

Auditing the inference processes of medical-image classifiers by leveraging generative AI and the expertise of physicians

Received: 30 October 2023

Accepted: 30 October 2023

Published online: 28 December 2023

 Check for updates


Alex J. DeGrave^{1,2}, Zhuo Ran Cai³, Joseph D. Janizek^{1,2}, Roxana Daneshjou^{4,5,6}  
& Su-In Lee^{1,6}  

The inferences of most machine-learning models powering medical artificial intelligence are difficult to interpret. Here we report a general framework for model auditing that combines insights from medical experts with a highly expressive form of explainable artificial intelligence. Specifically, we leveraged the expertise of dermatologists for the clinical task of differentiating melanomas from melanoma ‘lookalikes’ on the basis of dermoscopic and clinical images of the skin, and the power of generative models to render ‘counterfactual’ images to understand the ‘reasoning’ processes of five medical-image classifiers. By altering image attributes to produce analogous images that elicit a different prediction by the classifiers, and by asking physicians to identify medically meaningful features in the images, the counterfactual images revealed that the classifiers rely both on features used by human dermatologists, such as lesional pigmentation patterns, and on undesirable features, such as background skin texture and colour balance. The framework can be applied to any specialized medical domain to make the powerful inference processes of machine-learning models medically understandable.

Medical artificial intelligence (AI) classifiers have proliferated in recent years¹, but currently, the scientific and medical communities poorly understand what factors influence AI outputs and whether these factors could lead to failures and harm to patients when AI is deployed in practice. The reasoning processes of these high-stakes classifiers—namely, those that rely on neural networks and other complex ‘machine-learning’ techniques, which automatically learn statistical patterns in large datasets—remain opaque to all stakeholders, including patients, medical providers, regulators and even the developers of these AI systems. In principle, a detailed understanding of the reasoning processes of these AI classifiers could help us predict and prevent AI failures, help to improve AI models and offer scientific

value by contributing to the community’s knowledge of AI reasoning processes or their underlying training data. However, we lack a thorough medically interpretable picture of the reasoning processes of machine-learning-based medical-image classifiers. Previous efforts provided extremely limited peeks at medical-AI reasoning processes^{2,3}, typically via techniques that ‘sanity check’ whether a model is looking in the correct place^{4–7}, and both these and more expressive techniques^{8,9} typically suffer from a lack of principled and medically informed analysis, precluding a thorough understanding. Indeed, despite technical developments in these explainable AI (XAI) tools, the gap between XAI tool output and a pragmatic understanding of an AI classifier, particularly for image analysis and other ‘representation learning’ AI systems,

¹Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, WA, USA. ²Medical Scientist Training Program, University of Washington, Seattle, WA, USA. ³Program for Clinical Research and Technology, Department of Dermatology, Stanford University School of Medicine, Stanford, CA, USA. ⁴Department of Dermatology, Stanford University School of Medicine, Stanford, CA, USA. ⁵Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA, USA. ⁶These authors contributed equally: Roxana Daneshjou, Su-In Lee.

 e-mail: roxanad@stanford.edu; suinlee@cs.washington.edu

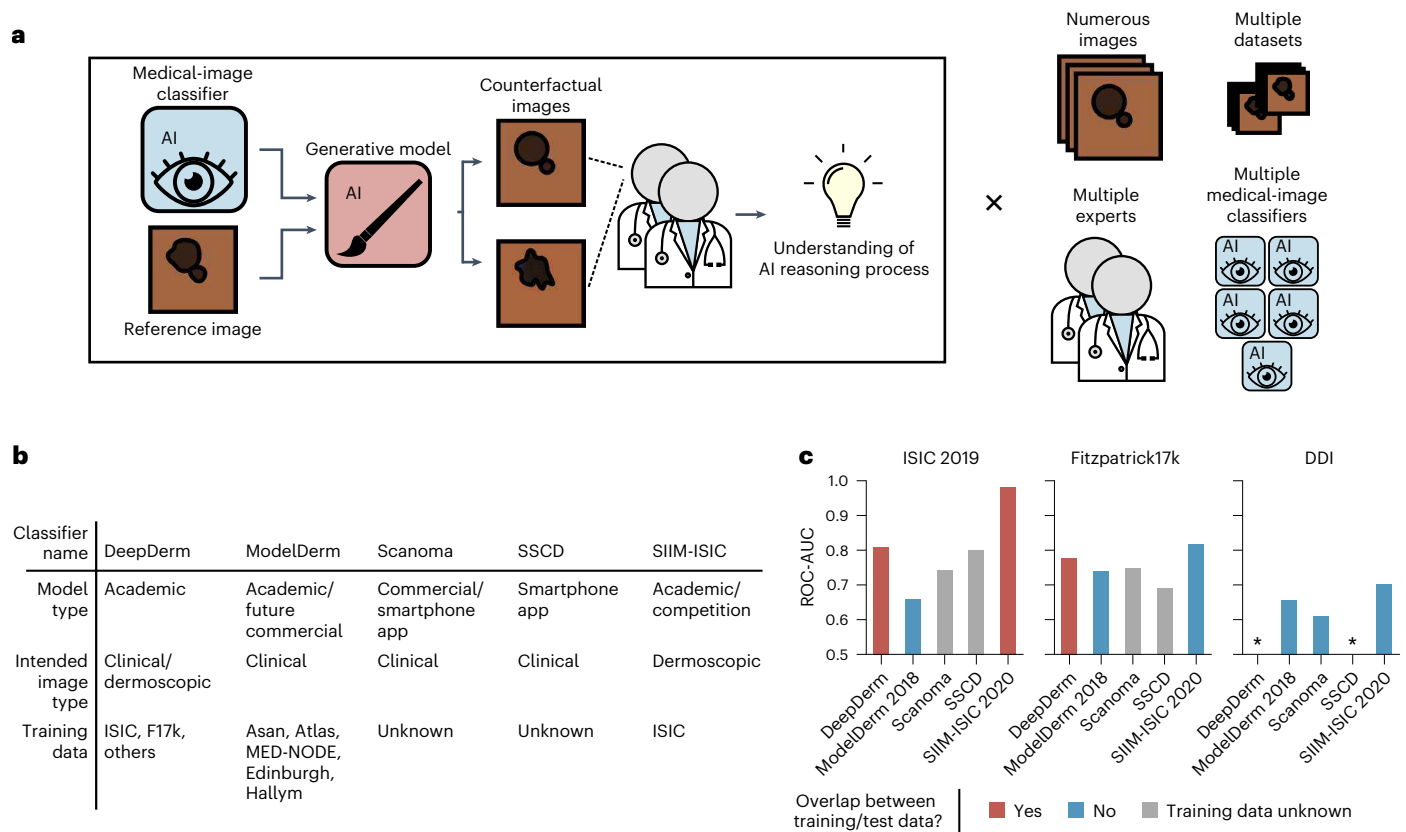


Fig. 1 | Overview of joint expert, XAI auditing procedure and audited AI classifiers. a, Our auditing procedure unites XAI with analysis by human experts to understand medical AI classifiers. Specifically, we leverage generative models to create counterfactual images that alter the prediction of a medical AI classifier; analysis of the counterfactuals by human experts (dermatologists) reveals the medical AI classifier's reasoning processes. We perform the analysis on numerous images from each of multiple datasets, gathering insights from two experts, for each of five different dermatology AI classifiers. **b**, Key details of dermatology AI classifiers audited in this study. **c**, Performance of the dermatology AI classifiers

on three datasets, including a dataset (DDI) external to the training data of every classifier. We examine the area under the receiver operating characteristic curve (ROC-AUC) to focus on the model's internal reasoning processes rather than emphasize the authors' original choices of model calibration. Asan, Atlas and Hallym datasets are described in ref. 22; MED-NODE is described in ref. 58; Edinburgh is available at <https://licensing.edinburgh-innovations.ed.ac.uk/product/dermofit-image-library> (*ROC-AUC < 0.5; that is, worse than random performance).

remains so large that efforts to apply XAI often miss severe faults in an AI-classifier's logic^{10–13}, such as strong dependence on spurious 'shortcut' features^{4,14}.

In exploring the reasoning processes of medical image AI, dermatology AI classifiers serve as a particularly impactful use case, for several reasons: numerous academic papers report high performance^{15–17}; the first handful of companies have received Conformité Européenne (CE) approval to deploy their AI classifiers on patients in the European Economic Area^{18,19}, and multiple developers are working on approval from the US Food and Drug Administration²⁰. Dermatology AI classifiers, often targeted directly at consumers, may pose particular risks due to the lack of involvement from healthcare providers, potential for bias on skin tone²¹ and other sensitive attributes, and heterogeneity of user-acquired images, resulting from variability in lighting conditions, image acquisition devices and digital processing procedures, none of which are standardized. Simultaneously, the de facto standard⁵ XAI modality to analyse image models—saliency maps, which highlight the regions of an image that most influence a model's prediction—appear poorly suited to understand dermatology AI classifiers, which may be best explained in terms of dermatological concepts (such as 'multiple colours of pigment' or 'atypical pigment networks') that spatially overlap or manifest diffusely throughout an image (Extended Data Fig. 1). Explanation of even a single prediction involves simultaneously high levels of technical AI knowledge and

dermatology expertise, impeding a global understanding of the AI classifier's behaviour.

In this work, we scrutinized numerous high-profile dermatology AI models to obtain a medically interpretable picture of medical-image AI reasoning processes. In the process, we showcase our workflow, which combines XAI with human domain expertise (Fig. 1a). We demonstrate solutions to severe practical issues with XAI in the imaging domain, including (1) conceptualizing AI behaviour in medically meaningful terms, (2) addressing sampling challenges to form robust conclusions, and (3) scaling from explanations of individual predictions to a global understanding of an AI classifier's reasoning processes. At a high level, our workflow involves the generative-AI-based synthesis of counterfactual images, which circumvent limitations of the de facto standard XAI modality (saliency maps) in medical-image analysis. Here we define counterfactuals as images that answer the question 'what realistic alterations elicit a different prediction from the AI?' We constrain the alterations to appear realistic, such that the differences between counterfactuals may be interpreted by medical experts (Methods). Our workflow continues with the analysis of thousands of such counterfactual images by dermatology experts, to characterize AI classifiers in human-understandable medical terms. Throughout the process, we emphasize rigour by mitigating problems of sampling and bias, via examination of numerous images, consideration of multiple datasets

and solicitation of insights independently from two dermatologists via a randomized and blinded analysis.

Results

Overview of dermatology AI classifier selection and reproduction

Aiming to best represent the current state of the art in dermatology AI classifiers, we explored the scientific literature and commercial market, ultimately choosing five AI classifiers to audit (Fig. 1b). These classifiers span the spectrum of academic and commercial classifiers and include classifiers already distributed for use by consumers. The five classifiers are: (1) DeepDerm, a previously developed reproduction²¹—using the original training data—of the classifier from a seminal academic publication¹⁵, which hailed the classifier for its ‘dermatologist-level’ performance; (2) ModelDerm 2018²², an academic classifier for which a later version (which we were unable to obtain) was CE approved for use in the European economic zone; (3 and 4) Scanoma and Smart Skin Cancer Detection (SSCD), two consumer-facing, smartphone apps; and (5) a ‘competition-style’ classifier, designed to mimic the key design decisions of the winning model²³ from the 2020 Society for Imaging Informatics in Medicine and International Skin Imaging Collaboration (SIIM-ISIC) Melanoma Classification Kaggle challenge²⁴ while circumventing that model’s prohibitive computational burden. Authors of additional AI classifiers declined to make available their full models (particularly the model weights), preventing us from analysing other high-profile classifiers^{16,17}.

As these diverse AI classifiers were trained on highly varied training data, we hypothesize they may show a wide range of internal reasoning processes, for instance, focusing on varied dermatological features or spurious signals. The training data include both dermoscopic images (taken through a specialized dermatological tool that magnifies and enables visualization of deeper layers of the skin) and clinical images (acquired with a digital camera, without the use of a dermatoscope). Dermoscopic and clinical images feature unique profiles of potential signals for AI systems to learn: for instance, dermoscopic images better reveal a lesion’s fine details, such as pigmentation patterns, and show unique artefacts, such as ruler markings and dark corner artefacts; clinical images likewise may provide more information on a lesion’s context (location, surrounding lesions), in addition to their own characteristic artefacts, such as presence of markings or patient clothing. Dermoscopic images from the ISIC database^{24–26} were used to train both DeepDerm and SIIM-ISIC, although the particular subsets of data used for each model differed. DeepDerm also included clinical images in its training set, gathered from numerous online sources. ModelDerm trained on only clinical images, including publicly available images as well as images that were never made publicly available. The training procedures for the smartphone app AI classifiers have not been published, but based on the wide public availability of dermatology image datasets, we speculate they could have trained at least in part on images from the ISIC archive, the Fitzpatrick17k database of clinical dermatology images²⁷ or other sources. Beyond the variability introduced by differences in training data, additional variation between the models may also arise from their diverse architectures, pre-processing schemes, ensembling and other computational differences.

We frame our analysis around the clinical task of differentiating melanomas from melanoma lookalikes (such as benign nevi, seborrheic keratoses or dermatofibromas), which has historically received great attention within the AI community and which aligns with the intended use cases of the AI classifiers. Four of the five AI classifiers explicitly predict melanoma, while the remaining AI classifier (DeepDerm) provides a more general prediction of ‘benign’ or ‘malignant’. To model this clinical task, we construct our test data to contain only melanomas and melanoma lookalikes; in this setting, DeepDerm effectively functions as a melanoma classifier, although the DeepDerm’s training for a more general task could still impart variation relative to the other classifiers.

