

Comparing Four Explanation Methods for Melanoma Classification: Fidelity, Utility and Trust

Okoi Michael Obeten

michealobeten@auchipoly.edu.ng

Auchi Polytechnic

Christiana Uchenna Ezeanya

National Open University of Nigeria

Research Article

Keywords: Explainable artificial intelligence, Melanoma, Dermoscopy, Saliency maps, Faithfulness, Trust calibration

Posted Date: October 27th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11323559/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Comparing Four Explanation Methods for Melanoma Classification: Fidelity, Utility and Trust

Okoi Michael Obeten^{1,*} and Christiana Uchenna Ezeanya²

¹Department of Artificial Intelligence, Auchi Polytechnic, Auchi, Edo State, Nigeria

²Department of Information Systems & Technology, Faculty of Computing, National Open University of Nigeria, Abuja, Nigeria

michealobeten@auchipoly.edu.ng; cezeanya@noun.edu.ng;

*Corresponding author: Okoi Michael Obeten, michealobeten@auchipoly.edu.ng

Abstract

Post-hoc explanation methods are used to make skin-lesion classifiers interpretable but are usually judged on one criterion, most often whether the heatmap looks plausible. We asked whether technical fidelity, a clinical utility proxy and computational trust calibration rank explanation methods the same way. An EfficientNet-B0 melanoma classifier was trained on the 10,015 HAM10000 dermoscopic images with a lesion-level split, and four methods (Grad-CAM, Integrated Gradients, LIME and occlusion sensitivity) were applied to 800 stratified test images. Explanations were compared with lesion masks using intersection-over-union under four binarisation rules, pointing-game accuracy, deletion and insertion curves, and a cascading parameter-randomisation sanity check. Utility was measured by lesion hit rate, centroid distance and artefact attribution, and trust by the fidelity gap between correct and incorrect predictions. The classifier reached a test area under the receiver operating characteristic curve of 0.9295. LIME and occlusion led on localisation and could not be separated statistically (Holm-corrected $p = 0.465$). Integrated Gradients ranked second on pointing game and centroid distance but failed the sanity check, keeping a rank correlation of 0.33 with its original map after seven of nine randomisation stages. All 12 trust comparisons were negative; stratifying by predicted class showed that fidelity tracked model confidence, and only Grad-CAM kept a significant positive difference, which may itself partly reflect confidence. Agreement between methods peaked at a rank correlation of 0.22. The axes disagree enough to change which method would be chosen, so multi-axis evaluation with a faithfulness gate should become standard.

Keywords: Explainable artificial intelligence; Melanoma; Dermoscopy; Saliency maps; Faithfulness; Trust calibration

1 Introduction

Deep convolutional neural networks (CNNs) can classify dermoscopic images at or near the level of dermatologists [1]. The clinical stakes are high. Cutaneous melanoma causes most skin-cancer deaths, and its global incidence and mortality are projected to rise substantially by 2040 [2]. Automated triage could be most valuable where specialist dermatologists are scarce. Yet CNNs remain opaque: they give a prediction without an account of why, and this opacity is widely cited as a barrier to clinical adoption and regulatory approval [3-5].

Post-hoc explanation methods respond to this opacity by producing a saliency map for each prediction. Gradient-based methods such as Gradient-weighted Class Activation Mapping (Grad-CAM) [6] and Integrated Gradients [7] derive the map from the network's gradients. Perturbation-based methods such as Local Interpretable Model-agnostic Explanations (LIME) [8] and occlusion sensitivity [9] treat the network as a black box and change the input to see what matters. In dermatology, these maps are mostly used to show that a classifier "looks at" the lesion, and Grad-CAM is by far the most common choice [10, 11]. Recent HAM10000 studies continue this pattern [12-14], and 2026 reviews note that such maps are rarely tested quantitatively [15, 16].

Visual plausibility, however, is a weak test. Adebayo et al. showed that some saliency methods produce almost the same maps for a trained network and for a network with randomised weights [17], so a convincing map need not reflect what the model learned. Benchmarking in chest radiography found that common saliency methods fall short of expert localisation [18], and critical work on evaluation metrics has shown that the metrics themselves can disagree [19]. Human-subjects studies add a further dimension. Explanation type can change physicians' diagnostic accuracy and trust, sometimes for the worse [20], and the relationship between explanation and trust is not monotonic [21]; current guidance therefore aims at appropriate rather than maximal reliance [22]. It is therefore now common to argue that explanations should be judged on several axes at once: faithfulness to the model, usefulness to the clinician, and the appropriateness of the trust they produce [23, 24].

Few studies do this in practice. Comparative work in melanoma has focused on agreement between explanation maps and regions marked by dermatologists [25], without formal faithfulness tests or a trust measure. Faithfulness benchmarks rarely

67 consider trust, and trust studies rarely test faithfulness. As a result, it is unclear
68 whether the axes agree, and whether a practitioner who evaluates only one of them
69 would choose well.

70 This study evaluates four representative post-hoc methods, two from each family,
71 across three axes within one reproducible framework built on a single model, a single
72 lesion-level data partition and one set of lesion masks. Our hypothesis was that the
73 three axes would not produce the same ranking of methods. The contributions are: (i)
74 a fully computational three-axis protocol with a faithfulness gate, a localisation
75 threshold sweep and a confound-controlled trust measure; (ii) evidence that the axes
76 disagree enough to change which method would be selected, including a method that
77 performs well on localisation while failing a basic faithfulness test; (iii) a demonstration
78 that a fixed binarisation threshold can reverse method rankings on intersection-over-
79 union; (iv) evidence that a naive trust-calibration measure tracks model confidence
80 rather than correctness unless it is stratified by predicted class; and (v) practical
81 guidance that weighs explanation quality against computational cost.

82 **2 Materials and Methods**

83 **2.1 Study Design**

84 Figure 1 summarises the pipeline. The HAM10000 images and lesion masks were
85 partitioned by lesion (Stage 1). A CNN was fine-tuned for binary melanoma
86 classification (Stage 2). Four explanation methods were applied to a stratified subset
87 of the test partition (Stage 3). Every (method, image) explanation was then scored on
88 three axes (Stage 4), and the results were combined into inter-method agreement,
89 cross-axis rankings and cost estimates (Stage 5).

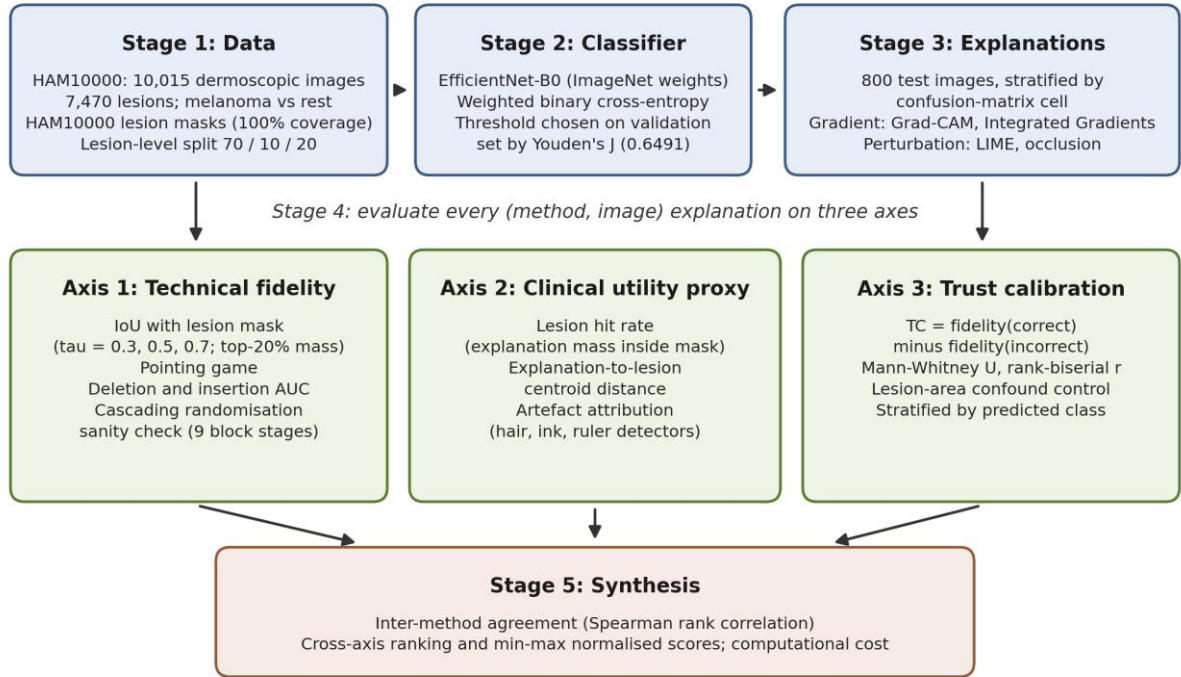


Fig. 1 Unified three-axis evaluation pipeline. All three axes are computed from the same classifier, the same test images and the same lesion masks

2.2 Data and Partitioning

All images came from HAM10000 [26], a public collection of 10,015 dermoscopic images at 600×450 pixels spanning seven diagnostic categories, which also forms the Task 3 classification set of the ISIC 2018 Challenge [27]. Melanoma (1,113 images, 11.11%) was the positive class, and the other six categories (8,902 images) formed the negative class. Diagnoses in HAM10000 were confirmed by histopathology for more than half of the lesions and by follow-up examination, expert consensus or in-vivo confocal microscopy for the rest [26]. Lesion masks came from the HAM10000 lesion segmentations [28], which give one binary mask per image. The masks were drafted by a fully convolutional network and then reviewed and corrected by a dermatologist. Because images and masks share identifiers and come from the same release, their correspondence is complete; coverage was verified at run time and was 100%. The collection is distributed under a CC BY-NC-SA 4.0 licence and contains no directly identifying patient information.

