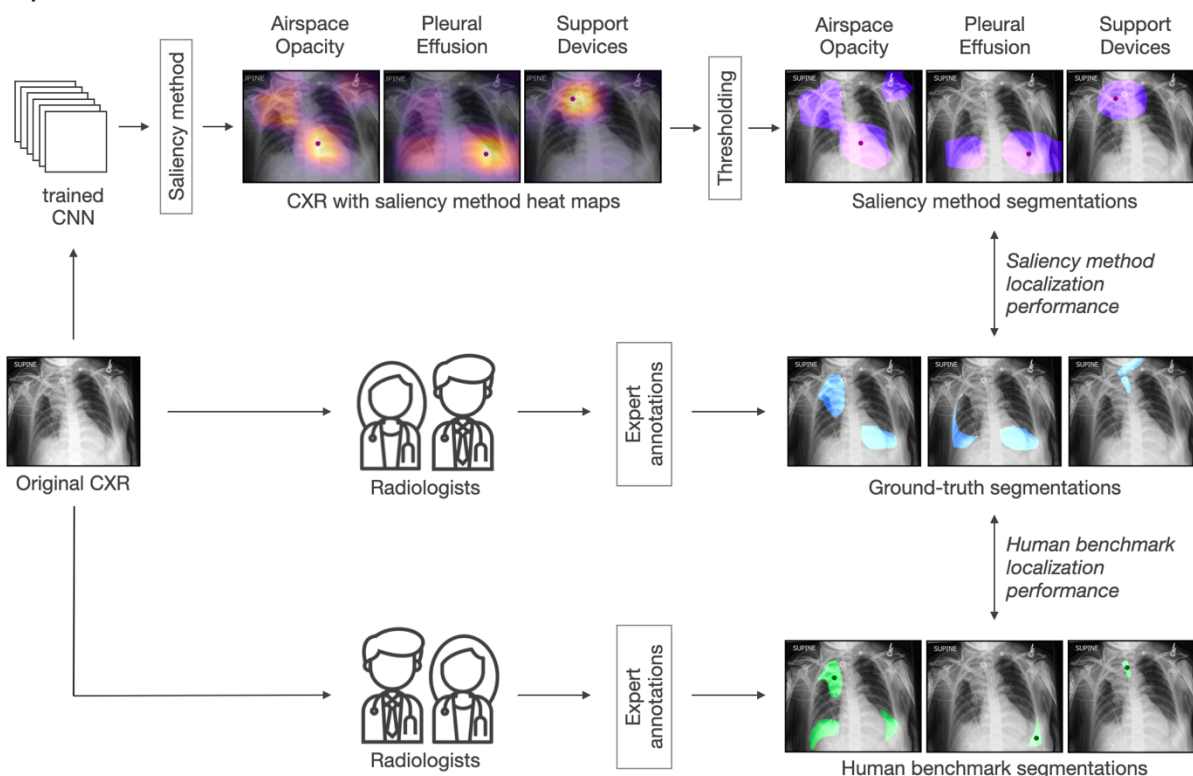
115

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

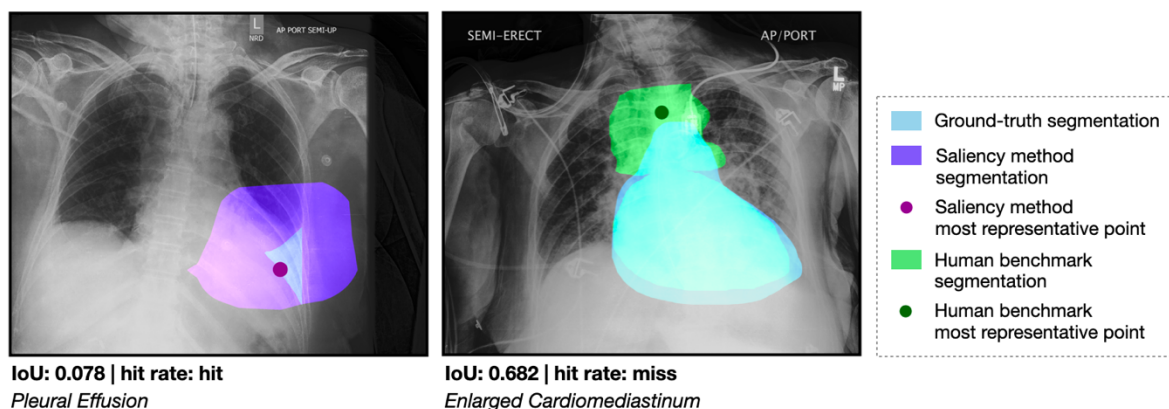
126

127 We used two evaluation metrics to compare segmentations (Fig. 1b). First, we used *mean*
128 *Intersection over Union* (mIoU), a stricter metric that measures how much, on average,
129 either the saliency method or benchmark segmentations overlapped with the ground-truth
130 segmentations. Second, we used *hit rate*, a less strict metric that does not require the
131 saliency method or benchmark annotators to locate the full extent of a pathology. Hit rate
132 is based on the pointing game setup³⁷, in which credit is given if the most representative
133 point identified by the saliency method or the benchmark annotators lies within the
134 ground-truth segmentation. A “hit” indicates that the correct region of the CXR was
135 located regardless of the exact bounds of the binary segmentations. Localization
136 performance is then calculated as the hit rate across the dataset³⁸.

a | Annotation and evaluation workflow



b | Evaluation metrics on example CXRs



137

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

148 ground-truth annotations to evaluate how well saliency method identifies clinically-relevant areas
149 of the input CXR (“saliency method localization performance”). Bottom row: Two board-certified
150 radiologists (separate from those in middle row) were also asked to segment the pathologies that
151 were present in the CXR as determined by the dataset’s ground-truth labels. In addition, these
152 radiologists were asked to locate the single point on the CXR that was most representative of
153 each present pathology. These benchmark annotations are compared to the ground-truth
154 annotations to determine a human benchmark (“human benchmark localization performance”).
155 **b**, Left: CXR with ground-truth and saliency method annotations for Pleural Effusion. The
156 segmentations have a low overlap (IoU is 0.078), but pointing game is a “hit” since the saliency
157 method’s most representative point is inside of the ground-truth segmentation. Right, CXR with
158 ground-truth and human benchmark annotations for Enlarged Cardiomeastinum. The
159 segmentations have a high overlap (IoU is 0.682), but pointing game is a “miss” since saliency
160 method’s most representative point is outside of the ground-truth segmentation.
161

162 **Evaluating localization performance of the saliency methods against the human** 163 **benchmark**

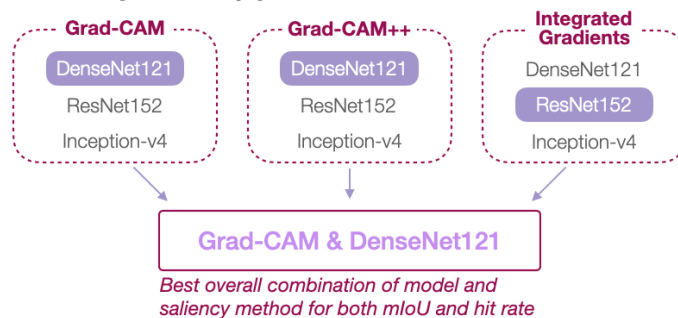
164 In order to compare the localization performance of the saliency methods with the human
165 benchmark, we ran eighteen experiments, one for each combination of saliency method
166 (Grad-CAM, Grad-CAM++, or Integrated Gradients) and CNN architecture
167 (DenseNet121, ResNet152, or Inception-v4) using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

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

181
182 We found that the saliency method pipeline demonstrated significantly worse localization
183 performance when compared with the human benchmark using both mIoU (Fig. 2b) and
184 hit rate (Fig. 2c) as an evaluation metric, regardless of model classification AUROC. For
185 each metric, we report the 95% confidence intervals using the bootstrap method with
186 1,000 bootstrap samples³⁹. For five of the 10 pathologies, the saliency method pipeline
187 had a significantly lower mIoU than the human benchmark. For example, the saliency
188 method pipeline had one of the highest AUROC scores of the 10 pathologies for Support
189 Devices (0.969), but had among the worst localization performance for Support Devices
190 when using both mIoU (0.163 [95% CI 0.154, 0.172]) and hit rate (0.357 [95% CI 0.303,
191 0.408]) as evaluation metrics. On two pathologies (Atelectasis and Consolidation) the
192 saliency method pipeline significantly outperformed the human benchmark. On average,
193 across all 10 pathologies, mIoU saliency method pipeline performance was 26.6% [95%
194 CI 18.1%, 35.0%] worse than the human benchmark, with Lung Lesion displaying the
195 largest gap in performance (76.2% [95% CI 59.1%, 87.5%] worse than the human
196 benchmark) (Supplementary Table 4). Consolidation was the pathology on which the
197 mIoU saliency method pipeline performance exceeded the human benchmark the most,
198 by 56.1% [95% CI 42.7%, 69.4%]. For seven of the 10 pathologies, the saliency method
199 pipeline had a significantly lower hit rate than the human benchmark. On average, hit rate
200 saliency method pipeline performance was 29.4% [95% CI 15.0%, 43.2%] worse than the

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

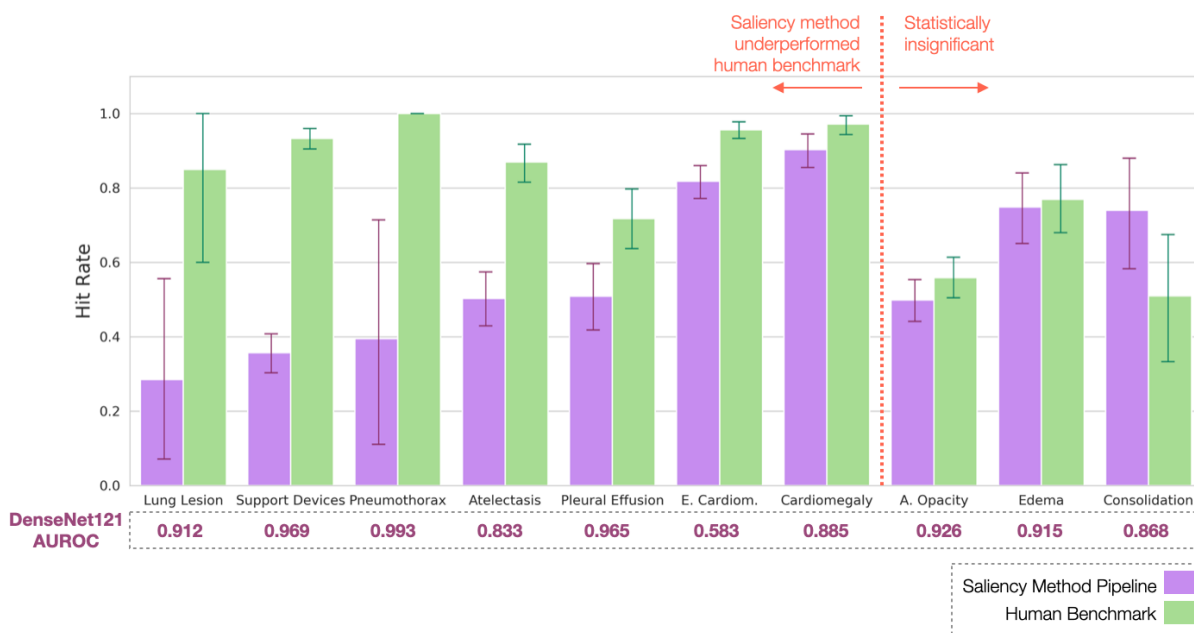
a | Selection strategy for saliency method pipeline



b | mIoU localization performances



c | Hit rate localization performances



211 **Fig. 2 | Evaluating the localization performance of the saliency methods against the human**
212 **benchmark. a,** The selection strategy for the mIoU saliency method pipeline and the hit rate
213 saliency method pipeline. For each saliency method, the best model architecture is selected
214 (highlighted in purple). Then, the best saliency method + model architecture pair, of the three, is
215 selected. The selection strategy was performed twice, once using mIoU as the evaluation metric
216 and once using hit rate as the evaluation metric. The best saliency method + model architecture
217 pair was the same for both mIoU and hit rate: Grad-CAM + DenseNet121. **b,** Comparing saliency
218 method and human benchmark localization performances under the overlap evaluation scheme
219 (mIoU). Pathologies, along with their DenseNet121 AUROCs, are sorted on the x-axis in
220 descending order of percentage decrease from human benchmark mIoU to saliency method
221 pipeline mIoU for each pathology. **c,** Comparing saliency method and human benchmark
222 localization performances under the hit rate evaluation scheme. Pathologies, along with their
223 DenseNet121 AUROCs, are sorted on the x-axis in descending order of percentage decrease
224 from human benchmark hit rate to saliency method pipeline hit rate for each pathology.
225

226 **Characterizing the underperformance of the saliency method pipeline localization**

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

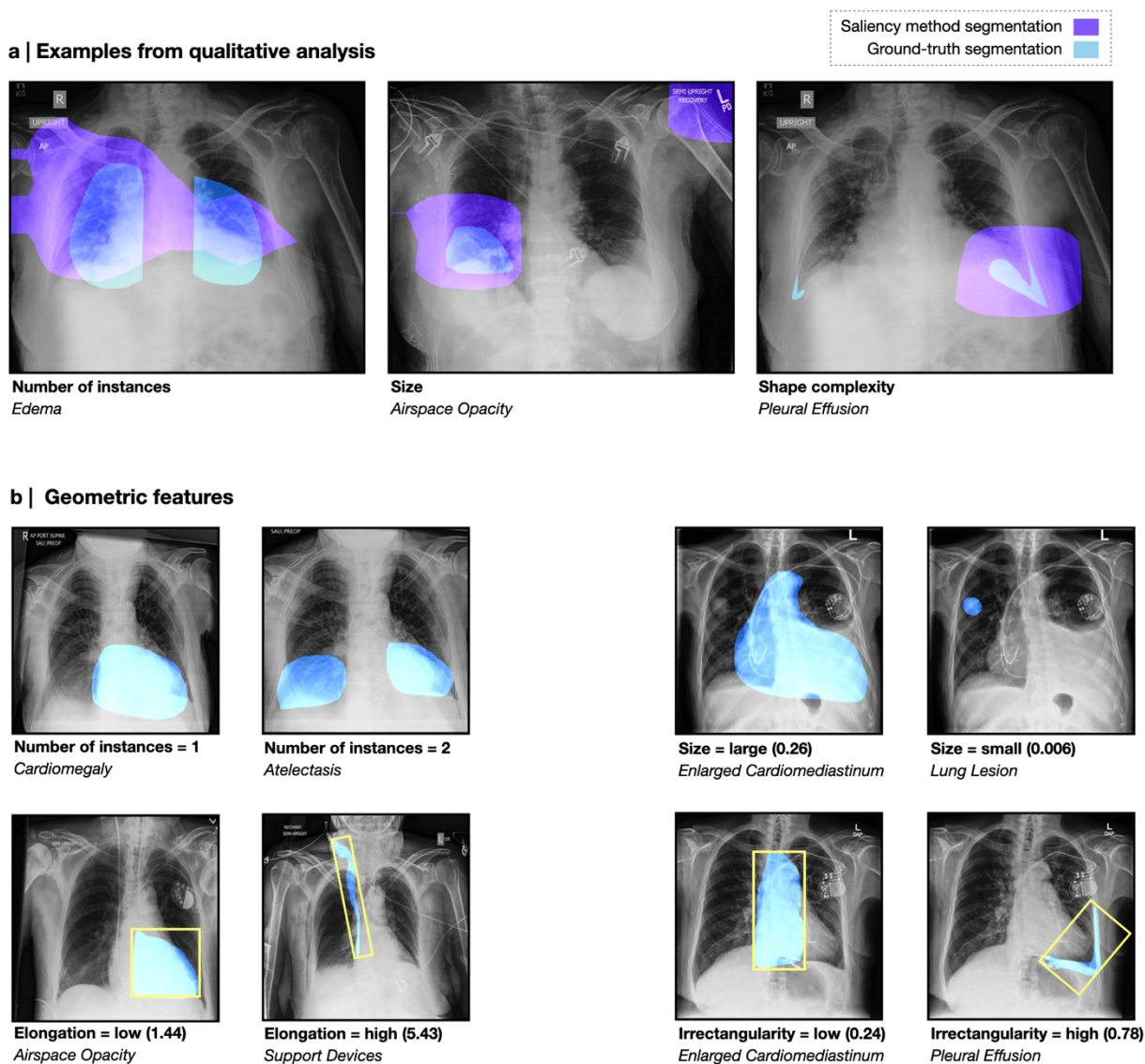
242

243 Informed by our qualitative analysis and previous work in histology⁴⁰, we defined four
244 geometric features for our quantitative analysis (Fig. 3b): (1) *number of instances* (for
245 example, bilateral Pleural Effusion would have two instances, whereas there is only one
246 instance for Cardiomegaly), (2) *size* (pathology area with respect to the area of the whole
247 CXR), (3) *elongation* and (4) *irrectangularity* (the last two features measure the complexity
248 of the pathology shape and were calculated by fitting a rectangle of minimum area
249 enclosing the binary mask). See Supplementary Fig. 17 for the distribution of the four
250 pathological characteristics across all 10 pathologies.

251

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

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



267
 268 **Fig. 3 | Characterizing the underperformance of saliency method localization.** **a**, Example
 269 CXRs that highlight the three pathological characteristics identified by our qualitative analysis:
 270 (1) Left, number of instances; (2) Middle, size; and (3) Right, shape complexity. **b**, Example CXRs
 271 with the four geometric features used in our quantitative analysis: (1) Top row left, number of
 272 instances; (2) Top row right, size = area of segmentation/area of CXR; (3) Bottom row left,
 273 elongation; and (4) Bottom row right, irrectangularity. Elongation and irrectangularity were
 274 calculated by fitting a rectangle of minimum area enclosing the binary mask (as indicated by the
 275 yellow rectangles). Elongation = $\text{maxAxis}/\text{minAxis}$. Irrectangularity = $1 - (\text{area of}$
 276 $\text{segmentation}/\text{area of enclosing rectangle})$.
 277

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

290
291 We found that these same three pathological characteristics—larger size, and higher
292 elongation and irrectangularity—characterized the *gap* in mIoU localization performance
293 between saliency method and human benchmark. We observed that the *gap* in hit rate
294 localization performance was significantly characterized by all four geometric features
295 (number of instances, size, elongation, and irrectangularity). As the number of instances
296 increased, despite no significant change in hit rate localization performance itself, the *gap*
297 in hit rate localization performance between saliency method and the human benchmark
298 increased (0.470 [95% CI 0.114, 0.825]). This suggests that the saliency method performs
299 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		

300

301

302 **Effect of model confidence on saliency method localization performance**

303 We also conducted statistical analyses to determine whether there was any correlation
304 between the model's confidence in its prediction and saliency method pipeline
305 performance (Table 2). We first ran a simple regression for each pathology using the
306 model's probability output as the single independent variable and using the saliency
307 method IoU as the dependent 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 used the full dataset since the analysis of false positives and false
310 negatives was also of interest. In addition to the linear regression coefficients, we also
311 computed the Spearman correlation coefficients to capture any potential non-linear
312 associations.

313
 314 We found that for all pathologies, model confidence was positively correlated with mIoU
 315 saliency method pipeline performance. The p-values for all coefficients were below 0.001
 316 except for the coefficients for Pneumothorax (n=11) and Lung Lesion (n=50), the two
 317 pathologies for which we had the fewest positive examples. Of all the pathologies, model
 318 confidence for positive predictions of Enlarged Cardiomeastinum had the largest linear
 319 regression coefficient with mIoU saliency method pipeline performance (1.974, p-
 320 value<0.001). Model confidence for positive predictions of Pneumothorax had the largest
 321 Spearman correlation coefficient with mIoU saliency method pipeline performance (0.734,
 322 p-value<0.01), followed by Pleural Effusion (0.690, p-value<0.001). Combining all
 323 pathologies (n=2365), the linear regression coefficient was 0.109 (95% CI [0.083, 0.135]),
 324 and the Spearman correlation coefficient was 0.285 (95% CI [0.239, 0.331]). We also
 325 performed analogous experiments using hit rate as the dependent variable and found
 326 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 Cardiomeastinum	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

327

328

329 **Discussion**

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

344

345 While there are several public CXR datasets with image-level labels annotated by expert
346 radiologists, including MIMIC-CXR⁴¹ and ChestX-ray8⁴², and several datasets containing
347 segmentations for a single pathology, including SIIM-ACR Pneumothorax Segmentation⁴³
348 and RSNA Pneumonia Detection⁴⁴, to our knowledge there are no other publicly available
349 CXR datasets with multilabel expert segmentations. By publicly releasing a development
350 dataset, CheXplanation, of 234 images with 885 expert segmentations, and a competition

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

353

354 Our work has several potential implications for human-AI collaboration in the context of
355 medical decision-making. Heat maps generated using saliency methods are advocated
356 as clinical decision support in the hope that they not only improve clinical decision-
357 making, but also encourage clinicians to trust model predictions⁴⁵⁻⁴⁷. Many of the large
358 CXR vendors⁴⁸⁻⁵⁰ use localization methods to provide pathology visualization in their
359 computer-aided detection (CAD) products. In addition to being used for clinical
360 interpretation, saliency method heat maps are also used for the evaluation of CXR
361 interpretation models, for quality improvement (QI) and quality assurance (QA) in clinical
362 practices, and for dataset annotation⁵¹. However, we found that saliency method
363 localization performance, on balance, performed worse than expert localization across
364 multiple analyses and across many important pathologies (our findings are consistent
365 with recent work focused on localizing a single pathology, Pneumothorax, in CXRs⁵²). If
366 used in clinical practice, heat maps that incorrectly highlight medical images may
367 exacerbate well documented biases (chiefly, automation bias) and erode trust in model
368 predictions (even when model output is correct), limiting clinical translation²².

369

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

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

393

394 Our work builds upon several studies investigating the validity of saliency maps for
395 localization^{56,57} and upon some early work on the trustworthiness of saliency methods to
396 explain DNNs in medical imaging⁵⁸. However, as recent work has shown³¹, evaluating

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

420

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

432

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

438

439 **Methods**

440 **Ethical and information governance approvals.**

441 This study does not involve human subject participants.

442

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.ai⁵⁹ (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 optimizer⁶⁰ was used with default β -

488 parameters of $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate was hyperparameter tuned for
489 the different model architectures. The best learning rate for each architecture was:
490 1×10^{-4} for DenseNet121, 1×10^{-5} for ResNet152, 1×10^{-5} for Inceptionv4. Batches were
491 sampled using a fixed batch size of 16 images.

492
493 *Ensembling.* We use an ensemble of checkpoints to create both predictions and saliency
494 maps to maximize model performance. In order to capture uncertainties inherent in
495 radiograph interpretation, we train our models using four uncertainty handling strategies
496 outlined in CheXpert: Ignoring, Zeroes, Ones, and 3-Class Classification. For each of the
497 four uncertainty handling strategies, we train our model three separate times, each time
498 saving the 10 checkpoints across the three epochs with the highest average AUC across
499 5 observations selected for their clinical importance and prevalence in the validation set:
500 Atelectasis, Cardiomegaly, Consolidation, Edema, and Pleural Effusion. In total, after
501 training, we have saved $4 \times 30 = 120$ checkpoints for a given model. Then, from the 120
502 saved checkpoints for that model, we select the top 10 performing checkpoints for each
503 pathology. For each CXR and each task, we compute the predictions and saliency maps
504 using the relevant checkpoints. We then take the mean both of the predictions and of the
505 saliency maps to create the final set of predictions and saliency maps for the ensemble
506 model. See Supplementary Table 3 for the performance of the model on each of the
507 pathologies.

508
509 **CNN interpretation strategy.** Saliency methods (Grad-CAM, Grad-CAM++, and
510 Integrated Gradients) were used to visualize the decision made by the classification

511 network. The saliency map was resized to the original image dimension using bilinear
512 interpolation. It was then normalized using max-min normalization and then converted
513 into a binary segmentation using binary thresholding (Otsu's method). We also reported
514 mIoU localization performance using different saliency map thresholding values. We first
515 applied max-min normalizations to the saliency maps so that each value gets transformed
516 into a decimal between 0 and 1. We then passed in a range of threshold values from 0.2
517 to 0.8 to create binary segmentations and calculated the mIoU score per pathology under
518 each threshold on the validation set. Then for the analysis with the full dataset (see
519 Supplementary Figure 16), we further ensure that the final binary segmentation is
520 consistent with model probability output by applying another layer of thresholding such
521 that the segmentation mask produced all zeros if the predicted probability was below a
522 chosen level. The probability threshold is searched on the interval of [0,0.8] with steps of
523 0.1. The exact value is determined per pathology by maximizing the mIoU on validation
524 set.

525
526 *Segmentation evaluation metrics.* Localization performance of each segmentation was
527 evaluated using Intersection over Union (IoU) score. The IoU is the ratio between the
528 area of overlap and the area of union between the ground truth and the predicted areas,
529 ranging from 0 to 1 with 0 signifying no overlap and 1 signifying perfectly overlapping
530 segmentation. Confidence intervals are calculated using bootstrapping with 1000
531 bootstrap samples. The variance in the width of CI across pathologies can be explained
532 by difference in sample sizes. For the percentage decrease from expert mIoU to AI mIoU,
533 we bootstrapped the difference between human benchmark and saliency method

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

536

537 **Statistical analysis.**

538 *Pathology Characteristics.* We used four features to characterize the pathologies. (1)

539 Number of instances is defined as the number of disjoint components in the

540 segmentation. (2) Size is the area of the pathology divided by the total image area. (3)

541 and (4) Elongation and irrectangularity are geometric features that measure shape

542 complexities. They were designed to quantify what radiologists qualitatively described as

543 focal or diffused. To calculate the metrics, a rectangle of minimum area enclosing the

544 contour is fitted to each pathology. Elongation is defined as the ratio of the rectangle's

545 longer side to short side. Irrectangularity = $1 - (\text{area of segmentation}/\text{area of enclosing}$

546 rectangle), with values ranging from 0 to 1 with 1 being very irrectangular. When there

547 are multiple instances within one pathology, we used the characteristics of the dominant

548 instance (largest in perimeter).

549

550 *Model Confidence.* We used the probability output of the DNN architecture for model

551 confidence. The probabilities were normalized using max-min normalization per

552 pathology before aggregation.

553

554 *Linear Regression.* For each evaluation scheme (overlap and hit rate), we ran two groups

555 of simple linear regressions, with AI evaluation metrics and their differences as the

556 response variables. Each group has four regressions using the above four pathological

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

561

562 **Acknowledgements**

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

566

567 **Data Availability**

568 CheXpert data is available at <https://stanfordmlgroup.github.io/competitions/chexpert/>.

569 The validation set and corresponding benchmark radiologist annotations can be
570 downloaded from <https://stanfordmlgroup.github.io/competitions/chexplanation/>.

571

572 **Code Availability**

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

576

577 **Competing Interests**

578 M.L. is an advisor for and/or has research funded by GE, Philips, Carestream, Nines
579 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-](https://jamanetwork-com.stanford.idm.oclc.org/journals/jamanetworkopen/fullarticle/2738349)
644 [com.stanford.idm.oclc.org/journals/jamanetworkopen/fullarticle/2738349](https://jamanetwork-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).

- 647 23. Rajpurkar, P. *et al.* AppendiXNet: Deep Learning for Diagnosis of Appendicitis
648 from A Small Dataset of CT Exams Using Video Pretraining. *Sci. Rep.* **10**, 3958
649 (2020).
- 650 24. Huang, S.-C. *et al.* PENet—a scalable deep-learning model for automated
651 diagnosis of pulmonary embolism using volumetric CT imaging. *Npj Digit. Med.* **3**, 1–
652 9 (2020).
- 653 25. Rudin, C. Stop explaining black box machine learning models for high stakes
654 decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215 (2019).
- 655 26. Eitel, F. & Ritter, K. Testing the robustness of attribution methods for
656 convolutional neural networks in MRI-based Alzheimer’s disease classification.
657 *ArXiv190908856 Cs Eess* (2019).
- 658 27. Young, K., Booth, G., Simpson, B., Dutton, R. & Shrapnel, S. Deep neural
659 network or dermatologist? *ArXiv190806612 Cs Eess Stat* **11797**, 48–55 (2019).
- 660 28. Reyes, M., Meier, R., Pereira, S., Silva, C. A. & Dahlweid, F.-M. On the
661 Interpretability of Artificial Intelligence in Radiology: Challenges and Opportunities. **2**,
662 12.
- 663 29. Selvaraju, R. R. *et al.* Grad-CAM: Visual Explanations from Deep Networks via
664 Gradient-based Localization. *Int. J. Comput. Vis.* **128**, 336–359 (2020).
- 665 30. Chattopadhyay, A., Sarkar, A., Howlader, P. & Balasubramanian, V. N. Grad-
666 CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional
667 Networks. in *2018 IEEE Winter Conference on Applications of Computer Vision*
668 (WACV) 839–847 (2018). doi:10.1109/WACV.2018.00097.

- 669 31. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic attribution for deep networks. in
670 *Proceedings of the 34th International Conference on Machine Learning - Volume 70*
671 3319–3328 (JMLR.org, 2017).
- 672 32. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely Connected
673 Convolutional Networks. in *2017 IEEE Conference on Computer Vision and Pattern*
674 *Recognition (CVPR)* 2261–2269 (IEEE, 2017). doi:10.1109/CVPR.2017.243.
- 675 33. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image
676 Recognition. in *2016 IEEE Conference on Computer Vision and Pattern Recognition*
677 *(CVPR)* 770–778 (IEEE, 2016). doi:10.1109/CVPR.2016.90.
- 678 34. Szegedy, C. *et al.* Going deeper with convolutions. in *2015 IEEE Conference on*
679 *Computer Vision and Pattern Recognition (CVPR)* 1–9 (2015).
680 doi:10.1109/CVPR.2015.7298594.
- 681 35. Irvin, J. *et al.* CheXpert: A Large Chest Radiograph Dataset with Uncertainty
682 Labels and Expert Comparison. *Proc. AAAI Conf. Artif. Intell.* **33**, 590–597 (2019).
- 683 36. Otsu, N. A Threshold Selection Method from Gray-Level Histograms. *IEEE Trans*
684 *Syst. Man Cybern.* 62–66 (1979).
- 685 37. Zhang, J., Lin, Z., Brandt, J., Shen, X. & Sclaroff, S. Top-down Neural Attention
686 by Excitation Backprop. *ArXiv160800507 Cs* (2016).
- 687 38. Kim, H.-E. *et al.* Changes in cancer detection and false-positive recall in
688 mammography using artificial intelligence: a retrospective, multireader study. *Lancet*
689 *Digit. Health* **2**, e138–e148 (2020).
- 690 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-](https://kaggle.com/c/siim-acr-pneumothorax-segmentation)
699 [pneumothorax-segmentation](https://kaggle.com/c/siim-acr-pneumothorax-segmentation).
- 700 44. RSNA Pneumonia Detection Challenge. [https://kaggle.com/c/rsna-pneumonia-](https://kaggle.com/c/rsna-pneumonia-detection-challenge)
701 [detection-challenge](https://kaggle.com/c/rsna-pneumonia-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).