We frame our analysis through this narrower problem, which has historically received great attention within the AI community and which models a well-defined clinical task. As some classifiers were designed to function on dermoscopic images, others on clinical images and at least one (DeepDerm) on both, we examine all classifiers in each context, using ISIC as our source of dermoscopic images and Fitzpatrick17k for clinical images (note that, as we are most interested in what alterations cause images to appear more benign or malignant and not benchmarking AI performance, we do not expect our XAI analysis to be sensitive to overlap between the training and test data)⁸.

We carefully adapted each AI classifier for use with our XAI tools, such that all analyses could be performed in a uniform software environment, thus eliminating a potential source of variation. Wherever feasible (that is, with the exception of SIIM-ISIC), we used the original model weights, to ensure that the original reasoning processes for that AI classifier could not change. While we suspect that the reasoning process of SIIM-ISIC should closely match the original 2020 SIIM-ISIC Kaggle competition winning model—we use the same training data, training procedure and test-time image augmentations/ensembling—we intend our audit of SIIM-ISIC to shed light on the influence of these common, performance-boosting techniques rather than to definitively comment on the reasoning process of that original model. We verified our adaptations against the original implementations and achieved close reproduction of the original results; only slight differences arose due to platform-dependent implementation differences in pre-processing or arithmetic (Supplementary Fig. 1).

Dermatology AI classifiers vary in melanoma-detection performance

As a first step toward understanding dermatology AI classifiers, we evaluated the performance of each classifier for differentiation between melanoma and melanoma lookalikes (Fig. 1c). While most AI classifiers detected melanomas in most datasets with at least limited success, performance was variable and often low. All failed to achieve satisfactory performance in the Diverse Dermatology Images (DDI) dataset, the only one of our three datasets known not to overlap with the training data of any AI classifier. This performance gap could come from the DDI’s inclusion of diverse skin tones and rare diseases but may also arise from other out-of-distribution features²¹. Despite training on no clinical images, SIIM-ISIC—which utilizes ensembling in conjunction with more modern neural network architectures—outperforms all other models on clinical images. Overall, our performance evaluation provides a sanity check that the dermatology AI classifiers likely rely in part on medically relevant attributes, given that most generalize, at least to a limited extent, to external datasets. In addition, our evaluation suggests that the five dermatology AI classifiers likely differ in their internal reasoning processes, as the pattern of performance gains or losses across the three datasets does not hold consistent among the AI classifiers. The findings from this retrospective analysis (which we do not intend as estimates of real-world performance as might be observed in deployment) motivate further analysis via XAI.

Counterfactual images reveal basis for AI decisions

To understand the reasoning processes of the AI classifiers, we examined each AI classifier via an XAI tool: generation of counterfactual images. Counterfactual images reveal the basis of an AI classifier’s decisions by altering attributes of a reference image to produce a similar image that elicits a different prediction from the AI classifier. For instance, consider the case that an AI classifier predicts a lesion is malignant, while a counterfactual predicted by the AI classifier to be benign differs in that it features lighter, more uniform pigmentation and fewer brown spots on the background skin; provided that we ensure all differences in the counterfactual push the AI classifier’s predictions in the desired direction (more benign), we may infer that the classifier uses

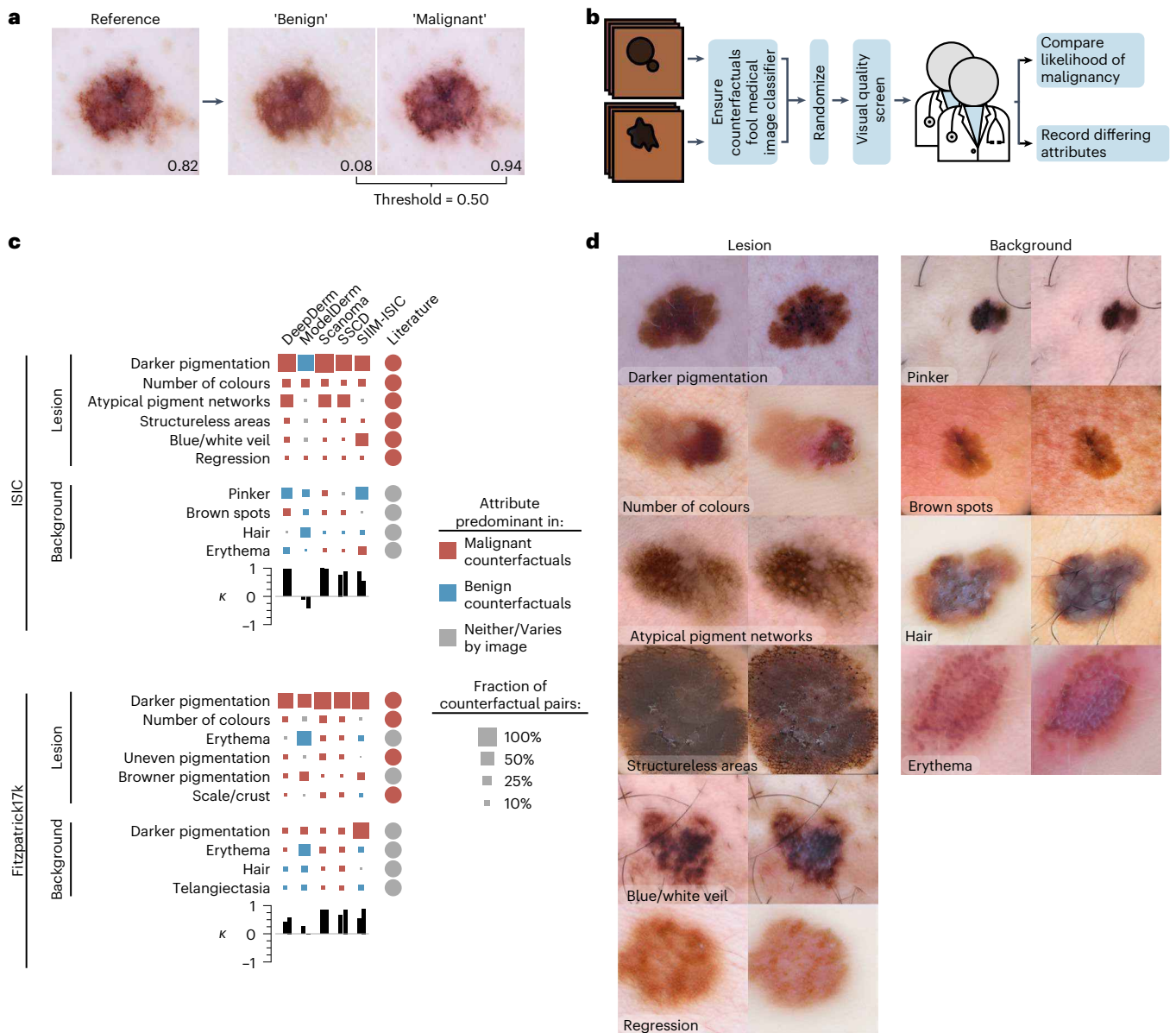


Fig. 2 | Joint expert and XAI auditing procedure reveals reasoning processes of dermatology AI classifiers.

a, Given a reference image and an AI classifier to investigate, our generative model produces ‘benign’ and ‘malignant’ counterfactuals, which resemble the reference image but differ in one or more attributes (such as pigmentation of the lesion and dots on the background skin). When evaluated by the AI classifier, the counterfactuals’ outputs lie on opposite sides of the decision threshold. Higher values indicate greater likelihood of malignancy, as predicted by an AI classifier (Scanoma). **b**, To obtain robust conclusions, dermatology experts evaluate numerous counterfactuals after pre-screening and randomization of the images. **c**, Attributes identified by our joint expert-XAI auditing procedure as key influences on the output of dermatology AI classifiers. For each attribute/classifier pair, we count the proportion of counterfactual pairs in which experts noted that attribute differs; we display the global top-10 attributes as determined by lowest rank-sum over all AI classifiers, then group by attribute category (‘lesion’ or ‘background’). Based on expert evaluation of whether the attribute was present to a greater extent in the malignant or benign counterfactual of each pair, we determine whether

that attribute was ‘predominant’ in benign or malignant counterfactuals, that is, present to a greater extent in benign (malignant) counterfactuals in at least twice as many images as malignant (benign) counterfactuals. The size of each square (the ‘fraction of counterfactual pairs’) is then determined as the proportion of counterfactual pairs with a difference noted in the predominant direction, averaged over both readers. For comparison, we specify how human dermatologists use each attribute (‘literature’), based on our review of the literature^{29–33,38,42} combined with expert opinion from two board-certified dermatologists; see ‘Discussion’ for additional information. Bar charts indicate Cohen’s κ values for agreement between each expert and the AI classifier, where each is asked which image in each counterfactual pair appeared more likely to be malignant. **d**, Examples of counterfactuals that differ in each of the top ten attributes identified in the ISIC data; the attribute is present to a greater extent in the right image of each pair. For conciseness, some attribute names were shortened; refer to Supplementary Table 1 for full names. Figure was adapted with permission from ref. 25, ViDIR Group, Department of Dermatology, Medical University of Vienna.

darker pigmentation of the lesion and brown spots on the background skin as part of its reasoning process (Fig. 2a).

To this end, we improved and applied a previously developed⁸ technique for generation of counterfactual images, Explanation by

Progressive Exaggeration, with updates to enable more rigorous conclusions. In the context of our dermatology AI classifiers, this technique enables the generation of both ‘benign’ and ‘malignant’ counterfactuals from a reference image (Fig. 2a). We can then learn from comparing

two opposing counterfactuals, which guards against potential misinterpretations, should the technique introduce any systematic changes to the counterfactuals. Explanation by Progressive Exaggeration trains a generative AI model in conjunction with an AI classifier, such that the generative model learns how to alter images to change the AI classifier's predictions. We train the generative model to create counterfactuals that are similar to the reference image and appear realistic but differ from the reference image to elicit the desired prediction from the AI classifier. Importantly, as the generated counterfactuals may alter more than one attribute, we updated the technique to ensure that we train the generative model to only change attributes when those changes elicit the desired effect on the AI classifier's output, whereas the previously published version of this technique may also alter attributes irrelevant to the classifier's output (Supplementary Fig. 2). Additional updates enabled generation of higher-quality images that retain fine details, such as hair, that might be important for dermatology AI classifiers (Supplementary Fig. 3). We separately trained such generative models for each AI classifier, for each of the ISIC and Fitzpatrick17k datasets, for a total of ten generative models (Methods and Supplementary Figs. 4 and 5); a uniform set of training parameters facilitates comparison between the AI classifiers (Supplementary Fig. 6).

While examination of a single counterfactual pair provides some information about an AI classifier's reasoning process, to obtain a more complete and rigorous understanding of the AI classifiers and enable direct comparisons between classifiers, we systematically interrogated thousands of counterfactual images, in a randomized and blinded fashion (Fig. 2b). We began our analysis by pre-screening the counterfactuals, to ensure we only examined high-quality counterfactuals and to facilitate comparisons between AI classifiers. We excluded counterfactuals that failed to produce the desired output from our AI classifiers (that is, we ensured that the 'malignant' and 'benign' counterfactuals lie on the correct sides of the decision threshold) or that contained visual artefacts (such as 'water-droplet-like' artefacts²⁸), as judged by dermatologists. Two dermatologists then independently annotated each counterfactual pair, which was randomized and blinded to reduce bias. To learn whether the dermatologists' general impressions of the counterfactuals agreed with each AI classifier regarding what appears more or less malignant, we first inquired, 'Which image appears most likely to represent a melanoma?' We then asked the dermatologists to record individual image attributes that differ between the 'benign' and 'malignant' counterfactuals, such that we could learn which attributes each AI classifier uses and how it uses them (Supplementary Fig. 7 and Supplementary Tables 1–3).

We aggregated the dermatologists' insights over thousands of counterfactuals to determine the reasoning process of each dermatology AI classifier. We conceptualize the reasoning process as swayed toward a benign or malignant prediction by key attributes identified as differing in counterfactual pairs; our analysis provides the typical direction of an attribute's effect, based on whether that attribute was predominant in the benign or malignant counterfactuals, as well as an approximate idea of the extent of the effect, based on the frequency with which dermatologists observed that attribute differing in counterfactuals. Note that we expect this frequency to depend on multiple factors, including the fraction of the dataset to which that attribute is relevant, inductive biases of our generative models and perhaps a combination of a dermatology AI system's sensitivity to an attribute and the sensitivity of our evaluators in detecting that attribute (which may be at odds, in the case of a visually subtle change that sizeably affects a prediction). Our analysis reveals that the AI classifiers focus on both medically relevant and putatively spurious attributes and show considerable heterogeneity in how they interpret those attributes (Fig. 2c).

A detailed view of medical AI reasoning

Our counterfactual analysis highlights the pigmentation of lesions as a key attribute in determining the predictions of all dermatology

AI classifiers examined, for both dermoscopic and clinical images. In all cases, 'darker pigmentation' surpassed all other attributes in frequency, with dermatologists noting this change in the majority of counterfactual pairs. Consistent with dermatologists' interpretation of more darkly pigmented lesions, dermatology AI classifiers typically associate darker pigmentation of lesions with increased likelihood of melanoma; the only exception is ModelDerm when evaluated on dermoscopic images—an image type upon which this model was never trained. Dermoscopic counterfactuals from a subset of the dermatology AI classifiers (DeepDerm, Scanoma, and SSCD) also showed atypical pigment networks, featuring these in the more 'malignant' images, in agreement with dermatologists' use of this attribute during pattern analysis of melanocytic lesions^{29,30}.

