

1 **Title**

2 Benchmarking saliency methods for chest X-ray interpretation

3

4 **Authors**

5 Adriel Saporta MS MBA^{1*}, Xiaotong Gui MS^{2*}, Ashwin Agrawal MS^{2*}, Anuj Pareek MD

6 PhD³, Steven QH Truong MBA⁴, Chanh DT Nguyen PhD^{4,5}, Van-Doan Ngo MD⁶, Jayne

7 Seekins DO⁷, Francis G. Blankenberg MD⁷, Andrew Y. Ng PhD², Matthew P. Lungren MD

8 MPH³, Pranav Rajpurkar PhD⁸

9

10 **Affiliations**

11 ¹Department of Computer Science, New York University, USA

12 ²Department of Computer Science, Stanford University, USA

13 ³Stanford Center for Artificial Intelligence in Medicine and Imaging, USA

14 ⁴VinBrain, Vietnam

15 ⁵VinUniversity, Vietnam

16 ⁶Vinmec International Hospital, Vietnam

17 ⁷Department of Radiology, Stanford University School of Medicine, USA

18 ⁸Department of Biomedical Informatics, Harvard University, USA

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

20

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

22

23 Current word count: 4438

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 seven saliency methods—
31 including Grad-CAM, Grad-CAM++, and Integrated Gradients—across multiple neural
32 network architectures using two evaluation metrics. We establish the first human
33 benchmark for chest X-ray segmentation in a multilabel classification set up, and examine
34 under what clinical conditions saliency maps might be more prone to failure in localizing
35 important pathologies compared to a human expert benchmark. We find that (i) while
36 Grad-CAM generally localized pathologies better than the other evaluated saliency
37 methods, all seven performed significantly worse compared with the human benchmark;
38 (ii) the gap in localization performance between Grad-CAM and the human benchmark
39 was largest for pathologies that had multiple instances, were smaller in size, and had
40 shapes that were more complex; (iii) model confidence was positively correlated with
41 Grad-CAM localization performance. While it is difficult to know whether poor localization
42 performance is attributable to the model or to the saliency method, our work demonstrates
43 that several important limitations of saliency methods must be addressed before we can
44 rely on them for deep learning explainability in medical imaging.

45

46 **Introduction**

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

56
57 One type of DNN interpretation strategy widely used in the context of medical imaging is
58 based on saliency (or pixel-attribution) methods⁹⁻¹². Saliency methods produce heat
59 maps highlighting the areas of the medical image that most influenced the DNN’s
60 prediction. Since saliency methods provide post-hoc interpretability of models that are
61 never exposed to bounding box annotations or pixel-level segmentations during training,
62 they are particularly useful in the context of medical imaging where ground-truth
63 segmentations can be especially time-consuming and expensive to obtain. The heat
64 maps help to visualize whether a DNN is concentrating on the same regions of a medical
65 image that a human expert would focus on, rather than concentrating on a clinically
66 irrelevant part of the medical image or even on confounders in the image¹³⁻¹⁵. Saliency
67 methods have been widely used for a variety of medical imaging tasks and modalities
68 including, but not limited to, visualizing the performance of a convolutional neural network
69 (CNN) in predicting (1) myocardial infarction¹⁶ and hypoglycemia¹⁷ from

70 electrocardiograms, (2) visual impairment¹⁸, refractive error¹⁹, and anaemia²⁰ from retinal
71 photographs, (3) long-term mortality²¹ and tuberculosis²² from chest X-ray (CXR) images,
72 and (4) appendicitis²³ and pulmonary embolism²⁴ on computed tomography scans.
73 However, recent work has shown that saliency methods used to validate model
74 predictions can be misleading in some cases and may lead to increased bias and loss of
75 user trust in high-stakes contexts such as healthcare^{25–28}. Therefore, a rigorous
76 investigation of the accuracy and reliability of these strategies is necessary before they
77 are integrated into the clinical setting²⁹.

78
79 In this work, we perform a systematic evaluation of seven common saliency methods in
80 medical imaging (Grad-CAM³⁰, Grad-CAM++³¹, Integrated Gradients³², Eigen-CAM⁴¹,
81 DeepLIFT⁴², Layer-Wise Relevance Propagation⁴³, and Occlusion⁴⁴) using three common
82 CNN architectures (DenseNet121³³, ResNet152³⁴, Inception-v4³⁵). In doing so, we
83 establish the first human benchmark in CXR segmentation by collecting radiologist
84 segmentations for 10 pathologies using CheXpert, a large publicly available CXR
85 dataset³⁶. To compare saliency method segmentations with expert segmentations, we
86 use two metrics to capture localization accuracy: (1) *mean Intersection over Union*, a
87 metric that measures the overlap between the saliency method segmentation and the
88 expert segmentation, and (2) *hit rate*, a less strict metric than mIoU that does not require
89 the saliency method to locate the full extent of a pathology. We find that (1) while Grad-
90 CAM generally localizes pathologies more accurately than the other evaluated saliency
91 methods, all seven perform significantly worse compared with a human radiologist
92 benchmark (although it is difficult to know whether poor localization performance is

93 attributable to the model or to the saliency method); (2) the gap in localization
94 performance between Grad-CAM and the human benchmark is largest for pathologies
95 that have multiple instances on the same CXR, are smaller in size, and have shapes that
96 are more complex; (3) model confidence is positively correlated with Grad-CAM
97 localization performance. We publicly release a development dataset of expert
98 segmentations, which we call CheXlocalize, to facilitate further research in DNN
99 explainability for medical imaging.

100

101 **Results**

102 **Framework for evaluating saliency methods**

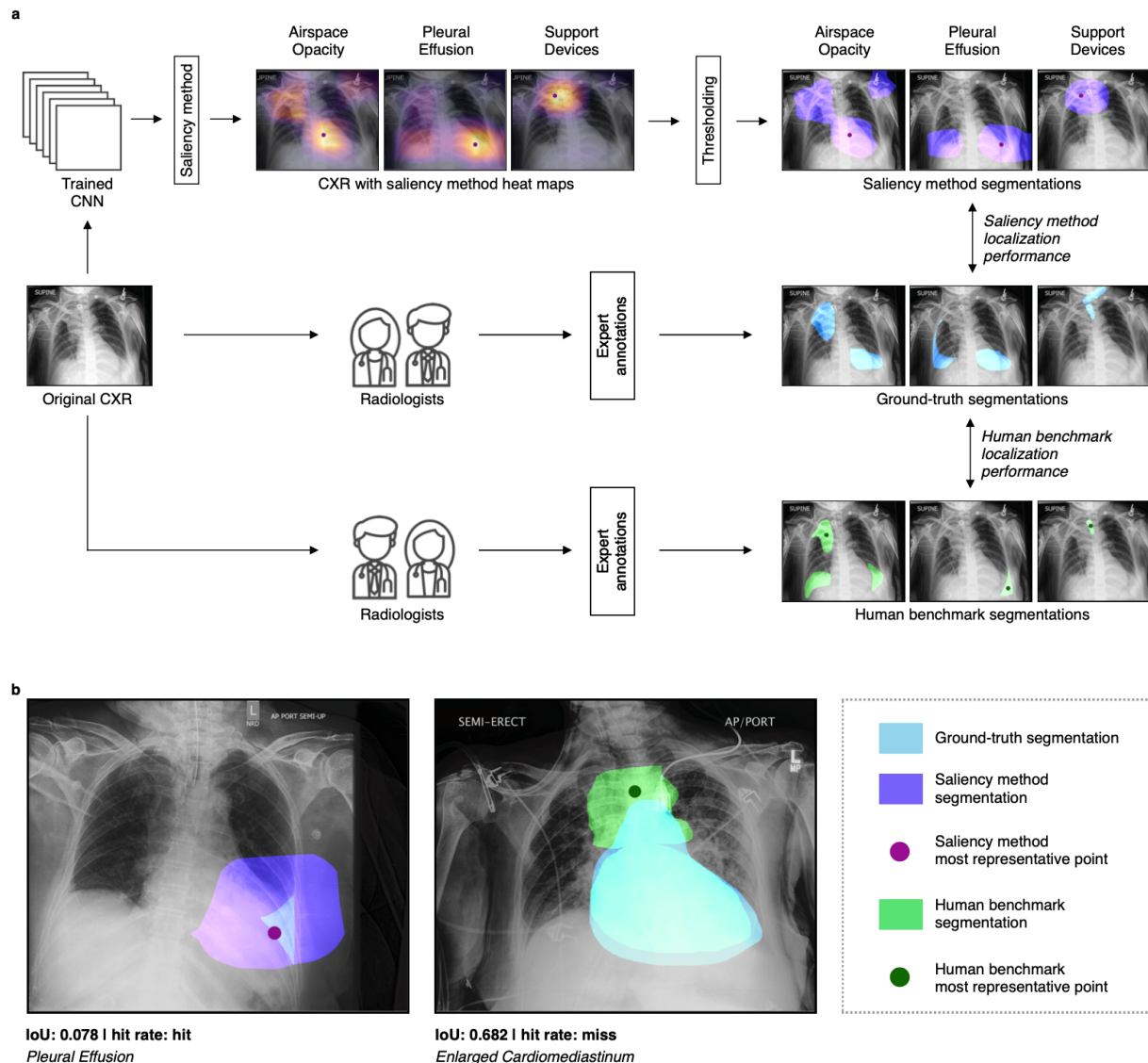
103 Seven methods were evaluated—Grad-CAM, Grad-CAM++, Integrated Gradients, Eigen-
104 CAM, DeepLIFT, Layer-Wise Relevance Propagation (LRP), and Occlusion—in a multi-
105 label classification setup on the CheXpert dataset (Fig. 1a). We ran experiments using
106 three CNN architectures previously used on CheXpert: DenseNet121, ResNet152, and
107 Inception-v4. For each combination of saliency method and model architecture, we
108 trained and evaluated an ensemble of 30 CNNs (see Methods for ensembling details).
109 We then passed each of the CXRs in the dataset’s holdout test set into the trained
110 ensemble model to obtain image-level predictions for the following 10 pathologies:
111 Atelectasis, Cardiomegaly, Consolidation, Edema, Enlarged Cardiomedastinum, Lung
112 Lesion, Lung Opacity, Pleural Effusion, Pneumothorax, and Support Devices. Of the 14
113 observations labeled in the CheXpert dataset: Fracture and Pleural Other were not
114 included in our analysis because they had low prevalence in our test set (fewer than 10
115 examples); Pneumonia was not included because it is a clinical (as opposed to a

116 radiological) diagnosis; and No Finding was not included because it is not applicable to
117 evaluating localization performance. For each CXR, we used the saliency method to
118 generate heat maps, one for each of the 10 pathologies, and then applied a threshold to
119 each heat map to produce binary segmentations (top row, Fig. 1a). Thresholding is
120 determined per pathology using Otsu's method³⁷, which iteratively searches for a
121 threshold value that maximizes inter-class pixel intensity variance. We also conducted a
122 second thresholding scheme in which we iteratively search for a threshold value that
123 maximizes per pathology mIoU on the validation set. Both thresholding schemes reported
124 similar findings (see Extended Data Fig. 1). The result shows that our evaluation of
125 localization performance is robust to different saliency map thresholding schemes.
126 Additionally, to calculate the hit rate evaluation metric (described below), we extracted
127 the pixel in the saliency method heat map with the largest value as the single most
128 representative point on the CXR for that pathology.