The 10,015 images correspond to 7,470 distinct lesions, so several images show the same lesion. To avoid leakage between partitions, a stratified grouped split keyed to lesion identifier assigned approximately 70%, 10% and 20% of lesions to training, validation and test partitions, with a fixed seed (Table 1). No lesion identifier appeared in more than one partition. HAM10000 does not release patient identifiers, so the split

prevents lesion-level leakage but cannot exclude different lesions from the same patient falling in different partitions. Data curation, reference standard, partitioning and training are reported following the SLIM best-practice checklist for classification papers [29].

Table 1 Lesion-level partition of HAM10000

Partition	Lesions	Images	Melanoma	Non-melanoma	Melanoma prevalence (%)
Training	5,229	7,019	806	6,213	11.48
Validation	747	1,001	100	901	9.99
Test	1,494	1,995	207	1,788	10.38
Total	7,470	10,015	1,113	8,902	11.11

Images were resized bilinearly to 224×224 pixels and normalised with ImageNet channel statistics. Masks were resized with nearest-neighbour interpolation and re-thresholded at 0.5.

2.3 Classifier

An ImageNet-pretrained EfficientNet-B0 [30] with a single output unit was fine-tuned with binary cross-entropy in which the positive class was weighted by the negative-to-positive ratio of the training partition ($w = 7.708$):

$$L = -\frac{1}{N} \sum_{i=1}^N [w y_i \log \sigma(\hat{y}_i) + (1 - y_i) \log(1 - \sigma(\hat{y}_i))] \quad (1)$$

where $\hat{y}_i$ is the logit, $y_i \in \{0,1\}$ the label and σ the logistic function. Optimisation used AdamW (learning rate 3×10^{-4} , weight decay 10^{-4}), batch size 32, mixed precision, and a learning rate halved when validation area under the receiver operating characteristic curve (AUC-ROC) did not improve for three epochs. Training augmentation comprised horizontal and vertical flips, 90-degree rotations, shift-scale-rotate transforms and mild colour jitter. Training ran for at most 20 epochs with early stopping after five epochs without improvement, and the checkpoint with the best validation AUC-ROC was kept.

Class weighting inflates the sigmoid output, so a threshold of 0.5 would over-call melanoma. The operating threshold was therefore chosen on the validation partition by maximising Youden’s J statistic [31] and applied unchanged to the test partition. This choice also defines which test images are correct and which are incorrect, and hence the two groups compared on the trust axis. The same validation partition was used for checkpoint selection and threshold selection. Test-set confidence intervals were obtained from 1,000 bootstrap resamples.

2.4 Explanation Methods

Let x be an input image of size $H \times W$, f the trained network, A the activation tensor of the final convolutional layer with channels indexed by k , and E an explanation map. Because the network has one output unit, every method explains the melanoma logit y^c . For a lesion that the model scores as benign, a map therefore shows which regions pushed the score towards melanoma, not which regions supported the benign decision.

Grad-CAM [6] weights each activation channel by its spatially averaged gradient and keeps positive evidence:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_k(i, j)}, \quad E_{\text{Grad-CAM}} = \text{ReLU} \left(\sum_k \alpha_k^c A_k \right) \quad (2)$$

where Z is the number of spatial positions. For EfficientNet-B0 at 224×224 input the final feature map is 7×7 , and the map was upsampled bilinearly to the input size.

Integrated Gradients [7] accumulates gradients along a straight path from a baseline x' to the input:

$$\text{IG}_i(x) = (x_i - x'_i) \int_0^1 \frac{\partial f(x' + \alpha(x - x'))}{\partial x_i} d\alpha \quad (3)$$

The integral was approximated with 50 steps. The baseline was the zero tensor in the normalised input space, which corresponds to the ImageNet mean image rather than a black image. Attributions were summed over colour channels and normalised by absolute value.

LIME [8] segments the image into superpixels (Quickshift), queries the network on perturbed images in which subsets of superpixels are removed, and fits a sparse weighted linear surrogate g :

$$g^* = \underset{g \in G}{\text{argmin}} \sum_z \pi_x(z) [f(z) - g(z)]^2 + \Omega(g) \quad (4)$$

where π_x is an exponential proximity kernel and Ω an L1 sparsity penalty. Quickshift used a kernel size of 4, a maximum distance of 200 and a ratio of 0.2, and 1,000 perturbation samples were drawn per image. Positive surrogate coefficients were projected back to their superpixels to form the map.

166 *Occlusion sensitivity* [9] replaces a $p \times p$ patch centred at (i, j) with a neutral value
 167 (the zero tensor in normalised space) and records the drop in the class score:

$$168 \quad E_{\text{occ}}(i, j) = f^c(x) - f^c(x^{(i, j)}) \quad (5)$$

169 with a 32-pixel window and 16-pixel stride. Grad-CAM (Captum LayerGradCam),
 170 Integrated Gradients and occlusion were implemented with Captum [32]; LIME used
 171 its reference package. Every map was min-max normalised to $[0, 1]$ before evaluation.

172 **2.5 Evaluation Subset and Artefact Regions**

173 LIME was too slow to run on the full test partition, so all four methods were run on a
 174 stratified subset of 800 test images. Stratification over the four confusion-matrix cells
 175 kept the proportions of the test partition and guaranteed enough misclassified images
 176 for the trust analysis.

177 HAM10000 has no annotation of imaging artefacts, so artefact regions were derived
 178 with three heuristic detectors at 224×224 resolution. Hair was detected with a
 179 morphological black-hat transform (17×17 rectangular structuring element), a
 180 threshold of 12 grey levels and a closing operation. Ink markings were detected by
 181 thresholding the hue-saturation-value colour space between $(100, 80, 30)$ and $(150,$
 182 $255, 255)$ in OpenCV units, which covers blue to violet. Rulers were detected by
 183 searching the outer 6% of each border for strips with mean intensity above 200 and
 184 standard deviation below 40. The union of the three masks formed the artefact mask.
 185 Air bubbles were not included because no reliable detector was available.

186 **2.6 Axis 1: Technical Fidelity**

187 *Localisation.* With M the lesion mask and E_τ the map binarised at threshold τ ,

$$188 \quad \text{IoU}(E, M) = \frac{|E_\tau \cap M|}{|E_\tau \cup M|} \quad (6)$$

189 The primary threshold was $\tau = 0.5$. Because any single threshold is arbitrary, IoU was
 190 also computed at $\tau = 0.3$ and $\tau = 0.7$, and under a mass-based rule that keeps the
 191 smallest set of pixels holding the top 20% of explanation mass. The pointing game
 192 scored 1 if the pixel of maximum explanation value lay inside the mask and 0
 193 otherwise.

194 *Deletion and insertion.* Following Petsiuk et al. [33], pixels were ordered by decreasing
 195 explanation value and, over $N = 50$ steps, either replaced by a Gaussian-blurred copy

of the image (kernel 51×51 , $\sigma = 11$) (deletion) or inserted into it (insertion). The deletion area under the curve (DAUC) and insertion area under the curve (IAUC) are the mean melanoma probabilities along each path. A faithful map gives a low DAUC and a high IAUC. Because class weighting leaves the probabilities uncalibrated, DAUC and IAUC are compared across methods on the same model rather than read as absolute probabilities.

Sanity check. A cascading model-parameter randomisation test [17] was applied at the level of EfficientNet-B0’s blocks rather than individual layers, since the backbone has more than 200 parameterised layers. Parameters were re-initialised cumulatively over nine stages, from the classifier head through the final convolution and blocks 6 to 0, and each method’s maps were recomputed on a fixed random subsample of 50 evaluation images (seed 42). To keep the test affordable, LIME used 250 perturbation samples during the sanity check. Original and new maps were resized to 56×56 and compared by Spearman rank correlation; when either map was constant, the correlation was recorded as zero. A faithful method should lose correlation as randomisation proceeds. We treated a method as passing if its mean correlation fell close to zero once the classifier head was randomised and stayed there through the later stages, and as failing if a substantial correlation persisted across several stages. This criterion was applied by inspection; no numerical threshold was fixed in advance. LIME used fewer samples in this test only because it was rerun on every randomised network.

2.7 Axis 2: Clinical Utility Proxy

Three measures approximated how helpful a map would be to a reader. The *lesion hit rate* is the fraction of total explanation mass inside the lesion mask. The *centroid distance* is the Euclidean distance, in pixels at 224×224 resolution, between the mass-weighted centroid of the map and the centroid of the mask. *Artefact attribution* is the fraction of extra-lesional explanation mass that falls on detected artefact pixels, reported for each artefact class and for their union.

2.8 Axis 3: Trust Calibration

Let C and W be the sets of correctly and incorrectly classified images and $S(E_i, M_i)$ a per-image fidelity score. The trust-calibration score is

227

$$TC = \frac{1}{|C|} \sum_{i \in C} S(E_i, M_i) - \frac{1}{|W|} \sum_{i \in W} S(E_i, M_i) \quad (7)$$

228

229

230

231

232

233

234

235

236

computed for IoU, hit rate and IAUC. A positive TC means explanations are stronger when the model is right, which is the property an explanation needs if it is to support appropriately calibrated trust [21, 34]. Two confounds were controlled. First, lesion area was compared between C and W , and the correlation between IoU and correctness was recomputed with lesion area partialled out. Second, because all methods explain the melanoma logit, TC was recomputed within each predicted class (true positives against false positives, and true negatives against false negatives), and the Spearman correlation between the predicted melanoma score and each fidelity measure was computed.

237

2.9 Synthesis and Statistics

238

239

240

241

242

243

244

245

Pairwise agreement between methods was measured by the mean per-image Spearman rank correlation between full maps. Each method was ranked on every criterion, and scores were also min-max normalised so that higher is better. Explanation generation time per image was recorded. Differences among the four methods on per-image IoU and hit rate were tested with the Friedman test, followed by pairwise Wilcoxon signed-rank tests with Holm correction [35]. Correct and incorrect groups were compared with two-sided Mann-Whitney U tests, with the rank-biserial correlation r as effect size. Significance was set at 0.05.

246

2.10 Implementation

247

248

249

250

251

252

253

The pipeline ran on Kaggle Notebooks with one NVIDIA Tesla T4 GPU, using Python 3.12.13, PyTorch 2.10.0 (CUDA 12.8), timm 1.0.26, Captum 0.9.0, lime 0.2, alumentations 1.4.24, scikit-learn 1.6.1, SciPy 1.16.3, NumPy 2.0.2 and OpenCV 4.13.0. The random seed was fixed at 42 for the partition, training and every stochastic step of explanation. Each stage saved its outputs to disk, and every table and figure was generated from these files through a manifest, so no reported value was transcribed by hand.

3 Results

3.1 Classifier Performance

Training stopped after 14 epochs; the checkpoint from epoch 9 (validation AUC-ROC 0.9128) was kept, after 12.6 min of GPU time. The validation-selected threshold was 0.6491. Table 2 reports performance.

Table 2 Classifier performance at the validation-selected threshold (0.6491). Test confidence intervals are 1,000-sample bootstrap percentiles

Metric	Validation	Test	Test 95% CI
AUC-ROC	0.9128	0.9295	0.9130 to 0.9463
AUC-PR	0.5482	0.6509	0.5820 to 0.7189
Accuracy	0.8282	0.8411	0.8246 to 0.8571
Balanced accuracy	0.8512	0.8494	0.8213 to 0.8762
Sensitivity	0.8800	0.8599	0.8079 to 0.9074
Specificity	0.8224	0.8389	0.8215 to 0.8558
Precision	0.3548	0.3820	0.3392 to 0.4289
Negative predictive value	0.9841	0.9810	0.9738 to 0.9881
F1	0.5057	0.5290	0.4818 to 0.5756

On the 1,995 test images there were 1,500 true negatives (TN), 288 false positives (FP), 29 false negatives (FN) and 178 true positives (TP). Errors were therefore overwhelmingly false positives (288 of 317, 90.9%), the expected profile of a high-sensitivity operating point on a low-prevalence task. The 800-image evaluation subset kept the same proportions: 602 TN, 115 FP, 12 FN and 71 TP, that is, 673 correct and 127 incorrect predictions.

3.2 Artefact Detection

Table 3 reports the prevalence of detected artefacts. The ink and ruler detectors fired on plausible fractions of images. The hair detector fired on every image and claimed almost a quarter of each frame. This is a property of the detector rather than of the images: a black-hat filter responds to any thin dark structure on a lighter background, including pigment network, dots and lesion borders.

Table 3 Heuristically detected artefacts in the 800-image evaluation subset

Artefact	Images with detection	Percent of subset	Mean coverage where present (%)
Hair	800	100.00	23.41
Ink	215	26.88	0.53
Ruler	84	10.50	13.65
Any	800	100.00	24.75

3.3 Axis 1: Technical Fidelity

Table 4 reports the fidelity measures. On IoU at $\tau = 0.5$, LIME (0.2487) and occlusion (0.2422) led, Grad-CAM reached 0.1150, and Integrated Gradients was near zero (0.0034). The Friedman test rejected equality of the four methods ($\chi^2 = 1283.92$, $df =$

3, $p = 4.54 \times 10^{-278}$). Five of the six pairwise Wilcoxon comparisons were significant after Holm correction (all $p < 10^{-51}$); the exception was LIME against occlusion (Holm-corrected $p = 0.465$), so their difference of 0.0065 cannot be distinguished from chance.

Table 4 Technical fidelity on the evaluation subset (mean $\pm$ standard deviation across 800 images; pointing game is a proportion)

Method	IoU ($\tau = 0.5$)	Pointing game	DAUC (lower is better)	IAUC (higher is better)
Grad-CAM	0.1150 ± 0.1748	0.3038	0.1250 ± 0.1749	0.1843 ± 0.2498
Integrated Gradients	0.0034 ± 0.0046	0.5337	0.1792 ± 0.2338	0.2617 ± 0.2912
LIME	0.2487 ± 0.1800	0.7475	0.0869 ± 0.1346	0.2965 ± 0.2913
Occlusion	0.2422 ± 0.1571	0.6587	0.1058 ± 0.1650	0.3182 ± 0.3010

The ranking depended on the measure. On the pointing game, Integrated Gradients (0.5337) moved well ahead of Grad-CAM (0.3038), reversing their IoU order. On the perturbation measures, LIME had the lowest DAUC and occlusion the highest IAUC. Grad-CAM was third on deletion and last on insertion, while Integrated Gradients was last on deletion and third on insertion. No method was best on all four measures. The mean deletion and insertion curves (Fig. 2) show LIME and occlusion causing the fastest early drop in melanoma probability, and Integrated Gradients causing a slight initial rise under deletion.

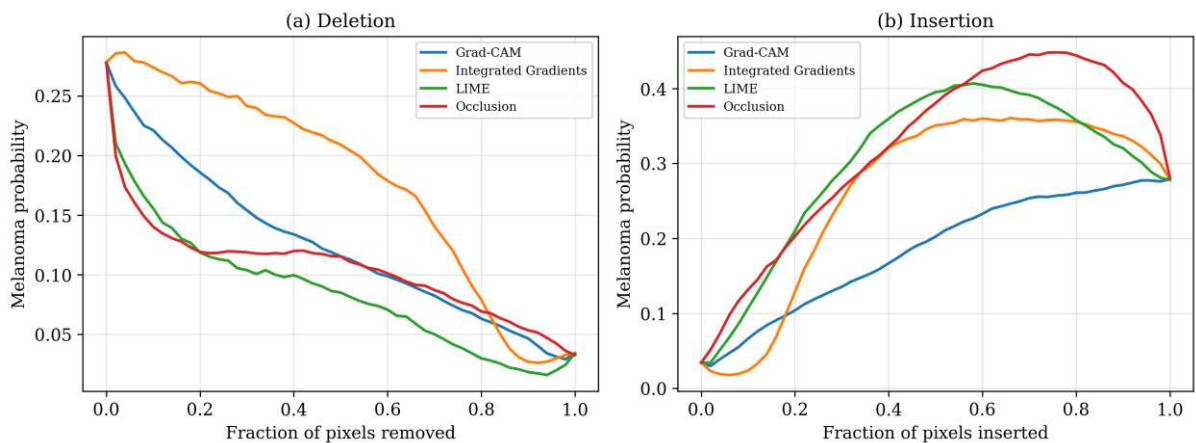


Fig. 2 Mean deletion (a) and insertion (b) curves on the evaluation subset. Faster falls under deletion and faster rises under insertion indicate more faithful maps

Table 5 shows how IoU depends on the binarisation rule. The order of LIME, occlusion and Grad-CAM was stable across fixed thresholds, but Integrated Gradients behaved differently. Its IoU fell from 0.0200 at $\tau = 0.3$ to 0.0007 at $\tau = 0.7$, yet under the top-20%-mass rule it recovered to 0.0712, close to Grad-CAM's 0.0879. Its near-zero IoU

at fixed thresholds is therefore an artefact of applying a fixed cut to a very sparse map,
not evidence that it fails to localise.

Table 5 Mean IoU under four binarisation rules

Method	$\tau = 0.3$	$\tau = 0.5$	$\tau = 0.7$	Top 20% of mass
Grad-CAM	0.1353	0.1150	0.0768	0.0879
Integrated Gradients	0.0200	0.0034	0.0007	0.0712
LIME	0.2719	0.2487	0.2104	0.1267
Occlusion	0.2594	0.2422	0.1575	0.1841

Figure 3 shows representative maps from each confusion-matrix cell. Grad-CAM maps are coarse and often spill beyond the lesion; Integrated Gradients maps are scattered single pixels; LIME maps follow superpixel boundaries; and occlusion maps are blocky, reflecting the 32-pixel window.

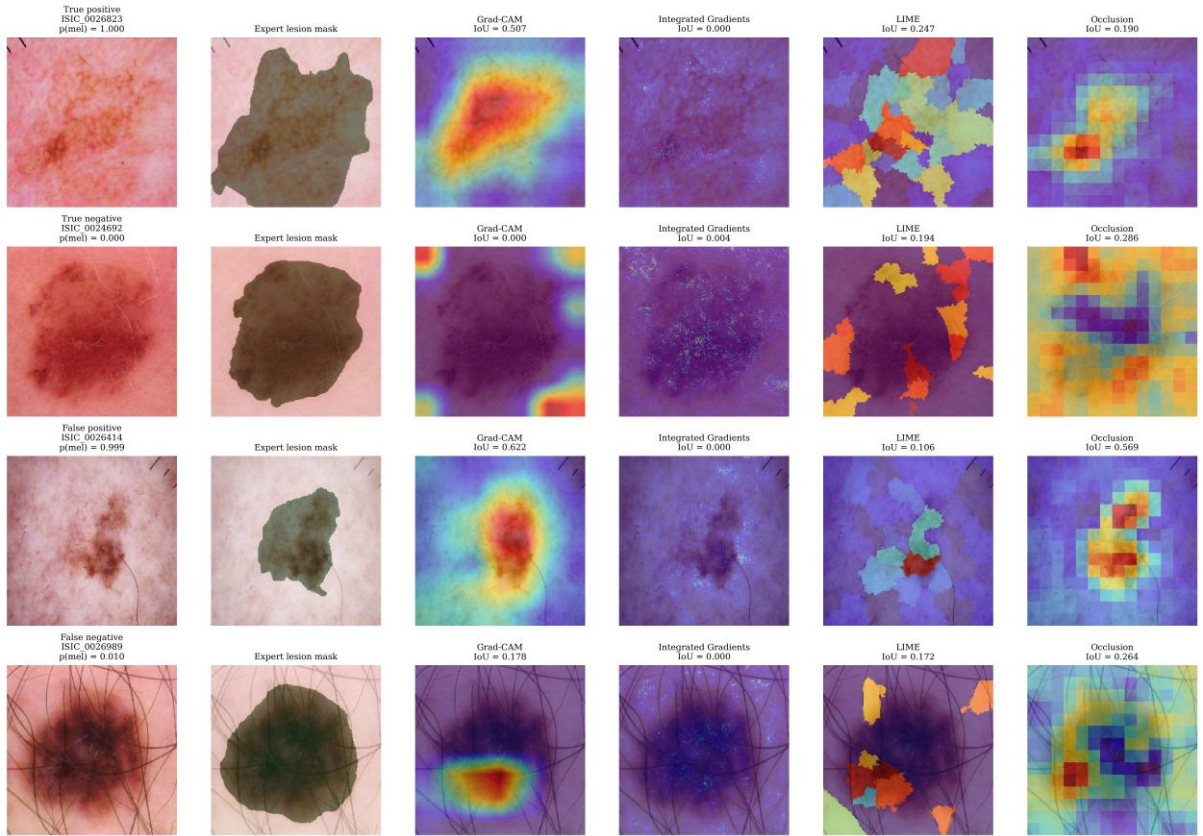


Fig. 3 Representative test images (true positive, true negative, false positive, false negative), their lesion masks and the four explanation maps, with per-image IoU. $p(mel)$ is the predicted melanoma probability

3.3.1 Sanity Check

Table 6 and Fig. 4 report the cascading randomisation test. Grad-CAM, LIME and occlusion lost their correlation with the original maps as soon as the classifier head was randomised; all their values stayed within about 0.12 of zero, and several were negative. Integrated Gradients did not. Its maps kept a rank correlation of 0.5538 after

the head was randomised, 0.5633 after the final convolution, and 0.3345 after seven of the nine stages. Only when the two earliest blocks were randomised, at which point the network's output degenerated, did the correlation reach zero. The exact zeros in the last rows of Table 6 arise because the degenerate network produced constant maps, whose correlation was recorded as zero by design.

Table 6 Mean Spearman correlation between original maps and maps after cumulative randomisation, from head to input (lower is better)

Stage	Grad-CAM	Integrated Gradients	LIME	Occlusion
Classifier head	0.0348	0.5538	-0.0115	0.0790
Final convolution	-0.0649	0.5633	0.1150	-0.0271
Block 6	0.0460	0.4960	-0.0164	0.1209
Block 5	-0.1164	0.4011	-0.0531	0.0021
Block 4	-0.0066	0.3745	-0.0131	0.0140
Block 3	-0.1159	0.2786	-0.0287	0.0291
Block 2	-0.0794	0.3345	0.0000	0.0000
Block 1	0.0000	0.0000	0.0000	0.0000
Block 0	0.0000	0.0000	0.0000	0.0000

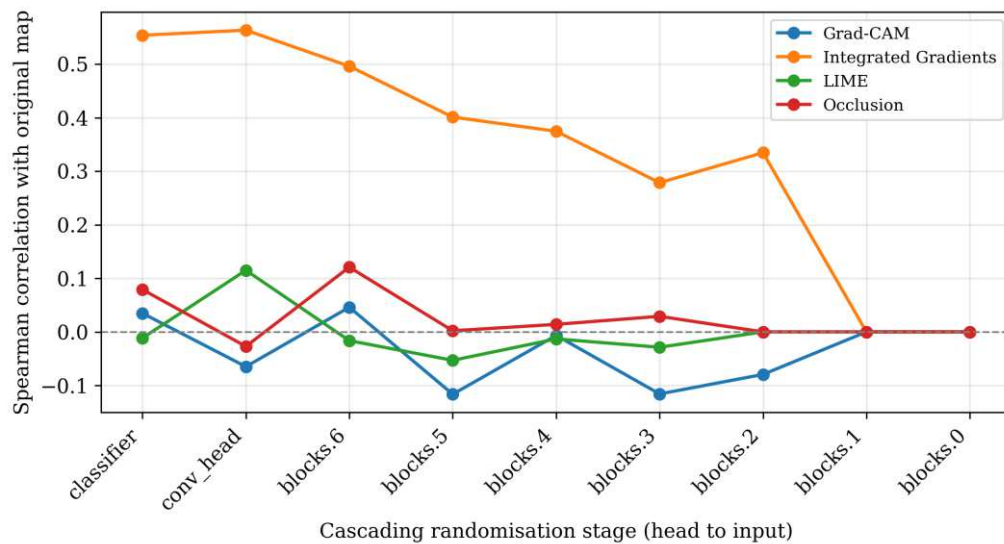


Fig. 4 Cascading model-parameter randomisation sanity check. Integrated Gradients keeps a substantial correlation with its original maps until the earliest blocks are randomised

3.4 Axis 2: Clinical Utility Proxy

Table 7 and Fig. 5 report the utility measures. LIME placed the largest share of its mass inside the lesion (0.4465), followed by Integrated Gradients (0.3151), occlusion (0.2823) and Grad-CAM (0.2045) (Friedman $\chi^2 = 897.12$, $p = 3.73 \times 10^{-194}$). Centroid distance gave a different order: occlusion was closest to the lesion centre (15.38 px), then Integrated Gradients (16.67 px), LIME (23.77 px) and Grad-CAM (36.61 px, standard deviation 32.27 px).

Table 7 Clinical utility proxy measures (mean $\pm$ standard deviation for hit rate, centroid distance and total artefact attribution)

Method	Lesion hit rate	Centroid distance (px)	Artefact attribution (any)	Hair	Ink	Ruler
Grad-CAM	0.2045 $\pm$ 0.2581	36.61 $\pm$ 32.27	0.1101 $\pm$ 0.1242	0.0946	0.0003	0.0171
Integrated Gradients	0.3151 $\pm$ 0.1822	16.67 $\pm$ 11.82	0.0962 $\pm$ 0.0998	0.0857	0.0002	0.0115
LIME	0.4465 $\pm$ 0.2232	23.77 $\pm$ 15.50	0.0897 $\pm$ 0.0941	0.0824	0.0001	0.0081
Occlusion	0.2823 $\pm$ 0.1951	15.38 $\pm$ 10.66	0.1218 $\pm$ 0.1195	0.1087	0.0003	0.0145

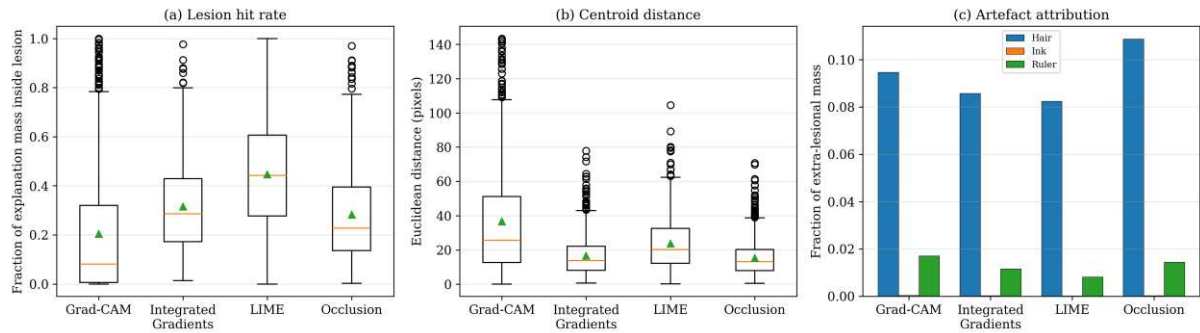


Fig. 5 Clinical utility proxy: (a) lesion hit rate, (b) explanation-to-lesion centroid distance, and (c) artefact attribution by artefact class

Total artefact attribution was similar across methods (0.0897 to 0.1218) and was dominated by the hair component. Given the behaviour of the hair detector (Table 3), this component measures attribution to dark, fine texture in general rather than to hair. The ink and ruler detectors are more selective. Attribution to ink was at most 0.0003 for every method, and attribution to rulers ranged from 0.0081 (LIME) to 0.0171 (Grad-CAM). On the two artefact classes that can be measured credibly, no method showed a serious susceptibility.

3.5 Axis 3: Trust Calibration

Every one of the 12 unstratified TC scores was negative, and 10 were significant (Table 8). On IAUC, TC ranged from -0.3073 to -0.4232 with rank-biserial effect sizes between -0.71 and -0.77 . Taken at face value, explanations were more faithful on the images the classifier got wrong.

Lesion area did not explain this. Mean lesion area was 0.2659 of the frame in the correct group and 0.2638 in the incorrect group (Mann-Whitney $p = 0.563$), and partialling out area changed the correlation between IoU and correctness by at most 0.021 for any method (Grad-CAM -0.3321 to -0.3525 ; Integrated Gradients -0.1221 to -0.1239 ; LIME -0.2225 to -0.2272 ; occlusion -0.1501 to -0.1615).

Predicted class did explain it. Explanation fidelity rose with the predicted melanoma score: the Spearman correlation with IAUC exceeded 0.90 for every method, and the correlation with IoU was positive for all four (Table 8). Because 90.9% of errors were false positives, the incorrect group consisted mostly of high-scoring images, while the correct group consisted mostly of low-scoring true negatives (602 of 673). Figure 6 shows IoU by confusion-matrix cell.

Table 8 Trust calibration. TC is mean fidelity on correct minus incorrect predictions (two-sided Mann-Whitney U); stratified TC uses IoU; ρ is the Spearman correlation between predicted melanoma score and fidelity

Method	TC, IoU (p)	TC, IAUC (p)	TC, TP vs FP (p)	TC, TN vs FN (p)	ρ (score, IoU)	ρ (score, IAUC)
Grad-CAM	-0.1456 (<0.001)	-0.3073 (<0.001)	0.1124 (<0.001)	-0.0834 (<0.001)	0.6853	0.9129
Integrated Gradients	-0.0012 (<0.001)	-0.4179 (<0.001)	-0.0009 (0.264)	-0.0010 (0.298)	0.1519	0.9032
LIME	-0.1009 (<0.001)	-0.4191 (<0.001)	0.0460 (0.112)	-0.1021 (0.077)	0.4679	0.9118
Occlusion	-0.0663 (<0.001)	-0.4232 (<0.001)	0.0022 (0.579)	-0.1101 (0.013)	0.2228	0.9056

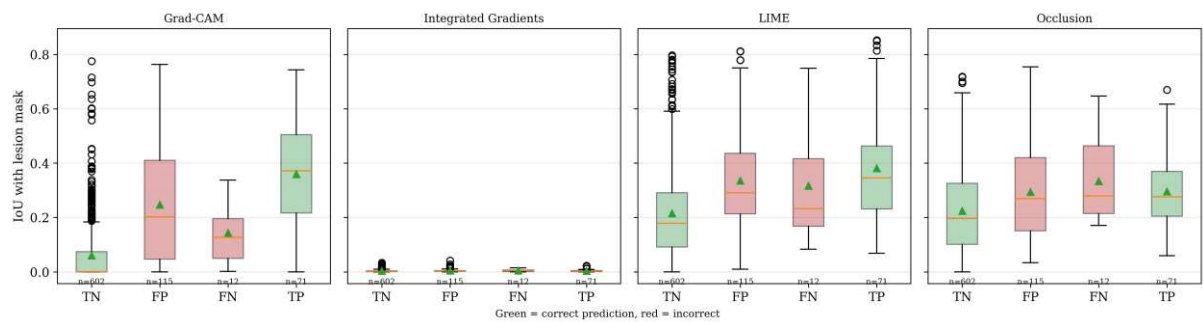


Fig. 6 IoU by confusion-matrix cell for each method (green, correct; red, incorrect; n, images per cell)

Within the melanoma-predicted group (TP against FP), the TC sign turned positive for three methods. Only Grad-CAM was significant (0.1124, $p = 4.2 \times 10^{-4}$, $r = 0.31$); LIME (0.0460) and occlusion (0.0022) were not, and Integrated Gradients stayed flat. Within the benign-predicted group (TN against FN), all four scores stayed negative, significantly so for Grad-CAM and occlusion, but this comparison rests on only 12 false negatives. Hit-rate TC scores for Integrated Gradients ($p = 0.363$) and occlusion ($p = 0.109$) were the two non-significant unstratified comparisons. With a Bonferroni threshold of $0.05/12 = 0.0042$, nine of the ten significant unstratified results remain significant; only the LIME hit-rate comparison ($p = 0.0044$) does not. The Grad-CAM TP-versus-FP result also remains significant under a Bonferroni threshold for the eight stratified tests ($0.05/8 = 0.0063$).

3.6 Agreement, Cross-Axis Ranking and Cost

Agreement between methods was low (Table 9, Fig. 7). The highest pairwise rank correlation was 0.2182 (Grad-CAM and occlusion) and the lowest -0.0724 (Grad-CAM and Integrated Gradients). LIME and occlusion, which were statistically indistinguishable on IoU, agreed at only 0.1365. Overlap between the top-20% regions of different methods was also low, between 0.13 and 0.21 (Fig. 7b).

Table 9 Mean per-image Spearman correlation between the maps of each pair of methods

	Grad-CAM	Integrated Gradients	LIME	Occlusion
Grad-CAM	1	-0.0724	0.1064	0.2182
Integrated Gradients	-0.0724	1	0.0647	-0.0473
LIME	0.1064	0.0647	1	0.1365
Occlusion	0.2182	-0.0473	0.1365	1

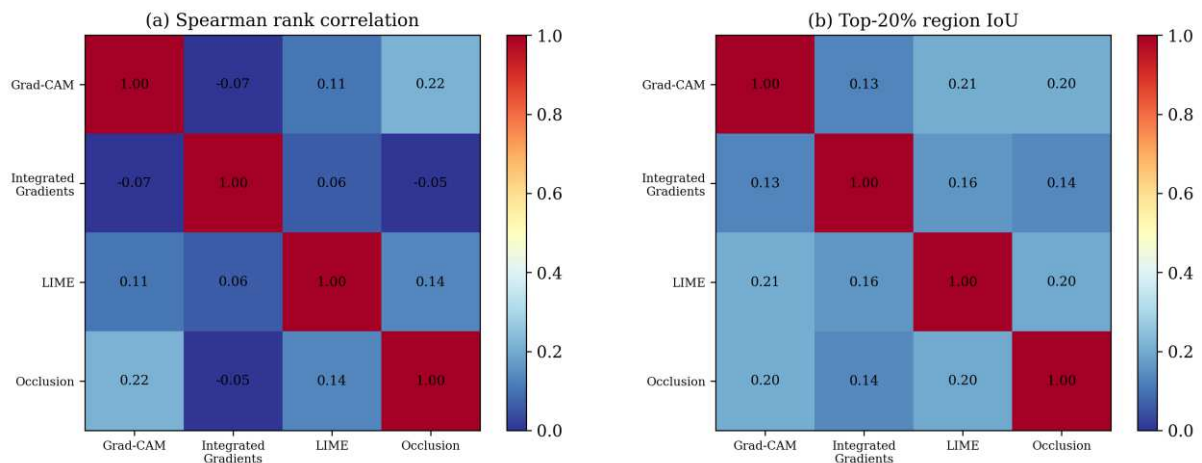


Fig. 7 Pairwise agreement between methods: (a) Spearman rank correlation of full maps; (b) IoU between the top-20% regions of each pair

Table 10 ranks each method on eight criteria taken directly from Tables 4, 7 and 8, and reports the sanity check separately as a pass-or-fail gate. Three choices differ from a naive ranking, and each follows from the results above. The sanity check is not ranked, because it is a necessary condition rather than a matter of degree. Artefact attribution uses rulers only, because the hair detector was unreliable and ink attribution was negligible for every method. Trust uses the stratified TP-versus-FP score, because the unstratified score is confounded and is degenerate for Integrated Gradients. LIME had the best mean rank (1.50), followed by occlusion (2.13), Integrated Gradients (3.00) and Grad-CAM (3.38). LIME was first on five criteria and never worse than third. Integrated Gradients, however, does not pass the gate.

397 **Table 10** Rank of each method on eight criteria (1 is best), mean rank, sanity-check gate,
398 and explanation cost on the 800-image subset

Criterion	Grad-CAM	Integrated Gradients	LIME	Occlusion
Fidelity: IoU ($\tau = 0.5$)	3	4	1	2
Fidelity: pointing game	4	3	1	2
Fidelity: DAUC	3	4	1	2
Fidelity: IAUC	4	3	2	1
Utility: lesion hit rate	4	2	1	3
Utility: centroid distance	4	2	3	1
Utility: ruler attribution	4	2	1	3
Trust: stratified TC (TP vs FP, IoU)	1	4	2	3
Mean rank (8 criteria)	3.38	3.00	1.50	2.13
Sanity check (gate)	Pass	Fail	Pass	Pass
Seconds per image	0.021	0.224	6.552	0.240
Total minutes (800 images)	0.3	3.0	87.4	3.2

399 Cost varied widely. LIME needed 6.552 s per image, 312 times Grad-CAM's 0.021 s
400 and 27 times occlusion's 0.240 s.

401 **4 Discussion**

402 The central finding is that the three axes disagree, and that the disagreement is large
403 enough to change which method a practitioner would pick. No method led on all three
404 axes. LIME led on overall rank and lesion coverage but was the slowest by far.
405 Occlusion matched LIME on localisation, led on insertion and centroid distance, and
406 cost a twenty-seventh as much. Integrated Gradients, with its mean-image baseline,
407 scored second on the pointing game and centroid distance while failing the one test
408 that asks whether an explanation depends on the model at all. Grad-CAM scored worst
409 on most measures yet was the only method with a significant positive trust-calibration
410 difference after stratification. An evaluation confined to any single axis would have
411 produced a confident recommendation, and in at least two cases (recommending
412 Integrated Gradients on localisation, or dismissing Grad-CAM entirely) that
413 recommendation would have been wrong.

414 **The classifier is adequate, not exceptional.** A test AUC-ROC of 0.9295 is
415 competent for binary melanoma classification on a lesion-level split. Reported results
416 on HAM10000 are often higher, but many use image-level splits, which place
417 photographs of the same lesion on both sides of the partition and inflate performance.
418 Here, 10,015 images came from 7,470 lesions, so this risk is real. For the present
419 purpose the key requirement is that the model has learned lesion-relevant structure,
420 so that the explanations explain something other than noise, and an AUC of 0.93
421 meets it. The low precision (0.3820) is the arithmetic result of a high-sensitivity

operating point at about 10% prevalence; the negative predictive value (0.9810) is the relevant figure for triage.

Fidelity is not one quantity. IoU, pointing game, deletion and insertion produced three different orderings of four methods. IoU and the pointing game compare the same maps with the same masks but ask different questions: how much of the thresholded map lies in the lesion, versus whether the single strongest pixel does. The threshold sweep showed the mechanism behind the largest reversal. Integrated Gradients concentrates attribution on scattered pixels, so after min-max normalisation almost nothing survives a fixed cut of 0.5 and IoU collapses regardless of where those pixels lie. Under the mass-based rule it localises about as well as Grad-CAM. This is the kind of metric fragility described by Tomsett et al. [19] and addressed by toolkits such as Quantus [36]. Any study that reports IoU for attribution maps at a single fixed threshold reports a number whose ranking may not survive a change of convention.

LIME and occlusion were statistically inseparable on IoU, which fits their shared design: both perturb contiguous regions and so tend to align with a segmentation mask. Yet their maps agreed at a rank correlation of only 0.14. Equal aggregate scores do not mean equal explanations, and a clinician shown one rather than the other would see materially different evidence for the same prediction.

The sanity check should be a gate. The randomisation test is the one result here that is not a matter of degree. Integrated Gradients retained about half of its structure after the classifier head was destroyed, which replicates the central finding of Adebayo et al. [17] in dermoscopy. One possible reason, which we did not test, is that dermoscopic images have strong edge structure at the lesion border, and an input-gradient method with a mean-image baseline registers that structure whether or not the network uses it. Because Integrated Gradients depends strongly on its baseline, this result applies to the mean-image baseline and the block-level cascade used here; other baselines may behave differently. This matters because Integrated Gradients looked good on two localisation-type measures. A map that passes localisation tests but fails a faithfulness test is more dangerous than one that looks poor, because it invites trust it has not earned [37]. We therefore recommend treating the sanity check as a gating condition applied before other scores, not as one criterion among many. For this reason Table 10 reports it as a gate rather than a rank.

454 **Utility measures separate aim from spread.** LIME placed more than twice as much
455 mass inside the lesion as Grad-CAM (0.4465 against 0.2045). Grad-CAM's weakness
456 is structural: its map comes from a 7×7 feature grid, so each native cell covers a 32
457 $\times 32$ pixel block and cannot trace a lesion boundary. Higher-resolution variants such
458 as Grad-CAM++ [38], Score-CAM [39] and Ablation-CAM [40] were designed partly to
459 address this. Hit rate and centroid distance also disagreed: occlusion was best aimed
460 (closest centroid) but diffuse (lower hit rate), whereas LIME was concentrated but
461 somewhat off-centre. These are different failure modes. A diffuse but well-aimed map
462 directs attention to the right region without saying much about it; a concentrated but
463 off-centre map can confidently point to the wrong part of the lesion.

464 The artefact analysis must be reported with more caution than planned. Two of three
465 detectors behaved as intended, and on those (ink and rulers) no method showed
466 shortcut behaviour. The hair detector did not: it fired on every image and covered about
467 a quarter of each frame, because the black-hat filter also responds to the pigment
468 network and other diagnostic texture. The hair column of Table 7 should therefore be
469 read as attribution to fine dark texture, much of which is diagnostically relevant, and
470 not as evidence that methods are misled by hair. HAM10000 contains no artefact
471 annotation, so this limitation could not be removed within the present design.

472 **Trust calibration measured confidence, not correctness.** Because all four
473 methods explain the melanoma logit, an image the model scores highly gives the
474 explanation a strong signal to attach to, and the map concentrates; an image scored
475 near zero gives a diffuse map. Fidelity therefore tracks the melanoma score closely (p
476 > 0.90 for IAUC). Combined with an error profile dominated by false positives, this
477 makes the correct-versus-incorrect comparison close to a low-score-versus-high-
478 score comparison, and the naive TC becomes negative for reasons unrelated to the
479 quality of the explanations. After stratification, only Grad-CAM produced significantly
480 better-localised maps on true melanomas than on false alarms. Stratification, however,
481 removes only part of the confound: true positives probably receive higher melanoma
482 scores than false positives, and Grad-CAM's IoU is the most score-dependent of the
483 four methods ($p = 0.6853$). The Grad-CAM difference may therefore be partly a
484 confidence effect as well; separating the two would require comparing true and false
485 positives matched on predicted score. Two lessons follow. Methodologically, any trust-
486 calibration measure that compares explanation quality between correct and incorrect

487 predictions should be stratified by predicted class, and should also control for the
488 model's score, when explanations target a single class logit and errors are unevenly
489 split, which is the usual case for a screening model. Practically, a crisp heatmap
490 signals that the model is confident, not that it is right. When most errors are confident
491 false positives, taking map quality as a sign of correctness gets the inference
492 backwards and could feed automation bias [20, 34].

493 **Cost is part of the decision.** At 6.55 s per image, LIME is acceptable in batch
494 research but awkward in interactive teledermatology, where an explanation is
495 expected alongside the prediction and hardware may be modest. Occlusion delivered
496 most of LIME's measured quality at 0.24 s per image and passed the sanity check,
497 which makes it the most defensible default when quality and cost are weighed
498 together.

499 **Relation to prior work.** The findings agree with Adebayo et al. [17] on the insensitivity
500 of some gradient-based attributions to model parameters, with Tomsett et al. [19] and
501 Hedström et al. [36] on the non-interchangeability of saliency metrics, and with
502 Saporta et al. [18] on the modest absolute localisation of saliency maps: the best mean
503 IoU recorded here was only 0.2487. They are also consistent with a recent meta-
504 analysis that found higher fidelity for LIME than for Grad-CAM [41]. They extend
505 comparative work in dermatology, which has mainly reported one axis at a time [25],
506 by showing directly how the axes diverge on one model and one dataset. They also
507 refine the recent observation that models with better lesion focus classify melanoma
508 better [42]: across images, focus appears to follow how strongly the model responds,
509 not whether it is right. Finally, they complement human-subjects trust research [20, 21,
510 43] by identifying a confound that a computational trust proxy must control before it
511 can be linked to human reliance.

512 **Implications for practice.** Four recommendations follow. First, apply a parameter-
513 randomisation sanity check as a gate before any localisation or utility score is
514 considered. Second, report localisation metrics under several binarisation rules,
515 including a mass-based rule. Third, do not present explanation quality as evidence of
516 correctness; present the calibrated probability and the operating characteristics of the
517 model beside any heatmap. Fourth, weigh cost explicitly: on this evidence, occlusion
518 is a sensible default, LIME is preferable when quality matters more than latency, Grad-
519 CAM is suitable when latency is critical and coarse localisation is acceptable, and

Integrated Gradients, with a mean-image baseline, is not recommended for this architecture and dataset.

Limitations. The trust axis is a computational proxy; the stratified analysis did not control for score differences within each predicted class; and any claim that Grad-CAM supports better-calibrated trust assumes that clinicians' reliance would follow fidelity as measured here; a reader study is needed to test this. The results rest on one architecture, one dataset and one explanation target. EfficientNet-B0's coarse final feature map shapes Grad-CAM's results, and all maps target the melanoma logit. The lesion masks were drafted by a network and corrected by a single dermatologist, so every fidelity and utility value is relative to that boundary definition and carries no inter-annotator variability. The hair detector did not work as intended, and artefact attribution for hair is not interpretable as stated. The benign-predicted trust stratum contained only 12 false negatives. All results come from one training run, one data split and one explanation seed, and fidelity scores and TC differences carry no confidence intervals. The split is lesion-level, not patient-level. The sanity check was applied at block level rather than layer level for computational reasons. Finally, HAM10000 is drawn mainly from lighter-skinned populations [44-46], so no claim is made about explanation behaviour on darker skin.

Future work. Priorities are bootstrap intervals and repeated seeds for all fidelity and trust measures; a score-matched trust analysis; other Integrated Gradients baselines; a clinician reader study that tests whether the stratified trust proxy predicts real reliance; a dermoscopic subset with pixel-level artefact annotation; repetition across backbones with different output strides and across higher-resolution CAM variants; an operating point or dataset with a more balanced error profile; a skin-tone-diverse evaluation set; and comparison with models that are interpretable by design, such as prototype networks [47] and concept-based models [48, 49].

5 Conclusions

Four widely used post-hoc explanation methods were evaluated on one melanoma classifier across technical fidelity, a clinical utility proxy and computational trust calibration. The axes disagreed materially, and single-axis evaluation would have led to wrong choices, most clearly by favouring a method that fails a basic faithfulness test. Two common measurement choices proved fragile: a fixed IoU threshold penalises sparse maps for reasons unrelated to their quality, and an unstratified trust

553 measure reports model confidence rather than correctness. Multi-axis evaluation, with
554 a faithfulness gate, threshold sweeps and predicted-class stratification, should
555 become standard practice before explanation methods are placed in front of clinicians.

556 **Declarations**

557 **Funding** The authors did not receive support from any organisation for the submitted
558 work.

559 **Competing interests** The authors declare no competing interests.

560 **Ethics approval and consent** Not applicable. The study used only the publicly
561 available, de-identified HAM10000 dataset under its CC BY-NC-SA 4.0 licence.

562 **Author contributions** Okoi Michael Obeten: conceptualisation, methodology,
563 software, data curation, formal analysis, visualisation and writing (original draft).
564 Christiana Uchenna Ezeanya: supervision, guidance on study design and evaluation,
565 critical review of the methods and results, and writing (review and editing). Both
566 authors read and approved the final manuscript.

567 **Data availability** The HAM10000 images, metadata and lesion segmentations are
568 publicly available from the Harvard Dataverse [28].

569 **Code availability** The analysis code and derived result files are available from the
570 corresponding author upon reasonable request.

571 **Acknowledgements** This work is derived from the first author's MSc (Information
572 Technology) dissertation at the National Open University of Nigeria.

573 **References**

- 574 1. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-
575 level classification of skin cancer with deep neural networks. *Nature*.
576 2017;542:115-118. <https://doi.org/10.1038/nature21056>
- 577 2. Arnold M, Singh D, Laversanne M, Vignat J, Vaccarella S, Meheus F, et al. Global
578 burden of cutaneous melanoma in 2020 and projections to 2040. *JAMA*
579 *Dermatol*. 2022;158:495-503. <https://doi.org/10.1001/jamadermatol.2022.0160>
- 580 3. Salahuddin Z, Woodruff HC, Chatterjee A, Lambin P. Transparency of deep neural
581 networks for medical image analysis: a review of interpretability methods.
582 *Comput Biol Med*. 2022;140:105111.
583 <https://doi.org/10.1016/j.combiomed.2021.105111>
- 584 4. Bhati D, Neha F, Amiruzzaman M. A survey on explainable artificial intelligence
585 (XAI) techniques for visualizing deep learning models in medical imaging. *J*
586 *Imaging*. 2024;10:239. <https://doi.org/10.3390/jimaging10100239>

5. Shevtsova D, Ahmed A, Boot IWA, Sanges C, Hudecek M, Jacobs JJJ, et al. Trust in and acceptance of artificial intelligence applications in medicine: mixed methods study. *JMIR Hum Factors*. 2024;11:e47031. <https://doi.org/10.2196/47031>
6. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: *Proceedings of the IEEE International Conference on Computer Vision*; 2017. p. 618-26. <https://doi.org/10.1109/ICCV.2017.74>
7. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. In: *Proceedings of the 34th International Conference on Machine Learning*. PMLR; 2017;70:3319-28.
8. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?" Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; 2016. p. 1135-44. <https://doi.org/10.1145/2939672.2939778>
9. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: *Computer Vision, ECCV 2014. Lecture Notes in Computer Science*, vol 8689. Cham: Springer; 2014. p. 818-33. https://doi.org/10.1007/978-3-319-10590-1_53
10. Fiaz M, Shoaib Khan MB, Khan AH, Bilal A, Abdullah M, Darem AA, et al. An explainable hybrid deep learning framework for precise skin lesion segmentation and multi-class classification. *Front Med*. 2025;12:1681542. <https://doi.org/10.3389/fmed.2025.1681542>
11. Gamage L, Isuranga U, Meedeniya D, De Silva S, Yogarajah P. Melanoma skin cancer identification with explainability utilizing mask guided technique. *Electronics*. 2024;13:680. <https://doi.org/10.3390/electronics13040680>
12. AbuAlkebash H, Saleh RAA, Ertunç HM. Explainable depth-wise and channel-wise fusion models for multi-class skin lesion classification. *PLoS One*. 2026;21:e0340901. <https://doi.org/10.1371/journal.pone.0340901>
13. Gabani V, Navamani TM, Shyamala K, Rajpal VKV. Multimodal skin lesion classification for early cancer diagnosis using deep learning. *Front Physiol*. 2026;17:1717517. <https://doi.org/10.3389/fphys.2026.1717517>
14. Alabduljabbar A, Akram T, Altherwy YN, Akram MA, Ashraf I. XAI-MedNet: a next-generation explainable AI framework for contrast-enhanced skin lesion classification via entropy-controlled optimization. *Bioengineering*. 2026;13:506. <https://doi.org/10.3390/bioengineering13050506>
15. Whewell J, Peters R, Kulon J. Explainable artificial intelligence for skin lesion classification: a comprehensive review of methods and challenges. *Technologies*. 2026;14:391. <https://doi.org/10.3390/technologies14070391>
16. Ud Din S, Kemna R, Ket JCF, Iqbal M, Bohoudi O, Hoogendoorn M, et al. Which explainable AI methods in medical imaging are clinically impactful? A systematic literature review addressing the clinician's perspective. *Front Artif Intell*. 2026;9:1819422. <https://doi.org/10.3389/frai.2026.1819422>
17. Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. Sanity checks for saliency maps. In: *Advances in Neural Information Processing Systems 31*; 2018. <https://arxiv.org/abs/1810.03292>

18. Saporta A, Gui X, Agrawal A, Pareek A, Truong SQH, Nguyen CDT, et al. Benchmarking saliency methods for chest X-ray interpretation. *Nat Mach Intell.* 2022;4:867-878. <https://doi.org/10.1038/s42256-022-00536-x>
19. Tomsett R, Harborne D, Chakraborty S, Gurram P, Preece A. Sanity checks for saliency metrics. *Proc AAAI Conf Artif Intell.* 2020;34:6021-6029. <https://doi.org/10.1609/aaai.v34i04.6064>
20. Prinster D, Mahmood A, Saria S, Jeudy J, Lin CT, Yi PH, et al. Care to explain? AI explanation types differentially impact chest radiograph diagnostic performance and physician trust in AI. *Radiology.* 2024;313(2):e233261. <https://doi.org/10.1148/radiol.233261>
21. Rosenbacke R, Melhus Å, McKee M, Stuckler D. How explainable artificial intelligence can increase or decrease clinicians' trust in AI applications in health care: systematic review. *JMIR AI.* 2024;3:e53207. <https://doi.org/10.2196/53207>
22. Davis SE, Salwei ME. Explainability in context: calibrating appropriate trust and reliance in artificial intelligence. *J Am Med Inform Assoc.* 2026;33:1554-1563. <https://doi.org/10.1093/jamia/ocag082>
23. Nauta M, Trienes J, Pathak S, Nguyen E, Peters M, Schmitt Y, et al. From anecdotal evidence to quantitative evaluation methods: a systematic review on evaluating explainable AI. *ACM Comput Surv.* 2023;55(13s):1-42. <https://doi.org/10.1145/3583558>
24. Lago MA, Zamzmi G, Eich B, Delfino JG. Evaluating explainability: a framework for systematic assessment of explainable AI features in medical imaging. *Bioengineering.* 2026;13:111. <https://doi.org/10.3390/bioengineering13010111>
25. Giavina-Bianchi M, Vitor WG, Fornasiero de Paiva V, Okita AL, Sousa RM, Machado B. Explainability agreement between dermatologists and five visual explanations techniques in deep neural networks for melanoma AI classification. *Front Med.* 2023;10:1241484. <https://doi.org/10.3389/fmed.2023.1241484>
26. Tschandl P, Rosendahl C, Kittler H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci Data.* 2018;5:180161. <https://doi.org/10.1038/sdata.2018.161>
27. Codella N, Rotemberg V, Tschandl P, Celebi ME, Dusza S, Gutman D, et al. Skin lesion analysis toward melanoma detection 2018: a challenge hosted by the International Skin Imaging Collaboration (ISIC). *arXiv:1902.03368 [Preprint].* 2019. <https://arxiv.org/abs/1902.03368>
28. Tschandl P. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions [dataset]. *Harvard Dataverse;* 2020. <https://doi.org/10.7910/DVN/DBW86T>
29. Kline TL, Kitamura F, Warren D, Pan I, Korchi AM, Tenenholtz N, et al. Best practices and checklist for reviewing artificial intelligence-based medical imaging papers: classification. *J Imaging Inform Med.* 2025. <https://doi.org/10.1007/s10278-025-01548-w>
30. Tan M, Le Q. EfficientNet: rethinking model scaling for convolutional neural networks. In: *Proceedings of the 36th International Conference on Machine Learning.* PMLR; 2019;97:6105-14.
31. Youden WJ. Index for rating diagnostic tests. *Cancer.* 1950;3:32-35.

32. Kokhlikyan N, Miglani V, Martin M, Wang E, Alsallakh B, Reynolds J, et al. Captum: a unified and generic model interpretability library for PyTorch. arXiv:2009.07896 [Preprint]. 2020. <https://arxiv.org/abs/2009.07896>
33. Petsiuk V, Das A, Saenko K. RISE: randomized input sampling for explanation of black-box models. In: Proceedings of the British Machine Vision Conference; 2018. <https://arxiv.org/abs/1806.07421>
34. Sakamoto T, Harada Y, Shimizu T. Facilitating trust calibration in artificial intelligence-driven diagnostic decision support systems for determining physicians' diagnostic accuracy: quasi-experimental study. JMIR Form Res. 2024;8:e58666. <https://doi.org/10.2196/58666>
35. Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat. 1979;6:65-70.
36. Hedström A, Weber L, Krakowczyk D, Bareeva D, Motzkus F, Samek W, et al. Quantus: an explainable AI toolkit for responsible evaluation of neural network explanations and beyond. J Mach Learn Res. 2023;24(34):1-11.
37. Zhang J, Chao H, Dasegowda G, Wang G, Kalra MK, Yan P. Overlooked trustworthiness of saliency maps. In: Medical Image Computing and Computer Assisted Intervention, MICCAI 2022. Cham: Springer; 2022. p. 451-61. https://doi.org/10.1007/978-3-031-16437-8_43
38. Chattopadhyay A, Sarkar A, Howlader P, Balasubramanian VN. Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV); 2018. p. 839-47. <https://doi.org/10.1109/WACV.2018.00097>
39. Wang H, Wang Z, Du M, Yang F, Zhang Z, Ding S, et al. Score-CAM: score-weighted visual explanations for convolutional neural networks. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2020. p. 111-9. <https://doi.org/10.1109/CVPRW50498.2020.00020>
40. Desai S, Ramaswamy HG. Ablation-CAM: visual explanations for deep convolutional network via gradient-free localization. In: 2020 IEEE Winter Conference on Applications of Computer Vision (WACV); 2020. p. 972-80. <https://doi.org/10.1109/WACV45572.2020.9093360>
41. Singh Y, Hathaway QA, Keishing V, Salehi S, Wei Y, Horvat N, et al. Beyond post hoc explanations: a comprehensive framework for accountable AI in medical imaging through transparency, interpretability, and explainability. Bioengineering. 2025;12:879. <https://doi.org/10.3390/bioengineering12080879>
42. Donde T. The impact of lesion focus on the performance of AI-based melanoma classification. arXiv:2601.00355 [Preprint]. 2026. <https://arxiv.org/abs/2601.00355>
43. Ewals LJS, Heesterbeek LJJ, Yu B, van der Wulp K, Mavroeidis D, Funk M, et al. The impact of expectation management and model transparency on radiologists' trust and utilization of AI recommendations for lung nodule assessment on computed tomography: simulated use study. JMIR AI. 2024;3:e52211. <https://doi.org/10.2196/52211>
44. Adamson AS, Smith A. Machine learning and health care disparities in dermatology. JAMA Dermatol. 2018;154:1247. <https://doi.org/10.1001/jamadermatol.2018.2348>

- 727 45. Daneshjou R, Vodrahalli K, Novoa RA, Jenkins M, Liang W, Rotemberg V, et
728 al. Disparities in dermatology AI performance on a diverse, curated clinical
729 image set. *Sci Adv.* 2022;8:eabq6147. <https://doi.org/10.1126/sciadv.abq6147>
- 730 46. Groh M, Badri O, Daneshjou R, Koochek A, Harris C, Soenksen LR, et al. Deep
731 learning-aided decision support for diagnosis of skin disease across skin
732 tones. *Nat Med.* 2024;30:573-583. [https://doi.org/10.1038/s41591-023-02728-](https://doi.org/10.1038/s41591-023-02728-3)
733 3
- 734 47. Chen C, Li O, Tao C, Barnett AJ, Su J, Rudin C. This looks like that: deep
735 learning for interpretable image recognition. In: *Advances in Neural*
736 *Information Processing Systems* 32; 2019.
- 737 48. Patrício C, Neves JC, Teixeira LF. Explainable deep learning methods in medical
738 image classification: a survey. *ACM Comput Surv.* 2024;56(4):1-41.
739 <https://doi.org/10.1145/3625287>
- 740 49. Hou J, Liu S, Bie Y, Wang H, Tan A, Luo L, et al. Self-eXplainable AI for medical
741 image analysis: a survey and new outlooks. *arXiv:2410.02331 [Preprint].*
742 2024. <https://arxiv.org/abs/2410.02331>