Our counterfactual analysis suggested that dermatology AI classifiers also depend on a variety of other attributes of the lesion, many of which dermatologists also consider when analysing melanocytic lesions. In both dermoscopic and clinical images, counterfactuals from all AI classifiers varied the number of colours in a lesion, typically associating a greater number of colours with predictions of malignancy³¹. Some AI classifiers, most prominently SIIM-ISIC, also elicited counterfactuals with blue/white veils, which has previously been reported as a specific finding for melanoma^{32,33}. Other attributes of the lesion that may factor into the AI classifiers' decisions include presence of structureless areas or regression in dermoscopic images, and uneven pigmentation or erythema in clinical images. Aside from erythema, which varies between a benign or malignant signal depending on the AI classifier, these attributes typically associate with the malignant counterfactuals. Their frequency, however, varies considerably between classifiers, pointing out heterogeneity in the classifiers' reasoning processes.

Analysis of each AI classifier's top attributes (Extended Data Figs. 2 and 3) revealed additional lesional attributes highlighted by counterfactuals from only a subset of the AI classifiers. In dermoscopic images, these attributes included patchiness (DeepDerm and SSCD), strawberry pattern (ModelDerm), white spots (SSCD), prominence of follicles or pores (SSCD), white striae (SIIM-ISIC) and scale (SIIM-ISIC). In clinical images, these attributes included erosion or ulceration (DeepDerm and Scanoma), nodular or papular appearance (ModelDerm), uneven borders (ModelDerm) and the shininess of a lesion (SIIM-ISIC).

Typically, inter-reader variability did not result in conflicting conclusions about the presence or direction of an attribute's effect (Extended Data Fig. 4).

Our counterfactuals indicate that attributes of the background skin also influence the dermatology AI classifiers; also, in comparison to attributes of the lesion, those of the background often elicit more diverse responses among the classifiers: Counterfactuals for multiple classifiers show brown spots on the background skin, and these variably associate with either malignant or benign predictions, depending on the classifier. Hair typically associates with benign counterfactuals in dermoscopic images but can also associate with malignant counterfactuals in clinical images. Reticulation of the background skin associates with the benign counterfactuals of Scanoma and ModelDerm (Extended Data Fig. 2) but is rarely highlighted by the counterfactuals of other classifiers. Erythema or telangiectasias of the background skin also feature prominently in the results of our counterfactual analysis, and the effects of these attributes vary both between AI classifiers and within an AI classifier, depending on whether an image is clinical or dermoscopic. Finally, counterfactuals highlighted the 'pinkness' of background skin as influencing AI classifiers' decisions, particularly in dermoscopic images. In contrast to erythema, this attribute often applies uniformly across an image (Fig. 2d), consistent with effects of lighting or an image's colour balance. Similarly, we recorded overall darker images and cooler colour temperatures as influential for one classifier (SIIM-ISIC). Similar to other background skin attributes,

lighting or colour balance changes may sway an AI classifier toward a more benign or more malignant prediction depending on the classifier. Aside from brown spots on the background skin, which could be interpreted as sun damage³⁴, we were unable to identify dermatological literature that establishes these attributes of the background skin as signals commonly used by dermatologists.

Darker pigmentation of the background skin, which stands out as the overall second most frequently recorded difference in our clinical counterfactuals, consistently associates with malignant counterfactuals. We observed that the darker pigmentation sometimes localized to discrete areas of the background skin, for instance to the immediate periphery of a lesion (effectively enlarging the lesion), or alternatively to areas of the image in shadow. In other instances, darker pigmentation extended more uniformly throughout the background skin. Among the classifiers, SIIM-ISIC featured this attribute most prominently in its counterfactuals.

In general, AI classifiers and human dermatologists agreed on which image in the counterfactual pair most likely depicted a malignancy. The exception, ModelDerm, showed negative Cohen Kappa values compared to dermatologists on dermoscopic images, in alignment with the unique profile of attributes highlighted in our analysis. This classifier also agreed poorly on clinical images, again coinciding with its focus on a unique profile of attributes. Curiously, Scanoma achieved the best agreement with dermatologists on both datasets, despite other AI classifiers achieving higher predictive performance (even when that performance was on external data and therefore not inflated by train-test overlap, as in the SIIM-ISIC with Fitzpatrick17k; Fig. 1c).

Validation of insights from counterfactuals

While we engineered our counterfactual generation procedure to ensure that detected attributes indeed influence AI classifiers' predictions, we performed additional analyses to verify these conclusions. Ideally, we may confirm our findings by performing a targeted intervention to experimentally modify a single attribute of an image, in a well-defined fashion, then monitor the intervention's effect on each AI classifier's prediction. While existing techniques such as CycleGANs (a type of generative adversarial network)³⁵ or manual image editing do not enable reliable modification of most attributes detected in our analysis (such as the addition or removal of atypical pigment networks without altering other attributes), transformation to a suitable colour space (in our case, the International Commission on Illumination's CIE 1976 L*, u*, v* color space, abbreviated CIELUV)³⁶ enables programmatic modification of the colour of an image, permitting us to experimentally produce images that are more or less 'pink', an attribute detected as influential to most classifiers (Figs. 2c and 3a). We shifted the colour (that is, the u' and v' chromaticity coordinates in the CIELUV colour space³⁶) of each image in the ISIC dataset, then monitored how each AI classifier's prediction changed for a range of colours (Fig. 3b).

These experimental modifications of image colour and their impact on the predictions of the AI classifiers recapitulates the trend observed in our previous analysis of counterfactual images (Fig. 3c compared to Fig. 3a): for example, pinker images elicit more benign predictions from DeepDerm and more malignant predictions from Scanoma. Multiple factors including the 'sensitivity' of an AI classifier to changes in an attribute determine the relative frequency of an attribute among counterfactuals (Fig. 3a); thus, magnitudes are not directly comparable (Results: 'Counterfactual images reveal basis for AI decisions'). This experiment validates that the attributes identified in our previous analysis of counterfactual images indeed influence the output of the AI classifiers in the direction described by the counterfactual analysis. In addition, this experiment validates our interpretation of 'pinker background skin' as a global change in lighting or colour balance. Indeed, our experimental procedure mirrors computational techniques used to perform white balancing (correction for chromatic adaptation) in digital cameras³⁷ and highlights how changes to lighting

or camera settings might affect AI dermatology classifiers' predictions in undesirable ways.

Counterfactuals explain failure cases

To reinforce the core findings from our systematic analysis of counterfactuals, we also present counterfactual explanations of cases in which the AI classifiers failed to correctly predict whether a lesion was malignant or benign.

The reliance of dermatology AI models on the pigmentation of a lesion can lead to failures that are 'reasonable', in that they might also be expected from human dermatologists (Fig. 4a): for instance, while presence of atypical pigment networks and darker pigmentation leads one AI classifier to predict a lesion was malignant, it turned out to be benign; indeed, authors of this present study who practice dermatology find this lesion concerning for the same reason and would have opted to biopsy the lesion.

In other cases, dermatology AI models rely on potentially relevant attributes of an image but use these attributes incorrectly. ModelDerm misclassified a malignant lesion as benign, and examination of the corresponding counterfactuals revealed attributes such as darker pigmentation of the lesion and absence of erythema as influential for this decision (Fig. 4b). However, dermatologists would not typically associate darker pigmentation with decreased likelihood of melanoma, and the distribution of erythema does not match the 'pink rim' sometimes associated with melanoma³⁸.

Dermatology AI classifiers also utilize likely irrelevant attributes in their reasoning process, including associating hair on background skin with benign lesions (Fig. 4b). In another example (Fig. 4c), a classifier misclassifies a benign lesion as melanoma in part due to an attribute of the background skin, namely, lack of prominent reticulation.

Discussion

Relative to previous techniques for the analysis of medical-image AI classifiers, our framework provides numerous advantages, which together enable us to present a detailed view of the reasoning processes of AI systems for medical images. Whereas saliency maps—the de facto standard XAI technique for image models—best reveal the importance of localizable attributes, our discovery of dependencies on numerous overlapping, textural and tonal changes to an image showcases the importance of our use of XAI based on counterfactual images, and highlights limitations of previous work that relied only on saliency maps⁵. In fact, we surmise that most attributes identified by our framework, such as darker pigmentation of lesions, number of colours in a lesion and the presence of erythema and pigmentation patterns, would be unlikely to be identified by saliency maps. Our framework also improves upon previous efforts^{8,9} to analyse medical-image AI systems via counterfactual images. In contrast with other generative techniques^{8,9} for counterfactual generation (including the original Explanation by Progressive Exaggeration) or with the simple comparison of real images predicted as benign and malignant, our method enables the inference that each attribute that differs in a benign–malignant pair is indeed important for the predictions of the AI classifier (Supplementary Fig. 2). Our method also offers a more detailed reproduction of fine-grained features such as hair (Supplementary Fig. 3), which we discovered to influence some AI classifiers. Perhaps more importantly, our framework introduces a means to translate XAI outputs to a human-understandable medically meaningful form, namely, via systematic, randomized and blinded analysis by medical experts. Particularly for a high-stakes application such as medical decision making, we contend that such a medically grounded understanding offers the greatest potential for actionability.

We find that dermatology AI classifiers leverage a number of medically meaningful attributes found within lesions, including attributes related to a lesion's pigmentation, in a manner consistent with human experts. Dermatology AI classifiers also rely on numerous attributes with debatable medical relevance and unclear desirability. Brown spots

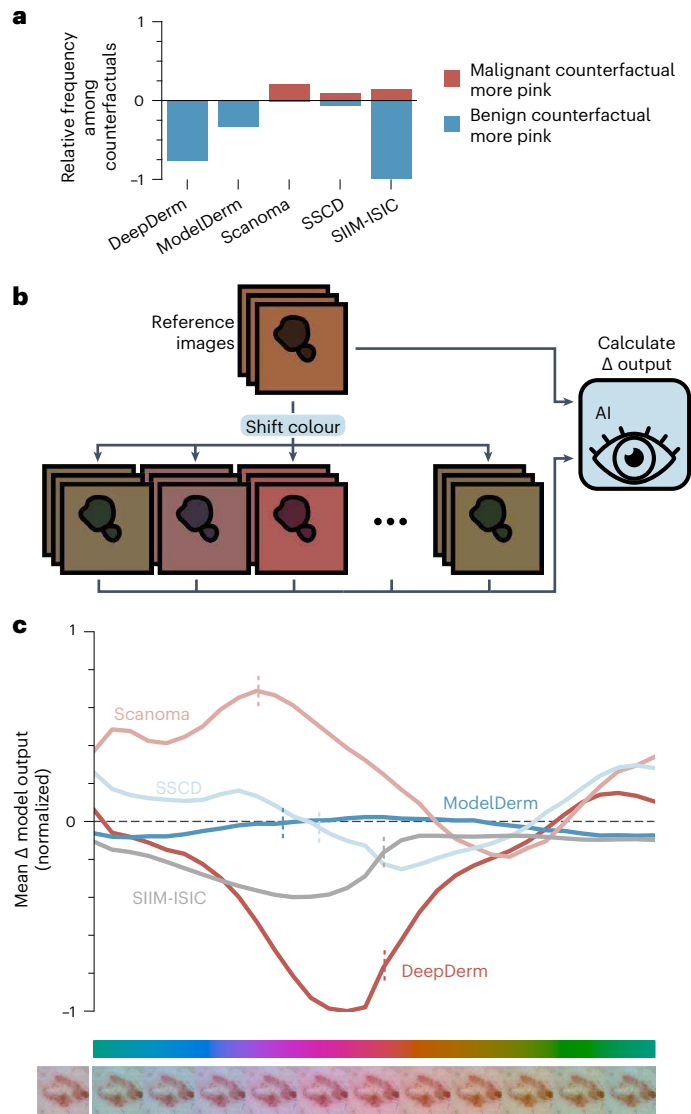


Fig. 3 | Experimental validation of findings from expert analysis of counterfactual images. **a**, The frequency with which experts noted that either the benign or malignant image in a pair of counterfactuals showed a pinker background; this view details our observations from the ISIC dataset summarized in Fig. 2c, in the row ‘pinker’ (ISIC dataset, ‘Background’ section). The vertical axis is normalized relative to the maximum observed frequency, that is, 42% of counterfactual pairs from SIIM-ISIC. **b**, Experimental set-up used to verify the importance of a pink tint to the AI classifiers’ predictions. We programmatically colour-shifted each image in the ISIC dataset ($n = 20,260$) by modifying its chromaticity coordinates in the CIELUV colour space (Methods), then compared each AI classifier’s predictions between the original and colour-shifted images. **c**, Sensitivity of each AI classifier to programmatic colour shifts, mirroring observations from our counterfactual experiments regarding the effect of pinker tints on the AI classifiers’ predictions. The vertical axis is normalized relative to the maximum change in AI classifier output, that is, a decrease of 0.17 with DeepDerm. Vertical dashed lines indicate the mean change in chromaticity (colour) among counterfactual pairs annotated as differing in their pink tone. Example colour-shifted images (below colour bar) show the extent of the colour shift; the reference image⁵⁰ appears at far left.

on the background skin may signify a patient’s age or history of sun exposure (a risk factor for melanoma³⁴) but are not in any established melanoma-diagnosis guidelines. Our observation of this attribute is consistent with previous works that suggested that AI classifiers may rely in part on perilesional sun damage when examining actinic keratoses or Bowen disease^{39,40}; a later study further corroborated that

perilesional sun damage also enhances the human diagnosis of actinic keratoses and, more directly relevant to the prediction task in our study, provided evidence that melanomas also show perilesional sun damage more frequently than benign nevi⁴¹. Erythema, particularly in a ‘pink rim’ distribution around a lesion³⁸, has been associated with melanoma, but also with benign melanoma lookalikes such as irritated seborrheic keratoses⁴². Hair may suggest a lesion’s location on the body, while skin grooves may provide clues on a lesion’s location (for example, acral) or the patient’s age or history of sun exposure. Lighting conditions or colour balance also influence many dermatology AI classifiers, and we surmise these almost certainly undesirable dependencies arise from spurious differences in image acquisition or pre-processing. The examined AI classifiers show considerable variability in their reasoning processes, especially with respect to their use of background attributes. While such variability might be partially explained by one model (DeepDerm) differing in its intended task (differentiation between malignant and benign lesions in general, as opposed to melanomas and benign melanoma lookalikes), the remaining models differ in reasoning processes despite sharing a common task. Beyond the fundamental scientific interest of this detailed characterization of AI reasoning processes, our approach could be used by AI developers to improve their models and to inform stakeholders on the trustworthiness of medical AI classifiers.