129
130 We obtained two independent sets of pixel-level CXR segmentations on the holdout test
131 set: ground-truth segmentations drawn by two board-certified radiologists (middle row,
132 Fig. 1a) and human benchmark segmentations drawn by a separate group of three board-
133 certified radiologists (bottom row, Fig. 1a). The human benchmark segmentations and the
134 saliency method segmentations were compared with the ground-truth segmentations to
135 establish the human benchmark localization performance and the saliency method
136 localization performance, respectively. Additionally, for the hit rate evaluation metric, the
137 radiologists who drew the benchmark segmentations were also asked to locate a single
138 point on the CXR that was most representative of the pathology at hand (see

139 Supplementary Figs. S1 through S11 for detailed instructions given to the radiologists).
140 Note that the human benchmark localization performance demonstrates interrater
141 variability, and we use it as a reference when evaluating saliency method pipelines.

142
143 We used two evaluation metrics to compare segmentations (Fig. 1b). First, we used *mean*
144 *Intersection over Union* (mIoU), a metric that measures how much, on average, either the
145 saliency method or benchmark segmentations overlapped with the ground-truth
146 segmentations. Second, we used *hit rate*, a less strict metric that does not require the
147 saliency method or benchmark annotators to locate the full extent of a pathology. Hit rate
148 is based on the pointing game setup³⁸, in which credit is given if the most representative
149 point identified by the saliency method or the benchmark annotators lies within the
150 ground-truth segmentation. A “hit” indicates that the correct region of the CXR was
151 located regardless of the exact bounds of the binary segmentations. Localization
152 performance is then calculated as the hit rate across the dataset³⁹. In addition, we report
153 the sensitivity and specificity values of the saliency method pipeline and the human
154 benchmark in Extended Data Fig. 2.



155

156 **Fig. 1 | Framework for evaluating saliency methods.** **a**, Top row left: a CXR image
 157 from the holdout test set is passed into an ensemble CNN trained only on CXR images
 158 and their corresponding pathology task labels. Saliency method is used to generate 10
 159 heat maps for the example CXR, one for each task. The pixel in the heat map with the
 160 largest value is determined to be the single most representative point on the CXR for
 161 that pathology. Top row middle: there are three pathologies present in this CXR (Airspace
 162 Opacity, Pleural Effusion, and Support Devices). Top row right: a threshold is applied to
 163 the heat maps to produce binary segmentations for each present pathology. Middle row:
 164 Two board-certified radiologists were asked to segment the pathologies that were
 165 present in the CXR as determined by the dataset's ground-truth labels. Saliency method
 166 annotations are compared to these ground-truth annotations to determine "saliency
 167 method localization performance". Bottom row: Two board-certified radiologists
 168 (separate from those in middle row) were also asked to segment the pathologies that
 169 were present in the CXR as determined by the dataset's ground-truth labels. In addition,

170 these radiologists were asked to locate the single point on the CXR that was most
171 representative of each present pathology. These benchmark annotations are compared
172 to the ground-truth annotations to determine “human benchmark localization
173 performance”. **b**, Left: CXR with ground-truth and saliency method annotations for
174 Pleural Effusion. The segmentations have a low overlap (IoU is 0.078), but pointing game
175 is a “hit” since the saliency method’s most representative point is inside of the ground-
176 truth segmentation. Right, CXR with ground-truth and human benchmark annotations
177 for Enlarged Cardiomeastinum. The segmentations have a high overlap (IoU is 0.682),
178 but pointing game is a “miss” since saliency method’s most representative point is
179 outside of the ground-truth segmentation.

180

181 **Evaluating localization performance**

182 In order to compare the localization performance of the saliency methods with the human
183 benchmark, we first used Grad-CAM, Grad-CAM++, and Integrated Gradients to run
184 eighteen experiments, one for each combination of saliency method (Grad-CAM, Grad-
185 CAM++, or Integrated Gradients) and CNN architecture (DenseNet121, ResNet152, or
186 Inception-v4) using one of the two evaluation metrics (mIoU or hit rate) (see Extended
187 Data Fig. 3). We also ran experiments to evaluate the localization performances of
188 DenseNet121 with Eigen-CAM, DeepLIFT, LRP, and Occlusion. We found that Grad-
189 CAM with DenseNet121 generally demonstrated better localization performance across
190 pathologies and evaluation metrics than the other combinations of saliency method and
191 architecture (see Table 1 for localization performance on the test set of all seven saliency
192 methods using DenseNet121). Accordingly, we compared Grad-CAM with DenseNet121
193 (“saliency method pipeline”) with the human benchmark using both mIoU and hit rate. The
194 localization performance for each pathology is reported on the true positive slice of the
195 dataset (CXR that contain saliency method, human benchmark, and ground-truth
196 segmentations). Localization performance was calculated this way so that saliency
197 methods were not penalized by DNN classification error: while the benchmark radiologists

198 were provided with ground-truth labels when annotating the dataset, saliency method
 199 segmentations were created based on labels predicted by the model. (See Extended Data
 200 Fig. 4 for saliency method pipeline localization performance on the full dataset using
 201 mIoU.)

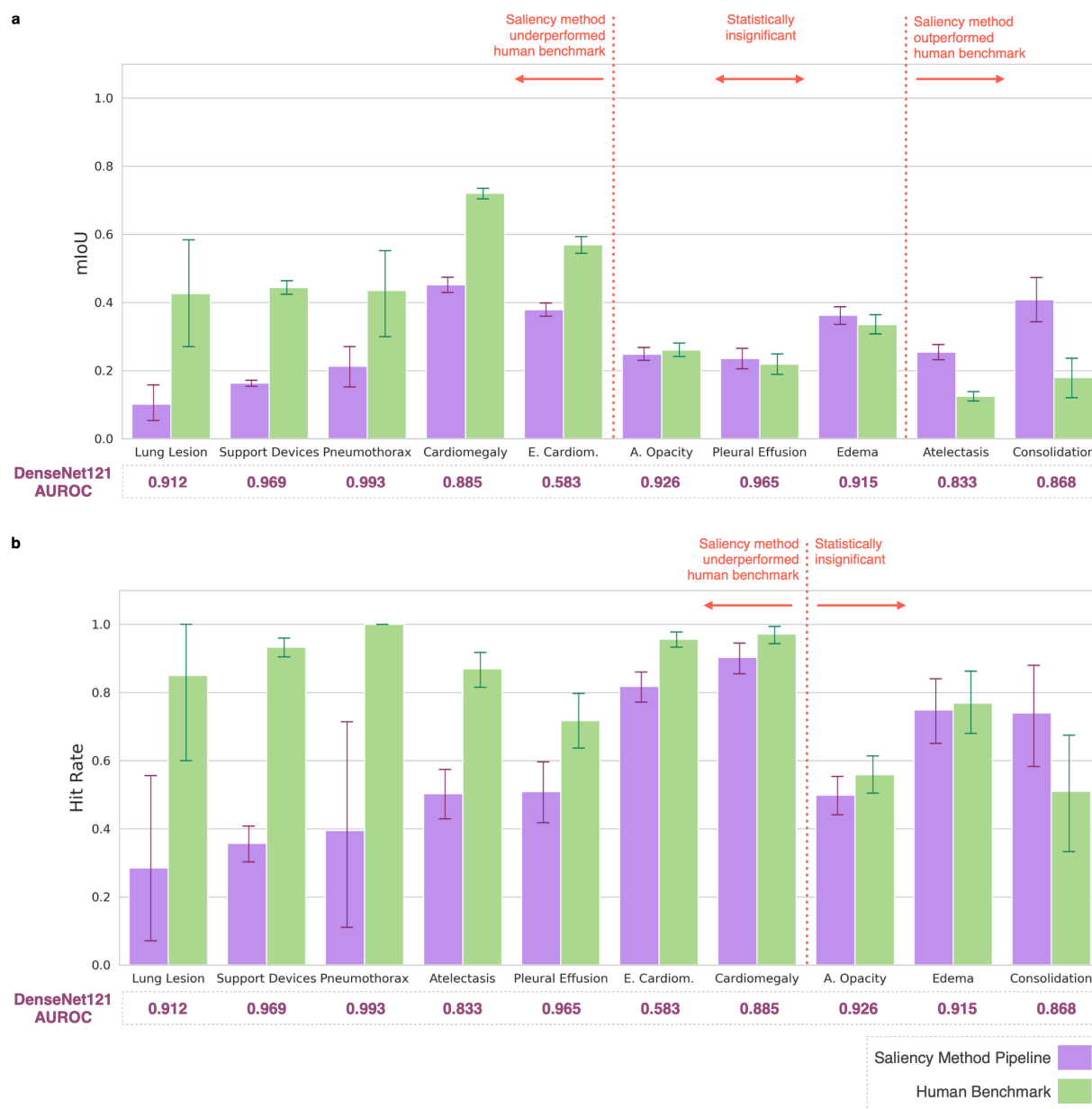
Table 1 | Localization performance of saliency methods using DenseNet121

Pathology	Grad-CAM	Grad-CAM++	Integrated Gradients	Eigen-CAM	DeepLIFT	LRP	Occlusion
mIoU							
Airspace Opacity	0.248	0.234	0.123	0.293	0.111	0.112	0.242
Atelectasis	0.254	0.245	0.116	0.267	0.126	0.109	0.250
Cardiomegaly	0.452	0.346	0.160	0.379	0.167	0.150	0.312
Consolidation	0.408	0.297	0.177	0.332	0.088	0.099	0.212
Edema	0.362	0.388	0.073	0.370	0.059	0.047	0.347
Enlarged Cardiom.	0.379	0.400	0.154	0.372	0.109	0.117	0.363
Lung Lesion	0.101	0.089	0.107	0.089	0.072	0.088	0.087
Pleural Effusion	0.235	0.195	0.088	0.249	0.090	0.082	0.215
Pneumothorax	0.213	0.218	0.077	0.218	0.084	0.066	0.214
Support Devices	0.163	0.133	0.099	0.116	0.086	0.052	0.126
Hit rate							
Airspace Opacity	0.498	0.558	0.606	0.566	0.528	0.566	0.367
Atelectasis	0.501	0.621	0.520	0.530	0.415	0.468	0.343
Cardiomegaly	0.903	0.732	0.697	0.709	0.610	0.644	0.515
Consolidation	0.738	0.708	0.624	0.626	0.571	0.283	0.338
Edema	0.746	0.781	0.300	0.758	0.468	0.156	0.469
Enlarged Cardiom.	0.818	0.630	0.704	0.612	0.469	0.594	0.767
Lung Lesion	0.290	0.290	0.423	0.146	0.497	0.356	0.072
Pleural Effusion	0.507	0.347	0.332	0.439	0.408	0.283	0.291
Pneumothorax	0.392	0.489	0.801	0.195	0.801	0.697	0.297
Support Devices	0.355	0.364	0.491	0.216	0.598	0.264	0.189

202 We found that the saliency method pipeline demonstrated significantly worse localization
 203 performance when compared with the human benchmark using both mIoU (Fig. 2a) and
 204 hit rate (Fig. 2b) as an evaluation metric, regardless of model classification AUROC. For
 205 each metric, we report the 95% confidence intervals using the bootstrap method with
 206 1,000 bootstrap samples⁴⁰. For five of the 10 pathologies, the saliency method pipeline
 207

208 had a significantly lower mIoU than the human benchmark. For example, the saliency
209 method pipeline had one of the highest AUROC scores of the 10 pathologies for Support
210 Devices (0.969), but had among the worst localization performance for Support Devices
211 when using both mIoU (0.163 [95% CI 0.154, 0.172]) and hit rate (0.357 [95% CI 0.303,
212 0.408]) as evaluation metrics. On two pathologies (Atelectasis and Consolidation) the
213 saliency method pipeline significantly outperformed the human benchmark. On average,
214 across all 10 pathologies, mIoU saliency method pipeline performance was 26.6% [95%
215 CI 18.1%, 35.0%] worse than the human benchmark, with Lung Lesion displaying the
216 largest gap in performance (76.2% [95% CI 59.1%, 87.5%] worse than the human
217 benchmark) (Extended Data Fig. 5). Consolidation was the pathology on which the mIoU
218 saliency method pipeline performance exceeded the human benchmark the most, by
219 56.1% [95% CI 42.7%, 69.4%]. For seven of the 10 pathologies, the saliency method
220 pipeline had a significantly lower hit rate than the human benchmark. On average, hit rate
221 saliency method pipeline performance was 29.4% [95% CI 15.0%, 43.2%] worse than the
222 human benchmark (Extended Data Fig. 6), with Lung Lesion again displaying the largest
223 gap in performance (65.9% [95% CI 35.3%, 91.7%] worse than the human benchmark).
224 The hit rate saliency method pipeline did not significantly outperform the human
225 benchmark on any of the 10 pathologies; for the remaining three of the 10 pathologies,
226 the hit rate performance differences between the saliency method pipeline and the human
227 benchmark were not statistically significant. Therefore, while the saliency method pipeline
228 significantly underperformed the human benchmark regardless of evaluation metric used,
229 the average performance gap was larger when using hit rate as an evaluation metric than
230 when using mIoU as an evaluation metric.

231
 232 We compared saliency method pipeline localization performance using an ensemble
 233 model to localization performance using the top performing single checkpoint for each
 234 pathology. We found that the single model has worse localization performance than the
 235 ensemble model for all pathologies when using mIoU and for six of the 10 pathologies
 236 when using hit rate (see Extended Data Fig. 7).



238 **Fig. 2 | Evaluating localization performance. a**, Comparing saliency method pipeline
239 and human benchmark localization performances under the overlap evaluation scheme
240 (mIoU). **b**, Comparing saliency method pipeline and human benchmark localization
241 performances under the hit rate evaluation scheme. For both **a** and **b**, pathologies, along
242 with their DenseNet121 AUROCs, are sorted on the x-axis first by statistical significance
243 of percentage decrease from human benchmark mIoU/hit rate to saliency method
244 pipeline mIoU/hit rate (high to low), and then by percentage decrease from human
245 benchmark mIoU/hit rate to saliency method pipeline mIoU/hit rate (high to low).
246

247 **Characterizing underperformance of saliency method pipeline**

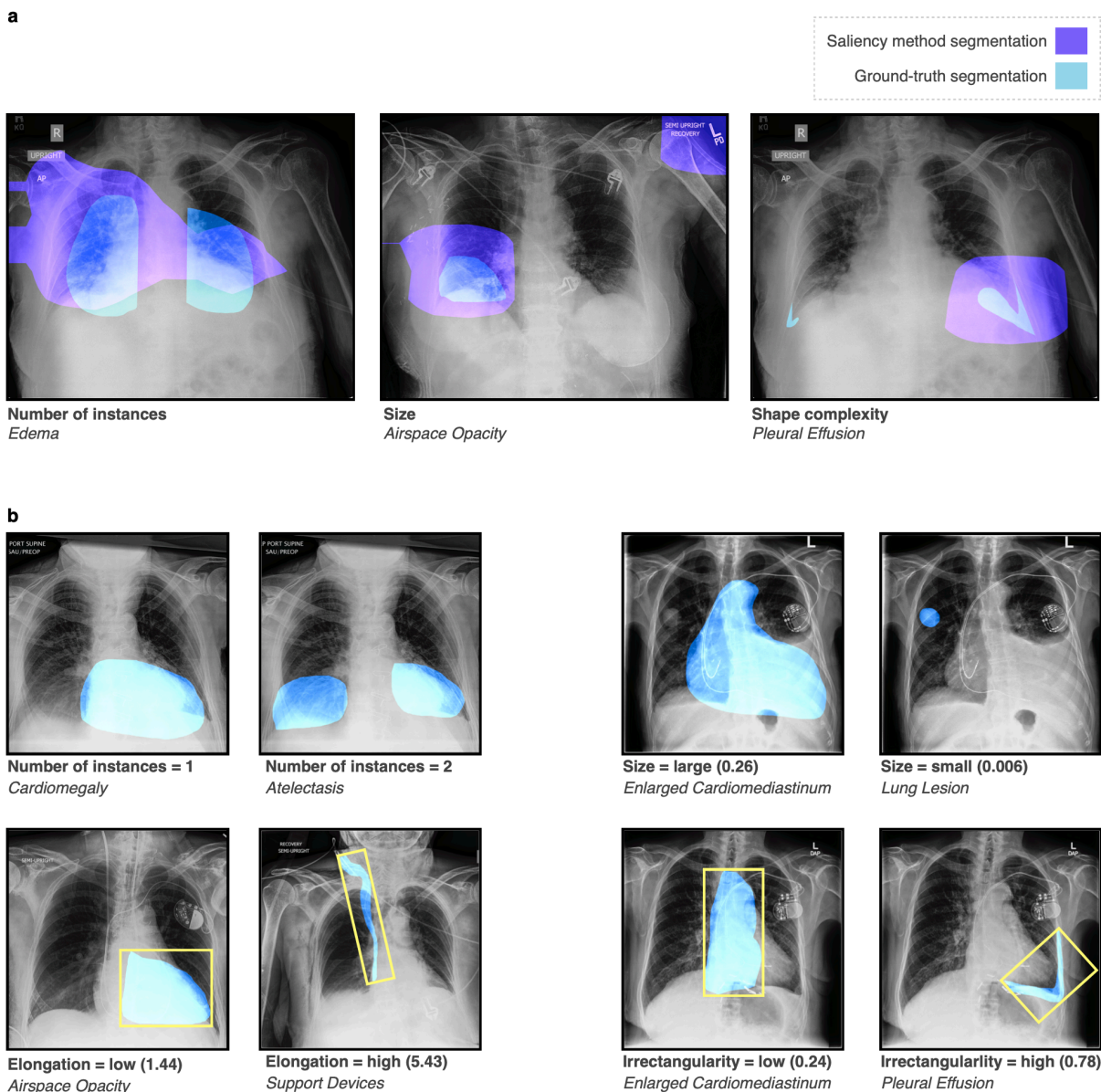
248 In order to better understand the underperformance of the saliency method pipeline
249 localization, we first conducted a qualitative analysis with a radiologist by visually
250 inspecting both the segmentations produced by the saliency method pipeline (Grad-CAM
251 with DenseNet121) and the human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

263

264 Informed by our qualitative analysis and previous work in histology⁴⁵, we defined four
265 geometric features for our quantitative analysis (Fig. 3b): (1) *number of instances* (for
266 example, bilateral Pleural Effusion would have two instances, whereas there is only one
267 instance for Cardiomegaly), (2) *size* (pathology area with respect to the area of the whole
268 CXR), (3) *elongation* and (4) *irrectangularity* (the last two features measure the complexity
269 of the pathology shape and were calculated by fitting a rectangle of minimum area
270 enclosing the binary mask). See Extended Data Fig. 8 for the distribution of the four
271 pathological characteristics across all 10 pathologies.

272
273 For each evaluation metric, we ran 8 simple linear regressions: four with the evaluation
274 metric (IoU or hit rate) of the saliency method pipeline (Grad-CAM with DenseNet121) as
275 the dependent variable (to understand the relationship between the geometric features of
276 a pathology and saliency method localization performance), and four with the difference
277 between the evaluation metrics of the saliency method pipeline and the human
278 benchmark as the dependent variable (to understand the relationship between the
279 geometric features of a pathology and the gap in localization performance between the
280 saliency method pipeline and the human benchmark). Each regression used one of the
281 four geometric features as a single independent variable, and only the true positive slice
282 was included in each regression. Each feature was normalized using z-score
283 normalization and the regression coefficient can be interpreted as the effect of that
284 geometric feature on the evaluation metric at hand. See Table 2 for coefficients from the
285 regressions using both evaluation metrics, where we also report the 95% confidence

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



288

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

297 maxAxis/minAxis. Irrectangularity = $1 - (\text{area of segmentation}/\text{area of enclosing}$
298 rectangle).
299

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

312
313 We found that these same three pathological characteristics—larger size, and higher
314 elongation and irrectangularity—characterized the *gap* in mIoU localization performance
315 between saliency method and human benchmark. We observed that the *gap* in hit rate
316 localization performance was significantly characterized by all four geometric features
317 (number of instances, size, elongation, and irrectangularity). As the number of instances
318 increased, despite no significant change in hit rate localization performance itself, the *gap*
319 in hit rate localization performance between saliency method and the human benchmark

320 increased (0.470 [95% CI 0.114, 0.825]). This suggests that the saliency method performs
321 especially poorly in the face of a multi-instance diagnosis.

Table 2 | 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	-0.244 (-0.346, -0.051)	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

322
323 **Effect of model confidence on localization performance**

324 We also conducted statistical analyses to determine whether there was any correlation
325 between the model's confidence in its prediction and saliency method pipeline
326 performance (Table 3). We first ran a simple regression for each pathology using the
327 model's probability output as the single independent variable and using the saliency
328 method IoU as the dependent variable. We then performed a simple regression that uses
329 the same approach as above, but that includes all 10 pathologies. For each of the 11
330 regressions, we used the full dataset since the analysis of false positives and false
331 negatives was also of interest. In addition to the linear regression coefficients, we also

332 computed the Spearman correlation coefficients to capture any potential non-linear
333 associations.

334

335 We found that for all pathologies, model confidence was positively correlated with mIoU
336 saliency method pipeline performance. The p-values for all coefficients were below 0.001
337 except for the coefficients for Pneumothorax (n=11) and Lung Lesion (n=50), the two
338 pathologies for which we had the fewest positive examples. Of all the pathologies, model
339 confidence for positive predictions of Enlarged Cardiomeastinum had the largest linear
340 regression coefficient with mIoU saliency method pipeline performance (1.974, p-
341 value<0.001). Model confidence for positive predictions of Pneumothorax had the largest
342 Spearman correlation coefficient with mIoU saliency method pipeline performance (0.734,
343 p-value<0.01), followed by Pleural Effusion (0.690, p-value<0.001). Combining all
344 pathologies (n=2365), the linear regression coefficient was 0.109 (95% CI [0.083, 0.135]),
345 and the Spearman correlation coefficient was 0.285 (95% CI [0.239, 0.331]). We also
346 performed analogous experiments using hit rate as the dependent variable and found
347 comparable results (Extended Data Fig. 9).

