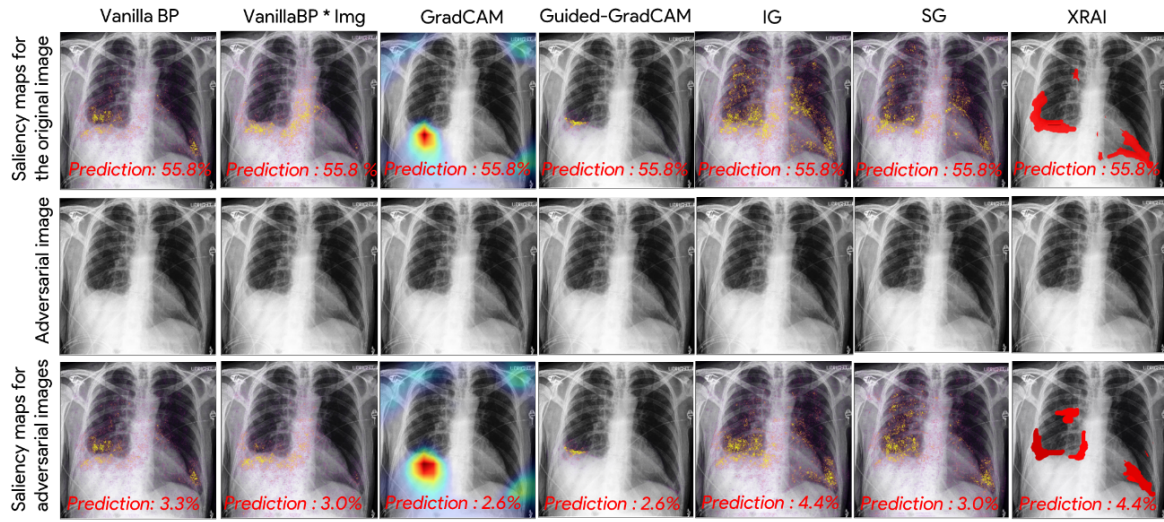


SUPPLEMENTARY MATERIALS

Supplementary section I: Additional evaluation details on the relevance and resistance of each individual disease class.

1. Relevance evaluation of each individual class

1) Atelectasis

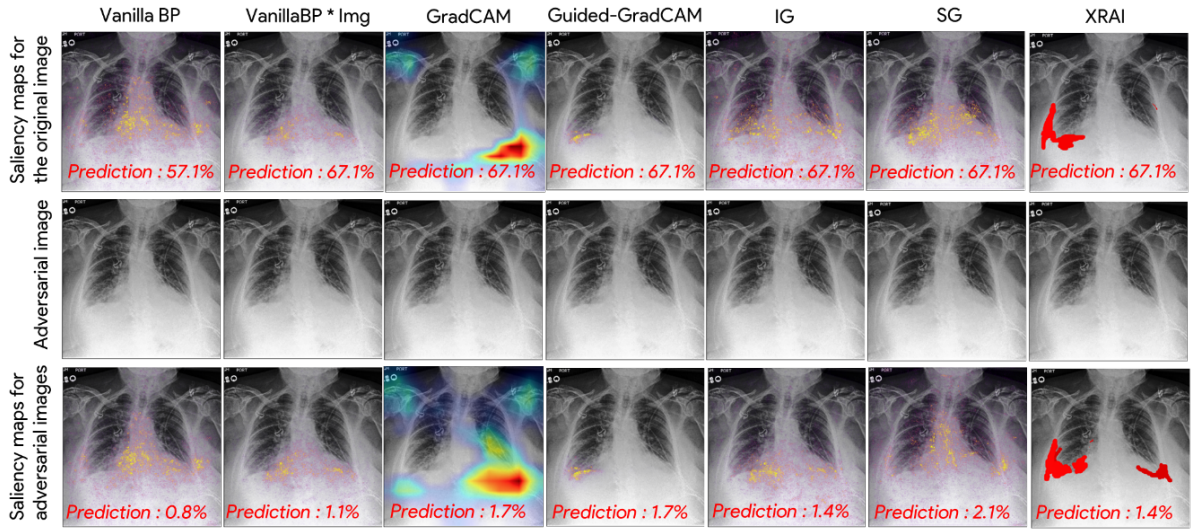


Supplementary Figure S1. Saliency relevance evaluation on atelectasis. Examples show that saliency maps remain unchanged with altered model predictions.

Supplementary Table S1. Saliency relevance quantification on atelectasis. The specific evaluation result of ResNet-152 and DenseNet-121 on atelectasis shows the lack of relevance between the saliency maps and model predictions.

Model	Evaluation metrics	Vanilla BP	Vanilla BP × Image	GradCAM	Guided GradCAM	IG	SG	XRAI
ResNet-152	AUC (dropped from 0.83)	0.01	0.02	0.00	0.00	0.00	0.00	0.00
	Saliency map similarity (SSIM)	0.77	0.78	0.82	0.87	0.84	0.79	0.84
	PSC	0.21	0.22	0.25	0.23	0.04	0.13	0.12
DenseNet-121	AUC (dropped from 0.83)	0.01	0.02	0.01	0.01	0.01	0.00	0.01
	Saliency map similarity (SSIM)	0.77	0.78	0.82	0.77	0.84	0.80	0.78
	PSC	0.25	0.26	0.25	0.21	0.25	0.02	0.26

2) Cardiomegaly

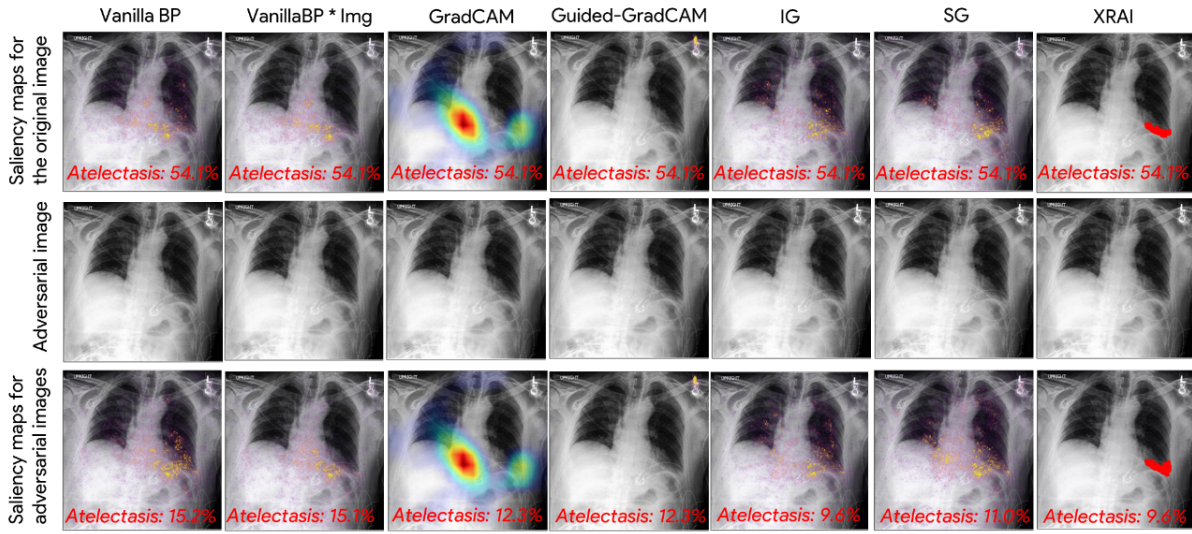


Supplementary Figure S2. Saliency relevance evaluation on cardiomegaly. Examples show that saliency maps remain unchanged with altered model predictions.

Supplementary Table S2. Saliency relevance quantification on cardiomegaly. The specific evaluation result of ResNet-152 and DenseNet-121 on cardiomegaly shows the lack of relevance between the saliency maps and model predictions.