This methodology can help uncover idiosyncratic failure modes of AI, with implications for its regulation and medical use. We expect distributional shifts in medical AI to be common—especially in dermatology AI, given the diversity of image acquisition devices, lighting conditions, skin appearances across demographics and lack of implemented image standards. Our findings suggest that common distributional shifts, such as changes in lighting or colour balance, will alter AI performance. Thus, we caution potential users of such classifiers that a classifier’s advertised performance, which is often estimated in a well-circumscribed setting, may not be achieved in real-world use²¹. Our findings also imply that regulators should scrutinize the distribution of data on which a classifier is evaluated, with particular attention toward (1) ensuring it well reflects the intended deployment distribution, and (2) considering differential performance across subgroups (such as varied acquisition devices or regions or key potential dependencies such as lighting and skin tone). For AI developers, we envision that our methodology may enable more tractable debugging of AI classifiers before more expensive and time-consuming multi-site performance evaluations⁴³. Finally, our framework might directly assist physicians by revealing new attributes that they could subsequently use to improve their diagnostic skills, as was previously exemplified with perilesional sun damage as a diagnostic clue⁴¹. By contrast, while use of XAI outputs to support the case-by-case decision-making of physicians as part of a human-AI team has received attention within the XAI community, our framework is not directly applicable to this task but focuses more on large-scale auditing, and additional studies would be required to ascertain the utility of the underlying counterfactuals for verifying AI decisions^{44,45}.

In light of a recent study that highlighted how dermatology AI classifiers perform worse on darker skin tones²¹, we considered how our analysis might detail the underpinnings of this behaviour; that is, which aspects of the classifiers’ reasoning processes might lead to inequitable performance across skin tones. In some malignant counterfactuals, particularly those of SIIM-ISIC, annotators noted diffusely darker background skin compared to the benign counterfactual. As multiple real-world variations, including differences in skin tone or lighting conditions, might recapitulate this effect, the precise explanation remains unclear, but either case may be concerning. To the extent that real-world variations in skin tone may mirror this difference between the counterfactuals, a dermatology AI model may depend directly on skin tone. To the extent that real-world variation in lighting conditions or camera settings might mirror this difference, there is also potential

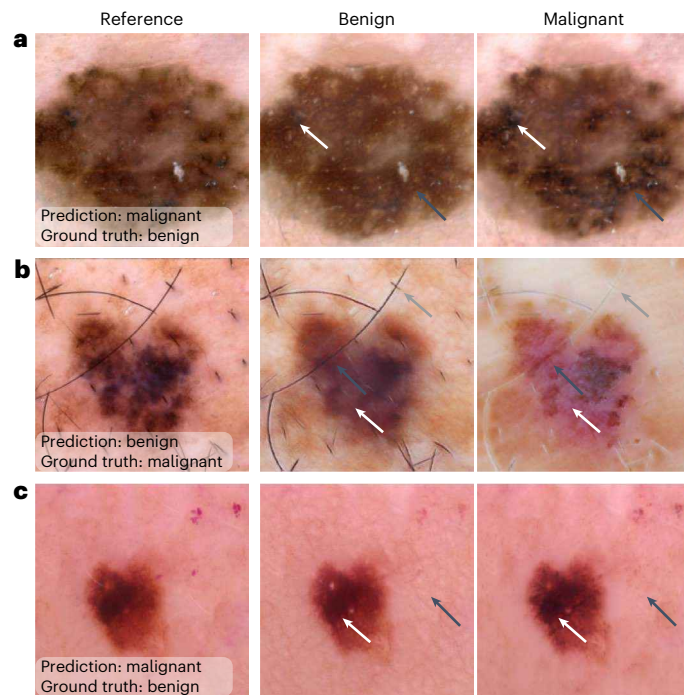


Fig. 4 | Explanations of failure cases of dermatology AI classifiers, illustrating key findings from our systematic analysis. a, Presence of atypical pigment networks (black arrows) and darker pigmentation (white arrows) contributed to a false-positive prediction from Scanoma. **b**, Lack of more colours of pigment may have contributed to a false-negative prediction from ModelDerm. Curiously, ModelDerm may have also required lighter pigmentation (black arrows), increased erythema (white arrows) and less hair on background skin (grey arrows) to correctly predict this image pictures a melanoma. **c**, Lack of prominent reticulation on the background skin (black arrows), alongside darker pigmentation of the lesion (white arrows), contributed to another false-positive prediction from Scanoma. Figure adapted, with permission, from ref. 25, ViDIR Group, Department of Dermatology, Medical University of Vienna.

for an indirect dependence of dermatology AI models on skin tone: camera designs are often biased toward ensuring appropriate exposure and colour in light skin tones, but not dark skin tones⁴⁶, implying that an AI classifier that depends on lighting and colour balance may as a result perform inequitably across skin tones. While we performed additional experiments modifying image brightness in hopes of better disentangling effects of lighting and skin tone, conclusions (Extended Data Fig. 5) varied considerably with the methodology used (in contrast to our experiments with image chromaticity; Extended Data Fig. 6). Finally, our counterfactuals occasionally highlighted reflections as influential, which could systematically bias predictions in images of dark skin acquired with suboptimal lighting (as with the use of camera flash)⁴⁷. Thus, our study suggests multiple potential avenues by which inequitable performance of dermatology AI classifiers may arise from a mechanistic point of view, although future studies would be required to alleviate ambiguity and verify potential links between skin tone and variations in image acquisition on a dataset-by-dataset basis.

While our framework provides a detailed picture of the reasoning processes of medical AI classifiers, limitations remain. First, we aimed to characterize the classifiers in medically meaningful, human-derived terms, but AI reasoning processes a priori need not coincide with human concepts. For instance, AI classifiers can predict sex from fundoscopic images⁴⁸, a challenging task for ophthalmologists, and the struggle to conceptualize these decisions in terms simple to humans⁴⁹ suggests the existence of peculiar, AI-specific abilities to detect certain attributes. While our use of an expressive XAI technique in combination with free-text annotations may improve our chances of capturing such

AI-specific attributes, human biases may nonetheless prevent their detection or description. Second, while our counterfactual generation technique is highly expressive (Extended Data Fig. 1), inductive biases may still limit detection of some attributes. For instance, considering similarities between our generative models and the CycleGAN, which struggles to produce large-scale geometric changes³⁵, our models may similarly be less likely to produce certain alterations in the counterfactuals, such as changes to the size of a lesion. Third, our approach does not provide information on the relationships between multiple attributes (for example, on the extent of any ‘interactions’ between attributes). Fourth, while we examine multiple modalities of dermatological images (clinical and dermoscopic), our analysis provides limited information on out-of-distribution features or features that rarely appear in the examined images (such as sutures). Fifth, the use of human annotators introduces variability, both due to stochastic effects of whether an annotator notices an attribute in a given image and due to variation in the background and training of experts. We found that our annotators typically agreed on the presence and direction of an attribute’s effect, but the frequency with which they noted that attribute was not quantitatively consistent (Extended Data Fig. 4). Thus, while the ‘fraction of counterfactual pairs’ in which an attribute was noted may help gauge our confidence in the attribute’s effect or enable approximate comparisons, granular comparisons of the ‘extent’ of an attribute’s effect are likely not meaningful. Moreover, domain experts from varied backgrounds may tend to focus on different attributes (for example, a dermoscopy expert may focus on traditional features of pattern analysis). Finally, while our use of free-text entry likely improves the expressiveness of our framework, there is no uniquely correct way to distil these responses into a uniform taxonomy, implying that another set of domain experts may have chosen different levels of granularity. Despite these limitations, we believe our use of an expressive XAI technique, expert annotators and free text entry together enable detailed, medically meaningful inferences on the AI classifiers’ reasoning processes and how they could lead to desirable or undesirable behaviour in deployment.

In addition to the immediate value of our analysis to understanding dermatology AI classifiers, the analysis provides a general framework for auditing complex AI systems that require specialized domain knowledge to best understand. Based on the success of our framework in multiple image modalities (dermoscopy and clinical images), for each of five AI classifiers, all in a particularly heterogeneous medical domain (dermatology), we anticipate that investigators could apply the framework towards understanding a variety of other AI systems: perhaps other AI medical-image analysis tools, such as the numerous AI-based medical image-analysis systems that have been deployed clinically, as well as non-medical computer-vision tasks such as facial recognition, scene classification in autonomous vehicles and industrial or agricultural monitoring. The modest number of images (less than 1,000) with which we were able to train a counterfactual generation model further bodes well for the broad applicability of this analysis. In addition, our framework for querying experts and compiling responses could be applied in conjunction with other XAI techniques to understand AI systems outside the image domain, in cases where input features still lack stable semantics, such as systems that operate on time-series data. More generally, our study offers a template for the rigorous application of XAI by addressing key issues that may have imperilled previous XAI analyses: insufficient sampling, potential for bias, lack of expert involvement and failure to examine AI systems in multiple contexts.

Methods

Image selection and pre-processing

To interrogate the performance of AI-based dermatological classifiers, we collected images of melanomas and melanoma lookalike lesions from multiple sources. We focus on this specific task for multiple

reasons, including (1) the substantial attention it has received within the machine-learning community^{24,50}, (2) alignment between this task and the intended use cases of the five AI classifiers, to enable comparison between classifiers, and (3) the improved likelihood of generating interesting information on the reasoning processes of dermatology AI classifiers, compared to simpler tasks that feature more visually salient signals.

Our first source, Fitzpatrick17k²⁷, consists of clinical (rather than dermoscopic) images previously aggregated from online dermatology atlases. We filtered Fitzpatrick17k to include only melanomas, benign melanocytic lesions, seborrheic keratoses and dermatofibromas. We additionally excluded diagrammatic and histopathological images and images that could be clearly identified as paediatric; after exclusions, the dataset consisted of 889 images. Advantages of Fitzpatrick17k include closer approximation of the expected inputs to consumer-facing dermatology AI tools (compared to dermoscopic images, which require specialized tools) and inclusion of a variety of skin tones. Disadvantages include its relatively small size after filtering and noise in the diagnosis labels, which may not have been acquired via histopathological analysis or other gold-standard means.

Our second source, the ISIC 2019 challenge dataset^{25,26,50}, consists of dermoscopic images from a variety of primary sources, including the ‘Human Against Machine with 10,000 training images’ (HAM10000) dataset²⁵ and the BCN20000 dataset²⁶. Like Fitzpatrick17k, we filtered the dataset to include melanomas, as well as melanoma lookalikes: benign melanocytic lesions, seborrheic keratoses and dermatofibromas. After filtering, the ISIC dataset consisted of 20,260 images. Most lesions were confirmed via histopathology ($n = 13,072$) or serial imaging showing no change ($n = 3,704$), while a smaller number were confirmed by single-image expert consensus ($n = 1,207$), confocal microscopy with consensus dermoscopy ($n = 712$) or unspecified means ($n = 1,565$). Compared to Fitzpatrick17k, ISIC thus offers more reliable diagnoses, but it lacks diversity in skin tones, featuring predominately light skin.

Finally, our third source, DDI²¹, consists of clinical images gathered from Stanford Clinics. Like other datasets, we filtered DDI to include only melanomas and melanoma lookalikes. In the case of DDI, which contains more granular and varied diagnoses, we included the following labels in our ‘melanoma’ category: acral lentiginous melanoma, melanoma in situ, nodular melanoma and the general tag ‘melanoma’. As melanoma lookalikes, we included the following labels: acral melanotic macule, atypical spindle cell nevus of reed, benign keratosis, blue nevus, dermatofibroma, dysplastic nevus, epidermal nevus, hyperpigmentation, keloid, inverted follicular keratosis, melanocytic nevi, nevus lipomatosus superficialis, pigmented spindle cell nevus of reed, seborrheic keratosis, irritated seborrheic keratosis and solar lentigo. After filtering, DDI included 282 images; due to the comparatively high volume of data required for training our generative models, DDI was used only for performance evaluation (Fig. 1) rather than for our in-depth analysis of medical AI reasoning processes. However, DDI offers a number of desirable characteristics for evaluation purposes: (1) its images were not publicly available until after we obtained the five audited dermatology AI classifiers, precluding train–test overlap; (2) DDI images have diverse skin tones, including enrichment for Fitzpatrick skin types V and VI; (3) DDI contains a wide variety of skin conditions, including uncommon conditions; and (4) the lesions are histopathologically proven, guaranteeing label accuracy. We note also that DDI is likely enriched for challenging lesions, as these are the lesions likely to require a biopsy.