Table 3 | 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, 0.610) ***
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 (0.470, 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, 0.250) ***	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

348

349 Discussion

350 The purpose of this work was to evaluate the performance of some of the most used
351 saliency methods for deep learning explainability using a variety of model architectures.
352 We establish the first human benchmark for CXR segmentation in a multilabel
353 classification setup and demonstrate that saliency maps are consistently worse than
354 expert radiologists regardless of model classification AUROC. We use qualitative and
355 quantitative analyses to establish that saliency method localization performance is most
356 inferior to expert localization performance when a pathology has multiple instances, is
357 smaller in size, or has shapes that are more complex, suggesting that deep learning
358 explainability as a clinical interface may be less reliable and less useful when used for
359 pathologies with those characteristics. We also show that model assurance is positively
360 correlated with saliency method localization performance, which could indicate that

361 saliency methods are safer to use as a decision aid to clinicians when the model has
362 made a positive prediction with high confidence.

363
364 Because ground-truth segmentations for medical imaging are time-consuming and
365 expensive to obtain, the current norm in medical imaging—both in research and in
366 industry—is to use classification models on which saliency methods are applied post-hoc
367 for localization, highlighting the need for investigations into the reliability of these methods
368 in clinical settings^{46,47}. There are public CXR datasets containing image-level labels
369 annotated by expert radiologists (e.g., the CheXpert validation set), multilabel bounding
370 box annotations (e.g., ChestX-ray8⁴⁸ and VinDr-CXR⁴⁹), and segmentations for a single
371 pathology (e.g., SIIM-ACR Pneumothorax Segmentation⁵⁰). To our knowledge, however,
372 there are no other publicly available CXR datasets with multilabel pixel-level expert
373 segmentations. By publicly releasing a development dataset, CheXlocalize, of 234
374 images with 885 expert segmentations, and a competition with a test set of 668 images,
375 we hope to encourage the further development of saliency methods and other
376 explainability techniques for medical imaging.

377
378 Our work has several potential implications for human-AI collaboration in the context of
379 medical decision-making. Heat maps generated using saliency methods are advocated
380 as clinical decision support in the hope that they not only improve clinical decision-
381 making, but also encourage clinicians to trust model predictions⁵¹⁻⁵³. Many of the large
382 CXR vendors⁵⁴⁻⁵⁶ use localization methods to provide pathology visualization in their
383 computer-aided detection (CAD) products. In addition to being used for clinical

384 interpretation, saliency method heat maps are also used for the evaluation of CXR
385 interpretation models, for quality improvement (QI) and quality assurance (QA) in clinical
386 practices, and for dataset annotation⁵⁷. Explainable AI is critical in high-stakes contexts
387 such as health care, and saliency methods have been used successfully to develop and
388 understand models generally. Indeed, we found that the saliency method pipeline
389 significantly outperformed the human benchmark on two pathologies when using mIoU
390 as an evaluation metric. However, our work also suggests that saliency methods are not
391 yet reliable enough to validate individual clinical decisions made by a model. We found
392 that saliency method localization performance, on balance, performed worse than expert
393 localization across multiple analyses and across many important pathologies (our findings
394 are consistent with recent work focused on localizing a single pathology, Pneumothorax,
395 in CXRs⁵⁸). We hypothesize that this could be an algorithmic artifact of saliency methods,
396 whose relatively small heat maps (14x14 for Grad-CAM) are interpolated to the original
397 image dimensions (usually 2000x2000), resulting in coarse resolutions. If used in clinical
398 practice, heat maps that incorrectly highlight medical images may exacerbate well
399 documented biases (chiefly, automation bias) and erode trust in model predictions (even
400 when model output is correct), limiting clinical translation²².

401
402 Since IoU computes the overlap of two segmentations but pointing game hit rate better
403 captures diagnostic attention, we suggest using both metrics when evaluating localization
404 performance in the context of medical imaging. While IoU is a commonly used metric for
405 evaluating semantic segmentation outputs, there are inherent limitations to the metric in
406 the pathological context. This is indicated by our finding that even the human benchmark

407 segmentations had low overlap with the ground truth segmentations (the highest expert
408 mIoU was 0.720 for Cardiomegaly). One potential explanation for this consistent
409 underperformance is that pathologies can be hard to distinguish, especially without
410 clinical context. Furthermore, whereas many people might agree on how to segment, say,
411 a cat or a stop sign in traditional computer vision tasks, radiologists use a certain amount
412 of clinical discretion when defining the boundaries of a pathology on a CXR. There can
413 also be institutional and geographic differences in how radiologists are taught to
414 recognize pathologies, and studies have shown that there can be high interobserver
415 variability in the interpretation of CXRs^{59–61}. We sought to address this with the hit rate
416 evaluation metric, which highlights when two radiologists share the same diagnostic
417 intention, even if it is less exact than IoU in comparing segmentations directly. The human
418 benchmark localization using hit rate was above 0.9 for four pathologies (Pneumothorax,
419 Cardiomegaly, Support Devices, and Enlarged Cardiomediatinum); these are
420 pathologies for which there is often little disagreement between radiologists about where
421 the pathologies are located, even if the expert segmentations are noisy. Further work is
422 needed to demonstrate which segmentation evaluation metrics, even beyond overlap and
423 hit rate, are more appropriate for certain pathologies and downstream tasks when
424 evaluating saliency methods for the clinical setting.

425
426 Our work builds upon several studies investigating the validity of saliency maps for
427 localization^{62,63,64} and upon some early work on the trustworthiness of saliency methods
428 to explain DNNs in medical imaging⁴⁷. However, as recent work has shown³², evaluating
429 saliency methods is inherently difficult given that they are post-hoc techniques. To

430 illustrate this, consider the following models and saliency methods as described by some
431 oracle: (1) a model M_{bad} that has perfect AUROC for a given image classification task,
432 but that we know does *not* localize well (i.e. because the model picks up on confounders
433 in the image); (2) a model M_{good} that also has perfect AUROC, but that we know *does*
434 localize well (i.e. is looking at relevant regions of the image); (3) a saliency method S_{bad}
435 that does *not* properly reflect the model's attention; and (4) a saliency method S_{good}
436 that *does* properly reflect the model's attention. Let us say that we are evaluating the
437 following pipeline: we first classify an image and we then apply a saliency method post
438 hoc. Imagine that our evaluation reveals poor localization performance as measured by
439 mIoU or hit rate (as was the case in our findings). There are three possible pipelines
440 (combinations of model and saliency method) that would lead to this scenario: (1) M_{bad}
441 + S_{good} ; (2) M_{good} + S_{bad} ; and (3) M_{bad} + S_{bad} . The first scenario (M_{bad} +
442 S_{good}) is the one for which saliency methods were originally intended: we have a
443 working saliency method that properly alerts us to models picking up on confounders. The
444 second scenario (M_{good} + S_{bad}) is our nightmare scenario: we have a working model
445 whose attention is appropriately directed, but we reject it based on a poorly localizing
446 saliency method. Because all three scenarios result in poor localization performance, it is
447 difficult—if not impossible—to know whether poor localization performance is attributable
448 to the model or to the saliency method (or to both). While we cannot say whether models
449 or saliency methods are failing in the context of medical imaging, we can say that we
450 should not rely on saliency methods to evaluate model localization. Future work should
451 explore potential techniques for localization performance attribution.
452

453 There are several limitations of our work. First, we did not investigate the impact of
454 pathology prevalence in the training data on saliency method localization performance.
455 Second, some pathologies, such as effusions and cardiomegaly, are in similar locations
456 across frontal view CXRs, while others, such as lesions and opacities, can vary in
457 locations across CXRs. Future work could investigate how the location of pathologies on
458 a CXR in the training/test data distribution, and the consistency of those locations, affect
459 saliency method localization performance. Third, while we compared saliency method-
460 generated pixel-level segmentations to human expert pixel-level segmentations, future
461 work might explore how saliency method localization performance changes when
462 comparing bounding-box annotations, instead of pixel-level segmentations. Fourth, we
463 explored post-hoc interpretability methods given their prevalence in the context of medical
464 imaging, but we hope that by publicly releasing our development dataset of pixel-level
465 expert segmentations we can facilitate the development of models that make use of
466 ground-truth segmentations during training⁵⁷. Fifth, the lack of a given finding can in
467 certain cases inform clinical diagnoses. A common example of this is the lack of normal
468 lung tissue pattern towards the edges of the thoracic cage, which is used to detect
469 pneumothorax. For any characteristic pattern, both the absence and the presence provide
470 diagnostic information to the radiologist. For example, the absence of a pleural effusion
471 pattern is also used to rule out pleural effusion. For any characteristic radiological pattern,
472 both the presence as well as the absence contributes to the final radiology report. Future
473 work can explore counterfactual visual explanations that are similar to the counterfactual
474 diagnostic process of a radiologist. Sixth, future work should further explore the potentially
475 confounding effect of model calibration on the evaluation of saliency methods, especially

476 when using segmentation, as opposed to classification, models. Finally, the impact of
477 saliency methods on the trust and efficacy of users is underexplored.

478

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

484

485 **Methods**

486 **Ethical and information governance approvals.**

487 A formal Stanford IRB review was conducted for the original collection of the CheXpert
488 dataset. The IRB waived the requirement to obtain informed consent as the data were
489 retrospectively collected and fully anonymized.

490

491 **Dataset and clinical taxonomy.** *Dataset description.* The localization experiments were
492 performed using CheXpert, a large public dataset for chest X-ray interpretation. The
493 CheXpert dataset contains 224,316 chest X-rays for 65,240 patients labeled for the
494 presence of 14 observations (13 pathologies and an observation of “No Finding”) as
495 positive, negative, or uncertain. The CheXpert validation set consists of 234 chest X-rays
496 from 200 patients randomly sampled from the full dataset and was labeled according to
497 the consensus of three board-certified radiologists. The test set consists of 668 chest X-
498 rays from 500 patients not included in the training or validation sets and was labeled

499 according to the consensus of five board-certified radiologists. See Extended Data Fig.
500 10 for test set summary statistics.

501
502 *Ground-truth segmentation.* The chest X-rays in our validation set and test set were
503 manually segmented by two board-certified radiologists with 18 and 27 years of
504 experience, using the annotation software tool MD.ai⁶⁵ (see Supplementary Figs. S12
505 through S14). The radiologists were asked to contour the region of interest for all
506 observations in the chest X-rays for which there was a positive ground truth label in the
507 CheXpert dataset. For a pathology with multiple instances, all the instances were
508 contoured. For Support Devices, radiologists were asked to contour any implanted or
509 invasive devices including pacemakers, PICC/central catheters, chest tubes,
510 endotracheal tubes, feeding tubes and stents and ignore ECG lead wires or external
511 stickers visible in the chest X-ray.

512
513 *Evaluating the expert performance using benchmark segmentation.* To evaluate the
514 expert performance on the test set using the IoU evaluation method, three radiologists,
515 certified in Vietnam with 9, 10, and 18 years of experience, were asked to segment the
516 regions of interest for all observations in the chest X-rays for which there was a positive
517 ground truth label in the CheXpert dataset. These radiologists were also provided the
518 same instructions for contouring as were provided to the radiologists drawing the
519 reference segmentations. To extract the “maximally activated” point from the benchmark
520 segmentations, we asked the same radiologists to locate each pathology present on each
521 CXR using only a single most representative point for that pathology on the CXR (see

522 Supplementary Figs. S1 through S11 for the detailed instructions given to the
523 radiologists). There was no overlap between these three radiologists and the two who
524 drew the reference segmentations.

525

526 **Classification network architecture and training protocol.** *Multi-label classification*
527 *model.* The model takes as input a single-view chest X-ray and outputs the probability for
528 each of the 14 observations. In case of availability of more than one view, the models
529 output the maximum probability of the observations across the views. Each chest X-ray
530 was resized to 320×320 pixels and normalized before it was fed into the network. We
531 used the same image resolutions as CheXpert³⁶ and CheXnet², which demonstrated
532 radiologist-level performance on external test sets with 320x320 images. There are
533 models that are commercially deployed and have similar dimensions. For example, the
534 architecture used by a medical AI software vendor Annalise.ai⁶⁶ is based on
535 EfficientNet⁶⁷, which takes input of 224x224. Chest X-rays were normalized prior to being
536 fed into the network by subtracting the mean of all images in the CheXpert training set
537 and then dividing by the standard deviation of all images in the CheXpert training set. The
538 model architectures (DenseNet121, ResNet152, and Inception-v4) were used. Cross-
539 entropy loss was used to train the model. The Adam optimizer⁶⁸ was used with default β -
540 parameters of $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate was hyperparameter tuned for
541 the different model architectures. Grid search was used to tune the learning rates. We
542 searched over learning rates of 1e-3, 1e-4, and 1e-5. The best learning rate for each
543 architecture was: 1×10^{-4} for DenseNet121, 1×10^{-5} for ResNet152, 1×10^{-5} for
544 Inceptionv4. Batches were sampled using a fixed batch size of 16 images.

545
546 *Ensembling.* We use an ensemble of checkpoints to create both predictions and saliency
547 maps to maximize model performance. In order to capture uncertainties inherent in
548 radiograph interpretation, we train our models using four uncertainty handling strategies
549 outlined in CheXpert: Ignoring, Zeroes, Ones, and 3-Class Classification. For each of the
550 four uncertainty handling strategies, we train our model three separate times, each time
551 saving the 10 checkpoints across the three epochs with the highest average AUC across
552 5 observations selected for their clinical importance and prevalence in the validation set:
553 Atelectasis, Cardiomegaly, Consolidation, Edema, and Pleural Effusion. In total, after
554 training, we have saved $4 \times 30 = 120$ checkpoints for a given model. Then, from the 120
555 saved checkpoints for that model, we select the top 10 performing checkpoints for each
556 pathology. For each CXR and each task, we compute the predictions and saliency maps
557 using the relevant checkpoints. We then take the mean both of the predictions and of the
558 saliency maps to create the final set of predictions and saliency maps for the ensemble
559 model. See Supplementary Table S1 for the performance of each model architecture
560 (DenseNet121, ResNet152, and Inception-v4) on each of the pathologies.

561
562 **CNN interpretation strategy.** Saliency methods were used to visualize the decision
563 made by the classification network. The saliency map was resized to the original image
564 dimension using bilinear interpolation. It was then normalized using max-min
565 normalization and then converted into a binary segmentation using binary thresholding
566 (Otsu's method). We also reported mIoU localization performance using different saliency
567 map thresholding values. We first applied max-min normalizations to the saliency maps

568 so that each value gets transformed into a decimal between 0 and 1. We then passed in
569 a range of threshold values from 0.2 to 0.8 to create binary segmentations and calculated
570 the mIoU score per pathology under each threshold on the validation set. Then for the
571 analysis on the full dataset (see Extended Data Fig. 4), we further ensure that the final
572 binary segmentation is consistent with model probability output by applying another layer
573 of thresholding such that the segmentation mask produced all zeros if the predicted
574 probability was below a chosen level. The probability threshold is searched on the interval
575 of [0,0.8] with steps of 0.1. The exact value is determined per pathology by maximizing
576 the mIoU on validation set.

577

578 For Occlusion, we used a window size of 40 and a stride of 40 for each CXR.

579

580 *Segmentation evaluation metrics.* Localization performance of each segmentation was
581 evaluated using Intersection over Union (IoU) score. The IoU is the ratio between the
582 area of overlap and the area of union between the ground truth and the predicted areas,
583 ranging from 0 to 1 with 0 signifying no overlap and 1 signifying perfectly overlapping
584 segmentation. Confidence intervals are calculated using bootstrapping with 1000
585 bootstrap samples. The variance in the width of CI across pathologies can be explained
586 by difference in sample sizes. For the percentage decrease from expert mIoU to AI mIoU,
587 we bootstrapped the difference between human benchmark and saliency method
588 localization and created the 95% confidence intervals. The confidence intervals for hit
589 rates were calculated in the same fashion. For the evaluation of Integrated Gradients
590 using IoU, we applied box filtering of kernel size 100 to smooth the pixelated map. For

591 DeepLIFT, we applied box filtering of kernel size 50. For LRP, we used a kernel size of
592 80. The kernel sizes are tuned on the validation set. The noisy map is not a concern for
593 hit rate because a single max pixel is extracted for the entire image.

594

595 **Statistical analysis.**

596 *Pathology Characteristics.* We used four features to characterize the pathologies. (1)
597 Number of instances is defined as the number of disjoint components in the
598 segmentation. (2) Size is the area of the pathology divided by the total image area. (3)
599 and (4) Elongation and irrectangularity are geometric features that measure shape
600 complexities. They were designed to quantify what radiologists qualitatively described as
601 focal or diffused. To calculate the metrics, a rectangle of minimum area enclosing the
602 contour is fitted to each pathology. Elongation is defined as the ratio of the rectangle's
603 longer side to short side. Irrectangularity = $1 - (\text{area of segmentation}/\text{area of enclosing}$
604 $\text{rectangle})$, with values ranging from 0 to 1 with 1 being very irrectangular. When there
605 are multiple instances within one pathology, we used the characteristics of the dominant
606 instance (largest in perimeter).

607

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

611

612 *Linear Regression.* For each evaluation scheme (overlap and hit rate), we ran two groups
613 of simple linear regressions, with AI evaluation metrics and their differences as the

614 response variables. Each group has four regressions using the above four pathological
615 characteristics as the regressions' single attribute, respectively, and only the true positive
616 slice was included in each regression. All features are normalized using min-max
617 normalization so that they are comparable on scales of magnitudes. We report the 95%
618 confidence interval and Bonferroni adjusted p-value of the regression coefficients.

619

620 **Data Availability**

621 The CheXlocalize dataset is available here:

622 <https://stanfordaimi.azurewebsites.net/datasets/abfb76e5-70d5-4315-badc->

623 [c94dd82e3d6d](https://stanfordaimi.azurewebsites.net/datasets/abfb76e5-70d5-4315-badc-c94dd82e3d6d). The CheXpert dataset is available here

624 <https://stanfordmlgroup.github.io/competitions/chexpert/>.

625

626 **Code Availability**

627 The code used to generate segmentations from saliency method heat maps, fine-tune
628 segmentation thresholds, generate segmentations from human annotations, and evaluate
629 localization performance is available in the following public repository under the MIT
630 License: <https://github.com/rajpurkarlab/cheXlocalize>. The version used for this
631 publication is available at <https://doi.org/10.5281/zenodo.6816288>⁶⁹.

632

633 **Acknowledgements**

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

637

638 **Author Contributions**

639 Conceptualization: P.R. and A.P. Design: P.R., A.P., A.S., X.G. and A.A. Data analysis
640 and interpretation: A.S., X.G., A.A., P.R., A.P., S.T., C.N., V.N., J.S., and F.B. Drafting of
641 the manuscript: A.S., X.G., A.A., and P.R. Critical revision of the manuscript for important
642 intellectual content: A.P, S.T., C.N., V.N., J.S., F.B, A.N., and M.L. Supervision: A.N.,
643 M.L., and P.R. Research was primarily performed while A.S. was at Stanford University.
644 M.L. and P.R. contributed equally.

645

646 **Competing Interests**

647 M.L. is an advisor for and/or has research funded by GE, Philips, Carestream, Nines
648 Radiology, Segmed, Centaur Labs, Microsoft, BunkerHill, and Amazon Web Services
649 (none of the funded research was relevant to this project). A.P. is a medical associate at
650 Cerebriu. The remaining authors declare no competing interests.

651

652 **References**

- 653 1. Rajpurkar, P. *et al.* Deep learning for chest radiograph diagnosis: A retrospective
654 comparison of the CheXNeXt algorithm to practicing radiologists. *PLOS Med.* **15**,
655 e1002686 (2018).
- 656 2. Rajpurkar, P. *et al.* CheXNet: Radiologist-Level Pneumonia Detection on Chest X-
657 Rays with Deep Learning. *ArXiv171105225 Cs Stat* (2017).

- 658 3. Bien, N. *et al.* Deep-learning-assisted diagnosis for knee magnetic resonance
659 imaging: Development and retrospective validation of MRNet. *PLOS Med.* **15**,
660 e1002699 (2018).
- 661 4. Baselli, G., Codari, M. & Sardanelli, F. Opening the black box of machine learning in
662 radiology: can the proximity of annotated cases be a way? *Eur. Radiol. Exp.* **4**, 30
663 (2020).
- 664 5. Litjens, G. *et al.* A survey on deep learning in medical image analysis. *Med. Image*
665 *Anal.* **42**, 60–88 (2017).
- 666 6. Wang, F., Kaushal, R. & Khullar, D. Should Health Care Demand Interpretable
667 Artificial Intelligence or Accept “Black Box” Medicine? *Ann. Intern. Med.* **172**, 59–60
668 (2019).
- 669 7. Goodman, B. & Flaxman, S. European Union regulations on algorithmic decision-
670 making and a ‘right to explanation’. *AI Mag.* **38**, 50–57 (2017).
- 671 8. Venugopal, V. K., Takhar, R., Gupta, S., Saboo, A. & Mahajan, V. *Clinical*
672 *Explainability Failure (CEF) & Explainability Failure Ratio (EFR) – changing the way*
673 *we validate classification algorithms?* *J Med Syst* **46**, 20 (2022).
- 674 9. Pasa, F., Golkov, V., Pfeiffer, F., Cremers, D. & Pfeiffer, D. Efficient Deep Network
675 Architectures for Fast Chest X-Ray Tuberculosis Screening and Visualization. *Sci.*
676 *Rep.* **9**, 6268 (2019).
- 677 10. Simonyan, K., Vedaldi, A. & Zisserman, A. Deep Inside Convolutional Networks:
678 Visualising Image Classification Models and Saliency Maps. *ArXiv13126034 Cs*
679 (2014).