Model	Evaluation metrics	Vanilla BP	Vanilla BP × Image	GradCAM	Guided GradCAM	IG	SG	XRAI
ResNet-152	AUC (dropped from 0.84)	0.02	0.05	0.09	0.09	0.01	0.01	0.01
	Saliency map similarity (SSIM)	0.74	0.79	0.92	0.85	0.81	0.81	0.85
	PSC	0.21	0.22	0.25	0.23	0.04	0.13	0.12
DenseNet-121	AUC (dropped from 0.84)	0.02	0.04	0.07	0.07	0.01	0.00	0.02
	Saliency map similarity (SSIM)	0.74	0.79	0.82	0.78	0.82	0.83	0.79
	PSC	0.25	0.26	0.25	0.21	0.25	0.02	0.26

3) Consolidation

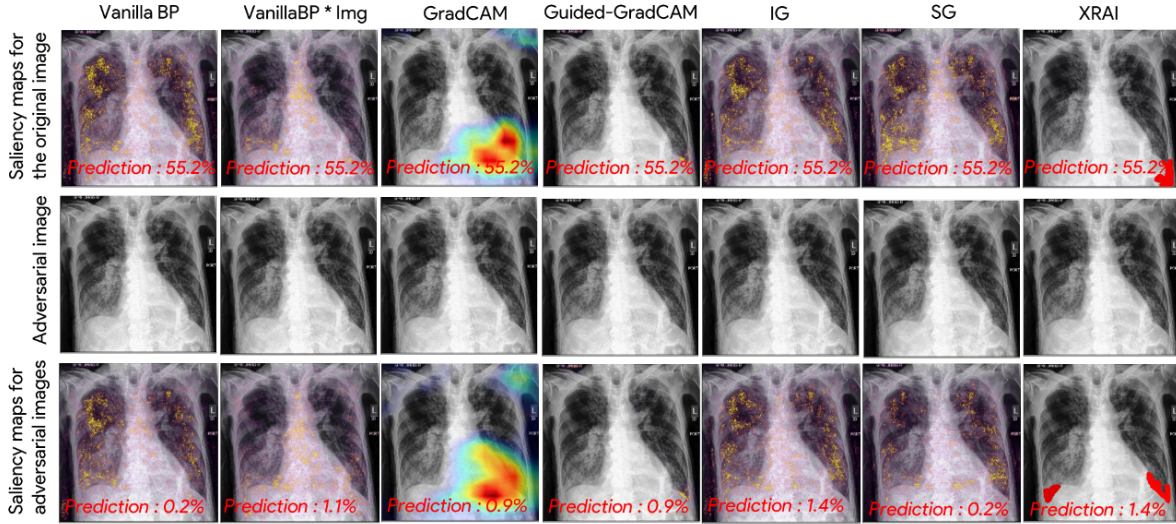


Supplementary Figure S3. Saliency relevance evaluation on consolidation. Examples show that saliency maps remain unchanged with altered model predictions.

Supplementary Table S3. Saliency relevance quantification on consolidation. The specific evaluation result of ResNet-152 and DenseNet-121 on cardiomegaly shows the lack of relevance between the saliency maps and model predictions.

Model	Evaluation metrics	Vanilla BP	Vanilla BP × Image	GradCAM	Guided GradCAM	IG	SG	XRAI
ResNet-152	AUC (dropped from 0.88)	0.02	0.04	0.02	0.02	0.01	0.02	0.01
	Saliency map similarity (SSIM)	0.74	0.78	0.89	0.85	0.83	0.86	0.85
	PSC	0.21	0.22	0.25	0.23	0.04	0.13	0.12
DenseNet-121	AUC (dropped from 0.88)	0.02	0.01	0.04	0.01	0.02	0.01	0.02
	Saliency map similarity (SSIM)	0.76	0.78	0.78	0.78	0.87	0.85	0.77
	PSC	0.25	0.26	0.25	0.21	0.25	0.02	0.26

4) Edema

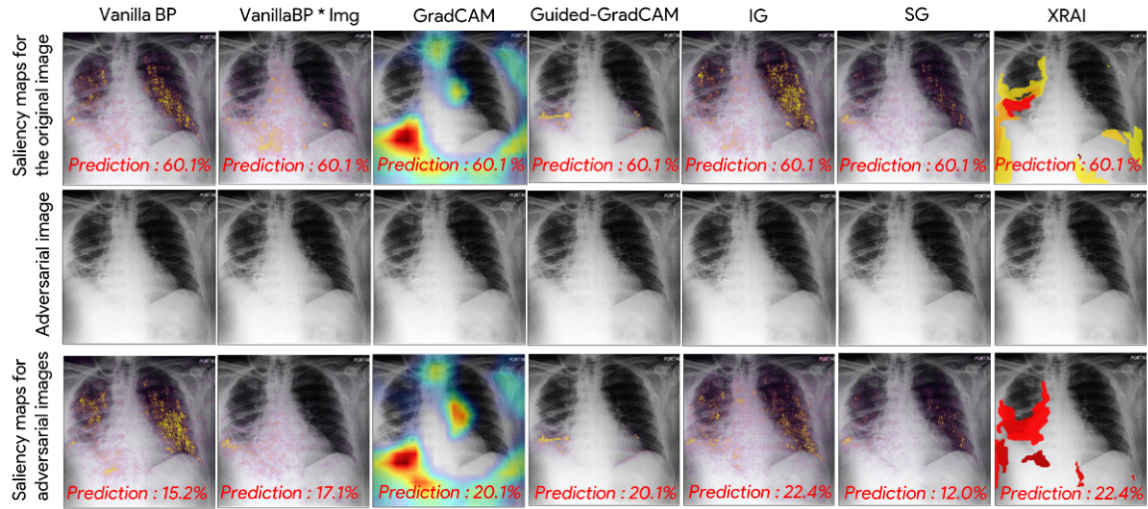


Supplementary Figure S4. Saliency relevance evaluation on edema. Examples show that saliency maps remain unchanged with altered model predictions.

Supplementary Table S4. Saliency relevance quantification on edema. The specific evaluation result of ResNet-152 and DenseNet-121 on cardiomegaly shows the lack of relevance between the saliency maps and model predictions.

Model	Evaluation metrics	Vanilla BP	Vanilla BP × Image	GradCAM	Guided GradCAM	IG	SG	XRAI
ResNet-152	AUC (dropped from 0.91)	0.02	0.05	0.01	0.01	0.02	0.02	0.02
	Saliency map similarity (SSIM)	0.77	0.75	0.83	0.85	0.84	0.84	0.85
	PSC	0.21	0.22	0.25	0.23	0.04	0.13	0.12
DenseNet-121	AUC (dropped from 0.91)	0.02	0.05	0.02	0.02	0.02	0.01	0.02
	Saliency map similarity (SSIM)	0.77	0.75	0.82	0.78	0.84	0.83	0.76
	PSC	0.25	0.26	0.25	0.21	0.25	0.02	0.26

5) Pleural Effusion



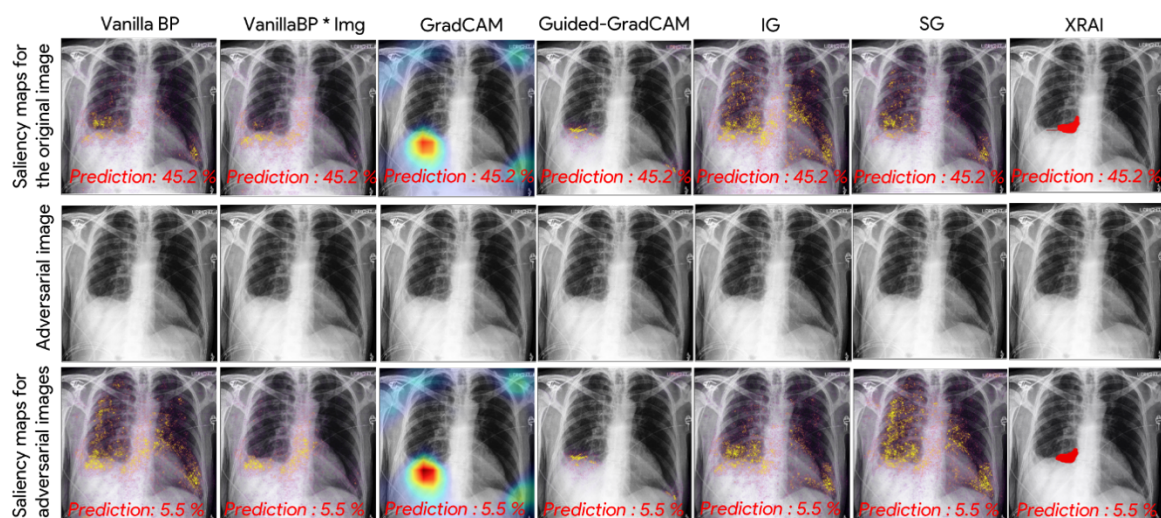
Supplementary Figure S5. Saliency relevance evaluation on pleural effusion. Examples show that saliency maps remain unchanged with altered model predictions.

Supplementary Table S5. Saliency relevance quantification on pleural effusion. The specific evaluation result of ResNet-152 and DenseNet-121 on cardiomegaly shows the lack of relevance between the saliency maps and model predictions.

Model	Evaluation metrics	Vanilla BP	Vanilla BP × Image	GradCAM	Guided GradCAM	IG	SG	XRAI
ResNet-152	AUC (dropped from 0.93)	0.05	0.08	0.03	0.03	0.01	0.02	0.01
	Saliency map similarity (SSIM)	0.82	0.86	0.82	0.82	0.88	0.87	0.85
	PSC	0.21	0.22	0.25	0.23	0.04	0.13	0.12
DenseNet-121	AUC (dropped from 0.93)	0.06	0.07	0.12	0.12	0.01	0.01	0.101
	Saliency map similarity (SSIM)	0.82	0.86	0.81	0.78	0.88	0.84	0.79
	PSC	0.25	0.26	0.25	0.21	0.25	0.02	0.26

2. Transferable relevance evaluation across different XAI

1) Atelectasis

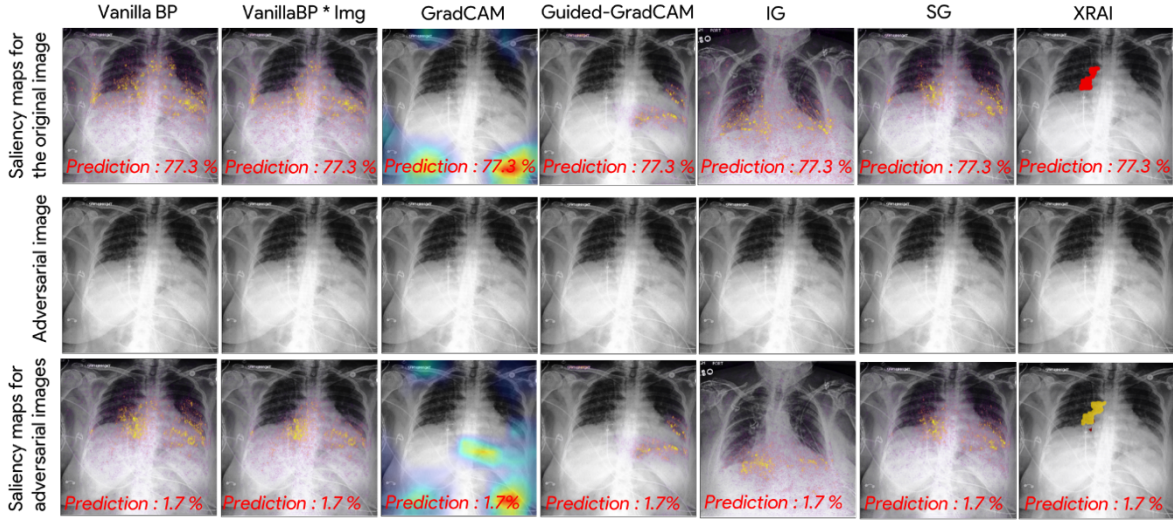


Supplementary Fig. S6. The lack of relevance is transferable across different saliency methods. An adversarial chest radiograph is generated via attacking IG on atelectasis first, then evaluated on all other saliency methods.

Supplementary Table S6. Relevance quantification across saliency methods on atelectasis. Relevance evaluation result on ResNet-152 and DenseNet-121 are reported by transferring images generated for IG to other saliency methods.

Model	Evaluation metrics	IG	Transfer to other XAI					
			Vanilla BP	Vanilla BP × Image	GradCAM	Guided GradCAM	SG	XRAI
ResNet-152	AUC (dropped from 0.83)	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Saliency map similarity (SSIM)	0.84	0.75	0.7	0.76	0.85	0.75	0.76
	PSC	0.02	0.32	0.34	0.13	0.09	0.30	0.27
DenseNet-121	AUC (dropped from 0.83)	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Saliency map similarity (SSIM)	0.84	0.75	0.73	0.62	0.83	0.83	0.81
	PSC	0.23	0.34	0.37	0.32	0.24	0.27	0.25

2) Cardiomegaly

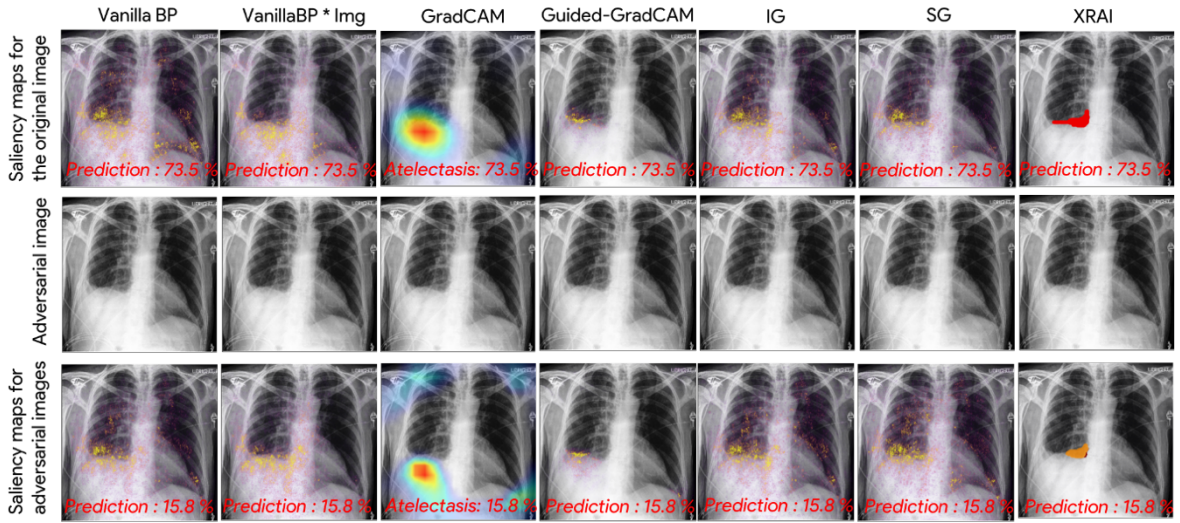


Supplementary Fig. S7. The lack of relevance is transferable across different saliency methods. An adversarial chest radiograph is generated via attacking IG on cardiomegaly first, then evaluated on all other saliency methods.

Supplementary Table S7. Relevance quantification across saliency methods on cardiomegaly. Relevance evaluation result on ResNet-152 and DenseNet-121 are reported by transferring images generated for IG to other saliency methods.

Model	Evaluation metrics	IG	Transfer to other XAI					
			Vanilla BP	Vanilla BP × Image	GradCAM	Guided GradCAM	SG	XRAI
ResNet-152	AUC (dropped from 0.84)	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	Saliency map similarity (SSIM)	0.81	0.76	0.77	0.74	0.88	0.73	0.78
	PSC	0.05	0.33	0.37	0.10	0.07	0.33	0.29
DenseNet-121	AUC (dropped from 0.84)	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Saliency map similarity (SSIM)	0.82	0.75	0.77	0.56	0.84	0.81	0.82
	PSC	0.25	0.35	0.32	0.33	0.21	0.25	0.22

3) Consolidation

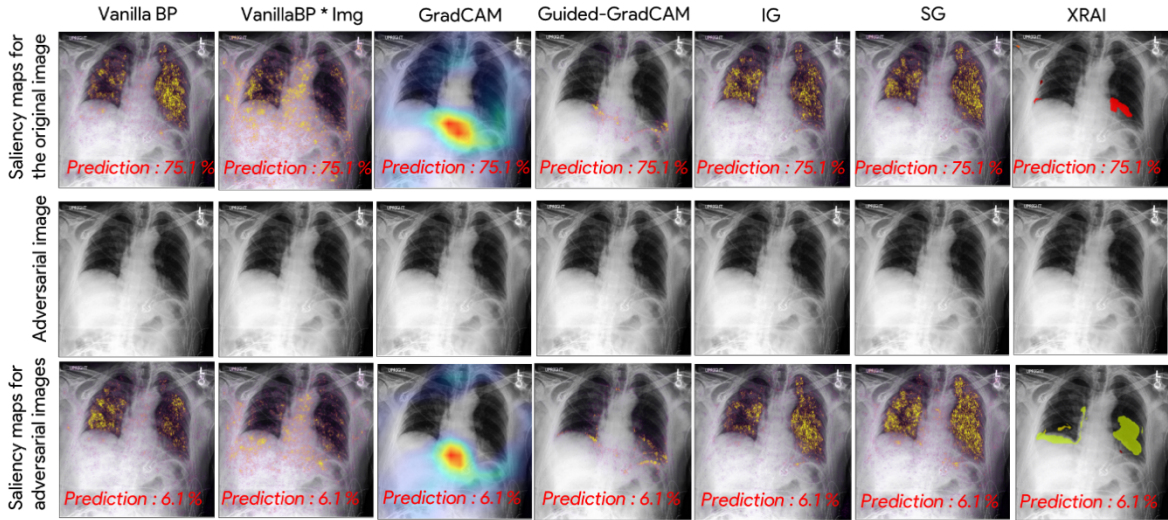


Supplementary Fig. S8. The lack of relevance is transferable across different saliency methods. An adversarial chest radiograph is generated via attacking IG on consolidation first, then evaluated on all other saliency methods.

Supplementary Table S8. Relevance quantification across saliency methods on consolidation. Relevance evaluation result on ResNet-152 and DenseNet-121 are reported by transferring images generated for IG to other saliency methods.

Model	Evaluation metrics	IG	Transfer to other XAI					
			Vanilla BP	Vanilla BP × Image	GradCAM	Guided GradCAM	SG	XRAI
ResNet-152	AUC (dropped from 0.88)	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	Saliency map similarity (SSIM)	0.83	0.77	0.78	0.78	0.86	0.72	0.77
	PSC	0.04	0.36	0.35	0.11	0.07	0.31	0.27
DenseNet-121	AUC (dropped from 0.88)	0.02	0.02	0.02	0.02	0.02	0.02	0.02
	Saliency map similarity (SSIM)	0.87	0.72	0.73	0.61	0.87	0.78	0.81
	PSC	0.27	0.35	0.35	0.33	0.21	0.25	0.22

4) Edema

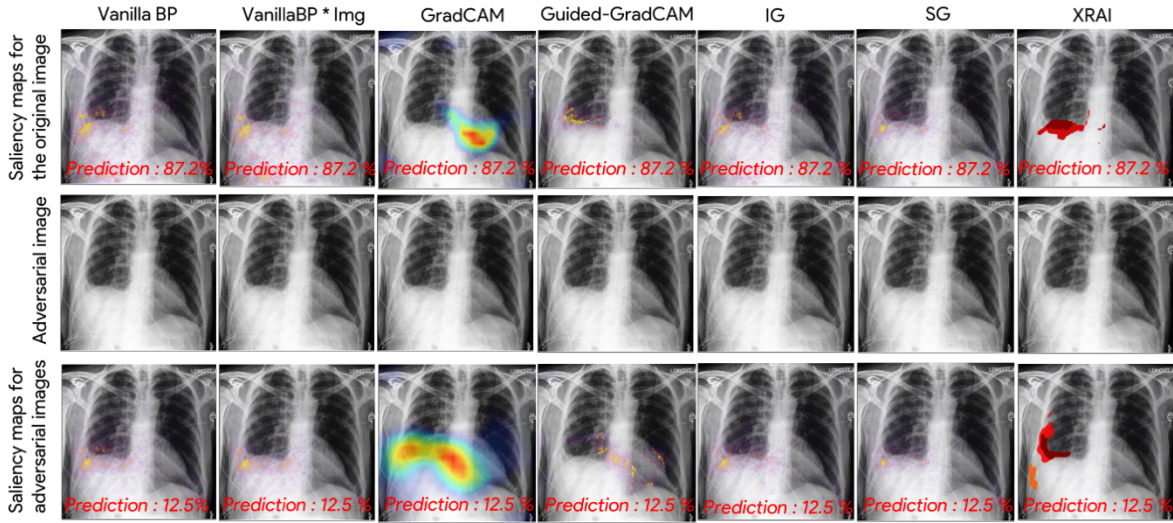


Supplementary Fig. S9. The lack of relevance is transferable across different saliency methods. An adversarial chest radiograph is generated via attacking IG on edema first, then evaluated on all other saliency methods.

Supplementary Table S9. Relevance quantification across saliency methods on edema. Relevance evaluation result on ResNet-152 and DenseNet-121 are reported by transferring images generated for IG to other saliency methods.

Model	Evaluation metrics	IG	Transfer to other XAI					
			Vanilla BP	Vanilla BP × Image	GradCAM	Guided GradCAM	SG	XRAI
ResNet-152	AUC (dropped from 0.91)	0.02	0.02	0.02	0.02	0.02	0.02	0.02
	Saliency map similarity (SSIM)	0.84	0.78	0.79	0.76	0.87	0.77	0.75
	PSC	0.04	0.35	0.35	0.11	0.07	0.31	0.27
DenseNet-121	AUC (dropped from 0.91)	0.02	0.02	0.02	0.02	0.02	0.02	0.02
	Saliency map similarity (SSIM)	0.84	0.73	0.75	0.61	0.83	0.80	0.82
	PSC	0.26	0.35	0.34	0.33	0.21	0.26	0.23

5) Pleural Effusion



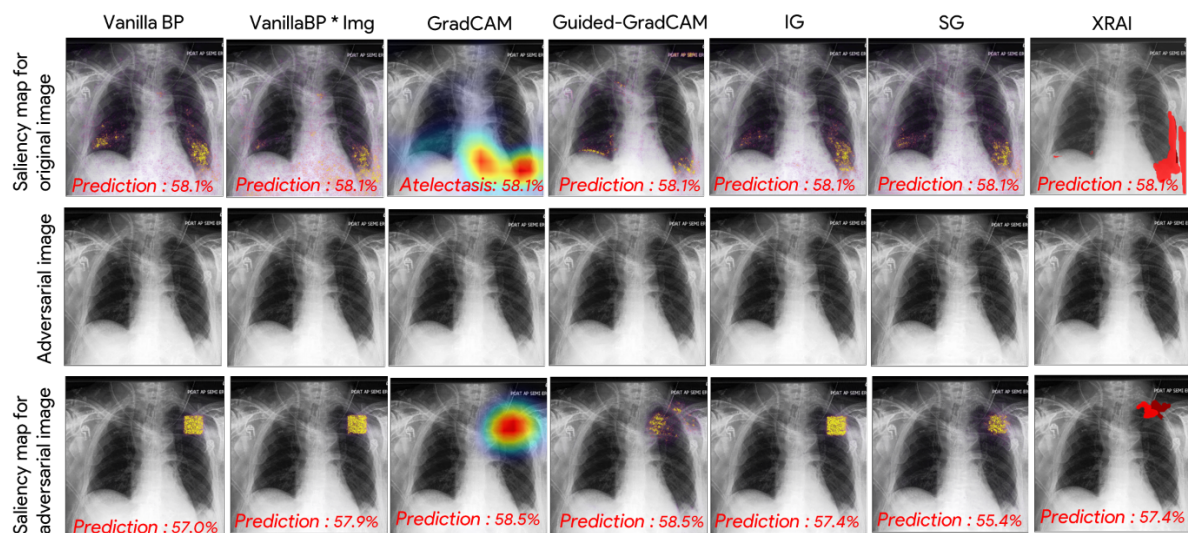
Supplementary Fig. S10. The lack of relevance is transferable across different saliency methods. An adversarial chest radiograph is generated via attacking IG on pleural effusion first, then evaluated on all other saliency methods.

Supplementary Table S10. Relevance quantification across saliency methods on pleural effusion. Relevance evaluation result on ResNet-152 and DenseNet-121 are reported by transferring images generated for IG to other saliency methods.

Model	Evaluation metrics	IG	Transfer to other XAI					
			Vanilla BP	Vanilla BP × Image	GradCAM	Guided GradCAM	SG	XRAI
ResNet-152	AUC (dropped from 0.93)	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	Saliency map similarity (SSIM)	0.88	0.76	0.77	0.77	0.86	0.74	0.74
	PSC	0.04	0.34	0.35	0.11	0.07	0.31	0.27
DenseNet-121	AUC (dropped from 0.93)	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	Saliency map similarity (SSIM)	0.88	0.74	0.75	0.60	0.86	0.80	0.84
	PSC	0.24	0.35	0.33	0.33	0.21	0.24	0.21

3. Resistance evaluation of each individual class

1) Atelectasis

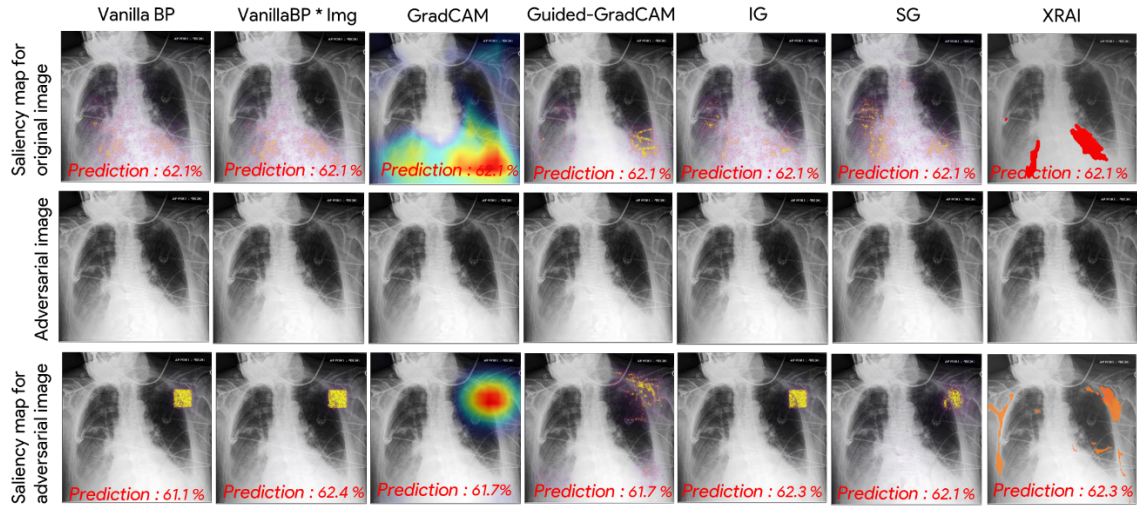


Supplementary Fig. S11. Saliency maps lacking resistance to saliency tampering on atelectasis. An example of manipulating saliency maps towards a target region while the original prediction probability of the atelectasis remains unaffected.

Supplementary Table S11. Quantification results of saliency resistance on atelectasis. Resistance evaluation on ResNet-152 and DenseNet-121 are reported. $SSIM_{org}$ and $SSIM_{tgt}$ denote the SSIMs of the new saliency map with the original saliency map and the target saliency map, respectively.

Models		Vanilla BP	Vanilla BP * Image	GradCAM	Guided GradCAM	IG	SG	XRAI
ResNet-152	AUC remains at 0.83	0.83	0.83	0.84	0.84	0.83	0.83	0.83
	Saliency map similarity	$SSIM_{org}$	0.45	0.49	0.23	0.38	0.44	0.47
		$SSIM_{tgt}$	0.94	0.94	0.72	0.74	0.94	0.85
	PSC	0.01	0.01	0.04	0.06	0.03	0.06	0.04
DenseNet-121	AUC remains at 0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83
	Saliency map similarity	$SSIM_{org}$	0.44	0.49	0.23	0.40	0.44	0.51
		$SSIM_{tgt}$	0.94	0.94	0.71	0.89	0.94	0.86
	PSC	0.01	0.01	0.11	0.12	0.07	0.10	0.09

2) Cardiomegaly

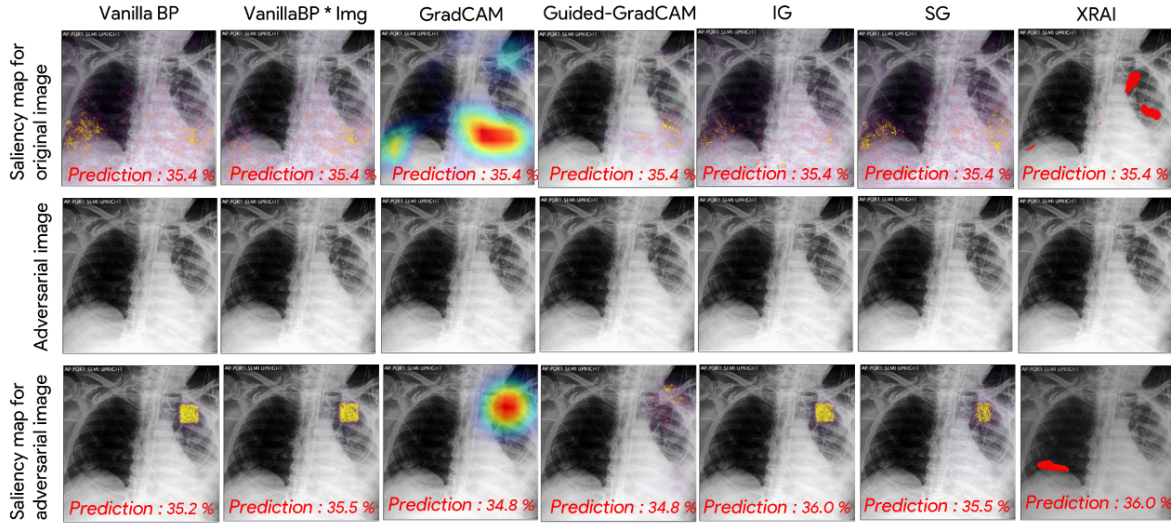


Supplementary Fig. S12. Saliency maps lacking resistance to saliency tampering on cardiomegaly. An example of manipulating saliency maps towards a target region while the original prediction probability of the atelectasis remains unaffected.

Supplementary Table S12. Quantification results of saliency resistance on cardiomegaly. Resistance evaluation on ResNet-152 and DenseNet-121 are reported. $SSIM_{org}$ and $SSIM_{tgt}$ denote the SSIMs of the new saliency map with the original saliency map and the target saliency map, respectively.

Models		Vanilla BP	Vanilla BP * Image	GradCAM	Guided GradCAM	IG	SG	XRAI
ResNet-152	AUC remains at 0.84	0.85	0.85	0.85	0.85	0.85	0.85	0.85
	Saliency map similarity	$SSIM_{org}$	0.46	0.56	0.39	0.38	0.43	0.51
		$SSIM_{tgt}$	0.94	0.93	0.67	0.74	0.93	0.79
	PSC	0.01	0.01	0.04	0.06	0.03	0.06	0.04
DenseNet-121	AUC remains at 0.84	0.85	0.85	0.85	0.85	0.85	0.85	0.85
	Saliency map similarity	$SSIM_{org}$	0.46	0.57	0.38	0.40	0.43	0.51
		$SSIM_{tgt}$	0.94	0.93	0.66	0.89	0.93	0.79
	PSC	0.01	0.01	0.09	0.12	0.07	0.08	0.10

3) Consolidation

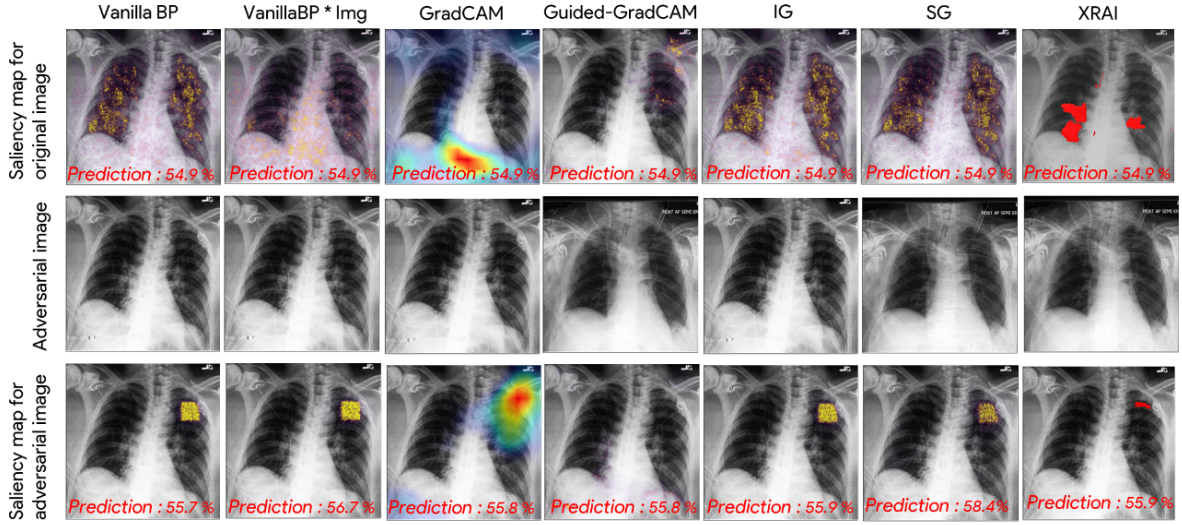


Supplementary Fig. S13. Saliency maps lacking resistance to saliency tampering on consolidation. An example of manipulating saliency maps towards a target region while the original prediction probability of the atelectasis remains unaffected.

Supplementary Table S13. Quantification results of saliency resistance on consolidation. Resistance evaluation on ResNet-152 and DenseNet-121 are reported. $SSIM_{org}$ and $SSIM_{tgt}$ denote the SSIMs of the new saliency map with the original saliency map and the target saliency map, respectively.

Models		Vanilla BP	Vanilla BP * Image	GradCAM	Guided GradCAM	IG	SG	XRAI
ResNet-152	AUC remains at 0.88	0.89	0.89	0.88	0.88	0.89	0.89	0.89
	Saliency map similarity	$SSIM_{org}$	0.40	0.51	0.33	0.38	0.42	0.41
		$SSIM_{tgt}$	0.94	0.93	0.62	0.74	0.93	0.78
	PSC	0.01	0.01	0.04	0.06	0.03	0.06	0.04
DenseNet-121	AUC remains at 0.88	0.89	0.89	0.89	0.89	0.89	0.89	0.89
	Saliency map similarity	$SSIM_{org}$	0.40	0.52	0.32	0.40	0.42	0.46
		$SSIM_{tgt}$	0.94	0.93	0.63	0.89	0.93	0.77
	PSC	0.01	0.01	0.13	0.12	0.07	0.12	0.09

4) Edema

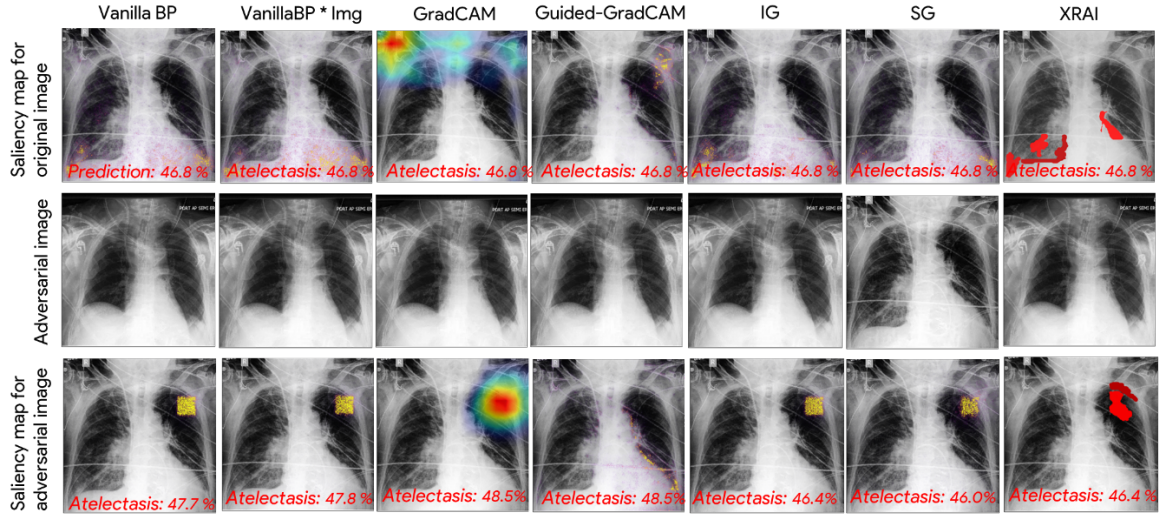


Supplementary Fig. S14. Saliency maps lacking resistance to saliency tampering on edema. An example of manipulating saliency maps towards a target region while the original prediction probability of the atelectasis remains unaffected.

Supplementary Table S14. Quantification results of saliency resistance on edema. Resistance evaluation on ResNet-152 and DenseNet-121 are reported. $SSIM_{org}$ and $SSIM_{tgt}$ denote the SSIMs of the new saliency map with the original saliency map and the target saliency map, respectively.

Models		Vanilla BP	Vanilla BP * Image	GradCAM	Guided GradCAM	IG	SG	XRAI
ResNet-152	AUC remains at 0.91	0.92	0.92	0.92	0.92	0.92	0.92	0.92
	Saliency map similarity	$SSIM_{org}$	0.37	0.40	0.33	0.38	0.40	0.43
		$SSIM_{tgt}$	0.94	0.94	0.61	0.74	0.94	0.85
	PSC	0.01	0.01	0.04	0.06	0.03	0.06	0.04
DenseNet-121	AUC remains at 0.91	0.92	0.92	0.92	0.92	0.92	0.92	0.92
	Saliency map similarity	$SSIM_{org}$	0.37	0.40	0.32	0.40	0.40	0.51
		$SSIM_{tgt}$	0.94	0.94	0.54	0.89	0.93	0.84
	PSC	0.01	0.01	0.14	0.12	0.07	0.14	0.08

5) Pleural Effusion



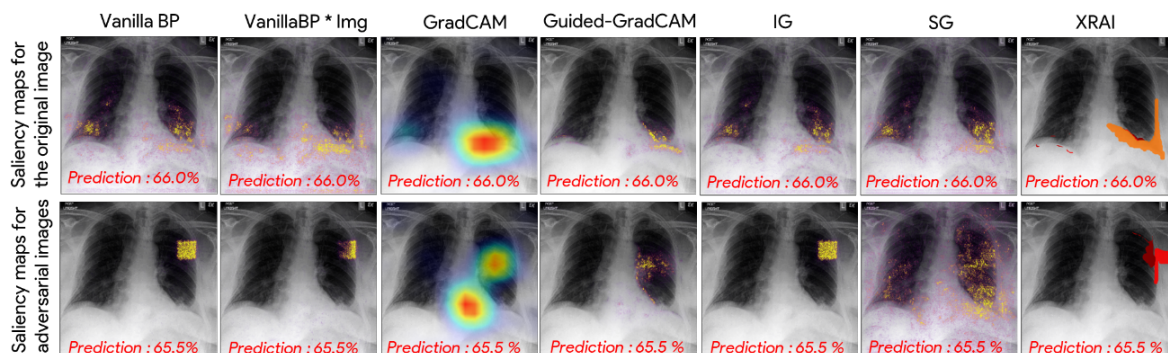
Supplementary Fig. S15. Saliency maps lacking resistance to saliency tampering on pleural effusion. An example of manipulating saliency maps towards a target region while the original prediction probability of the atelectasis remains unaffected.

Supplementary Table S15. Quantification results of saliency resistance on pleural effusion. Resistance evaluation on ResNet-152 and DenseNet-121 are reported. $SSIM_{org}$ and $SSIM_{tgt}$ denote the SSIMs of the new saliency map with the original saliency map and the target saliency map, respectively.

Models		Vanilla BP	Vanilla BP * Image	GradCAM	Guided GradCAM	IG	SG	XRAI
ResNet-152	AUC remains at 0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93
	Saliency map similarity	$SSIM_{org}$	0.68	0.72	0.33	0.38	0.47	0.47
		$SSIM_{tgt}$	0.94	0.93	0.73	0.74	0.93	0.85
	PSC	0.01	0.01	0.04	0.06	0.03	0.06	0.04
DenseNet-121	AUC remains at 0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93
	Saliency map similarity	$SSIM_{org}$	0.69	0.70	0.33	0.40	0.47	0.48
		$SSIM_{tgt}$	0.94	0.93	0.72	0.89	0.93	0.83
	PSC	0.01	0.01	0.10	0.12	0.07	0.12	0.09

4. Transferable resistance evaluation across different XAIs

1) Atelectasis

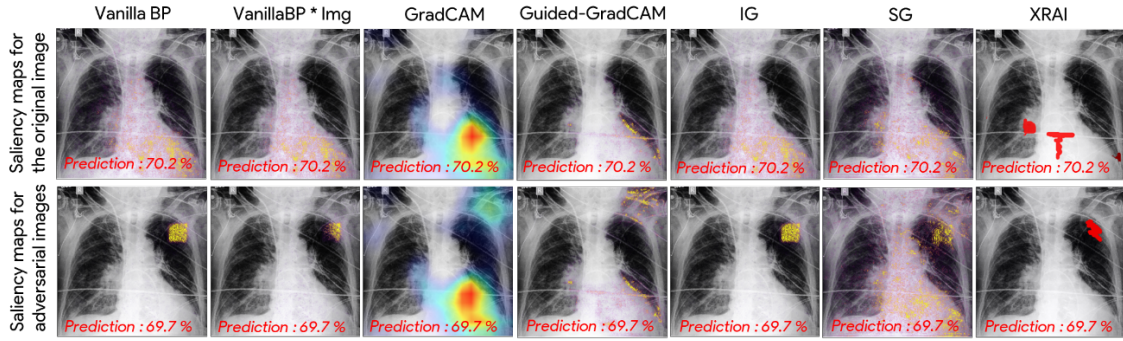


Supplementary Fig. S16. The lack of resistance is transferable across different saliency methods. An adversarial image is generated via attacking the IG on atelectasis first, then evaluated on all other saliency methods.

Supplementary Table S16. Resistance quantification across saliency methods on atelectasis. Resistance evaluation result on ResNet-152 and DenseNet-121 by transferring images generated for IG to other saliency methods.

Models	Evaluation metrics		IG	Transfer to other XAI					
				Vanilla BP	Vanilla BP * Image	GradCAM	Guided GradCAM	SG	XRAI
ResNet-152	AUC remains at 0.88		0.88	0.88	0.88	0.88	0.88	0.88	0.88
	Saliency map similarity	SSIM _{org}	0.47	0.44	0.50	0.75	0.64	0.58	0.46
		SSIM _{igt}	0.93	0.89	0.88	0.60	0.62	0.27	0.76
	PSC		0.03	0.01	0.02	0.27	0.21	0.34	0.04
DenseNet-121	AUC remains at 0.88		0.88	0.88	0.88	0.88	0.88	0.88	0.88
	Saliency map similarity	SSIM _{org}	0.47	0.50	0.57	0.70	0.59	0.62	0.47
		SSIM _{igt}	0.93	0.90	0.88	0.28	0.56	0.34	0.72
	PSC		0.07	0.07	0.18	0.17	0.07	0.16	0.17

2) Cardiomegaly

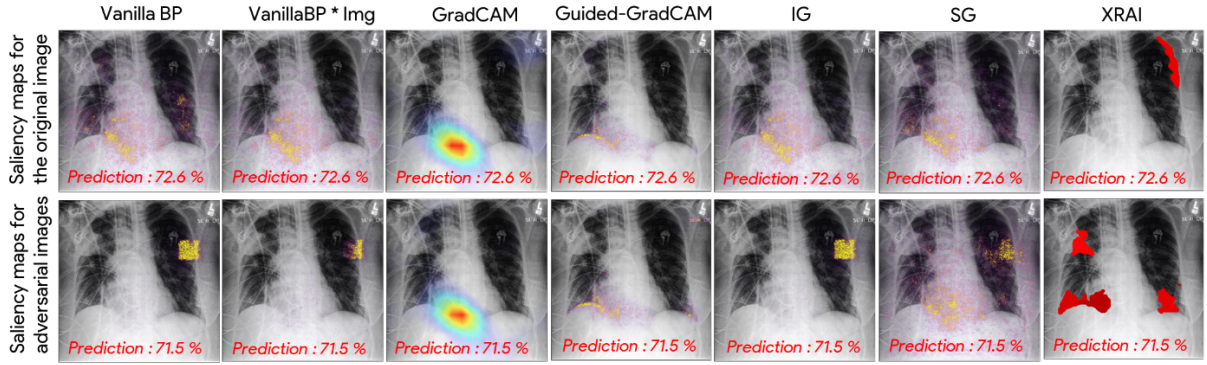


Supplementary Fig. S17. The lack of resistance is transferable across different saliency methods. An adversarial image is generated via attacking the IG on cardiomegaly first, then evaluated on all other saliency methods.

Supplementary Table S17. Resistance quantification across saliency methods on cardiomegaly. Resistance evaluation result on ResNet-152 and DenseNet-121 by transferring images generated for IG to other saliency methods .

Models	Evaluation metrics		IG	Transfer to other XAI					
				Vanilla BP	Vanilla BP * Image	GradCAM	Guided GradCAM	SG	XRAI
ResNet-152	AUC remains at 0.88		0.88	0.88	0.88	0.88	0.88	0.88	0.88
	Saliency map similarity	SSIM _{org}	0.47	0.44	0.50	0.75	0.64	0.58	0.46
		SSIM _{tgt}	0.93	0.89	0.88	0.60	0.62	0.27	0.76
	PSC		0.03	0.01	0.02	0.27	0.21	0.34	0.04
DenseNet-121	AUC remains at 0.88		0.88	0.88	0.88	0.88	0.88	0.88	0.88
	Saliency map similarity	SSIM _{org}	0.47	0.50	0.57	0.70	0.59	0.62	0.47
		SSIM _{tgt}	0.93	0.90	0.88	0.28	0.56	0.34	0.72
	PSC		0.07	0.07	0.18	0.17	0.07	0.16	0.17

3) Consolidation

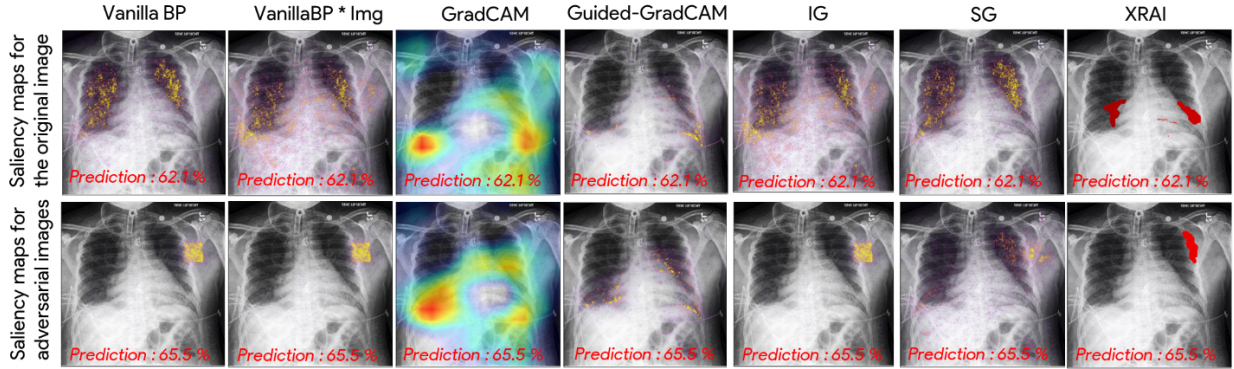


Supplementary Fig. S18. The lack of resistance is transferable across different saliency methods. An adversarial image is generated via attacking the IG on consolidation first, then evaluated on all other saliency methods.

Supplementary Table S18. Resistance quantification across saliency methods on consolidation. Resistance evaluation result on ResNet-152 and DenseNet-121 by transferring images generated for IG to other saliency methods.

Models	Evaluation metrics		IG	Transfer to other XAI					
				Vanilla BP	Vanilla BP * Image	GradCAM	Guided GradCAM	SG	XRAI
ResNet-152	AUC remains at 0.88		0.88	0.88	0.88	0.88	0.88	0.88	0.88
	Saliency map similarity	SSIM _{org}	0.47	0.44	0.50	0.75	0.64	0.58	0.46
		SSIM _{tgt}	0.93	0.89	0.88	0.60	0.62	0.27	0.76
	PSC		0.03	0.01	0.02	0.27	0.21	0.34	0.04
DenseNet-121	AUC remains at 0.88		0.88	0.88	0.88	0.88	0.88	0.88	0.88
	Saliency map similarity	SSIM _{org}	0.47	0.50	0.57	0.70	0.59	0.62	0.47
		SSIM _{tgt}	0.93	0.90	0.88	0.28	0.56	0.34	0.72
	PSC		0.07	0.07	0.18	0.17	0.07	0.16	0.17

4) Edema

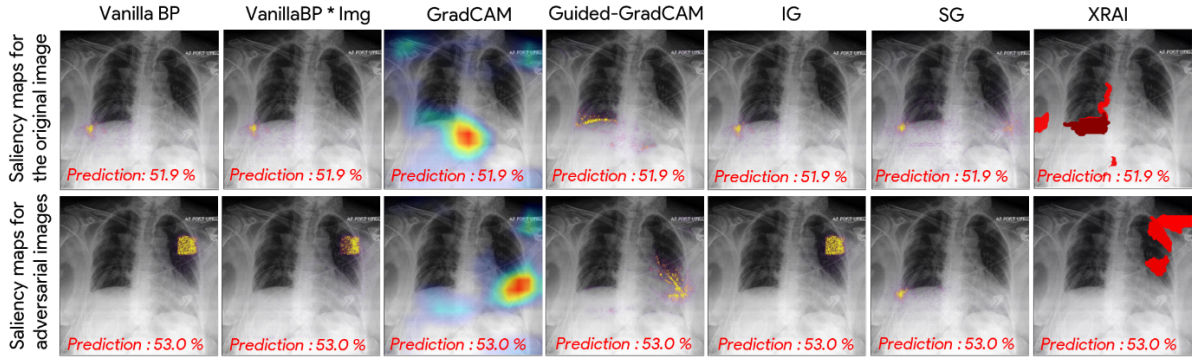


Supplementary Fig. S19. The lack of resistance is transferable across different saliency methods. An adversarial image is generated via attacking the IG on edema first, then evaluated on all other saliency methods.

Supplementary Table S19. Resistance quantification across saliency methods on edema. Resistance evaluation result on ResNet-152 and DenseNet-121 by transferring images generated for IG to other saliency methods.

Models	Evaluation metrics		IG	Transfer to other XAI					
				Vanilla BP	Vanilla BP * Image	GradCAM	Guided GradCAM	SG	XRAI
ResNet-152	AUC remains at 0.88		0.88	0.88	0.88	0.88	0.88	0.88	0.88
	Saliency map similarity	SSIM _{org}	0.47	0.44	0.50	0.75	0.64	0.58	0.46
		SSIM _{igt}	0.93	0.89	0.88	0.60	0.62	0.27	0.76
	PSC		0.03	0.01	0.02	0.27	0.21	0.34	0.04
DenseNet-121	AUC remains at 0.88		0.88	0.88	0.88	0.88	0.88	0.88	0.88
	Saliency map similarity	SSIM _{org}	0.47	0.50	0.57	0.70	0.59	0.62	0.47
		SSIM _{igt}	0.93	0.90	0.88	0.28	0.56	0.34	0.72
	PSC		0.07	0.07	0.18	0.17	0.07	0.16	0.17

5) Pleural Effusion



Supplementary Fig. S20. The lack of resistance is transferable across different saliency methods. An adversarial image is generated via attacking the IG on pleural effusion first, then evaluated on all other saliency methods.

Supplementary Table S20. Resistance quantification across saliency methods on pleural effusion. Resistance evaluation result on ResNet-152 and DenseNet-121 by transferring images generated for IG to other saliency methods.

Models	Evaluation metrics		IG	Transfer to other XAI					
				Vanilla BP	Vanilla BP * Image	GradCAM	Guided GradCAM	SG	XRAI
ResNet-152	AUC remains at 0.88		0.88	0.88	0.88	0.88	0.88	0.88	0.88
	Saliency map similarity	SSIM _{org}	0.47	0.44	0.50	0.75	0.64	0.58	0.46
		SSIM _{tgt}	0.93	0.89	0.88	0.60	0.62	0.27	0.76
	PSC		0.03	0.01	0.02	0.27	0.21	0.34	0.04
DenseNet-121	AUC remains at 0.88		0.88	0.88	0.88	0.88	0.88	0.88	0.88
	Saliency map similarity	SSIM _{org}	0.47	0.50	0.57	0.70	0.59	0.62	0.47
		SSIM _{tgt}	0.93	0.90	0.88	0.28	0.56	0.34	0.72
	PSC		0.07	0.07	0.18	0.17	0.07	0.16	0.17

Supplementary Section II: Why would the saliency method lack of trustworthiness? -- A theoretical explanation of saliency maps lacking relevance and resistance.

It is fascinating to further analyze why adversaries exist and how they decouple saliency maps from model predictions. Let $\mathbf{x}$ and $\mathbf{x}_{adv}$ denote an image and its adversary with a small perturbation $\boldsymbol{\delta}$, respectively,

$$\mathbf{x}_{adv} = \mathbf{x} + \boldsymbol{\delta}, \|\boldsymbol{\delta}\|_p \leq \varepsilon_i \ll 1,$$

where ε is a very small value. Without loss of generality, we denote the saliency map as $\mathbf{S}(\mathbf{x}) = \nabla_{\mathbf{x}} g(\mathbf{x})$ and the Hessian matrix as $\mathbf{H}(\mathbf{x}) = \nabla_{\mathbf{x}\mathbf{x}}^2 g(\mathbf{x})$, where $g(\mathbf{x})$ is the network output logit associated with the annotated class label. For a pair of images, an original image $\mathbf{x}$ and its adversarial image $\mathbf{x}_{adv}$, the difference between model prediction losses can be formulated by applying Taylor expansion as

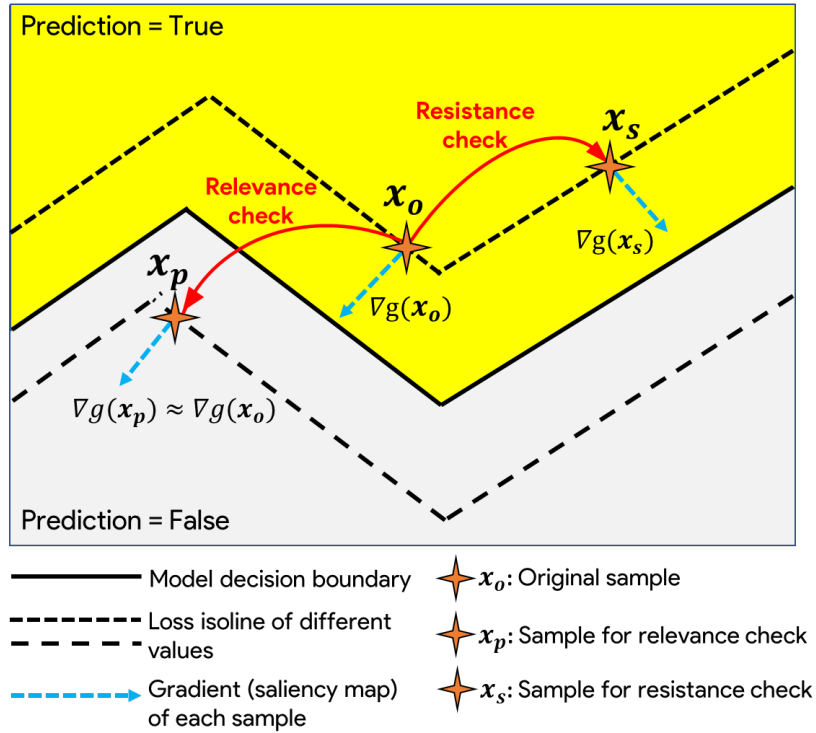
$$g(\mathbf{x}_{adv}) - g(\mathbf{x}) = \boldsymbol{\delta}^T \nabla_{\mathbf{x}} g(\mathbf{x}) + \boldsymbol{\delta}^T \mathbf{H}(\mathbf{x}) \boldsymbol{\delta} + o(\boldsymbol{\delta}^2),$$

and the difference between the L_p -norm saliency map changes is defined as

$$\|\mathbf{S}(\mathbf{x}_{adv}) - \mathbf{S}(\mathbf{x})\|_p = \|\nabla_{\mathbf{x}} g(\mathbf{x} + \boldsymbol{\delta}) - \nabla_{\mathbf{x}} g(\mathbf{x})\|_p = \|\boldsymbol{\delta}^T \mathbf{H}(\mathbf{x}) + o(\boldsymbol{\delta}^2)\|_p.$$

Comparing the above two equations, we can see that the main difference is the term $\boldsymbol{\delta}^T \nabla_{\mathbf{x}} g(\mathbf{x})$, which describes the first-order local smoothness of the loss landscape regarding the input image space. If the loss landscape is not first-order smooth, this term can have large values, which will cause the discrepancy between the output logits and the saliency maps. Thus, **the lack of relevance** is caused by the lack of first-order local smoothness of a regularly trained deep neural network in the image space. As presented in Eq. (2), the adversarial image for relevance evaluation is generated via maximizing the prediction loss $L(\mathbf{x}_{adv}, \mathbf{y})$, while remaining the similarity between $\mathbf{S}(\mathbf{x}_{adv})$ and $\mathbf{S}(\mathbf{x})$. Optimizing Eq. (2) leads to $\|\mathbf{S}(\mathbf{x}_{adv}) - \mathbf{S}(\mathbf{x})\|_p \leq \varepsilon_s \ll 1$. We thus have $g(\mathbf{x}_{adv}) - g(\mathbf{x}) > \boldsymbol{\delta}^T \nabla_{\mathbf{x}} g(\mathbf{x}) - \|\boldsymbol{\delta}^T \mathbf{H}(\mathbf{x}) + o(\boldsymbol{\delta}^2)\|_p \cdot \|\boldsymbol{\delta}\|_p > \boldsymbol{\delta}^T \nabla_{\mathbf{x}} g(\mathbf{x}) - \varepsilon_i \cdot \varepsilon_s$, where $\varepsilon_i \cdot \varepsilon_s$ is sufficiently small to be ignorable. As discovered by several

previous studies on adversarial attacks^{24,25}, $\delta^T \nabla_x g(\mathbf{x})$ can be large enough to flip the original prediction within such imperceptible perturbation range ε_i . **Supplementary Fig. S21** illustrates this problem. The adversarial sample $\mathbf{x}_p$ is intentionally generated for evaluating the relevance of saliency method. It can be seen that $\mathbf{x}_p$ may be generated by pushing the original image $\mathbf{x}_o$ across the model decision boundary along the gradient ascent direction. Thus, the model will produce an opposite prediction result while keeping the saliency map unchanged, since the gradient orientations are not modified ($\nabla g(\mathbf{x}_p) \approx \nabla g(\mathbf{x}_o)$).



Supplementary Figure S21. Illustration of the lack of relevance and resistance. The decision boundary and contours corresponding to a loss function landscape for a binary classification task are presented. Neural networks with many parameters have decision boundaries that are roughly piecewise linear with many transitions²⁴. Consider an original sample $\mathbf{x}_o$ with that is slightly perturbed to two new positions $\mathbf{x}_p$ and $\mathbf{x}_s$. The former will belong to a different prediction, but the gradient ($\nabla g(\mathbf{x}_p)$) remains the same as $\nabla g(\mathbf{x}_o)$. The latter stays in the same class but with very different gradient $\nabla g(\mathbf{x}_s)$.

The lack of resistance can be attributed to the fact that deep neural networks lack the second order (Hessian) local smoothness in the image space. As shown in **Eqn. (4)**, adversarial images

for evaluating resistance can be generated via minimizing the difference between the adversarial saliency map $\mathbf{S}(\mathbf{x}_{adv})$ and the target saliency map $\mathbf{S}_t$ while keeping the similarity between the model output logits $g(\mathbf{x}_{adv})$ and $g(\mathbf{x})$. The setting of $\|\delta\|_p \leq \varepsilon_i \ll 1$ indicates that the magnitude of the perturbation δ is an imperceptible value. Thus, we have $\|\delta^T \nabla_x g(\mathbf{x})\|_p \gg \|\delta^T \mathbf{H}(\mathbf{x}) \delta\|_p$. When minimizing the logit difference $g(\mathbf{x}_{adv}) - g(\mathbf{x})$, the optimizer will focus more on minimizing $\|\delta^T \nabla_x g(\mathbf{x})\|_p$ (the first-order local smoothness) rather than the second-order smoothness term $\|\delta^T \mathbf{H}(\mathbf{x}) \delta\|_p$. This results in maximizing the saliency difference between $\mathbf{S}(\mathbf{x}_{adv})$ and $\mathbf{S}(\mathbf{x}_o)$, while maintaining the model prediction output unchanged. An intuitive case is shown in **Supplementary Fig. S21**, where an adversarial sample $\mathbf{x}_s$ is generated for evaluating the resistance. Previous works²⁴ observed that deep neural networks with many parameters have decision boundaries that are roughly piecewise linear with many transitions. The drastic change of the saliency map at local data point can be intrinsically caused by the inconsistency of the non-linear activation functions such as ReLU. It can be seen in **Fig. S1** that $\mathbf{x}_s$ may be generated by pushing the original image $\mathbf{x}$ to a new location close to another transition edge within the same decision region but having a different gradient orientation. Although the model will still give $\mathbf{x}_s$ the same prediction, the saliency map can be completely different from the original sample due to the change of gradients. The observation also implies that a potential solution to mitigate the revealed issues could be using second-order Hessian smoothing with adversarial training.