For all evaluations, we pre-process the images to match the native input resolution of the AI classifier, which is 299×299 pixels for DeepDerm and 224×224 pixels for all other classifiers. When evaluating AI classifier performance or generating counterfactuals (after generator training is complete), we resize the image via bilinear interpolation such that its shorter edge matches the input size of the AI classifier, then centre-crop to obtain a square image. When training our generative models, which benefit from image augmentation, we instead

resize the image such that its shorter edge is 120% of the input size of the corresponding AI classifier, then perform a random square crop matching the input size.

Classifier reproduction

We reproduced five AI-based dermatological classifiers, including prominent academically designed classifiers proposed for clinical use and classifiers currently in use by the public. Two of the classifiers, Scanoma and SSCD, are designed for use on mobile devices by the general public. The DeepDerm classifier is a previously published reproduction²¹ of a prominent academic model¹⁵, sharing its training data and architecture. The ModelDerm 2018 classifier is a publicly distributed academic model²², of which a later iteration (for which model weights are not publicly available) has been CE marked for use by the general public in Europe. The SIIM-ISIC classifier is a reproduction of the first-place classifier²³ in the 2020 SIIM-ISIC Kaggle competition²⁴. These models cover a broad range of architectures, pre-processing techniques and training data sources; as such we believe these models offer a thorough view of both current practices and the state of the art in dermatology AI.

Scanoma is a commercial software available for mobile platforms including iOS and Android; at the time of writing, the app’s AI classifier is free to use, while follow-up human evaluation is available for a fee. Architecturally, it is a custom convolutional neural network consistent with a MnasNet⁵¹, that is further optimized for use on mobile devices via quantization⁵². We obtained and unzipped the Scanoma Android package (APK) file (normally installed on Android devices) to examine its TensorFlow Lite (TFLite) file, which contains the model specification and weights. As our analysis tools are based on the PyTorch software library (version 1.9), we converted the network to the cross-library Open Neural Network Exchange (ONNX) format, which we then parsed in PyTorch. To maintain consistency with the original, quantized network while maintaining useful gradients, we implement the network using ‘fake quantization’⁵². We verified that our PyTorch re-implementation matches the TensorFlow Lite implementation by comparing a series of 1,000 test images, and we achieved nearly identical outputs ($r = 0.99$; Supplementary Fig. 1a). To account for the small discrepancy between the classifiers, we analysed the processing pipeline step by step and found slight differences in the bilinear rescaling pre-processing step, which may differ due to different anti-aliasing constants; the remaining differences were explained by sporadic single-bit differences in the quantized feature maps, likely resulting from numerical differences between TensorFlow Lite’s native integer arithmetic routines and the equivalent operations performed in floating point arithmetic followed by fake quantization.

Like Scanoma, SSCD is a publicly available app intended for use on mobile devices. The architecture is a MobileNetV1, evaluated using floating-point (non-quantized) arithmetic. We followed a similar process to re-implement the SSCD classifier in PyTorch: a TFLite file was obtained from the app’s APK package, then converted to ONNX before loading in PyTorch. We again verified our reproduction using a series of 1,000 images and found that our PyTorch re-implementation of the neural network exactly matched the original TensorFlow Lite network. However, to ease comparison between classifiers, we update the input image resizing routine (a pre-processing step, before the neural network) in our implementation relative to the original app. Whereas the original app asks a user to specify a bounding box and then scales this box to the 224×224 pixel input image (warping the aspect ratio), we use the same pre-processing routine for all other networks, in which we first centre-crop the image and then resize the image using a bilinear filter. To assess the impact of this change in image pre-processing, we compared our PyTorch implementation against (1) the original TFLite model accompanied by pre-processing with square centre-cropping and nearest-neighbour resizing and (2) the original TFLite model with variable aspect-ratio resizing using nearest-neighbour rescaling

(matching the original Android implementation, under the assumption that the uncropped image represents a user-defined bounding box), and we observed Pearson correlation coefficients of 0.97 and 0.92, respectively (Supplementary Fig. 1b,c). While evaluation of the entire processing pipeline including user selection of bounding boxes and choice of resampling filters is important for clinical evaluation of an AI system, our study instead focuses on the decision-making processes of the neural networks.

ModelDerm²² is an academic classifier that has undergone multiple iterations, some of which have been tested in clinical settings and one version of which has been approved for use in Europe via CE marking. We analyse the latest version for which model weights are publicly available, which we term ModelDerm 2018 based on the date of the accompanying publication²²; authors declined to provide weights for the latest version of the model due to commercialization plans. ModelDerm is a ResNet-152⁵³ that runs natively in PyCaffe, with pre-processing performed in OpenCV (in our reproduction efforts, we used version 3.4.2). We parse the model architecture and weights directly from Caffe Protocol Buffer files and reconstruct the model in PyTorch. While the majority of the processing pipeline is highly reproducible in PyTorch relative to the original implementation, the original implementation pre-processes images channel by channel using the histogram equalization function in OpenCV, which we could not exactly reproduce in PyTorch while maintaining meaningful gradients during backpropagation. Instead, we implemented a custom, differentiable analogue of histogram equalization, in which the empirical cumulative density function used in OpenCV's implementation is replaced with a piecewise linear approximation. Our PyTorch reimplementations of ModelDerm 2018, including the differentiable histogram equalization pre-processing step, retains close correspondence to the original PyCaffe/OpenCV implementation ($r = 0.96$; Supplementary Fig. 1d).

The SIIM-ISIC competition classifier is intended to represent key features responsible for the high performance of the first-place winning classifier from the 2020 SIIM-ISIC melanoma classification Kaggle challenge, while reducing the computational complexity to permit feasible analysis. The original classifier is an extremely large ensemble of 90 networks, comprising mostly neural networks of the 'EfficientNet' architecture⁵⁴ but also a few networks of the 'SE-ResNext 101' architecture⁵⁵ and 'ResNest101' architecture⁵⁶, all of which are evaluated at test time on 8 flips and rotations of the test image, for a total of 720 model evaluations per prediction. We reduced the computational complexity by retraining an ensemble of 3 EfficientNets (an EfficientNet-B5, -B6 and -B7), which comprise 80 of the 90 classifiers in the original ensemble, using a lower resolution of 224×224 pixels. To encourage similarity to the original model, we use the same training data, augmentation scheme and hyperparameters as the original classifiers. Our classifier additionally retains eightfold image augmentation at test time, which we suspected may reduce the classifier's sensitivity to subtle image variations. While not intended to be an exact reproduction of the original winning classifier, our classifiers attain only slightly lower classification performance in fivefold cross validation compared to the original classifier (area under the receiver operating characteristic curve of 0.966 versus 0.985).

The DeepDerm classifier is a previously published reproduction²¹ of an academically developed model that was acclaimed for performing similarly well to dermatologists¹⁵. DeepDerm shares the same architecture (Inception-V3 (ref. 57)) and, importantly, the same training data as the original model, which was not publicly released. As DeepDerm is distributed natively in PyTorch, no conversion steps were necessary for this classifier.

Counterfactual generation

To identify specific image factors responsible for each classifier's predictions, we generated counterfactual images using a variant of the technique 'explanation by progressive exaggeration'⁸. However, to

improve image quality, stabilize training and better restrict generated alterations to those that cause a classifier to output a different prediction, we introduce multiple updates. We begin with an overview of the technique, then explain our specific updates. Full details of our generative models, including a formal mathematical treatment and an explanation of training parameters, are described in Supplementary Methods.

At a high level, a counterfactual considers a scenario that did not occur, typically for the purpose of comparison to a scenario that did occur or to another counterfactual scenario. Such comparison may enable inferences about how a different AI classifier output may have been achieved or which factors lead to that outcome. To enable these inferences, a counterfactual must typically be sufficiently similar to allow comparison, while differing in a realistic manner and eliciting a different outcome. In our case, we consider counterfactual images, which are alternative versions of real images. To create these counterfactual images, we use a type of generative image AI based on generative adversarial networks. We train our generative AI models to produce counterfactuals by altering real, reference images, with the goal of eliciting different predictions from an AI classifier; we also constrain these differences to be realistic (Supplementary Table 4). Then, examination of the differences between counterfactuals thus enables inferences regarding which image attributes influence an AI classifier's predictions.

We updated an existing generative AI technique for counterfactual images, explanation by progressive exaggeration⁸ (Supplementary Figs. 8 and 9), to better suit our purposes: First, we found that the original formulation of this technique could alter attributes of an image upon which the classifier does not depend but which correlate with attributes upon which it does depend (Supplementary Fig. 2). We found that this behaviour, which could lead to misinterpretations about the reasoning processes of the AI classifiers, arose from the specification of the discriminator, a component of the generative model that helps ensure realism of the generated images, and thus we updated our discriminator to remove this behaviour (see Supplementary Methods for full details). Second, we also update the generator component of our model to use an architecture similar to that used in CycleGANs³⁵. This network is similar to the residual network-based autoencoder used in the original implementation of Explanation by Progressive Exaggeration, but we found it produced images of higher visual quality (Supplementary Fig. 3). Finally, we applied data augmentation, including random cropping and random brightness modifications, to improve training when only a modest number of images are available (as is the case for Fitzpatrick17k).

Expert evaluation of counterfactuals

To identify specific image factors upon which dermatological classifiers base their predictions, we asked two board-certified dermatologists, each with 6 years of experience, to analyse generated counterfactual images and determine which aspects of each image were altered, implying that they affect the classifiers' decisions. We queried these dermatologists on hundreds of pairs of counterfactuals for each of five classifiers and two image datasets, amounting to thousands of responses. Each pair of counterfactuals was generated from a common 'reference' image and consisted of an image that the classifier predicted to appear more benign and an image that the classifier predicted to appear more malignant, such that both images depicted the same lesion but showed differences that altered the output of a classifier.

To facilitate interpretation of the dermatologists' responses and comparison of the classifiers, we pre-screened the counterfactual images before analysis of the alterations within counterfactual pairs. Our pre-screening consisted of a 'classifier-consistency' criterion to ensure that the alterations between each pair of counterfactuals meaningfully changed the classifiers' predictions and a 'visual quality' criterion to mitigate the presence of artefacts, which could impede our ability to infer the importance of non-artefactual alterations. Our

classifier-consistency criterion required the ‘benign’ and ‘malignant’ images in a counterfactual pair lay on opposite sides of the decision threshold (that is, they were classified as benign and malignant). In the visual-quality pre-screening step, two board-certified dermatologists independently evaluated for artefacts each image that passed the classifier-consistency criterion, and we excluded images rejected by either evaluator. To ease comparison between classifiers, we included the same set of counterfactual pairs (modulo counterfactual alterations) for all classifiers; more precisely, for each reference image x_r , we included the corresponding counterfactual images $\{G_c(x_r)\}_{c \in C}$ (where C represents the set of classifiers, and G_c is the generative model trained to produce counterfactuals for classifier c), if and only if $G_c(x_r)$ passed the pre-screen for each classifier c . For subsequent analysis, we included the 92 images from Fitzpatrick17k that passed our pre-screening criteria, and we included 100 images from ISIC to achieve a similar quantity of images.

To learn which attributes differ between benign and malignant counterfactuals—and thus influence an AI classifier’s predictions—we developed a two-stage annotation approach. We designed the first stage of this approach to encourage discovery of a wide variety of attributes, which we then leverage in the second stage to more efficiently collect data. Both stages leverage a graphical interface that runs locally in a web browser; expert evaluators view a pair of benign and malignant counterfactuals, then answer questions regarding (1) which member of the pair appears most likely to be malignant and (2) what attributes differ, and how they differ, between the counterfactuals. In the first stage, evaluators enter attributes as free text (for example, ‘skin lines more prominent’), accompanied by a ‘direction’ specifying how the images differ (Supplementary Fig. 7). After the first 100 pairs were evaluated by each expert, we pooled and grouped the free text terms to determine ‘pre-set’ attributes (such as ‘skin lines more prominent’ and ‘more skin lines’ map to the pre-set ‘Prominence of skin grooves/dermatoglyphs’) that could be selected during the second stage of annotation. This stage also retained the option for free text entry in case a new attribute was discovered. To mitigate potential bias, we randomized and blinded evaluators to (1) the appearance order of a counterfactual pair (that is, whether the benign or malignant counterfactual appeared on the left or right) and (2) the overall order of the counterfactual pairs, including randomization of the corresponding reference images and shuffling counterfactual pairs from the various AI classifiers. Evaluators annotated the counterfactual pairs in sets of 20, which required approximately 30 min to complete.