- 680 11. Aggarwal, M. *et al.* Towards Trainable Saliency Maps in Medical Imaging.
681 *ArXiv201107482 Cs Eess* (2020).
- 682 12. Tjoa, E. & Guan, C. Quantifying Explainability of Saliency Methods in Deep
683 Neural Networks. *ArXiv200902899 Cs* (2020).
- 684 13. Badgeley, M. A. *et al.* Deep learning predicts hip fracture using confounding
685 patient and healthcare variables. *Npj Digit. Med.* **2**, 31 (2019).
- 686 14. Zech, J. R. *et al.* Variable generalization performance of a deep learning model
687 to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Med.* **15**,
688 e1002683 (2018).
- 689 15. DeGrave, A. J., Janizek, J. D. & Lee, S.-I. AI for radiographic COVID-19
690 detection selects shortcuts over signal. *Nat Mach Intell* **3**, 610-619 (2021).
- 691 16. Makimoto, H. *et al.* Performance of a convolutional neural network derived from
692 an ECG database in recognizing myocardial infarction. *Sci. Rep.* **10**, 8445 (2020).
- 693 17. Porumb, M., Stranges, S., Pescapè, A. & Pecchia, L. Precision Medicine and
694 Artificial Intelligence: A Pilot Study on Deep Learning for Hypoglycemic Events
695 Detection based on ECG. *Sci. Rep.* **10**, 1–16 (2020).
- 696 18. Tham, Y.-C. *et al.* Referral for disease-related visual impairment using retinal
697 photograph-based deep learning: a proof-of-concept, model development study.
698 *Lancet Digit. Health* **3**, e29–e40 (2021).
- 699 19. Varadarajan, A. V. *et al.* Deep Learning for Predicting Refractive Error From
700 Retinal Fundus Images. *Invest. Ophthalmol. Vis. Sci.* **59**, 2861–2868 (2018).
- 701 20. Mitani, A. *et al.* Detection of anaemia from retinal fundus images via deep
702 learning. *Nat. Biomed. Eng.* **4**, 18–27 (2020).

- 703 21. Deep Learning to Assess Long-term Mortality From Chest Radiographs |
704 Pulmonary Medicine | JAMA Network Open | JAMA Network. [https://jamanetwork-](https://jamanetwork-com.stanford.idm.oclc.org/journals/jamanetworkopen/fullarticle/2738349)
705 [com.stanford.idm.oclc.org/journals/jamanetworkopen/fullarticle/2738349](https://jamanetwork-com.stanford.idm.oclc.org/journals/jamanetworkopen/fullarticle/2738349).
- 706 22. Rajpurkar, P. *et al.* CheXaid: deep learning assistance for physician diagnosis of
707 tuberculosis using chest x-rays in patients with HIV. *Npj Digit. Med.* **3**, 1–8 (2020).
- 708 23. Rajpurkar, P. *et al.* AppendiXNet: Deep Learning for Diagnosis of Appendicitis
709 from A Small Dataset of CT Exams Using Video Pretraining. *Sci. Rep.* **10**, 3958
710 (2020).
- 711 24. Huang, S.-C. *et al.* PENet—a scalable deep-learning model for automated
712 diagnosis of pulmonary embolism using volumetric CT imaging. *Npj Digit. Med.* **3**, 1–
713 9 (2020).
- 714 25. Rudin, C. Stop explaining black box machine learning models for high stakes
715 decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215 (2019).
- 716 26. Eitel, F. & Ritter, K. Testing the Robustness of Attribution Methods for
717 Convolutional Neural Networks in MRI-Based Alzheimer’s Disease Classification.
718 *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal*
719 *Learning for Clinical Decision Support* (eds. Suzuki, K. *et al.*) 3-11 (Springer
720 International Publishing, 2019). doi:10.1007/978-3-030-33850-3_1.
- 721 27. Young, K., Booth, G., Simpson, B., Dutton, R. & Shrapnel, S. Deep Neural
722 Network or Dermatologist? in *Interpretability of Machine Intelligence in Medical Image*
723 *Computing and Multimodal Learning for Clinical Decision Support* (eds. Suzuki, K. *et*
724 *al.*) 48–55 (Springer International Publishing, 2019). doi:10.1007/978-3-030-33850-
725 3_6.

- 726 28. Ghassemi, M., Oakden-Rayner, L. & Beam, A. L. The false hope of current
727 approaches to explainable artificial intelligence in health care. *Lancet Digit. Health* 3,
728 e745–e750 (2021).
- 729 29. Reyes, M. *et al.* On the Interpretability of Artificial Intelligence in Radiology:
730 Challenges and Opportunities. *Radiol. Artif. Intell.* **2**, e190043 (2020).
- 731 30. Selvaraju, R. R. *et al.* Grad-CAM: Visual Explanations from Deep Networks via
732 Gradient-based Localization. *Int. J. Comput. Vis.* **128**, 336–359 (2020).
- 733 31. Chattopadhyay, A., Sarkar, A., Howlader, P. & Balasubramanian, V. N. Grad-
734 CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional
735 Networks. in *2018 IEEE Winter Conference on Applications of Computer Vision*
736 (WACV) 839–847 (2018). doi:10.1109/WACV.2018.00097.
- 737 32. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic attribution for deep networks. in
738 *Proceedings of the 34th International Conference on Machine Learning - Volume 70*
739 3319–3328 (JMLR.org, 2017).
- 740 33. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely Connected
741 Convolutional Networks. in *2017 IEEE Conference on Computer Vision and Pattern*
742 *Recognition (CVPR)* 2261–2269 (IEEE, 2017). doi:10.1109/CVPR.2017.243.
- 743 34. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image
744 Recognition. in *2016 IEEE Conference on Computer Vision and Pattern Recognition*
745 (CVPR) 770–778 (IEEE, 2016). doi:10.1109/CVPR.2016.90.
- 746 35. Szegedy, C. *et al.* Going deeper with convolutions. in *2015 IEEE Conference on*
747 *Computer Vision and Pattern Recognition (CVPR)* 1–9 (2015).
748 doi:10.1109/CVPR.2015.7298594.

- 749 36. Irvin, J. *et al.* CheXpert: A Large Chest Radiograph Dataset with Uncertainty
750 Labels and Expert Comparison. *Proc. AAAI Conf. Artif. Intell.* **33**, 590–597 (2019).
- 751 37. Otsu, N. A Threshold Selection Method from Gray-Level Histograms. *IEEE Trans*
752 *Syst. Man Cybern.* 62–66 (1979).
- 753 38. Zhang, J. *et al.* Top-down Neural Attention by Excitation Backprop. *Int. J.*
754 *Comput. Vis.* **126**, 1084-1102 (2018).
- 755 39. Kim, H.-E. *et al.* Changes in cancer detection and false-positive recall in
756 mammography using artificial intelligence: a retrospective, multireader study. *Lancet*
757 *Digit. Health* **2**, e138–e148 (2020).
- 758 40. Efron, B. & Tibshirani, R. J. *An Introduction to the Bootstrap*. (CRC Press, 1994).
- 759 41. Bany Muhammad, M. *et al.* Eigen-CAM: Visual Explanations for Deep
760 Convolutional Neural Networks. *SN COMPUT. SCI.* **2**, 47 (2021).
- 761 42. Shrikumar, A., Greenside, P. & Kundaje, A. Learning Important Features
762 Through Propagating Activation Differences. In *International conference on machine*
763 *learning* 3145-3153 (PMLR, 2017).
- 764 43. Bach, S. *et al.* On Pixel-Wise Explanations for Non-Linear Classifier Decisions by
765 Layer-Wise Relevance Propagation. *PLOS ONE* **10**, e0130140 (2015).
- 766 44. Zeiler, M. D. & Fergus, R. Visualizing and Understanding Convolutional
767 Networks. in *Computer Vision – ECCV 2014* (eds. Fleet, D., Pajdla, T., Schiele, B. &
768 Tuytelaars, T.) 818–833 (Springer International Publishing, 2014). doi:10.1007/978-3-
769 319-10590-1_53.
- 770 45. Vrabac, D. *et al.* DLBCL-Morph: Morphological features computed using deep
771 learning for an annotated digital DLBCL image set. *Sci Data* **8**, 135 (2021).

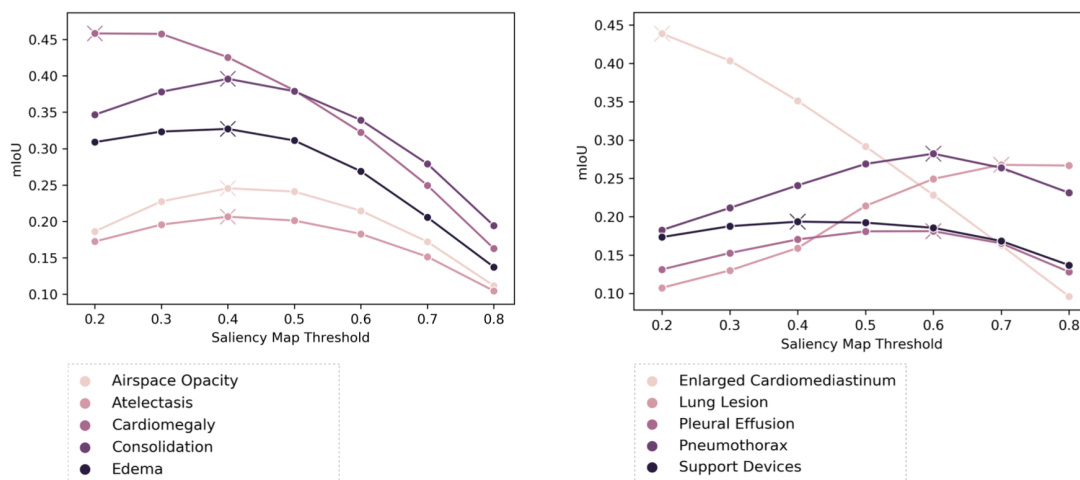
- 772 46. Ayhan, M. S. et al. Clinical validation of saliency maps for understanding deep
773 neural networks in ophthalmology. *Med. Image Anal.* 77, 102364 (2022).
- 774 47. Arun, N. et al. Assessing the Trustworthiness of Saliency Maps for Localizing
775 Abnormalities in Medical Imaging. *Radiol. Artif. Intell.* 3, e200267 (2021).
- 776 48. Wang, X. et al. ChestX-ray8: Hospital-Scale Chest X-Ray Database and
777 Benchmarks on Weakly-Supervised Classification and Localization of Common
778 Thorax Diseases. in *2017 IEEE Conference on Computer Vision and Pattern
779 Recognition (CVPR)* 2097–2106 (IEEE, 2017).
- 780 49. Nguyen, H. Q. et al. VinDr-CXR: An open dataset of chest X-rays with radiologist's
781 annotations. (2022) doi:10.48550/arXiv.2012.15029.
- 782 50. SIIM-ACR Pneumothorax Segmentation. [https://kaggle.com/c/siim-acr-](https://kaggle.com/c/siim-acr-pneumothorax-segmentation)
783 [pneumothorax-segmentation](https://kaggle.com/c/siim-acr-pneumothorax-segmentation).
- 784 51. Steiner, D. F. et al. Impact of Deep Learning Assistance on the Histopathologic
785 Review of Lymph Nodes for Metastatic Breast Cancer. *Am. J. Surg. Pathol.* 42,
786 1636–1646 (2018).
- 787 52. Uyumazturk, B. et al. Deep Learning for the Digital Pathologic Diagnosis of
788 Cholangiocarcinoma and Hepatocellular Carcinoma: Evaluating the Impact of a Web-
789 based Diagnostic Assistant. *ArXiv191107372 Eess* (2019).
- 790 53. Park, A. et al. Deep Learning–Assisted Diagnosis of Cerebral Aneurysms Using
791 the HeadXNet Model. *JAMA Netw. Open* 2, e195600 (2019).
- 792 54. Annalise.ai - Medical imaging AI, by clinicians for clinicians. *Annalise.ai*
793 <https://annalise.ai/>.
- 794 55. Lunit Inc. <https://www.lunit.io/en>.

