This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

A Supplementary Material

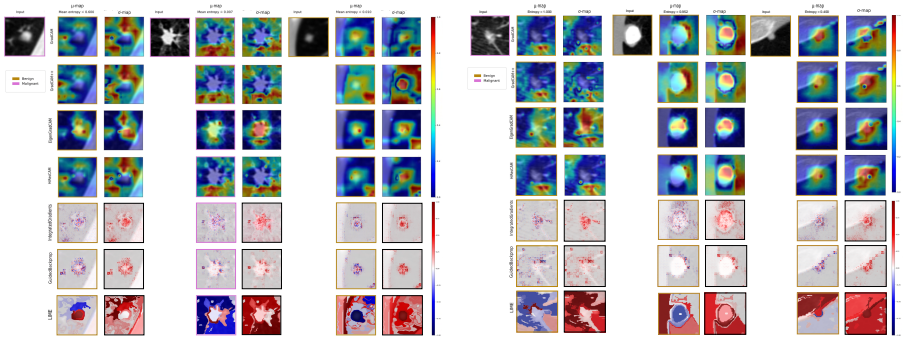

Fig. 9. Some interesting example nodules alongside the generated μ-maps and σ-maps per XAI method. See caption of Figure 4 for an explanation of the elements in this figure.

Method	F-value		p-value		Sanity check passed?
	μ-map	σ-map	μ-map	σ-map	
GradCAM	1.78	0.86	0.15	0.46	✓
GradCAM++	2.30	2.95	0.08	0.03*	✓
EigenGradCAM	18.40	15.80	$p \ll 0$**	$p \ll 0$**	✗
HiResCAM	2.19	4.35	0.09	0.01**	✗
Integrated Gradients	5.04	11.20	0**	$p \ll 0$**	✗
Guided Backpropagation	2.50	5.81	0.06	0**	✗
LIME	2.44	1.10	0.07	0.35	✓

*Statistical significance measured against the significance level $\alpha = 0.05$.
**Statistical significance measured against the significance level $\alpha = 0.01$; in these cases, we can reject H_0.

Table 2. Results of Repeated Measures ANOVA on each XAI method's SSIM scores after the five-crop sanity check. The results are given for the μ-maps and σ-maps.

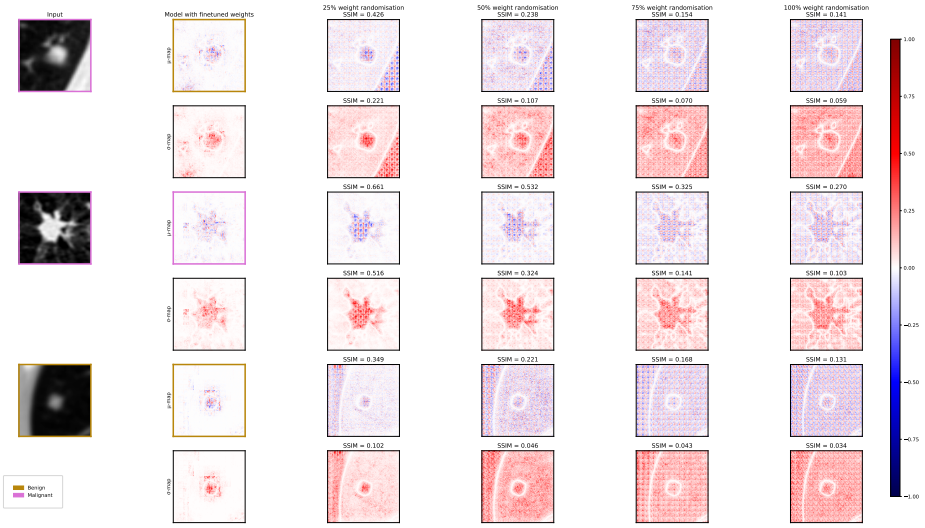

Fig. 10. 3 sample nodules with SSIM scores for each saliency map generated by the weight randomisation sanity check for Integrated Gradients. The first column shows the input to the model, with the border colour representing the true label: *benign* or *malignant*. The nodule in the second column is taken as the 'ground truth', with the border colour representing the predicted class. The third to sixth columns show the saliency maps per % weight randomisation. The odd-numbered rows show the μ-maps, and the even-numbered rows show the σ-maps. The colour scheme used to visualise the saliency maps is shown on the right.

Method	**Page's L test statistic**		**p-value**		**Sanity check passed?**
	μ-map	σ-map	μ-map	σ-map	
GradCAM	286	289	$p \ll 0^{**}$	$p \ll 0^{**}$	✓
GradCAM++	287	290	$p \ll 0^{**}$	$p \ll 0^{**}$	✓
EigenGradCAM	230	237	0.99	0.93	✗
HiResCAM	293	290	$p \ll 0^{**}$	$p \ll 0^{**}$	✓
Integrated Gradients	300	300	$p \ll 0^{**}$	$p \ll 0^{**}$	✓
Guided Backpropagation	289	284	$p \ll 0^{**}$	$p \ll 0^{**}$	✓
LIME	227	255	1.00	0.31	✗

**Statistical significance is measured against the significance level $\alpha = 0.01$.

Table 3. Results of Page's L test on each XAI method's SSIM scores after the weight randomisation sanity check for the μ-maps and σ-maps.