To infer general conclusions regarding which attributes influence the AI classifiers, we aggregated data from both evaluators and both stages of annotation. First, we mapped the free text attributes from the first stage of annotation to a common list of attributes, as agreed upon by the evaluators. We then filtered any counterfactual noted by either evaluator as ‘unable to assess’ due to the presence of substantial artefacts, which amounted to 4% of the total images. Finally, to obtain a global picture of each AI classifier, we tabulated the number of times an evaluator noted an attribute, along with the direction in which that attribute differed between the benign and malignant counterfactuals. Mathematically, we define an indicator function $s_{e,c,a,d,i}$ as 1 if evaluator e recorded for AI classifier c that attribute a differs in direction d in image i , and $s_{e,c,a,d,i} = 0$ otherwise. Then the score for an AI classifier is given by the mean of s over images $i \in I$ and evaluators $e \in E$, where I is the set of all examined images (that is, those that pass the pre-screen) and E is the set of evaluators:

$$\bar{s}_{c,a,d} := \sum_{i \in I} \sum_{e \in E} s_{e,c,a,d,i} / \sum_{i \in I} \sum_{e \in E} 1$$

To visualize the resulting values (Fig. 2), we further aggregated the ‘directions’ d which originally included five options: benign only, benign > malignant, different, benign < malignant and malignant only

(during data collection, which was blinded, these terms appears as A only, $A < B$ and so on, where images A and B were randomized to benign or malignant). We aggregated benign only and benign > malignant into a new category, benign, and likewise aggregated benign < malignant and malignant only into the new category malignant. Finally, for each pair of attribute and AI classifier, we determined the ‘predominate direction’ of that attribute which we defined as benign if $\bar{s}_{c,a,\text{benign}} > 2 \cdot \bar{s}_{c,a,\text{malignant}}$, we defined as malignant if $\bar{s}_{c,a,\text{malignant}} > 2 \cdot \bar{s}_{c,a,\text{benign}}$, and we defined as neither otherwise, where the cut-off factor of 2 was chosen to prevent emphasis on small differences in frequency between the benign and malignant directions. In Fig. 2, the size of the square is then proportional \bar{s} for the predominate direction or the average of the directions if neither was predominate.

Experimental validation of findings from counterfactuals via colour shifts

To validate the attributes identified as important for dermatology AI classifier’s predictions in our counterfactual experiments, we aimed to experimentally modify a single attribute and observe the effect on each AI classifier; we chose image colour as a test case, as existing mathematical tools³⁶ enable well-defined, unambiguous changes to this attribute. To alter the colour of each image, we converted from the sRGB colour space to the CIE 1976 L*, u*, v* colour space (CIELUV)³⁶, added an offset to the chromaticity coordinates (u^* , v^*), then converted back to sRGB. Different chromaticity shifts were generated by varying the offset along a circle centred at (u^* , v^*) = 0 with radius 20, where the factor 20 was chosen heuristically to produce colour changes that we deemed visible while remaining plausible.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The images used in this study were obtained from publicly available repositories. ISIC images are available at <https://challenge.isic-archive.com/data>. Fitzpatrick17k images are available at <https://github.com/mattgroh/fitzpatrick17k>. The DDI images are available at <https://stanfordaimi.azurewebsites.net/datasets/35866158-8196-48d8-87bf-50dca81df965>. Model weights for the DeepDerm classifier are available at <https://zenodo.org/record/6784279#.ZFrDc9LMK-Z>. The weights and model specification for the ModelDerm classifier are available at https://figshare.com/articles/Caffemodel_files_and_Python_Examples/5406223. Model weights for our retrained variant of the SIIM-ISIC competition classifier are available at <https://zenodo.org/doi/10.5281/zenodo.10049216>. Scanoma and Smart Skin Cancer Detection are third-party software for which we cannot redistribute model weights. At the time of writing, both are apps that are available for download with no fee from the Google Play store and from third-party APK-package download sites.

Code availability

Custom codes, including a PyTorch implementation of explanation by progressive exaggeration and of classes for loading datasets and classifiers, are available at https://github.com/suinleelab/derm_audit. The weights for the trained generative models and the re-trained SIIM-ISIC classifier are available at <https://zenodo.org/doi/10.5281/zenodo.10049216>.

References

1. Wu, E. et al. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. *Nat. Med.* **27**, 582–584 (2021).
2. Reddy, S. Explainability and artificial intelligence in medicine. *Lancet Digit. Health* **4**, E214–E215 (2022).

3. Young, A. T. et al. Stress testing reveals gaps in clinic readiness of image-based diagnostic artificial intelligence models. *npj Digit. Med.* **4**, 10 (2021).
4. DeGrave, A. J., Janizek, J. D. & Lee, S.-I. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nat. Mach. Intell.* **3**, 610–619 (2021).
5. Singh, N. et al. Agreement between saliency maps and human-labeled regions of interest: applications to skin disease classification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 3172–3181 (IEEE, 2020).
6. Bissoto, A., Fornaciali, M., Valle, E. & Avila, S. (De) constructing bias on skin lesion datasets. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2766–2774 (IEEE, 2019).
7. Winkler, J. K. et al. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA Dermatol.* **155**, 1135–1141 (2019).
8. Singla, S., Pollack, B., Chen, J. & Batmanghelich, K. Explanation by progressive exaggeration. In *International Conference on Learning Representations (ICLR, 2020)*.
9. Mertes, S., Huber, T., Weitz, K., Heimerl, A., & Andr, E. GANterfactual—counterfactual explanations for medical non-experts using generative adversarial learning. *Front. Artif. Intell.* **5**, 825565 (2022).
10. Ghoshal, B. & Tucker, A. Estimating uncertainty and interpretability in deep learning for coronavirus (COVID-19) detection. Preprint at arXiv:2003.10769 (2020).
11. Ozturk, T. et al. Automated detection of COVID-19 cases using deep neural networks with X-ray images. *Comput. Biol. Med.* **121**, 103792 (2020).
12. Brunese, L., Mercaldo, F., Reginelli, A. & Santone, A. Explainable deep learning for pulmonary disease and coronavirus COVID-19 detection from X-rays. *Comput. Methods Programs Biomed.* **196**, 105608 (2020).
13. Karim, M. et al. DeepCOVIDExplainer: explainable COVID-19 diagnosis from chest X-ray images. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 1034–1037 (IEEE, 2020).
14. Geirhos, R. et al. Shortcut learning in deep neural networks. *Nat. Mach. Intell.* **2**, 665–673 (2020).
15. Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
16. Liu, Y. et al. A deep learning system for differential diagnosis of skin diseases. *Nat. Med.* **26**, 900–908 (2020).
17. Han, S. S. et al. Augmented intelligence dermatology: deep neural networks empower medical professionals in diagnosing skin cancer and predicting treatment options for 134 skin disorders. *J. Invest. Dermatol.* **140**, 1753–1761 (2020).
18. Sun, M. D. et al. Accuracy of commercially available smartphone applications for the detection of melanoma. *Br. J. Dermatol.* **186**, 744–746 (2022).
19. Freeman, K. et al. Algorithm based smartphone apps to assess risk of skin cancer in adults: systematic review of diagnostic accuracy studies. *Br. Med. J.* **368**, m127 (2020).
20. Beltrami, E. J. et al. Artificial intelligence in the detection of skin cancer. *J. Am. Acad. Dermatol.* **87**, 1336–1342 (2022).
21. Daneshjou, R. et al. Disparities in dermatology AI performance on a diverse, curated clinical image set. *Sci. Adv.* **8**, eabq6147 (2022).
22. Han, S. S. et al. Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. *J. Invest. Dermatol.* **138**, 1529–1538 (2018).
23. Ha, Q., Liu, B. & Liu, F. Identifying melanoma images using EfficientNet ensemble: winning solution to the SIIM-ISIC melanoma classification challenge. Preprint at arXiv:2010.05351 (2020).
24. Rotemberg, V. et al. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Sci. Data* **8**, 34 (2021).
25. Tschandl, P., Rosendahl, C. & Kittler, H. The HAM10000 dataset, a large collection of multi-source dermoscopic images of common pigmented skin lesions. *Sci. Data* **5**, 180161 (2018).
26. Combalia, M. et al. BCN20000: dermoscopic lesions in the wild. Preprint at arXiv:1908.02288 (2019).
27. Groh, M. et al. Evaluating deep neural networks trained on clinical images in dermatology with the Fitzpatrick 17k dataset. In *Proceedings of the Computer Vision and Pattern Recognition (CVPR) Sixth ISIC Skin Image Analysis Workshop (IEEE, 2021)*.
28. Karras, T. et al. Analyzing and improving the image quality of StyleGAN. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 8107–8116 (IEEE, 2020).
29. Shi, K. et al. A retrospective cohort study of the diagnostic value of different subtypes of atypical pigment network on dermoscopy. *J. Am. Acad. Dermatol.* **83**, 1028–1034 (2020).
30. Yélamos, O. et al. Usefulness of dermoscopy to improve the clinical and histopathologic diagnosis of skin cancers. *J. Am. Acad. Dermatol.* **80**, 365–377 (2019).
31. Halpern, A. C., Marghoob, A. A. & Reiter, O. Melanoma Warning Signs: *What You Need to Know About Early Signs of Skin Cancer* (Skin Cancer Foundation, 2021); <https://www.skincancer.org/skin-cancer-information/melanoma/melanoma-warningsigns-and-images/>. Accessed April 2023.
32. Massi, D., De Giorgi, V., Carli, P. & Santucci, M. Diagnostic significance of the blue hue in dermoscopy of melanocytic lesions: a dermoscopic-pathologic study. *Am. J. Dermatopathol.* **23**, 463–469 (2001).
33. Marghoob, N. G., Liopyris, K. & Jaimes, N. Dermoscopy: a review of the structures that facilitate melanoma detection. *J. Osteopath. Med.* **119**, 380–390 (2019).
34. Oliveria, S. A., Saraiya, M., Geller, A. C., Heneghan, M. K. & Jorgensen, C. Sun exposure and risk of melanoma. *Arch. Dis. Child.* **91**, 131–138 (2006).
35. Zhu, J.-Y., Park, T., Isola, P. & Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)* 2223–2232 (IEEE, 2017).
36. Illumination, I. C. on. ISO/CIE 11664-5:2016(e) Colorimetry—part 5: CIE 1976 L*u*v* colour space and u', v' uniform chromaticity scale diagram (2016).
37. Deng, Z., Gijsenij, A. & Zhang, J. Source camera identification using auto-white balance approximation. In *2011 IEEE International Conference on Computer Vision* 57–64 (IEEE, 2011).
38. Rader, R. K. et al. The pink rim sign: location of pink as an indicator of melanoma in dermoscopic images. *J. Skin Cancer* **2014**, 719740 (2014).
39. Tschandl, P. et al. Human–computer collaboration for skin cancer recognition. *Nat. Med.* **26**, 1229–1234 (2020).
40. Tschandl, P. et al. Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based international, diagnostic study. *Lancet Oncol.* **20**, 938–947 (2019).
41. Weber, P., Sinz, C., Rinner, C., Kittler, H. & Tschandl, P. Perilesional sun damage as a diagnostic clue for pigmented actinic keratosis and Bowen's disease. *J. Eur. Acad. Dermatol. Venereol.* **35**, 2022–2026 (2021).
42. Fitzpatrick, J. E., High, W. A. & Kyle, W. L. *Urgent Care Dermatology: Symptom-Based Diagnosis*. 477–488 (Elsevier, 2018).

43. Wu, E. et al. *Toward Stronger FDA Approval Standards for AI Medical Devices* (Stanford University Human-centered Artificial Intelligence (2022).
44. Bansal, G. et al. Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (ACM, 2021).
45. Rok, R. & Weld, D. S. In search of verifiability: explanations rarely enable complementary performance in AI-advised decision making. Preprint at [arXiv:2305.07722v3](https://arxiv.org/abs/2305.07722v3) (2023).
46. Roth, L. Looking at Shirley, the ultimate norm: colour balance, image technologies, and cognitive equity. *Can. J. Commun.* **34**, 111–136 (2009).
47. Lester, J. C., Clark, L., Linos, E. & Daneshjou, R. Clinical photography in skin of colour: tips and best practices. *Br. J. Dermatol.* **184**, 1177–1179 (2021).
48. Poplin, R. et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat. Biomed. Eng.* **2**, 158–164 (2018).
49. Yamashita, T. et al. Factors in color fundus photographs that can be used by humans to determine sex of individuals. *Transl. Vis. Sci. Technol.* **9**, 4 (2020).
50. Codella, N. C. F. et al. Skin lesion analysis toward melanoma detection: a challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), hosted by the International Skin Imaging Collaboration (ISIC). In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI)*, 168–172 (IEEE, 2018).
51. Tan, M. et al. MnasNet: platform-aware neural architecture search for mobile. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 2820–2828 (IEEE, 2019).
52. Jacob, B. et al. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2704–2713 (IEEE, 2018).
53. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778 (IEEE, 2016).
54. Tan, M. & Le, Q. EfficientNet: rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)* 6105–6114 (PMLR, 2019).
55. Hu, J., Shen, L. & Sun, G. Squeeze-and-excitation networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 7132–7141 (IEEE, 2018).
56. Zhang, H. et al. ResNeSt: split-attention networks. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* 2735–2745 (IEEE, 2022).
57. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2818–2826 (IEEE, 2016).
58. Giotis, I. et al. MED-NODE: a computer-assisted melanoma diagnosis system using non-dermoscopic images. *Expert Syst. Appl.* **42**, 6578–6585 (2015).

Acknowledgements

A.J.D., J.D.J., and S.-I.L. were supported by the National Science Foundation (CAREER DBI-1552309 and DBI-1759487) and the National Institutes of Health (R35 GM 128638 and R01 AG061132). R.D. was supported by the National Institutes of Health (5T32 AR007422-38) and the Stanford Catalyist Program.

Author contributions

A.J.D., J.D.J., R.D. and S.-I.L. conceived the initial study. A.J.D. prepared data and developed software for the reproduction of dermatology AI classifiers, for their counterfactual analysis and for confirmatory experiments. A.J.D. and J.D.J. developed software for the generation of saliency maps. Z.R.C. and R.D. analysed counterfactual images and examined saliency maps. A.J.D., Z.R.C., J.D.J., R.D. and S.-I.L. analysed data and designed additional experiments. Z.R.C. and R.D. provided dermatological insights and clinical context. A.J.D., Z.R.C., J.D.J., R.D., and S.-I.L. wrote the manuscript. S.-I.L. secured funding, and R.D. and S.-I.L. jointly supervised the study.