- 795 56. Qure.ai | Artificial Intelligence for Radiology. <https://qure.ai/>.
- 796 57. Gadgil, S., Endo, M., Wen, E., Ng, A. Y. & Rajpurkar, P. CheXseg: Combining
797 Expert Annotations with DNN-generated Saliency Maps for X-ray Segmentation. In
798 *Proceedings of the Fourth Conference on Medical Imaging with Deep Learning* 190-
799 204 (PMLR, 2021).
- 800 58. Crosby, J., Chen, S., Li, F., MacMahon, H. & Giger, M. Network output
801 visualization to uncover limitations of deep learning detection of pneumothorax. in
802 *Medical Imaging 2020: Image Perception, Observer Performance, and Technology*
803 *Assessment* vol. 11316 113160O (International Society for Optics and Photonics,
804 2020).
- 805 59. Melbye, H. & Dale, K. Interobserver Variability in the Radiographic Diagnosis of
806 Adult Outpatient Pneumonia. *Acta Radiol.* **33**, 79–81 (1992).
- 807 60. Herman, P. G. *et al.* Disagreements in Chest Roentgen Interpretation. *CHEST*
808 **68**, 278–282 (1975).
- 809 61. Albaum, M. N. *et al.* Interobserver Reliability of the Chest Radiograph in
810 Community-Acquired Pneumonia. *CHEST* **110**, 343–350 (1996).
- 811 62. Arun, N. T. *et al.* Assessing the validity of saliency maps for abnormality
812 localization in medical imaging. In *Medical Imaging with Deep Learning*. (2020).
- 813 63. Graziani, M., Lompech, T., Müller, H. & Andrearczyk, V. *Evaluation and*
814 *Comparison of CNN Visual Explanations for Histopathology*. (2020).
- 815 64. Choe, J. *et al.* Evaluating Weakly Supervised Object Localization Methods Right.
816 In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern*
817 *Recognition* 3133-3142 (IEEE, 2020).

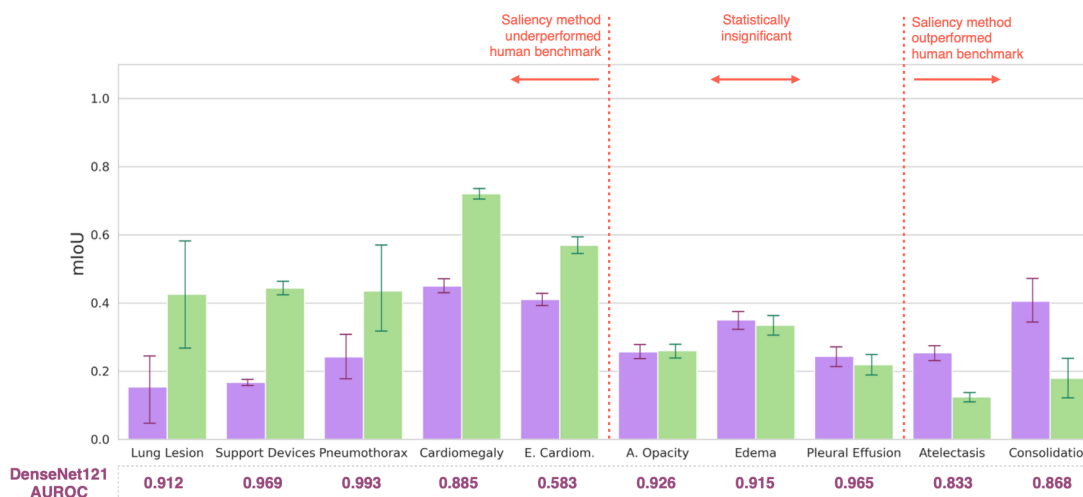
- 818 65. MD.ai. <https://www.md.ai/>.
- 819 66. Seah, J. C. Y. et al. Effect of a comprehensive deep-learning model on the
820 accuracy of chest x-ray interpretation by radiologists: a retrospective, multireader
821 multicase study. *Lancet Digit. Health* 3, e496–e506 (2021).
- 822 67. Tan, M. & Le, Q. V. EfficientNet: Rethinking Model Scaling for Convolutional
823 Neural Networks. In *International conference on machine learning* 6105-6114 (PMLR,
824 2020).
- 825 68 Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization.
826 *ArXiv14126980 Cs* (2017).
- 827 69. Saporta, A. et al. Code for 'Benchmarking saliency methods for chest X-ray
828 interpretation' (Zenodo, 2022); <https://doi.org/10.5281/zenodo.6816288>.
- 829

830 **Extended Data**
831

a



b



832 **Extended Data Fig. 1 | mIoU localization performance of the saliency method**
 833 **pipeline using threshold values tuned on the validation set. a,** We first applied max-
 834 min normalizations to the Grad-CAM saliency maps so that each value gets transformed
 835 into a decimal between 0 and 1. We then passed in a range of threshold values from 0.2
 836 to 0.8 to create binary segmentations and plotted the mIoU score per pathology under
 837 each threshold on the validation set. The threshold that gives the max mIoU for each
 838 pathology is marked with an “X”. Pathologies are sorted alphabetically and shown in two
 839 plots for readability. **b,** Comparing mIoU localization performances of the saliency
 840 method pipeline (using the best thresholds tuned on the validation set) and the human
 841 benchmark. We found that the saliency method pipeline outperformed the human
 842 benchmark on two pathologies and underperformed the human benchmark on five
 843 pathologies. For the remaining three pathologies, the performance differences were not
 844 statistically significant. This finding is consistent with what we report in the manuscript
 845 using Otsu's method.
 846

847

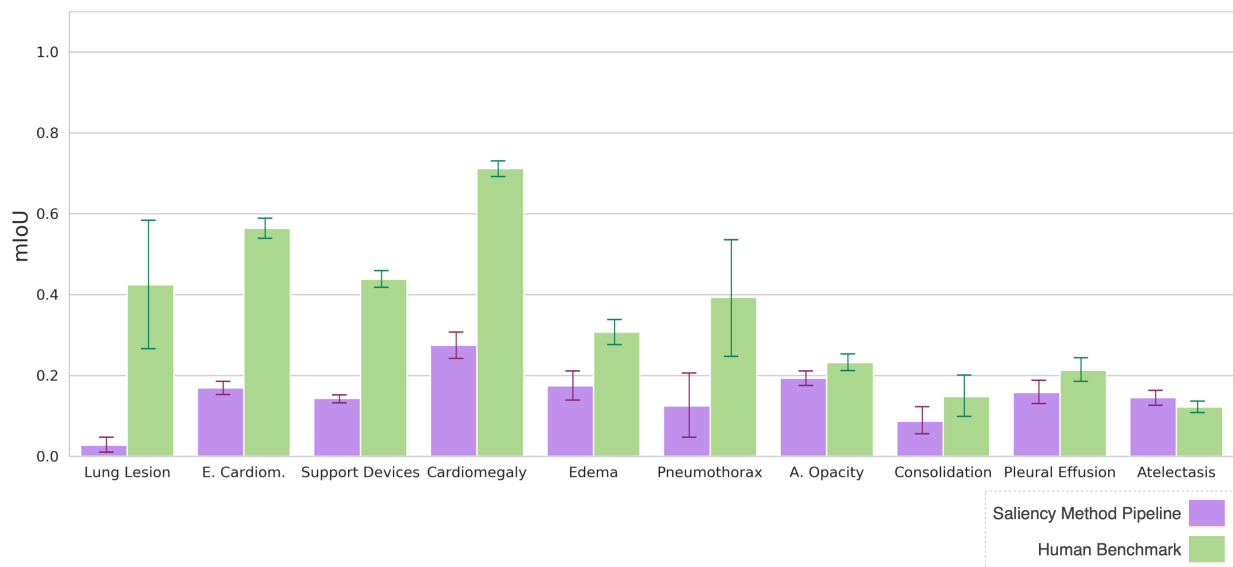
pathology	specificity (precision)		sensitivity (recall)	
	saliency method pipeline	human benchmark	saliency method pipeline	human benchmark
Airspace Opacity	0.844	0.982	0.975	0.961
Atelectasis	0.854	0.999	0.985	0.971
Cardiomegaly	0.914	0.998	0.978	0.986
Consolidation	0.916	1.000	0.998	0.995
Edema	0.851	0.997	0.988	0.980
Enlarged Cardiom.	0.935	0.993	0.938	0.958
Lung Lesion	0.887	1.000	1.000	1.000
Pleural Effusion	0.808	0.999	0.994	0.987
Pneumothorax	0.866	1.000	1.000	1.000
Support Devices	0.862	0.997	0.980	0.979

848
849
850
851
852

Extended Data Fig. 2 | Specificity and sensitivity values of the saliency method pipeline and human benchmark. For each pathology, we highlight the higher of the two metrics (saliency method pipeline or human benchmark) in **bold**.

pathology	Grad-CAM			Grad-CAM++			Integrated Gradients		
	DenseNet121	ResNet152	Inception-v4	DenseNet121	ResNet152	Inception-v4	DenseNet121	ResNet152	Inception-v4
mIoU									
Airspace Opacity	0.248	0.194	0.090	0.234	0.198	0.115	0.123	0.119	0.052
Atelectasis	0.254	0.221	0.115	0.245	0.210	0.106	0.116	0.115	0.064
Cardiomegaly	0.452	0.424	0.120	0.346	0.257	0.196	0.160	0.154	0.089
Consolidation	0.408	0.334	0.079	0.296	0.245	0.130	0.177	0.112	0.069
Edema	0.362	0.240	0.203	0.388	0.345	0.266	0.073	0.062	0.099
Enlarged Cardiom.	0.379	0.272	0.065	0.400	0.382	0.295	0.154	0.152	0.094
Lung Lesion	0.101	0.066	0.003	0.089	0.069	0.045	0.107	0.063	0.001
Pleural Effusion	0.235	0.204	0.120	0.195	0.176	0.090	0.088	0.091	0.067
Pneumothorax	0.213	0.171	0.088	0.216	0.184	0.124	0.077	0.070	0.078
Support Devices	0.163	0.147	0.116	0.133	0.126	0.099	0.099	0.074	0.066
hit rate									
Airspace Opacity	0.498	0.428	0.106	0.558	0.522	0.148	0.606	0.586	0.122
Atelectasis	0.501	0.490	0.062	0.621	0.621	0.118	0.520	0.453	0.187
Cardiomegaly	0.903	0.915	0.126	0.732	0.297	0.493	0.697	0.748	0.268
Consolidation	0.738	0.797	0.030	0.708	0.600	0.284	0.624	0.538	0.115
Edema	0.746	0.432	0.385	0.781	0.745	0.457	0.300	0.350	0.180
Enlarged Cardiom.	0.818	0.627	0.030	0.630	0.631	0.731	0.704	0.730	0.205
Lung Lesion	0.290	0.146	0.000	0.290	0.146	0.000	0.423	0.211	0.000
Pleural Effusion	0.507	0.499	0.133	0.347	0.473	0.107	0.332	0.400	0.182
Pneumothorax	0.392	0.600	0.000	0.489	0.698	0.097	0.801	0.498	0.097
Support Devices	0.355	0.287	0.133	0.364	0.334	0.150	0.491	0.442	0.324

853
 854 **Extended Data Fig. 3 | Test set localization performance for each combination of**
 855 **saliency method and CNN architecture.** For each pathology and saliency method, we
 856 highlight the highest performing CNN architecture in **bold**.
 857



858
859 **Extended Data Fig. 4 | Saliency method pipeline localization performance on the**
860 **full dataset using mIoU.** True negatives (CXR's whose ground-truth label is negative
861 for a given pathology and for which there were neither human benchmark nor saliency
862 method pipeline segmentations for that pathology) were excluded from the metric
863 calculation. To control for false positives, we ensure that the final binary segmentation is
864 consistent with model probability output by applying another layer of thresholding such
865 that the segmentation mask produced all zeros if the predicted probability was below a
866 chosen level. The probability threshold is searched on the interval of [0,0.8] with steps
867 of 0.1. The exact value is determined per pathology by maximizing the mIoU on the
868 validation set. We found that on the full dataset, for seven of the 10 pathologies, the
869 saliency method pipeline had a significantly lower mIoU than the human benchmark.
870

pathology	human benchmark mIoU	saliency method pipeline mIoU	% decrease (95% CI)
Lung Lesion	0.426	0.101	76.2 (59.1, 87.5)
Support Devices	0.444	0.163	63.3 (60.8, 65.8)
Pneumothorax	0.435	0.213	51.0 (14.6, 69.5)
Cardiomegaly	0.720	0.452	37.2 (34.0, 40.4)
Enlarged Cardiom.	0.569	0.379	33.4 (29.0, 37.4)
Airspace Opacity	0.260	0.248	4.6 (-5.8, 14.6)
Pleural Effusion	0.219	0.235	-6.8 (-25.6, 13.3)
Edema	0.335	0.362	-7.4 (-16.4, 2.6)
Atelectasis	0.124	0.254	-51.2 (-57.3, -43.9)
Consolidation	0.179	0.408	-56.1 (-69.4, -42.7)
Average	0.383	0.281	26.6 (18.1, 35.0)

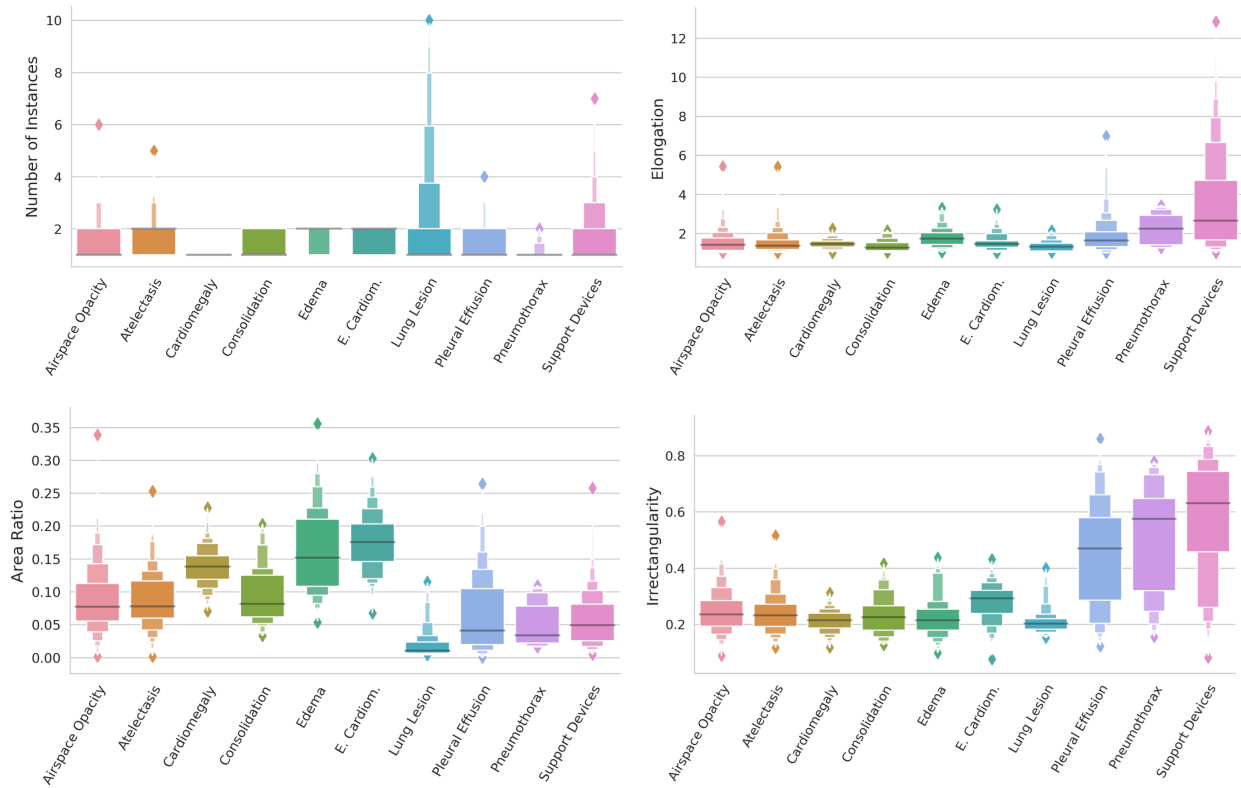
871
872 **Extended Data Fig. 5 | Percentage decrease from human benchmark mIoU to**
873 **saliency method pipeline mIoU.** Pathologies are sorted first by statistical significance
874 of percentage decrease from human benchmark mIoU to saliency method pipeline
875 mIoU (high to low), and then by percentage decrease from human benchmark mIoU to
876 saliency method pipeline mIoU (high to low). We use 95% bootstrap confidence interval.
877

pathology	human benchmark hit rate (%)	saliency method pipeline hit rate (%)	% decrease (95% CI)
Lung Lesion	0.850	0.290	65.9 (35.3, 91.7)
Support Devices	0.933	0.355	62.0 (56.2, 67.5)
Pneumothorax	1.000	0.392	60.8 (27.3, 92.3)
Atelectasis	0.870	0.501	42.4 (0.3, 0.5)
Pleural Effusion	0.718	0.507	29.4 (14.3, 42.5)
Enlarged Cardiom.	0.957	0.818	14.5 (9.6, 19.2)
Cardiomegaly	0.972	0.903	7.1 (2.1, 11.8)
Airspace Opacity	0.559	0.498	10.9 (-2.0, 23.1)
Edema	0.769	0.746	3.0 (-11.7, 18.5)
Consolidation	0.510	0.738	-44.7 (-56.5, 0.5)
Average	0.820	0.580	29.4 (15.0, 43.2)

878
879 **Extended Data Fig. 6 | Percentage decrease from human benchmark hit rate to**
880 **saliency method pipeline hit rate.** Pathologies are sorted first by statistical
881 significance of percentage decrease from human benchmark hit rate to saliency method
882 pipeline hit rate (high to low), and then by percentage decrease from human benchmark
883 hit rate to saliency method pipeline hit rate (high to low). We use 95% bootstrap
884 confidence interval.
885

pathology	ensemble model	single checkpoint
mIoU		
Airspace Opacity	0.248	0.241
Atelectasis	0.254	0.233
Cardiomegaly	0.452	0.419
Consolidation	0.408	0.369
Edema	0.362	0.360
Enlarged Cardiom.	<u>0.379</u>	0.297
Lung Lesion	0.101	0.099
Pleural Effusion	0.235	0.205
Pneumothorax	0.213	0.181
Support Devices	0.163	0.150
hit rate		
Airspace Opacity	0.498	0.534
Atelectasis	0.501	0.504
Cardiomegaly	0.903	0.846
Consolidation	0.738	0.711
Edema	0.746	0.749
Enlarged Cardiom.	<u>0.818</u>	0.704
Lung Lesion	0.290	0.286
Pleural Effusion	0.507	0.390
Pneumothorax	0.392	0.491
Support Devices	0.355	0.312

886
 887 **Extended Data Fig. 7 | Saliency method pipeline localization performance using**
 888 **an ensemble model vs. using the top performing single checkpoint for each**
 889 **pathology.** For each pathology, we highlight in **bold** the model (ensemble or single
 890 checkpoint) that has the higher metric, and we underline it if the difference is statistically
 891 significant (using 95% bootstrap confidence interval).
 892



893
894 **Extended Data Fig. 8 | Distribution of four geometric features across all 10**
895 **pathologies.** The black horizontal line in each box indicates the median feature value
896 for that pathology, and each successive level outward contains half of the remaining
897 data. The height of the box indicates the range of feature values in the quantile.
898

pathology	CXRs (n)	Linear regression coefficient	Spearman correlation coefficient
Airspace Opacity	381	0.498***	0.160**
Atelectasis	296	0.443**	0.126
Cardiomegaly	229	0.195*	0.185*
Consolidation	120	0.082	0.199
Edema	124	0.195	0.132
Enlarged Cardiom.	668	0.548**	0.253***
Lung Lesion	50	0.540	0.453
Pleural Effusion	159	0.654***	0.278**
Pneumothorax	11	0.210	0.142
Support Devices	327	-0.058	-0.029
All pathologies	2365	-0.411***	-0.239***

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

899
900
901
902

Extended Data Fig. 9 | Hit rate: Coefficients from regressions on model assurance.

sample size	
Number studies	500
Number CXRs	668
pathology	CXRs (n)
Airspace Opacity	309
Atelectasis	177
Cardiomegaly	175
Consolidation	35
Edema	83
Enlarged Cardiom.	297
Lung Lesion	14
Pleural Effusion	120
Pneumothorax	10
Support Devices	314
No pathology identified	169

903
904

Extended Data Fig. 10 | Test set summary statistics.