Competing interests

R.D. reports fees from L’Oreal, Frazier Healthcare Partners, Pfizer, DWA and VisualDx for consulting; stock options from MDAcne and Revea for advisory board; and research funding from UCB. The other authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41551-023-01160-9>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41551-023-01160-9>.

Correspondence and requests for materials should be addressed to Roxana Daneshjou or Su-In Lee.

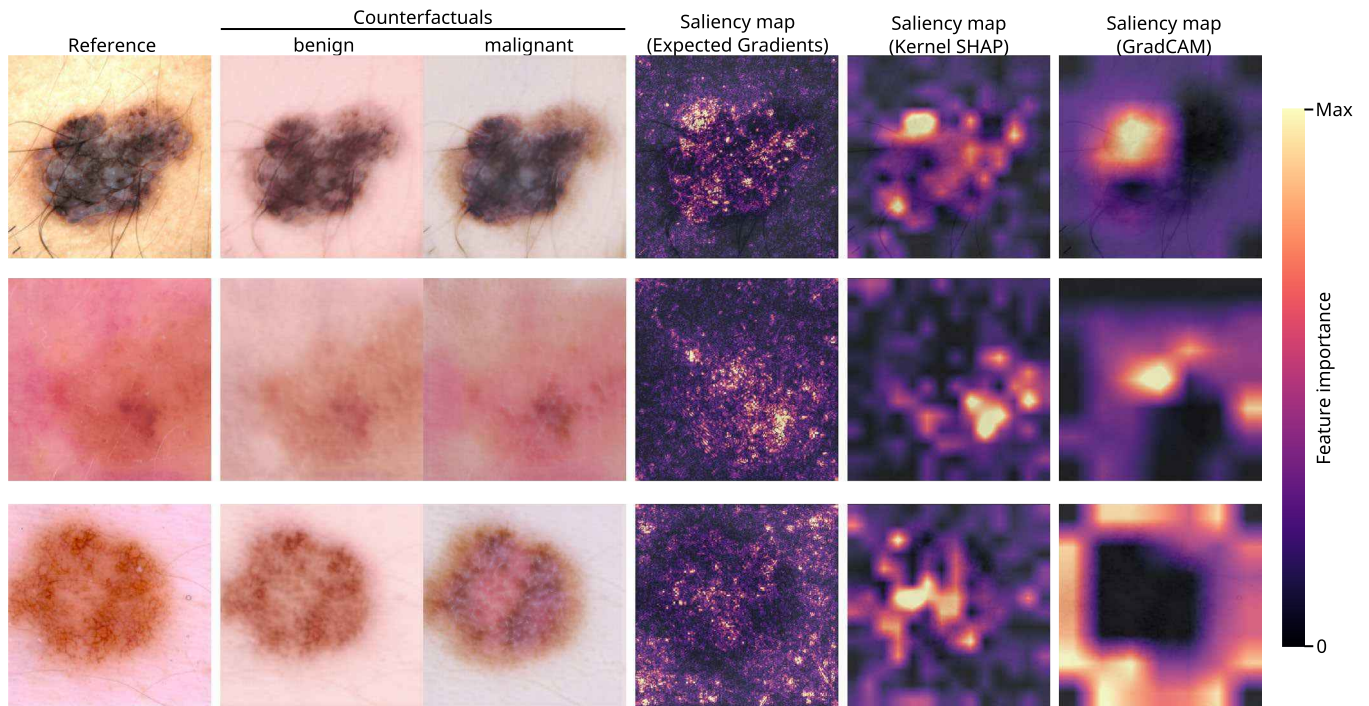
Peer review information *Nature Biomedical Engineering* thanks the anonymous reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

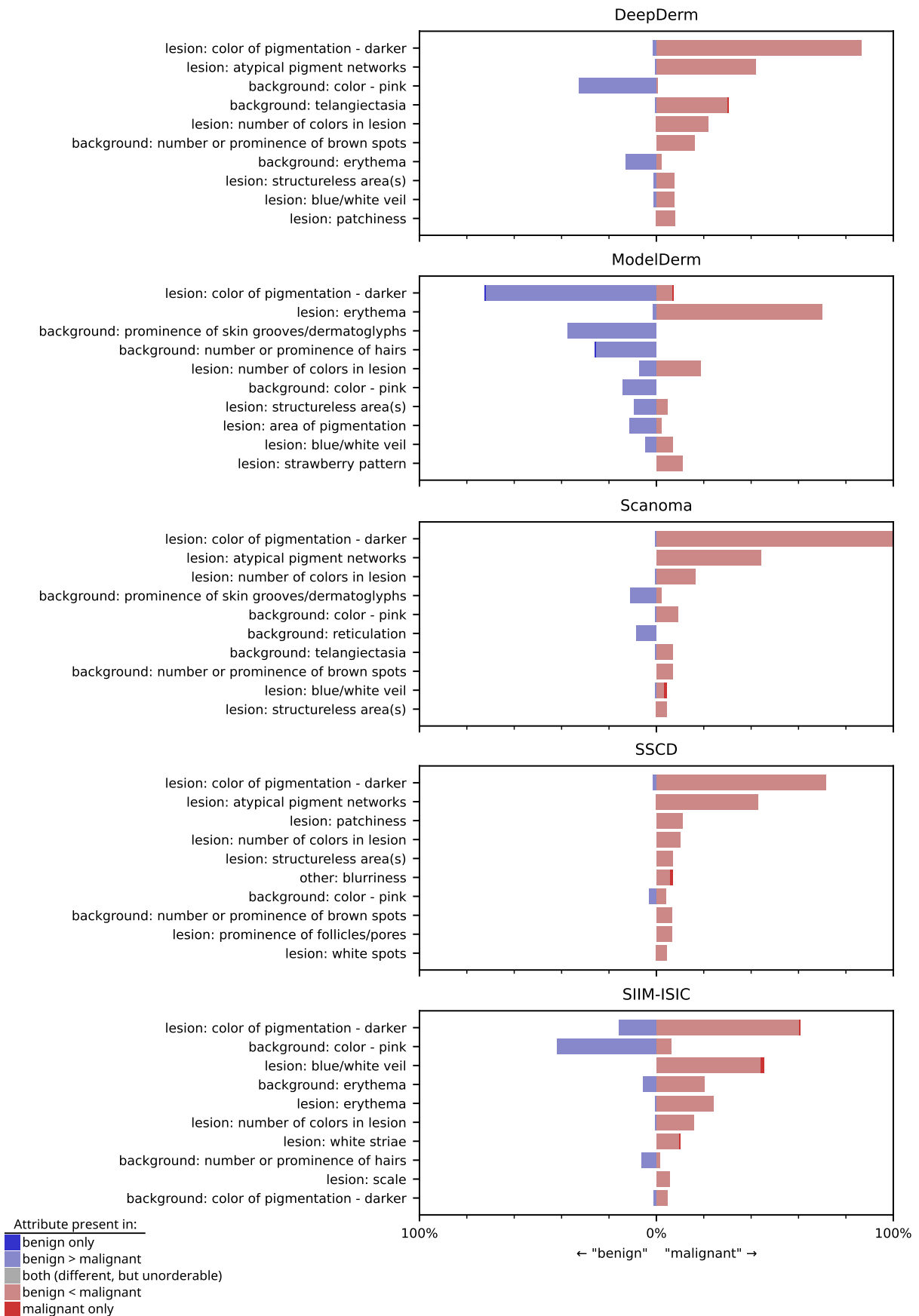
Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature Limited 2023



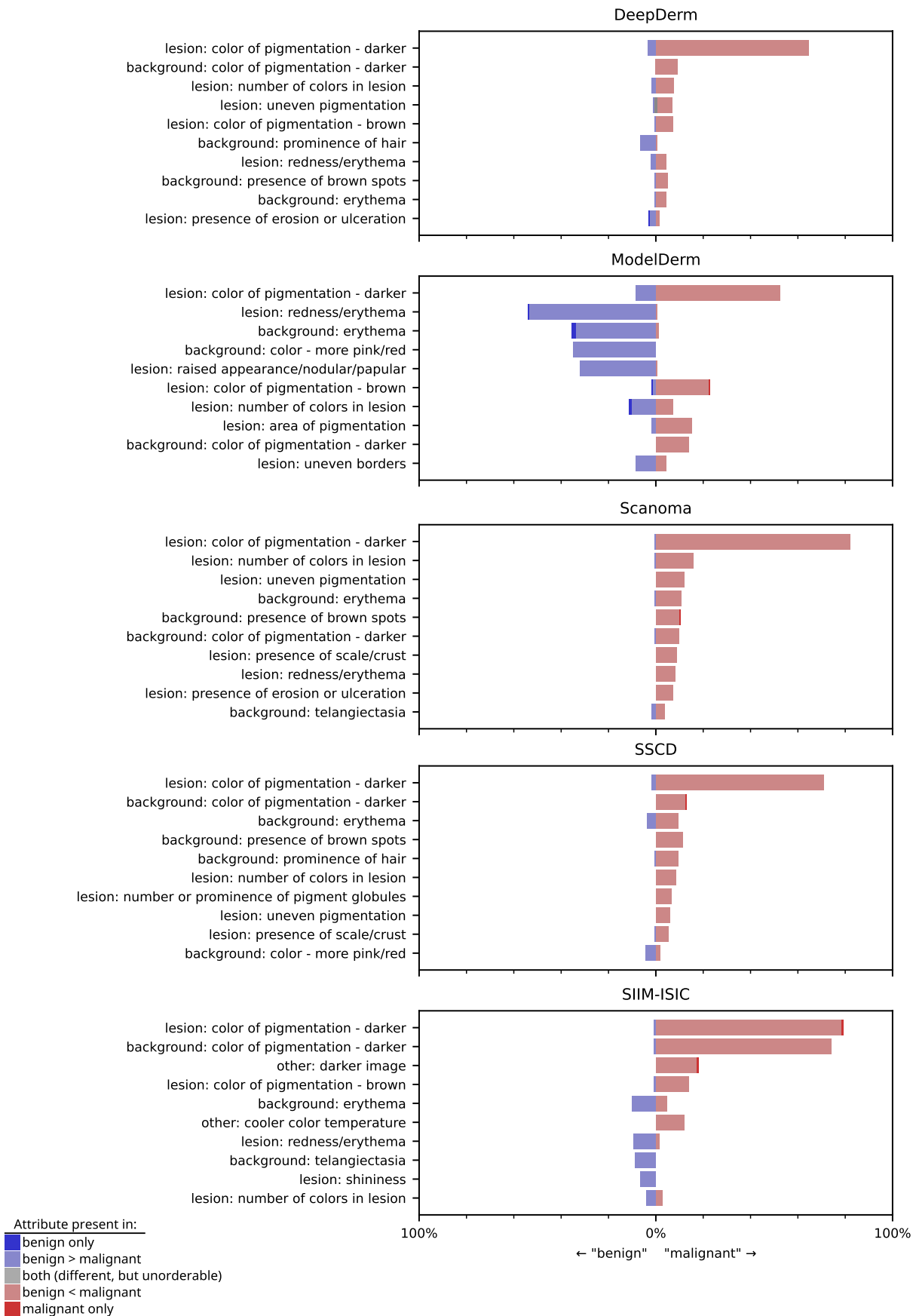
Extended Data Fig. 1 | Comparison of insights from counterfactuals and saliency maps. We calculated feature attributions using three popular techniques, Expected Gradients, Kernel SHAP, and GradCAM (see Supplementary Methods) and then produced our best-effort visualizations of the resulting saliency maps. We failed to gather insights from the saliency maps, except that the AI classifier may focus on the lesion (but perhaps not always, depending on the saliency technique). In contrast, the counterfactuals provided

more granular and medically interpretable insights: for instance, based on the malignant counterfactuals we inferred that multiple colors of pigment (top + bottom), erythema (middle + bottom), darker pigmentation (all), and blue-white veil (bottom) tend to elicit more malignant predictions. In this figure, all saliency maps and counterfactuals were generated in reference to our AI classifier 'SIIM-ISIC'. Figure adapted with permission from ref. 25, ViDIR Group, Department of Dermatology, Medical University of Vienna.



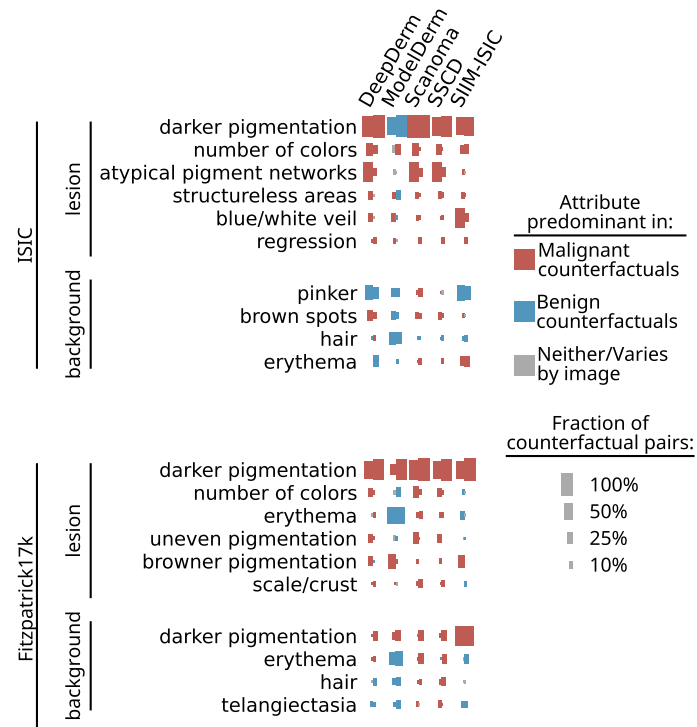
Extended Data Fig. 2 | Attributes identified by the joint expert-XAI auditing procedure as key influences on the output of individual dermatology AI classifiers, when evaluated on the ISIC dataset. In contrast to main text Fig. 2, attributes are ordered by the proportion of counterfactual pairs

from the specified AI classifier in which experts noted that attribute differs, enabling examination of attributes relevant to a particular AI classifier but not necessarily to most AI classifiers (for example, prominence of skin grooves or dermatoglyphs, which influences Scanoma and ModelDerm).



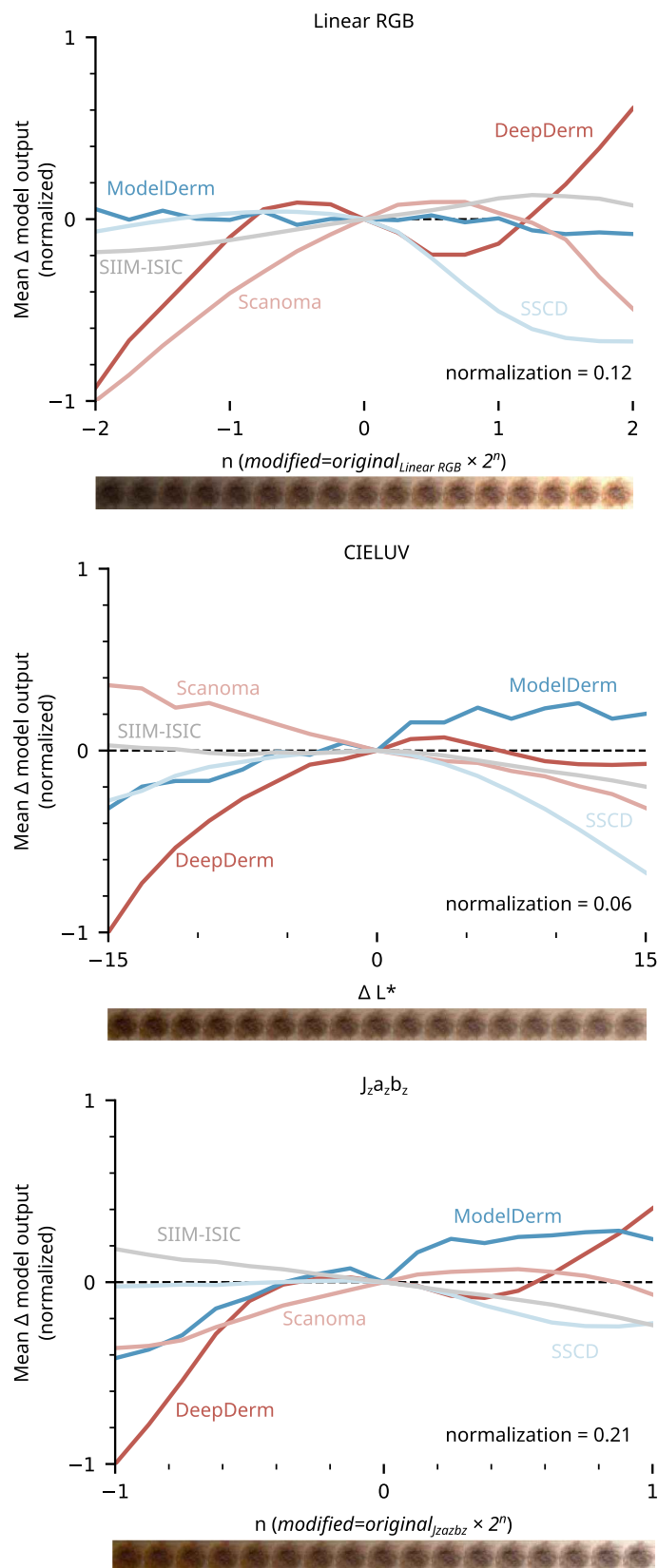
Extended Data Fig. 3 | Attributes identified by the join expert–XAI auditing procedure as key influences on the output of individual dermatology AI classifiers, when evaluated on the Fitzpatrick17k dataset. In contrast to main text Fig. 2, attributes are ordered by the proportion of counterfactual pairs from

the specified AI classifier in which experts noted that attribute differs, enabling examination of attributes relevant to a particular AI classifier but not necessarily to other AI classifiers.



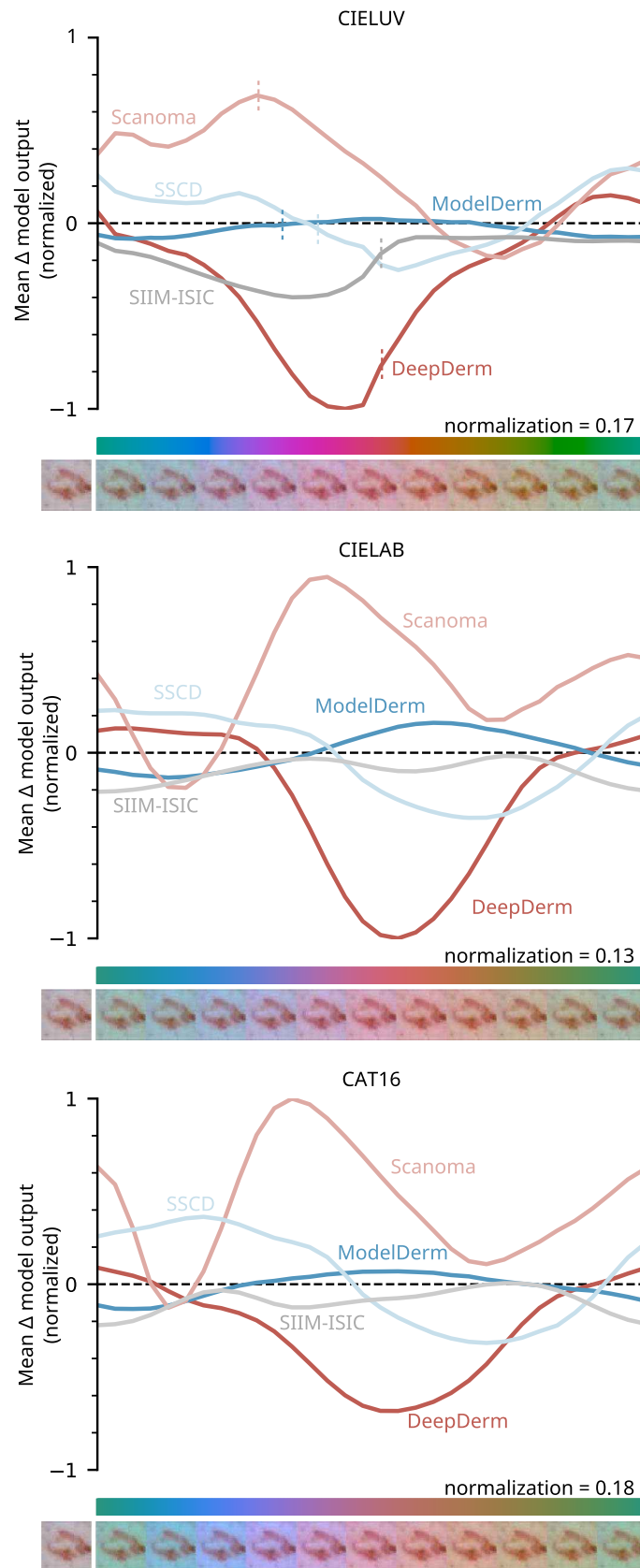
Extended Data Fig. 4 | Analysis of inter-reader variability, displaying the two readers' individual conclusions side-by-side for each attribute. For each reader, we separately determine whether that attribute was 'predominant' in benign or malignant counterfactuals, *that is*, present to a greater extent in benign (malignant) counterfactuals in at least twice as many images as malignant (benign) counterfactuals. The size of each rectangle (the 'fraction of

counterfactual pairs') is then determined as the proportion of counterfactual pairs with a difference noted in the predominant direction, for that reader alone. While readers typically do not attain quantitative agreement on the fraction of counterfactual pairs for a given attribute, the presence and direction of an attribute's effect typically remains consistent. For conciseness, attribute names are shortened as described in Supplementary Table 1.



Extended Data Fig. 5 | Effect of the programmatic modification of image brightness on the predictions of the AI classifier. We separately applied three methods of image brightness modification (see Supplementary Methods), then calculated the mean change in AI classifier output relative to the original, unaltered images. For modifications in linear RGB or $J_z a_z b_z$ space, we modified brightness by applying a multiplicative factor $B = 2^n$; we display AI classifier

responses as a function of n . For modifications in CIELUV space, we add a constant ΔL^* to the perceptual lightness L^* , where the maximum value of L^* is 100. To facilitate visualization, the vertical axis is normalized to the maximum absolute change in AI classifier output observed for a given method; the normalization factors are displayed at bottom right. Images indicate the effect of each given brightness modification.



Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Effect of the programmatic modification of image chromaticity on the predictions of the AI classifier. We separately applied three methods of image chromaticity modification (see Supplementary Methods), then calculated the mean change in AI classifier output relative to the original, unaltered images. Each method of chromaticity modification reflects the chromatic adaptation transform (white balancing method) provided by the corresponding color appearance model (CIE 1976 $L^*u^*v^*$, CIE 1976 L^*a^*

b^* , or CAM16). To facilitate visualization, the vertical axis is normalized to the maximum absolute change in AI classifier output observed for a given method; the normalization factors are displayed at bottom right. Images indicate the effect of each given chromaticity modification. Color bars indicate the hue to which a neutral color (white) is shifted by the chromaticity modification; colorfulness in the color bar (but not example images) is exaggerated for ease of viewing.

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- | n/a | Confirmed |
|-------------------------------------|--|
| <input type="checkbox"/> | <input checked="" type="checkbox"/> The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> The statistical test(s) used AND whether they are one- or two-sided
<i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A description of all covariates tested |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
<i>Give P values as exact values whenever suitable.</i> |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

Data analysis

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

The images used in this study were obtained from publicly available repositories. ISIC images are available at <https://challenge.isic-archive.com/data>. Fitzpatrick17k images are available at <https://github.com/mattgroh/fitzpatrick17k>. The DDI images are available at <https://stanfordaimi.azurewebsites.net/>

datasets/35866158-8196-48d8-87bf-50dca81df965. Model weights for the DeepDerm classifier are available at <https://zenodo.org/record/6784279#.ZFrdc9LMK-Z>. The weights and model specification for the ModelDerm classifier are available at https://figshare.com/articles/Caffemodel_files_and_Python_Examples/5406223. Model weights for our retrained variant of the SIIM-ISIC competition classifier are available at <https://zenodo.org/doi/10.5281/zenodo.10049216>. Scanoma and Smart Skin Cancer Detection are third-party software for which we cannot redistribute model weights. At the time of writing, both are apps that are available for download with no fee from the Google Play store and from third-party APK-package download sites.

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender

The study relied on previously published data for which sex and gender data were only partly available. Images derived from both male and female patients were included in this study. The method of determination of sex was not stated in the publications from which we derived data.

Reporting on race, ethnicity, or other socially relevant groupings

The study relied on previously published data, which did not make available information on race, ethnicity or other socially relevant groupings.

Population characteristics

Participants with the following conditions were included in the study: acral melanotic macule, atypical spindle cell nevus of reed, benign keratosis, blue nevus, dermatofibroma, dysplastic nevus, epidermal nevus, hyperpigmentation, keloid, inverted follicular keratosis, melanocytic nevus, nevus lipomatosus superficialis, pigmented spindle cell nevus of reed, seborrheic keratosis, irritated seborrheic karatosis, solar lentigo,acral lentiginous melanoma, melanoma in situ, nodular melanoma, and melanoma.

Recruitment

No participants were recruited for this study.

Ethics oversight

The study relied entirely on publicly available and previously published data.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

We determined sample sizes on the basis of the total number of images in each of the publicly available datasets. In the case of our counterfactual experiments, we included all images that fit our inclusion criteria for Fitzpatrick17k, and included a comparable number of images from the ISIC database (92 from Fitzpatrick17k and 100 from ISIC). The goal of the study was to infer the reasoning processes of AI classifiers with only qualitative comparisons between classifiers; hence, no sample-size calculation was necessary.

Data exclusions

We restricted our images to include only clinical or dermoscopic images of melanomas and melanoma look-alikes (Methods). In our counterfactual analysis, we excluded counterfactual pairs that did not straddle the decision boundary of each dermatology AI classifier, as well as counterfactual pairs that contained visual artifacts, as identified by at least one of two board certified dermatologists (Methods).

Replication

We independently retrained our generative models to ensure that they maintained the same attributes as differing between the benign and malignant counterfactual images. We observed that these attributes are indeed maintained.

Randomization

We randomized counterfactuals throughout our evaluation, from their screening phase to their annotation. We pooled counterfactual pairs from all AI classifiers and all reference images, then randomized their viewing order independently for each dermatologist evaluator. Within each counterfactual pair, we additionally randomized whether the benign or malignant counterfactual appeared on the left or right, to prevent evaluators from inferring their identity.

Blinding

The dermatologist evaluators were blinded to the identity of the counterfactuals, including to which classifier they corresponded and their identity as 'benign' or 'malignant', until all labelling of the counterfactuals was complete.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involvement in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern
<input checked="" type="checkbox"/>	<input type="checkbox"/> Plants

Methods

n/a	Involvement in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging