Mechanisms underlying AI prediction of protected attributes

April 2024

Introduction

Machine learning (ML) technologies have been widely applied to image classification problems in medicine, a trend that has led to the approval of hundreds of ML-based devices by the United States Food and Drug Administration (FDA).[1–3] As AI becomes increasingly embedded in clinical decision-making processes, ensuring the robustness and fairness of these models must become an integral part of the development process. This paper examines the mechanisms underlying AI predictions of protected attributes. Exposing these mechanisms provides developers with the opportunity to reduce model vulnerabilities that could lead to discriminatory performance in downstream clinical tasks.

To generate their predictions, ML models rely on *image attributes*, including medically significant details trusted by physicians. However, the use of image attributes is not without serious limitations. Some attributes include image acquisition artifacts or other medically irrelevant and likely undesirable by-products[4,5] that can degrade their predictive performance. Evidence also shows that ML-based medical image classifiers leverage protected demographic information to generate predictions;[6,7] inappropriate use of such attributes (e.g., race or gender) could lead to fragile predictive performance or discrimination due to domain shift in clinically relevant tasks. Perhaps more concerning, ML classifiers have unexpectedly displayed the ability to predict a range of demographic variables directly from medical images[8,9] even in the absence of demographic metadata. Examples include the prediction of a patient's sex from retinal fundus images[8] and the prediction of a patient's race from different forms of radiological imaging.[9] Prior work reveals that such classifiers can use these demographic shortcuts in disease classification, leading to biased predictions across subpopulations and fairness disparities across both in-distribution and external test sets.[10]

Significantly, physicians in relevant imaging fields cannot explain how ML classifiers predict certain demographic features from medical images. This raises the fundamental question that motivates this research: *What mechanisms underlie AI predictions of protected attributes?*

Prior work attempting to answer this question has only partially explained the high performance of these classifiers,[9,11,12] citing factors like correlations between protected demographic variables and diagnoses or other variables more visible in the images (e.g., age). However, it fails to quantify the degree to which each signal contributes to overall classification performance. Further, the unexplained performance of these devices has led to speculation about the existence of unique image attributes detectable only by machines, a hypothetical category of image attributes we refer to as *AI-specific signals.*

The postulated existence of AI-specific signals raises two sub-questions that we also address here: (1a) *How do medical image classifiers 'reason'?* That is, what signals do they use to generate predictions? (1b) *How might such classifiers produce disparate outcomes among protected classes*? By studying the existence and nature of AI-specific signals that could be used to predict a protected demographic variable, we address a gap in our knowledge of how medical image classifiers generate predictions, which currently includes only image attributes readily recognized by humans.[5] Simultaneously, we detail a mechanism by which classifiers might display undesirable characteristics across protected classes.

To answer the preceding questions, we *examine medical image classifiers trained to predict sex from dermoscopic lesions,* which offer a magnified view of a patient's skin lesion. To do so, we apply a range of methodologies from the field of explainable AI, namely, *clustering analysis* and *counterfactual image generation* to identify signals that may be instrumental to classifier predictions. We also introduce a technique called *removal via balancing*, based on propensity score reweighting, that quantifies the importance of each identified signal to the overall prediction. Figure 1 summarizes our investigative framework.

To our knowledge, we are the first to perform rigorous data auditing to (1) identify specific signals that contribute to the predictive performance of protected attributes from multiple modalities and (2) quantify the extent of the contribution.

In sum, this work motivates and introduces a comprehensive interpretability framework to identify and quantify important signals in medical images that lead to the prediction of protected attributes. Although this study focused on sex prediction from dermoscopic images, our framework is flexible and applicable to other modalities (e.g., chest X-rays, retinal fundus images) and protected attributes (e.g., race, age) that we anticipate could reveal similar vulnerabilities.

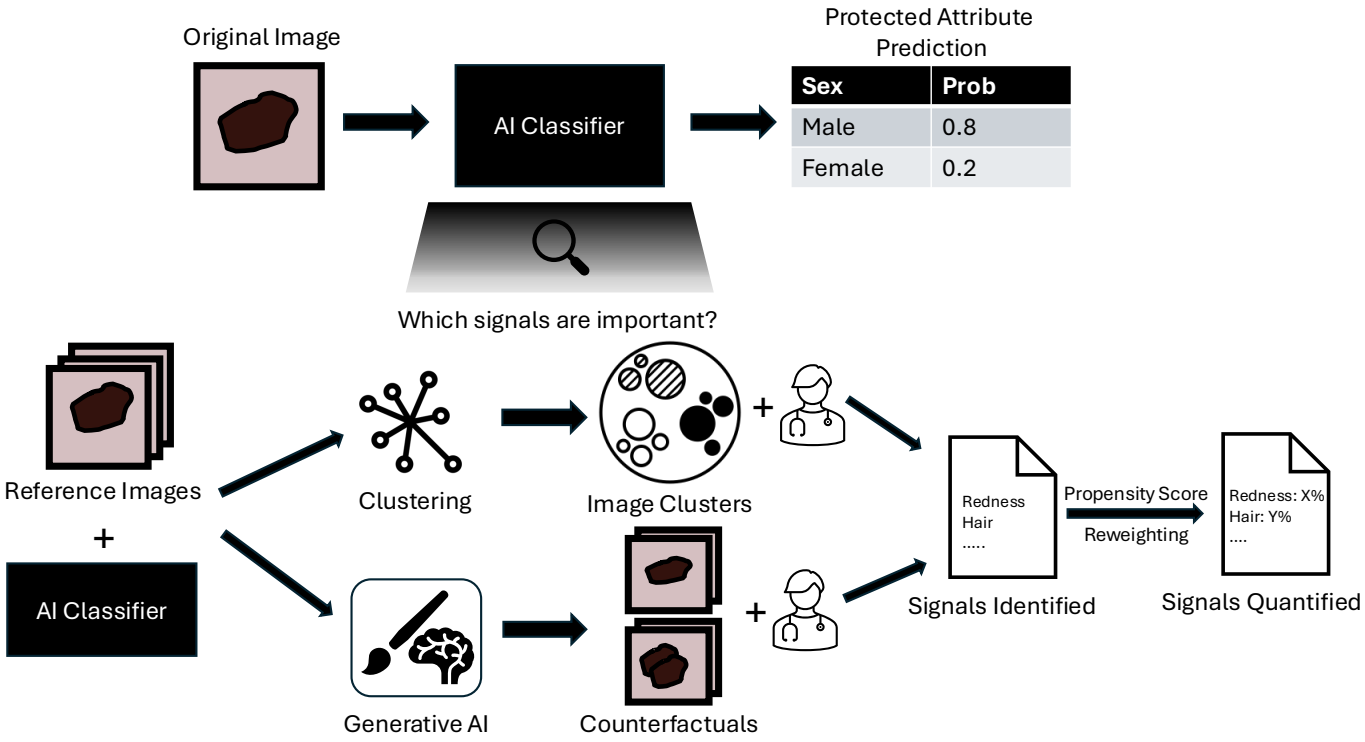


Fig. 1 | Framework overview. We first train a ViT-based AI classifier to predict the protected attribute. Then, using the explainable AI techniques of clustering analysis (for global artifacts) and counterfactual generative AI (for local artifacts), we identify signals of potential importance to classifier predictions by leveraging the expertise of board-certified physicians. Finally, using our novel *removal via balancing* technique based on propensity score reweighting, we quantify the importance of each identified signal to the overall prediction.

Results

ML classifiers successfully predict protected attributes

To investigate how ML classifiers predict protected attributes from medical images, we trained neural networks for the specific task of identifying a patient's sex based on a dermoscopic image of a skin lesion (Figure **??**a).

Dermoscopic images offer a magnified view of the skin and typically lack anatomical landmarks (eyes, nose, fingers, etc.) that could offer a route to the identification of a patient's sex. We trained our neural networks on images from the International Skin Imaging Collaboration (ISIC) archive. To mitigate the possibility that the networks would rely on source-specific confounding rather than signals that generalize across data sources, we partitioned the ISIC archive based on the image collection site and used a disjoint group of collection sites for training and testing. In this external test scenario, our classifiers predicted patient sex with substantial accuracy (area under the receiver operating characteristic curve, ROC-AUC, of 0.782 ± 0.012, mean $\pm$ standard deviation; Figure **??**b).

Prediction of protected features enables undesirable outcomes

Although we typically would not expect medical image classifiers to be used to predict protected attributes, we hypothesized that their ability to do so could lead to undesirable behavior in medically related prediction tasks. If true, this hypothesis would motivate the need for improved understanding of the mechanisms that underlie the classifier's prediction of protected attributes.

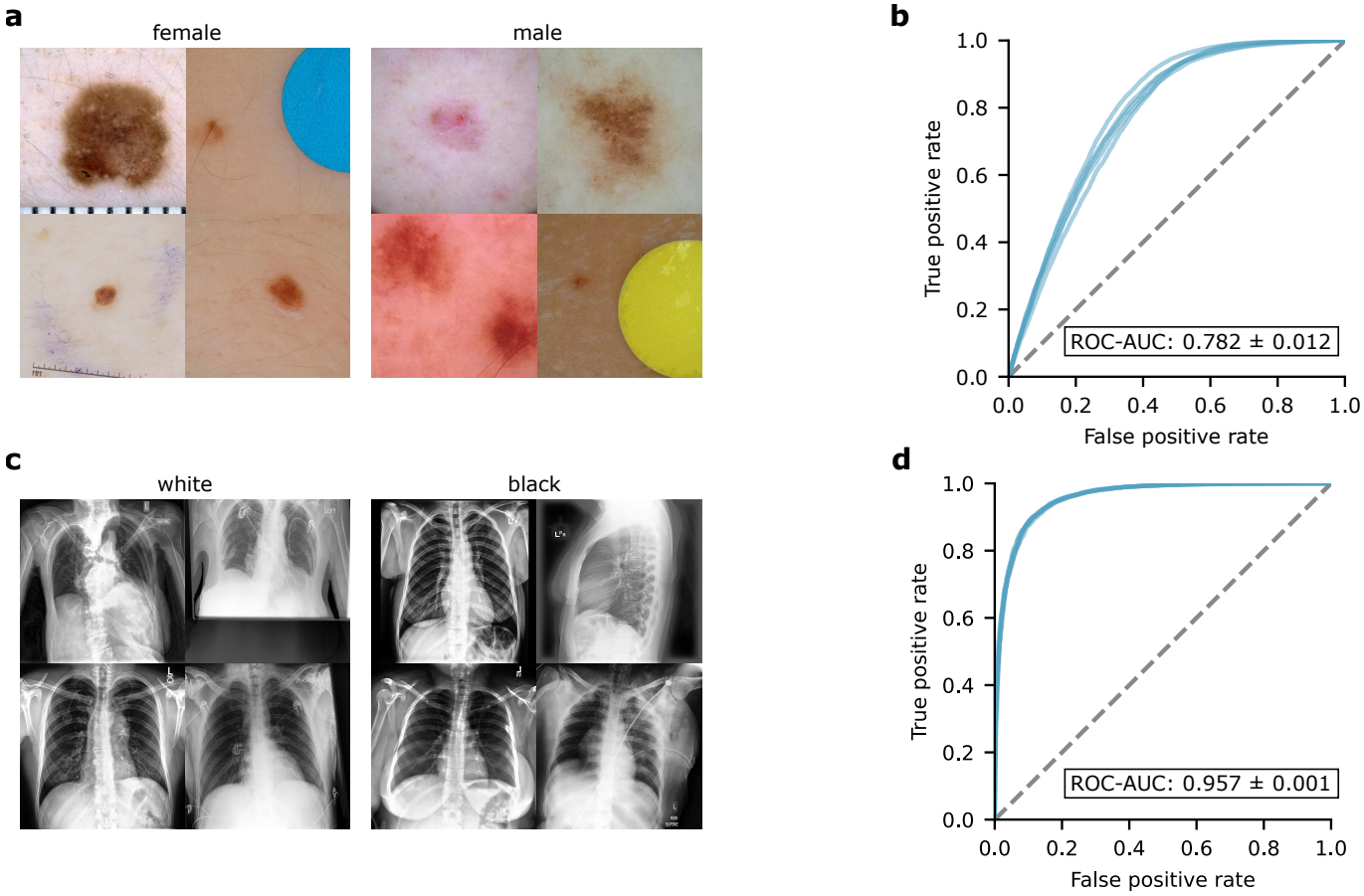


Fig. 2 | AI prediction of protected attributes. (a) A randomly selected set of dermoscopic images from female and male patients. **(b)** Performance of trained ML classifiers, based on a vision transformer architecture, at prediction of sex from dermoscopic images. **(c)** A randomly selected set of chest x-rays from patients from different races. **(d)** Performance of trained ML classifiers, based on a vision transformer architecture, at prediction of race from chest x-rays.

We conjectured that one mechanism by which ML prediction of protected attributes may degrade performance at medically relevant tasks is if the classifier learns to use the protected attribute as a 'shortcut'[13] for identification of a disease due to *an association between the attribute and disease in the training data*. Though some associations could reflect genuine medical differences (e.g., rates of breast cancer among females and males),[14] they could also reflect societal disparities or other spurious variations. If the association changes—or, in the worst case, *reverses*—at test-time (e.g., deployment), then a model that learned to leverage the association from the training data should decline in accuracy.

To test our hypothesis, we focused on the tasks of differentiating melanoma from look-alike lesions (benign nevi, seborrheic keratoses, solar lentigo, etc.; see Methods). We engineered datasets to exhibit an association between the protected attribute and the disease target (Figure 3a). We conducted our tests for a variety of odds ratios, in each case setting the ratio in the external test data to the inverse of that in the training data.

We observe that though performance on internal validation data remains high across scenarios (indeed, even improving with stronger associations between the protected attribute and the target), external test set performance declines as the odds ratios vary from unity (Figure 3b). Performance drops precipitously for extreme changes in odds ratios (in particular, when the protected attribute correlates perfectly with the prediction target). Performance changes more modestly for moderate differences in odds ratio (e.g., a drop of 6% in external test set ROC-AUC for melanoma from an odds ratio of 1 to an odds ratio of 0.5/2 in the train/test data, respectively). We also observed that among the data sources that comprise the ISIC archive, the odds ratio for 'female' as a predictor for 'melanoma' varies from 0.475 to 1.185, confirming that reversals in the association between protected attributes and a prediction target indeed occur naturally in medical data.

To further motivate this analysis, we also considered real-world melanoma classifiers, which are available for public use as smartphone apps. We analysed two classifiers, Scanoma and SSCD, to identify whether they relied on protected

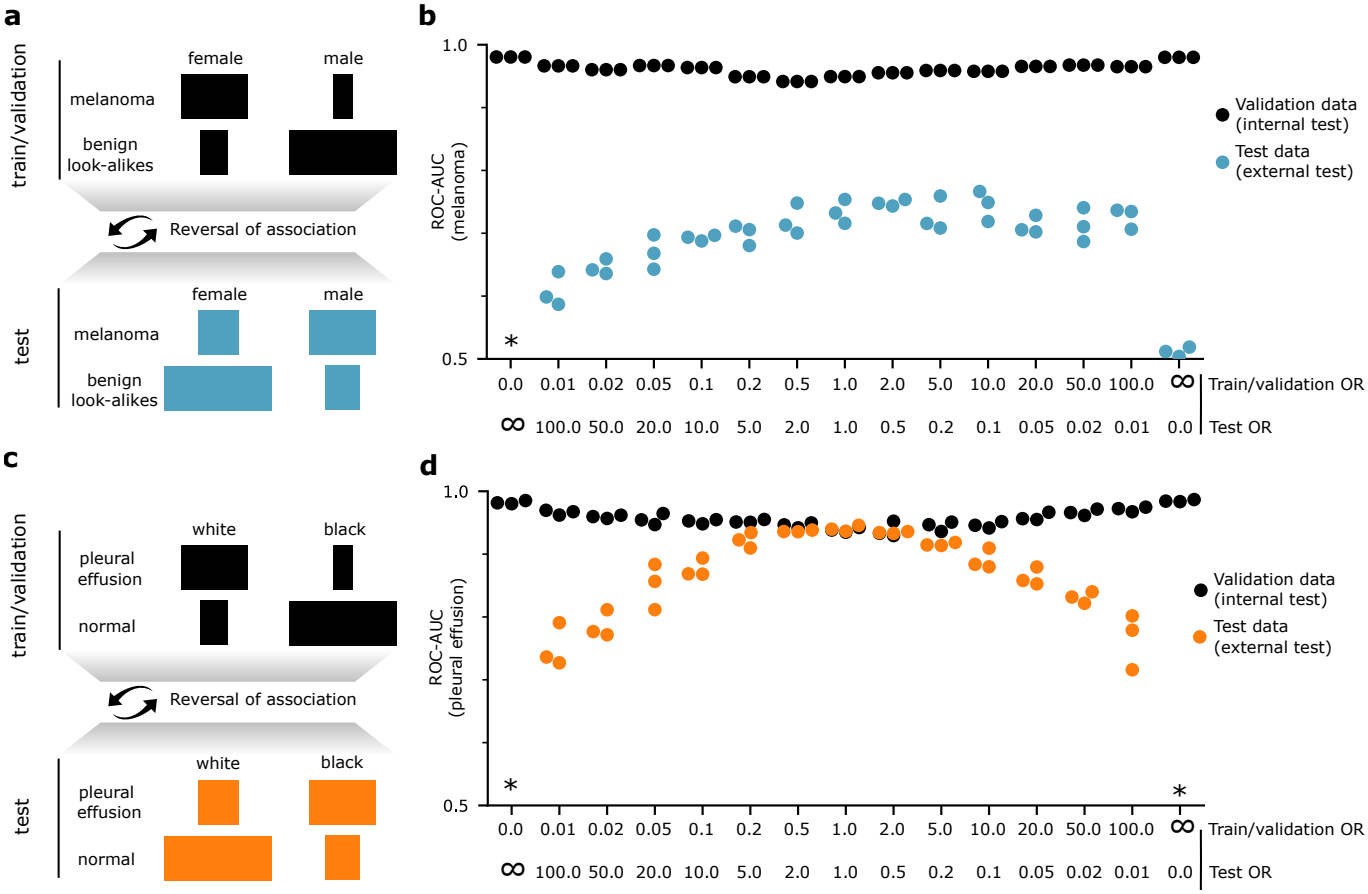


Fig. 3 | Spurious correlation with a protected attribute may result in performance deficit. (a, c) Overview of the experiment's setup. We engineered datasets (by sub-sampling the original data) such that the prediction target (melanoma/pleural effusion) correlates with a protected attribute (patient sex/race). The correlation is inverted in the test data. **(b, d)** Performance of the trained classifiers on the melanoma and pleural effusion prediction tasks prediction task. *Odds ratios* measure the association between the demographic 'female' or 'white' and the prediction target 'melanoma' or 'pleural effusion,' that is, an odds ratio greater than one indicates that a higher proportion of the corresponding protected attribute depict the disease. An asterisk (*) indicates ROC-AUC$<$ 0.5 (that is, worse than random performance). OR indicates odds ratio.

attributes for the melanoma prediction task. First, we first took the trained classifiers and tested them on a synthetic test set with a correlation between sex and melanoma with varying odds ratios. For both Scanoma and SSCD, we observed that performance fluctuates as the strength of the odds ratio deviates from unity. More specifically, we see a decreasing trend in performance as the odds ratio increases (ROC-AUC drops by 8.73% for Scanoma and by 11.07% for SSCD, Figure 4).

These declines indicate that both commercial classifiers also leverage demographic encodings for disease predictions. To confirm this hypotheses, we then retrained the classifiers using an equalized training set with no correlation between sex and melanoma, essentially removing the dependence on demographic encodings. For both Scanoma and SSCD, we observed that performance remained consistent on the test set across varying odds ratios; in fact, performance *increased* compared to the earlier setting. We suspect that the retraining data for both classifiers came from a distribution that better matched the test set distribution than the original training data; however, the fact that the performance fluctuations were mitigated indicates that the original classifiers did, in fact, rely on demographic encodings that degraded their downstream performance for the clinically relevant melanoma classification task.

Exploration of statistical associations with patient sex

As a first step toward understanding how the classifiers could identify sex from dermoscopic images, we examined statistical associations with the available metadata characteristics, finding a few characteristics that associate with sex, albeit weakly. Multiple diagnoses were weakly associated with sex, ane some of these associations persisted

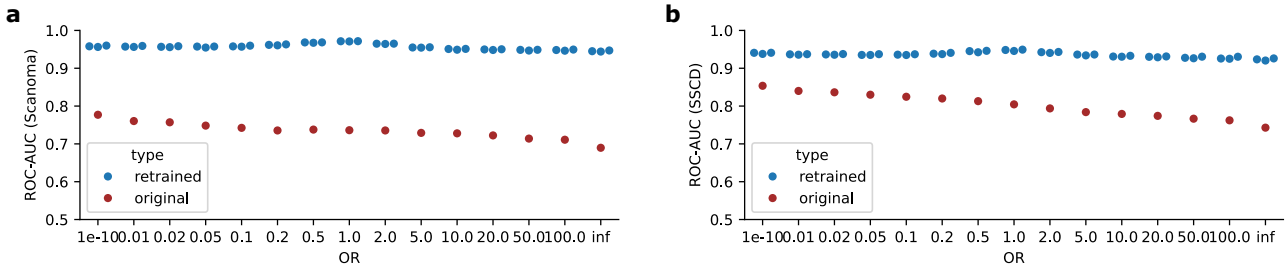


Fig. 4 | TODO

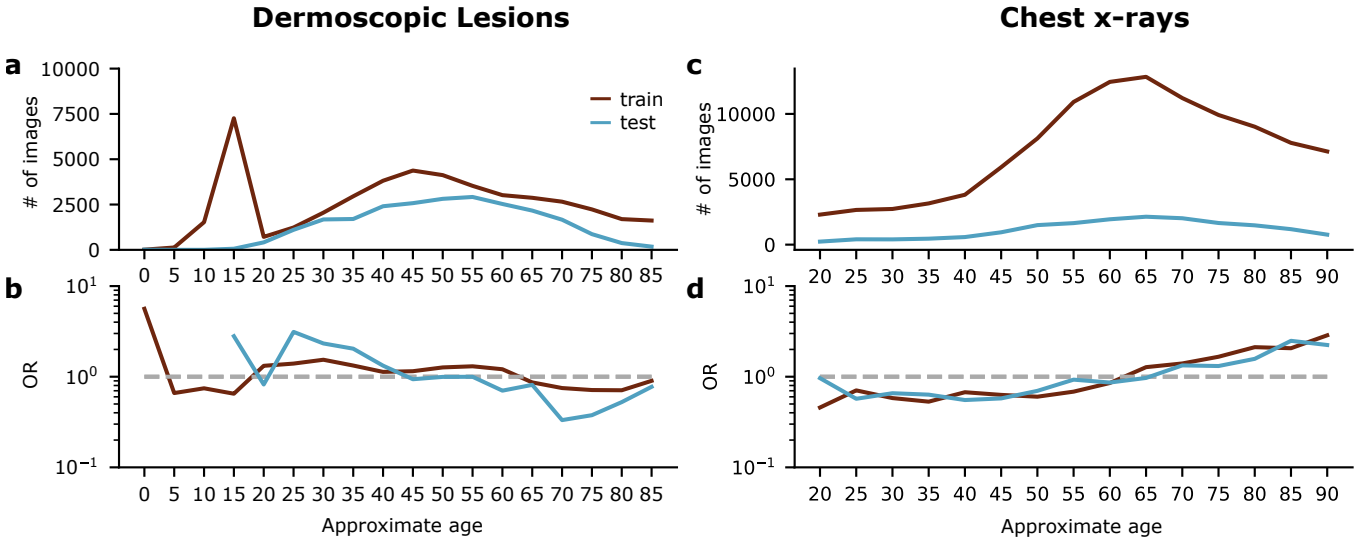


Fig. 5 | **Association of patients' sex and race with their age.** **(a)** Histogram of patients' ages in the training and test dermoscopic data. **(b)** Odds ratio (OR) for prediction of female sex on the basis of a patient's age. **(c)** Histogram of patients' ages in the training and test chest x-ray data. **(d)** Odds ratio (OR) for prediction of the white race on the basis of a patient's age.

from the training data, where an association must be present for the classifier to learn it, and the test data, where the association must persist to benefit performance. These include an association between female sex and solar lentigo, dermatofibromas or nevi, and an association between male sex and seborrheic keratoses. Since prior studies have successfully identified diagnoses from dermoscopic images,[15] a classifier could, in principle, then leverage this knowledge to help identify a patient's sex. Considering the dermoscopy method used for image acquisition, there was a weak association between non-contact polarized images and the male sex (Table 4). Overall, however, associations between diagnoses or dermoscopy type and sex appeared unlikely to account for the classifier's performance on external test data considering that many associations were weak and that some associations were reversed between training and external test data.

In contrast, we observed a more consistent association of sex with patient age. In the training data, patients aged 20-60 were enriched for females, while patients aged 5-15 and 65-85 were enriched for males (Figure 5). In the external test set, patients aged 60-85 were also enriched for males, suggesting that a correlation between older ages and patient sex may persist across data sources.

Clustering-based analysis helps identify global artifacts

Using clustering analysis (see Methods), we examined visually similar clusters of images in the training data that differed the most in terms of the ratio of males to females predicted by the trained classifier. Hair images differed strikingly between the two clusters, being highly prevalent in the cluster with more predicted males.

Since hair was a strong signal, to identify other signals, we equalized the training set by sub-sampling images so there was no statistical correlation between hair and sex. To label the images (that lack detailed annotation), we manually annotated 500 female and 500 male images for the presence of hair and applied these hand-labeled images to train a classifier (ViT-Base architecture) for this task, achieving ROC-AUC of 0.96 on a held-out test set (90-10

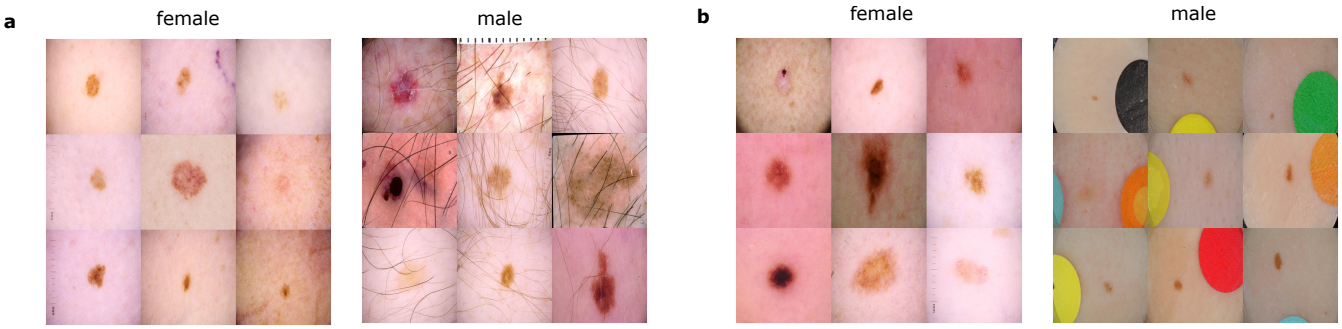


Fig. 6 | Clustering analysis of dermoscopic images. A sample of images from the visually similar clusters with the lowest and highest predicted male to female ratios for **(a)** the unequalized sex classifier, in which hair is identified as a potential signal, and **(b)** the sex classifier equalized for hair, in which the sticker is identified as a potential signal.

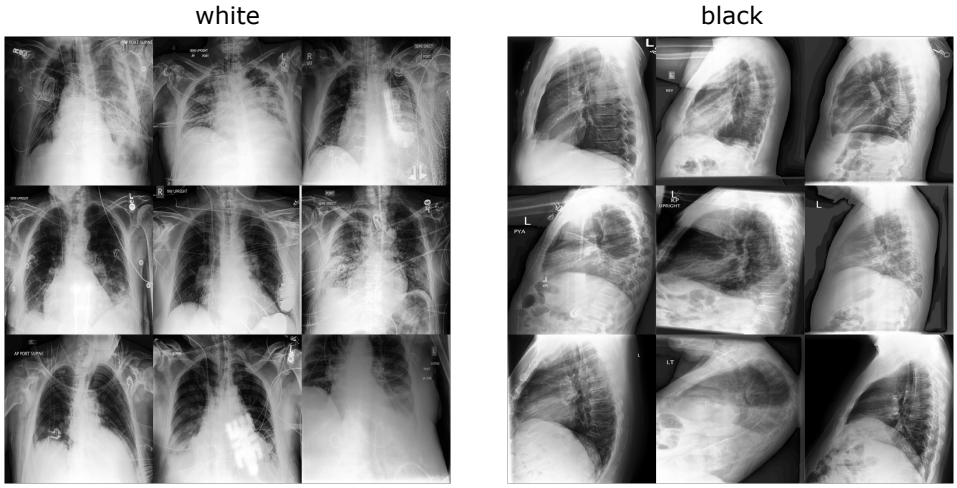


Fig. 7 | Clustering analysis of chest x-rays. A sample of images from the visually similar clusters with the lowest and highest predicted white to black ratios. The view position (frontal or lateral) is identified as a potential signal.

train-set split). We then used that classifier to label the rest of the dataset. After sub-sampling, the new training set contained 11190 images without hair and 9230 images with hair for each of the female and male sexes, resulting in an odds ratio of 1.

After retraining the sex classifier using the equalized training set, we performed the clustering analysis again; this time, stickers (i.e., small adhesive markers placed on the skin to indicate the location of lesions or areas of interest to guide biopsies, surgical excisions or other treatments) were identified as being more prevalent in the cluster with the highest number of predicted males, indicating that stickers could be a potential signal associated with males by the sex classifier (Figure 6b). The exact type and color of the stickers can vary by hospital site, making them an easy 'shortcut'[13] to learn for predicting protected attributes (like sex) that can vary between sites.

Generative image AI reveals local prediction mechanism

Our generative techniques (see Methods) produced realistic counterfactual images of dermoscopic lesions. The distribution of images produced by the final network differed from that of training images by a Fréchet Inception Distance (FID score)[16] of 6.29 for the latent space optimization technique and 10.32 for the Explanation by Progressive Exaggeration (EBPE) technique. Figure 10 shows samples of the generated counterfactuals. We generated 1000 pairs of counterfactuals from both the generative techniques that elicited a desired sex prediction from the classifier, i.e., a pair classified as 'female' and 'male.' We then manually analyzed the image pairs to identify signals in addition to hair and stickers that were prevalent in either of the sexes and could potentially be used by the classifier for predictions.

Table 5 lists the signals identified along with their prevalence in either the male or female sex. These signals were identified by a board-certified dermatologist in at least 100 pairs of images (for each generative technique) in the same direction (always in male or always in female), indicating a possible correlation with sex. To streamline the process of labeling images, we created a web app that displayed the counterfactual pairs and asked dermatologists to identify

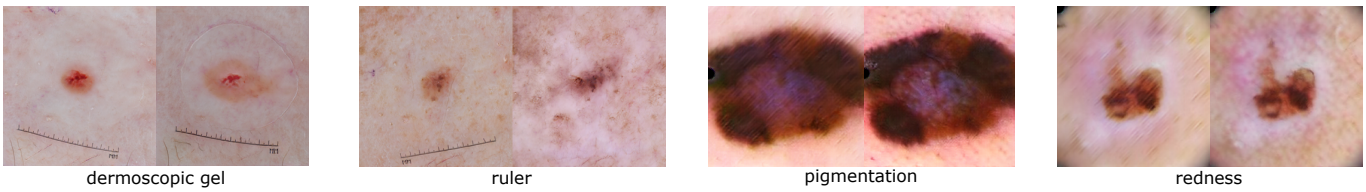


Fig. 8 | Signals identified from producing counterfactual pairs of dermoscopic lesions using the generative techniques. The two leftmost signal pairs are identified by the latent space optimization technique, and the two rightmost by EBPE. For each pair, the left lesion is the female counterfactual, and the right is the male.

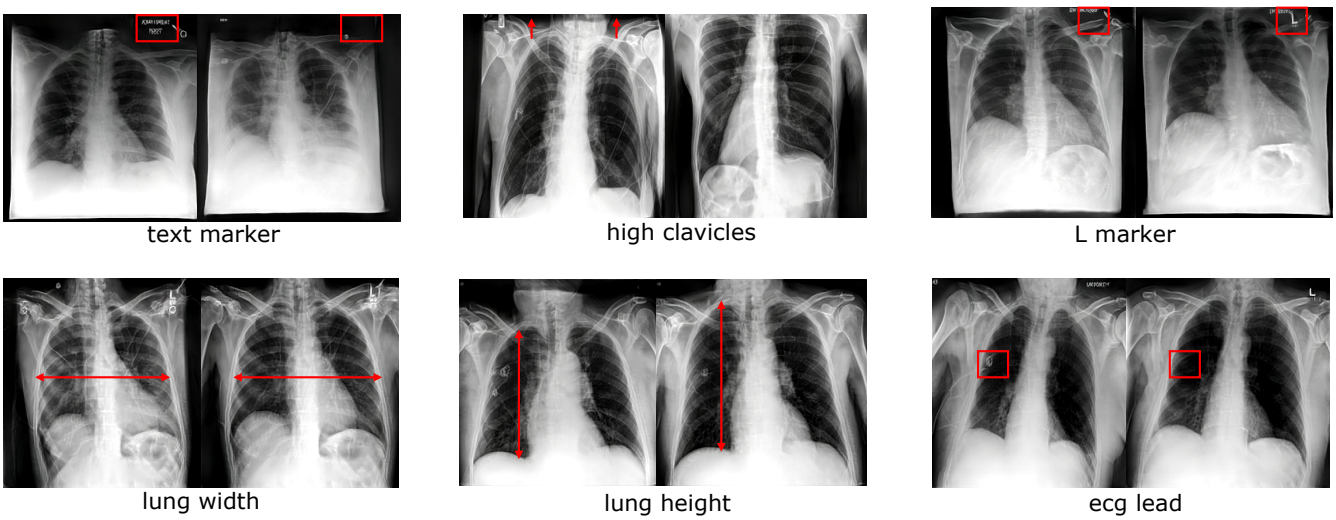


Fig. 9 | Signals identified from producing counterfactual pairs of chest x-rays using the generative techniques. The pairs in the top row are identified by the latent space optimization technique, and the ones in the bottom row by EBPE. For each pair, the left lesion is the black counterfactual, and the right is the white. The red markings highlight the differences based on the labeled signals.

relevant signals (see Methods). Since we hypothesized that signals important for the classifier would be present in multiple counterfactual pairs, we discarded any signals that appeared in less than 100 pairs as being insignificant for the sex prediction task. A qualitative visualization of the different signals is shown in Figure 8.

Table 1: First Table Example

Header A	Header B	Header C
Row 1A	Row 1B	Row 1C
Row 2A	Row 2B	Row 2C

Table 2: Second Table Example

Header X	Header Y	Header Z
Data 1X	Data 1Y	Data 1Z
Data 2X	Data 2Y	Data 2Z

'Removal via balancing' successfully quantifies classifier performance

After applying a range of techniques, including statistical association, clustering analyses, and generative modeling, to identify putatively meaningful signals, we confirmed and quantified their importance using *removal via balancing* (see Methods). This technique quantifies the importance of a putatively important *query signal* for a classifier's predictions in a particular test set. To implement this technique, we compared the model's performance in the original test data to its performance in an alternate version of that test data in which the query signal was statistically independent of the prediction target. In other words, under the hypothesis that a classifier depends on a particular query signal, we would expect the classifier's performance to decline when there is no difference in the query signal between target classes (female and male), and the degree by which the performance declines quantifies the importance of that signal.

The removal via balancing technique required a numerical representation for each signal in each image; since these annotations were absent from the original data, we produced annotations using a hybrid approach, similar to that

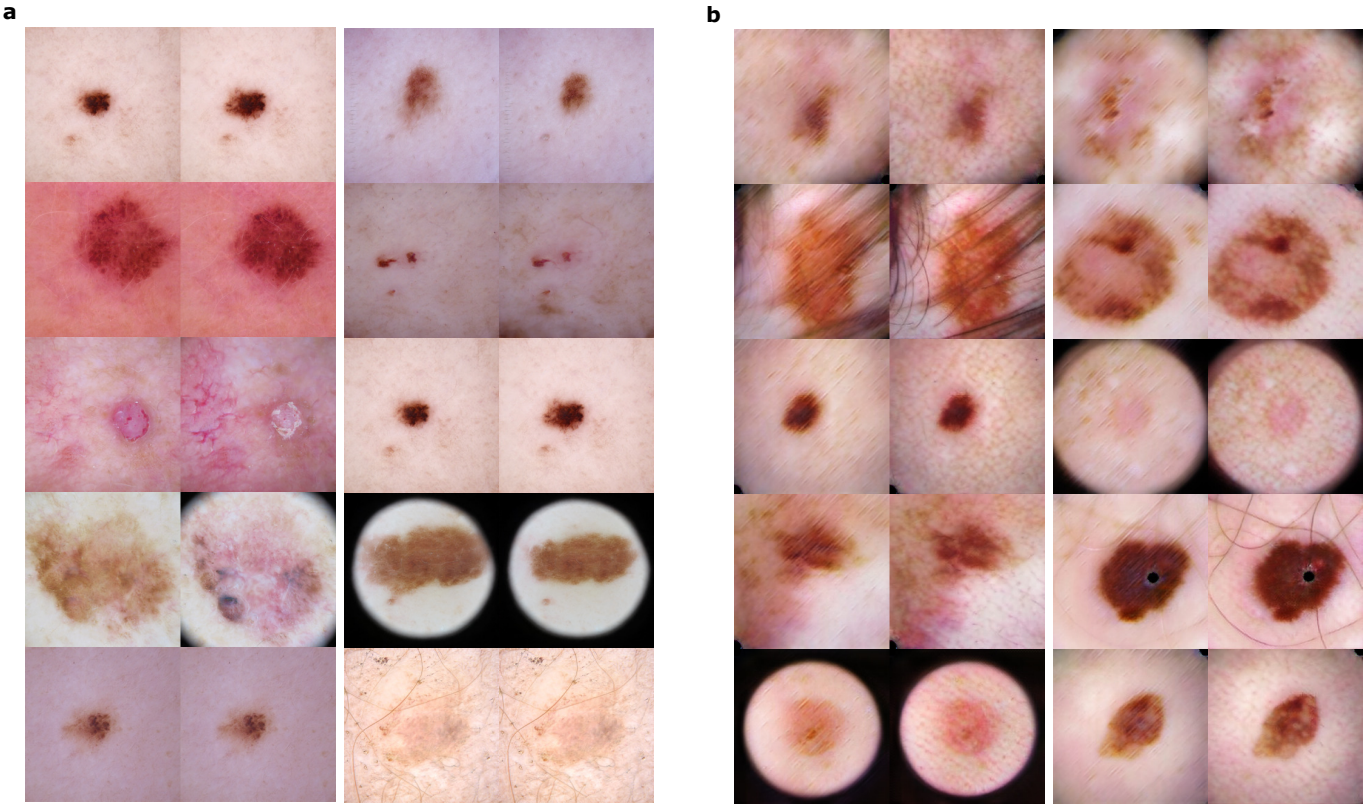


Fig. 10 | The counterfactual images produced by our generative techniques. In each pair of images, the left one is the female counterfactual, and the right the male counterfactual. **(a)** Latent space optimization. **(b)** Explanation by Progressive Exaggeration.

used to equalize hair in the clustering analysis. The dermatologists manually labeled 500 female and 500 male images from the test set for each identified signal. We developed another web app for this labeling task (see Methods). We then trained separate classifiers (ViT-Base architecture, ROC-AUC: 0.95) for each signal to label the rest of the test set.

After quantifying the importance of the signals we previously identified (age, hair, sticker, redness, ruler, pigmentation, and gel), we found that they collectively explained about 44% of the classifier's performance; the largest single contributor was hair, which alone explained about 37% of the classifier's performance. Other attributes, e.g., age, redness, and presence of rulers, explained a smaller but non-negligible proportion of the classifier's performance (Figure 11). Attributes like pigmentation, sticker, and gel explained minimal performance, suggesting that these attributes are either not leveraged by the classifier (false positives) or they are not prevalent in the test set. For example, the test set contains only two images showing stickers, indicating that the classifier cannot rely on this signal for prediction in the test set despite learning the signal during training. Some of the identified signals, such as hair, can be explained by physiological sex-based differences since males typically have more body hair than females. However, the other identified signals, e.g., age, redness, and ruler, do not conform to known biological insights and can be specific to the training data.

Discussion

Given the rapid advances in medical AI technologies, they may soon become an integral part of the clinical workflow. Thus, it becomes vital to ensure that these models work as designed and do not encode sensitive information that can be leveraged to cause discriminatory performance in downstream tasks, an unacceptable vulnerability.

It has previously been shown that AI models indeed [finish sentence.] The first step towards mitigating this behavior is understanding the extent to which this issue is prevalent and why it occurs. This study shows that we can train AI classifiers to predict protected demographic attributes such as sex directly from dermoscopic images, which is a cause of concern since this is a difficult task even for board-certified dermatologists.

Here, we use GANs for image generation. Although diffusion models have been shown to generate more realistic images, they have not had as much success in counterfactual generation. Prior work that attempted to use diffusion

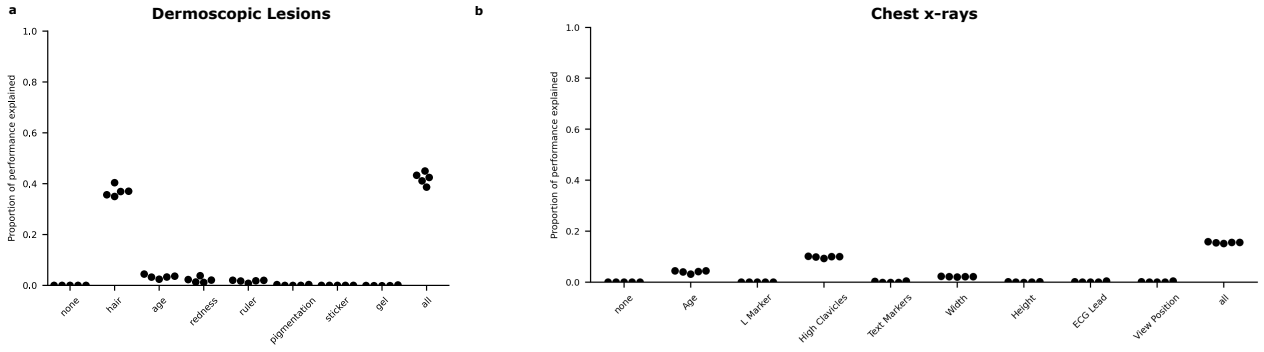


Fig. 11 | The proportion of performance explained by signals identified in our prior experiments. (a) The proportion of ROC-AUC explained across five replicates. **(b)** Quantification with 95% confidence intervals.

models for this task [17] focused on natural images, not medical ones, which are arguably more difficult to generate. Furthermore, diffusion model output is expensive to evaluate compared to GANs since they require multiple iterations of the reverse process to create one sample. GANs have been successfully used for counterfactual generation of medical images, and their quality in terms of FID scores is satisfactory for our analysis.

We first demonstrate that we can achieve unexpectedly high performance in predicting sex from dermoscopic images, which aligns with prior work [8, 9]. We then show how these high-performing predictions can be detrimental to the prediction of clinically relevant downstream tasks by disease classifiers[18, 19]. Furthermore, our analysis of commercial melanoma prediction models reveals that real-world systems already display this vulnerability, with marked performance drops under manipulated protected attribute correlations.

To shed light on the exact mechanisms these protected attribute classifiers use to achieve such high performance, a preliminary statistical analysis revealed a correlation between age and patient sex. In addition, by combining methods like clustering and counterfactual image generation with expert analysis, we were able to identify visual signals, including hair, pigmentation, redness, ruler presence, stickers, and dermoscopic gel, that the classifier could potentially leverage.

These findings indicate that some of the predictive cues models exploit reflect physiological differences (e.g., hair presence), while others reflect acquisition artifacts or dataset-specific biases that might not generalize across populations or settings. Notably, certain signals — such as age or ruler markings — are unlikely to have direct biological relevance and may serve as dataset-specific shortcuts that models opportunistically exploit. The fact that only $\sim$42% of model performance could be explained by these signals also suggests that there may exist residual non-visual features, i.e., subtle image patterns imperceptible to human observers but learnable by AI classifiers. The existence of such features underscores the need for greater scrutiny and transparency in model development and deployment.

Our results also generalize across model families: further analysis of classifiers trained with transfer learning (Supplementary Section 3.3) and with dermatology foundation model embeddings exhibited similar behaviors and relied on similar signals to predict protected attributes (Supplementary Section 3.4). This indicates that demographic encoding is not the result of a specific architecture or training paradigm, but rather a broader property of data-driven feature learning in high-dimensional visual domains.

In sum, this work motivates and introduces a comprehensive interpretability framework to identify and quantify important signals in medical images that lead to the prediction of protected attributes. Although this study focused on sex prediction from dermoscopic images, our framework is flexible and applicable to other modalities (e.g., chest X-rays, retinal fundus images) and protected attributes (e.g., race, age), which we anticipate could reveal similar vulnerabilities. As AI becomes increasingly embedded in clinical decision-making processes, ensuring robustness and fairness must be considered an integral part of the development process.

Methods

Data preparation

To study how AI predicts protected demographics from medical images, we focused on the prediction of a patient's sex from a dermoscopic image of their skin, leveraging data contained in the ISIC archive [20–23]. The ISIC archive consists primarily of dermoscopic images collected by global medical professionals along with associated metadata on diagnoses, demographic characteristics, and details on image acquisition. For our study, we excluded non-dermoscopic

images and images lacking metadata on the patient's sex. We then partitioned that data based on provenance as encoded by the 'attribution' metadata label, intended to credit images under Creative Commons Attribution licenses but often also used to provide information on the image acquisition site (Table 3). This partitioning scheme minimizes chance of overlap in patients between the training and test data; it additionally provides a more robust test scenario for domain shift analysis since spurious associations in training data are unlikely to persist in the different hospitals and geographic regions of the test data. After preprocessing and partitioning, the train set had 45924 images and the test set had 23461 images.

Sites for Train Data	Sites for Test Data
Hospital Clinic de Barcelona	Memorial Sloan Kettering Cancer Center
Department of Dermatology, Hospital Clinic de Barcelona	Sydney Melanoma Diagnostic Center at Royal Prince Alfred Hospital
Department of Dermatology, Medical University of Vienna	Sydney Melanoma Diagnostic Center at Royal Prince Alfred Hospital, Pascale Guitera
ViDIR Group, Department of Dermatology, Medical University of Vienna	The University of Queensland Diamantina Institute, The University of Queensland, Dermatology Research Center
Hospital Italiano de Beunos Aires	
Konstantinos Liopyris	

Table 3: Hospital sites used in the train and test sets.

Model training

We trained image transformers (ViT-Base architectures)[24] to predict a patient's sex based on a dermoscopic image. To do so, we started with classifiers pre-trained on ImageNet[25] and then replaced the 1000-class linear classification head with a new linear head suited for binary prediction. For training, we held out 10% of the training data (selected at random) as a validation set. We optimized the network using an Adam optimizer with learning rate 10^{-5}, $\beta_1 = 0.9$, and $\beta_2 = 0.999$, with a mini-batch size of 16 and cross-entropy loss as our optimization criterion. We optimized the models for 30 epochs, reducing the learning rate by a factor of 0.2 (i.e., $lr_{new} = 0.2 \times lr_{old}$) if the model's loss did not improve for 5 epochs. Finally, we used the epoch with the highest ROC-AUC in the validation data for all subsequent experiments. We repeated this procedure for 5 replicates.

Clustering-based analysis

To identify signals in the data that the classifier relies on to predict sex, we grouped images in the test dataset based on their visual similarity using the K-means clustering algorithm[26] implemented in the scikit-learn Python package (K=20). To obtain clustering features, we then used the first 50 principal components derived by running Principal Component Analysis (PCA) on the embeddings of the penultimate layer of the Resnet50 model[27] (pretrained on ImageNet[28]), a layer capable of capturing the visual and structural similarities between different dermoscopic images.

Once we had the clusters, we calculated the ratio $\frac{\#(\text{predicted males})}{\#(\text{predicted females})}$ for each cluster and chose the two with maximum and minimum ratios corresponding to the male-dominant and female-dominant clusters, respectively. Then, board-certified dermatologists analyzed a subset of 100 images selected uniformly at random from the two clusters to identify signals that differ. Visually clustering the images before comparison let us hypothesize that the signals identified would be significant to the model prediction and not just identified by chance.

Generation of counterfactuals

Another XAI tool we used to investigate the signals that potentially guide the prediction of AI classifiers is called *counterfactual image generation*.[29–31] Counterfactual images are synthetic images that reveal the basis of an AI classifier's decisions by altering attributes of a reference image to create a similar image that prompts a different prediction of a protected attribute from the classifier. For example, consider an AI classifier that predicts a lesion as belonging to a female and a counterfactual from the classifier that differs in some visual signals that predicts the lesion belongs to a male; assuming that we ensure all differences in the counterfactual push the AI classifier's predictions in the desired direction (more male), we may infer that the classifier uses those signals as part of its reasoning process.

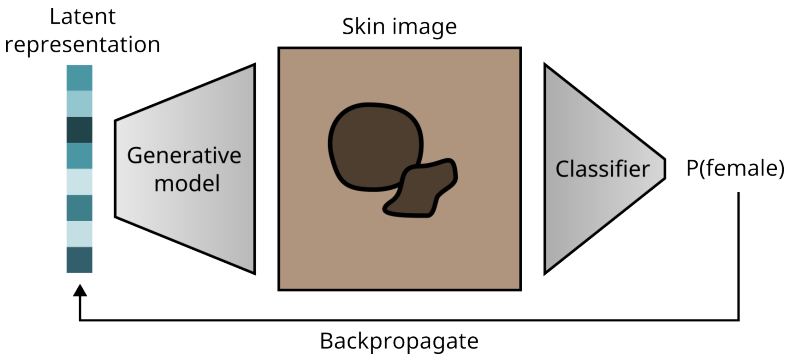


Fig. 12 | Latent space optimization for generating counterfactuals using a StyleGAN. Given a StyleGAN trained on dermoscopic images and a classifier trained to predict sex from lesions, we start with a random latent vector in the GAN's latent space. We then use the generator and the classifier's prediction to optimize the latent representation to generate an image that elicits the desired output probability from the classifier. Throughout the training process, only the latent representation is updated while keeping both the generator and the classifier frozen.

We want to create images that (1) appear "realistic" in the sense that they lie on the manifold of training images, (2) produce the desired target prediction from a classifier, such as a prediction on the opposite side of the decision threshold as the original image, and (3) are similar to the original image in the sense that the original image can be approximately reconstructed by passing an altered, generated image back through the generator. To generate counterfactual images, we employed two different methods to get a diverse range of counterfactuals and obtain a wide spread of potentially important signals for the sex prediction task, as follows.

Latent space optimization

In the first method, we followed prior work[32–34] to simply optimize the image using gradient descent to elicit a desired response from the classifier. To ensure the image remains realistic, we optimize the latent representation of that image in the latent space of a generative adversarial network rather than do so directly in pixel-space. (Figure 12) Since we are interested in broadly understanding the predictions of the classifier rather than explaining the classifier's predictions for a specific output, we utilize randomly generated images as our references, eliminating the need for an encoder network as was used in prior efforts[32].

We generated a pair of counterfactual images by first choosing a random latent vector $\boldsymbol{z} \sim \mathcal{N}(\mathbf{1}, I)$, where $\mathbf{1}$ is a d-dimensional vector of 1s and I is the $d \times d$ identity matrix. Given a generator $G : \mathcal{R}^d \rightarrow \mathcal{R}^{224\times224}$ and classifier $f : \mathcal{R}^{224\times224} \rightarrow [0, 1]$ that quantifies the probability of the image representing a female patient's lesion, we performed gradient descent on z to optimize $f(G(z))$. Based on any given z, we generated a female counterfactual by minimizing $-f(G(z))$ until $f(G(z')) > 0.95$ (where z' represents the updated latent vector) and a male counterfactual by minimizing $f(G(z))$ until $f(G(z')) < 0.05$. Since the optimization is deterministic and the optimization problem is not convex, a portion ($\sim$60%) of initial vectors z fail to produce either a female or male counterfactual based on the preceding criteria; we stop optimization after a maximum of 10 steps and exclude these from further analysis.

During optimization, we use a learning rate of 0.02. The generator and the classifier are kept frozen during the optimization procedure, and only the latent representation is updated. For the generator G, we chose a styleGAN2 (a generative adversarial network, or GAN) with adaptive discriminator augmentation.[35] Supplementary methods provide details about the training procedure.

Explanation by Progressive Exaggeration

In the second, more rigorous method, we generated counterfactual images using a variant of the technique *Explanation by Progressive Exaggeration* (EBPE),[36] as described in prior work.[5] These updates seek to improve image quality, stabilize training, and restrict the altered attributes to those that would drive a classifier toward a different prediction (e.g., avoiding any adversarial perturbations or altering confounding attributes).

Formally, let $\mathcal{X} \subset [0, 1]^{d^2}$ represent a set of images drawn from some data manifold $\mathcal{M}_\mathcal{X}$, where $d \in \mathbb{N}$ is the horizontal and vertical resolution of the (square) images, and let $f : [0, 1]^{d^2} \rightarrow [0, 1]$ be a classifier that outputs the probability that a given dermoscopic image belongs to a female. Our goal is to obtain a generator $G : [0, 1]^{d^2} \times \mathcal{C} \rightarrow [0, 1]^{d^2}$ that produces a counterfactual image $\tilde{x}$ when given an input image x and a condition $c \in \mathcal{C} \subset \mathbb{N}$, which indicates

the target output probability that the classifier should produce when evaluated on the counterfactual image $\tilde{x}$. (Note that for simplicity of notation, we condense the generator and encoder of the original paper into a single function G).

As in the original implementation of EBPE, our condition c is a discrete value that indexes a "bin" in the discretized output space of the classifier f; we chose $\mathcal{C} = \{0, 1, ..., 9\}$ with corresponding target outputs in the bins $\{[0, 0.1), [0.1, 0.2), ..., [0.9, 1]\}$. The three requirements listed above then mean that (1) the range of the generator $G(\mathcal{X}, \mathcal{C})$ is contained in the data manifold $\mathcal{M}_{\mathcal{X}}$, (2) the classifier's prediction for the generated image $f(G(x, c))$ is approximately equal to the target output (in our case, the bin's center at $c/10 + 0.05$), and (3) if $f(x)$ falls within the bin indexed by c, then $G(G(x, c'), c) \approx x$ for each $c' \in \mathcal{C}$.

To obtain a generator with these properties, we optimize the generator G in conjunction with a discriminator network $D : [0, 1]^{d^2} \to \mathbb{R}$, which attempts to distinguish real from generated images. Unlike the original implementation, we update the discriminator so that it does not depend on a condition c. The original discriminator implementation attempts to differentiate generated from real images that elicit a particular prediction from the classifier, which could encourage generated images to appear similar to the subset of real images that include potentially via changes that do not alter the output of the classifier. In contrast, our discriminator implementation instead attempts to differentiate generated images from any real image such that it encourages only the generated images to appear similar to real images (Figure 13).

To reflect this update, we choose the following functions for the loss of the discriminator L_D and the generator L_G. In the following equations, the random variables X and C take values in $\mathcal{X}$ and $\mathcal{C}$ and are distributed uniformly over $\mathcal{X}$ and $\mathcal{C}$; θ_D and θ_G are the parameters of the discriminator and generator, respectively; $b : [0, 1] \to \mathcal{C}$ returns the bin index $b(f(X))$ of the output of the classifier; $\tilde{b} : \mathcal{C} \to \{0.05, 0.15, ..., 0.95\}$ returns the center of the bin at index C; and D_{KL} is the Kullback–Leibler divergence.

$$L_D(\theta_D) = -\lambda_{GAN}\mathbb{E}_{X,C}[min(0, -1 + D_{\theta_D}(X)) + min(0, -1 - D_{\theta_D}(G_{\theta_G}(X, C))].$$

$$L_G(\theta_G) = \lambda_{GAN}L_{GAN}(\theta_G; \theta_D) + \lambda_{rec}L_{rec}(\theta_G) + \lambda_f L_f(\theta_G).$$

The individual components of L_G are:

$$L_{GAN}(\theta_G; \theta_D) = -\mathbb{E}_{X,C}[D_{\theta_D}(G_{\theta_G}(X, C))].$$

$$L_{rec}(\theta_G) = \mathbb{E}_{X,C}[\|X - G_{\theta_G}(X, b(f(X)))\|_1 + \|X - G_{\theta_G}(G_{\theta_G}(X, C), b(f(X)))\|_1].$$

$$L_f(\theta_G) = \mathbb{E}_{X,C}[D_{KL}(\tilde{b}(C)\|f(G_{\theta_G}(X, C)))].$$

In addition to our introduction of a non-conditional discriminator, we also update G to use an architecture similar to that used in CycleGANs[37] to improve image quality. Supplementary methods describe the optimization procedure and further training details.

Expert evaluation of counterfactuals

To identify specific image factors that the classifiers use to predict protected attributes, we asked multiple board-certified experts to analyze counterfactual images generated by the two preceding methods and determine which aspects of each image were altered, implying that they could be potential signals that affect the classifier's decisions.

To help the experts evaluate the counterfactuals, we developed a graphical interface in the form of a web app hosted on Amazon Web Services (AWS). Each web app page displays a single pair of counterfactual images, with both having opposite predictions from the classifier of the protected attribute class. The counterfactual pair to be displayed is generated by either EBPE or by the latent space optimization technique (selected at random), and the expert evaluators were not told which technique was used.

The experts first analyzed the pair of counterfactuals and then answered questions, including (1) Did the images look realistic? (2) How did the two counterfactuals differ from each other in terms of visual attributes? The evaluators could enter a free text response for the second question. They could also save their progress and leave then return to the app to pick up where they left off. The app provided a slider expert evaluators could use to linearly interpolate from one counterfactual to the other for examining the subtle changes that occur between the two.

After we analyzed all counterfactual pairs, we discarded the pairs for which either of the images looked unrealistic and grouped similar visual features that were recorded in multiple images in a given direction.

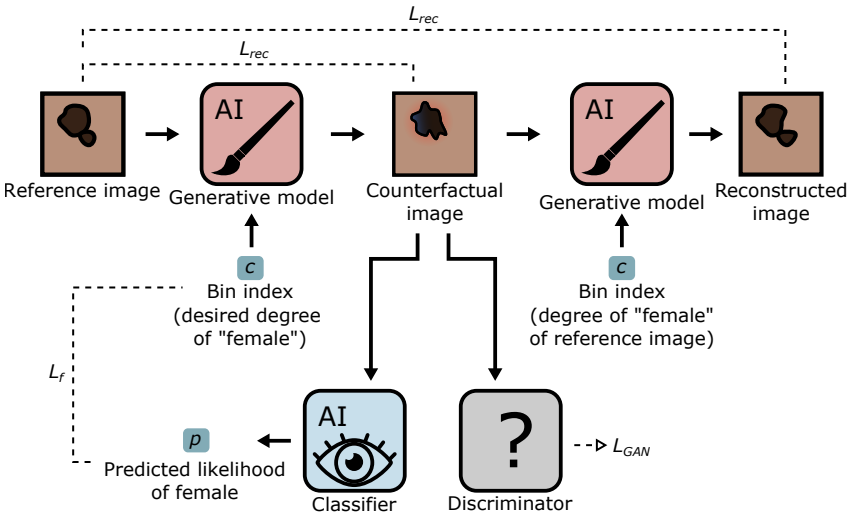


Fig. 13 | The GAN setup used during training with the Explanation by Progressive Exaggeration (EBPE) technique. Given a reference image and a bin index representing the desired prediction from the classifier, the generative model creates a counterfactual image. The loss term L_f enforces that the counterfactual, when evaluated by the classifier (which is frozen), generates the corresponding output as specified by the bin index. The counterfactual is also passed to the discriminator, which attempts to discern whether it represents a real or generated image and thus enforces realism of the counterfactuals (via L_{GAN}). The reconstruction L1 loss L_{rec} enforces that the counterfactual is similar to the reference image. It consists of two steps: (1) The counterfactual is passed back to the generative model, along with the bin index that corresponds to the classifier's prediction on the reference image, in an attempt to reconstruct the reference image (L_{rec}, top). (2) We also attempt to reconstruct the reference image with only a single pass through the generator (L_{rec}, lower) by again passing a bin index that matches the classifier's output on the reference image.

Quantification of explained performance with 'removal via balancing'

Once we identified the signals that could potentially exist, we quantified the amount of the ML classifier's performance that could be explained by a putatively important signal (or signals); we refer to these as the 'query signal(s)' as the drop in predictive performance when the signal(s) is 'removed' from the test data, in the sense that the query signal is balanced with respect to the prediction target. We term this technique *removal via balancing*. Specifically, to 'remove' a query signal A, we update the test data to form a new pseudo-population in which $A \perp Y$, where Y is the protected attribute being predicted (that is, a patient's sex). Our scheme requires that each signal be represented by a scalar or vector in $\mathbb{R}^n$ (where $n \in \mathbb{N}$), e.g., as a scalar quantification of that signal or a one-shot encoded vector. Our goal is to weight each sample by the reciprocal of its propensity, $1/P(Y|A)$, which, in alignment with prior work [38–41] on inverse probability treatment weighting (IPTW) in the field of treatment-effect estimation, provides a pseudo-population with the desired property that $A \perp Y$. Since the true propensity is not known, we followed prior work [40] and estimated via a logistic regression $\hat{p} : \mathbb{R}^n \to [0, 1]$; we assigned to sample i with vector of signals a_i the weight $1/\hat{p}(a_i)$.

We caution that despite borrowing balancing scores from the causal inference literature, our technique does not aim to infer causal relationships between query attributes and the model's predictions. In our view, no direct analogy can be drawn between our use of balancing scores and their use in treatment effect estimation. Importantly, when our technique removes a query signal via balancing, correlated signals may also be (partially) removed. For instance, if two signals correlate perfectly in the test data, our technique cannot differentiate which signal is important for a classifier. In this way, our technique defines signals on the basis of how they appear in the test data; e.g., in the extreme case, two semantically different but perfectly correlated signals are effectively defined as a single signal for the purposes of our analysis.

We quantified the *proportion of performance explained* as the ratio of the performance decline after balancing to the maximum possible performance decline (to random performance of ROC-AUC=0.5):

$$\text{Proportion explained} := \frac{\text{ROC-AUC}_{\text{original}} - \text{ROC-AUC}_{\text{balanced}}}{\text{ROC-AUC}_{\text{original}} - 0.5}$$

This method needs some form of signal labels for all images in the test data, and there are no labels for these signals in the metadata. To generate these labels, we created another web app hosted on AWS. This app shows an

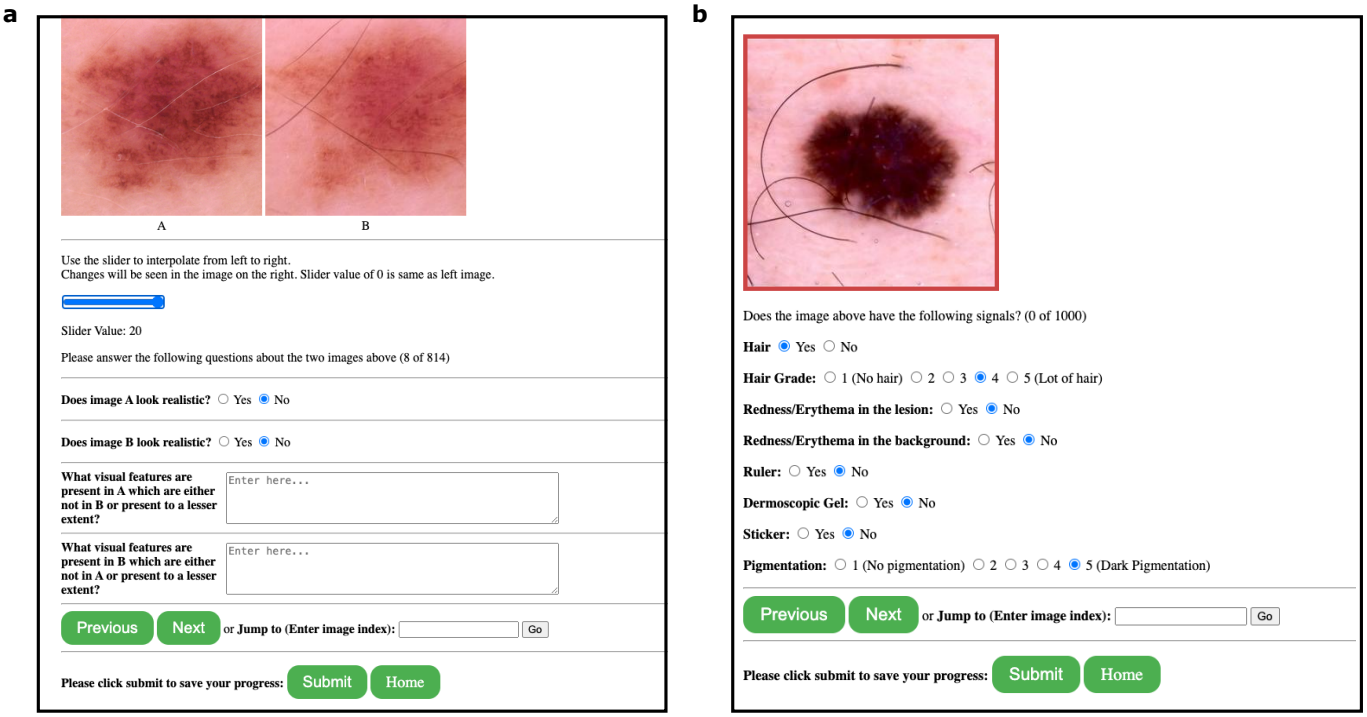


Fig. 14 | Web apps used to identify signals from the counterfactual images and label signals in the testing set for quantification. (a) Screenshot of the UI used to identify signals that are visually different between a pair of counterfactuals. The slider can be used to interpolate between the counterfactuals to assist in highlighting the differences. **(b)** Screenshot of the UI used to label the identified signals. The user labels the presence or absence of each signal. Specifically for hair and pigmentation, we also let the user select from a scale, with each number representing the grade of the signal present in the image.

image from the test set and asks the expert evaluators to label it with the identified signals. We did this for a subset of the images and trained a classifier to label the remaining test data.

Data availability

Images used in this study were obtained from publicly available repositories. ISIC images are available at `https://challenge.isic-archive.com/data/`.

Code availability

Author contributions

Funding

Ethics declarations

Competing interests

References

1. *Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices* `https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices`. Accessed: 2023-11-10.
2. Benjamens, S., Dhunnoo, P. & Meskó, B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *npj Digital Medicine* **3** (2020).

3. Wu, E. *et al.* How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. *Nature Medicine* **27,** 582–584 (2021).
4. DeGrave, A. J., Janizek, J. D. & Lee, S.-I. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence* **3,** 610–619 (2021).
5. DeGrave, A., Ran Cai, Z., Janizek, J. D., Daneshjou, R. & Lee, S.-I. Auditing the inference processes of medical-image classifiers by leveraging generative AI and the expertise of physicians. *accepted in principle at Nature Biomedical Engineering* (2023).
6. Buolamwini, J. & Gebru, T. *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification* in *Proceedings of Machine Learning Research Conference on Fairness, Accountability, and Transparency* **81** (2018), 1–15.
7. Daneshjou, R. *et al.* Disparities in dermatology AI performance on a diverse, curated clinical image set. *Science Advances* **8,** eabq6147 (2022).
8. Poplin, R. *et al.* Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering* **2,** 158–164 (2018).
9. Gichoya, J. W. *et al.* AI recognition of patient race in medical imaging: a modelling study. *The Lancet Digital Health* **4,** E406–E414 (6 2022).
10. Yang, Y., Zhang, H., Gichoya, J. W., Katabi, D. & Ghassemi, M. The limits of fair medical imaging AI in real-world generalization. *Nature Medicine* **30,** 2838–2848 (2024).
11. Yamashita, T. *et al.* Factors in color fundus photographs that can be used by humans to determine sex of individuals. *Transitional Vision Science and Technology* **9** (4 2020).
12. Lang, O. *et al.* Using generative AI to investigate medical imagery models and datasets. *EBioMedicine* **102** (2024).
13. Geirhos, R. *et al.* Shortcut learning in deep neural networks. *Nature Machine Intelligence* **2,** 665–673 (2020).
14. Ly, D., Forman, D., Ferlay, J., Brinton, L. A. & Cook, M. B. An international comparison of male and female breast cancer incidence rates. *International Journal of Cancer* **132,** 1918–1926 (8 2013).
15. 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).
16. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B. & Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* **30** (2017).
17. Jeanneret, G., Simon, L. & Jurie, F. *Diffusion models for counterfactual explanations* in *Proceedings of the Asian conference on computer vision* (2022), 858–876.
18. Glocker, B., Jones, C., Bernhardt, M. & Winzeck, S. Algorithmic encoding of protected characteristics in image-based models for disease detection. *arXiv preprint arXiv:2110.14755* (2021).
19. Zou, J., Gichoya, J. W., Ho, D. E. & Obermeyer, Z. Implications of predicting race variables from medical images. *Science* **381,** 149–150 (2023).
20. 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). *arXiv:1710.05006* (2018).
21. Tschandl, P., Rosendahl, C. & Kittler, H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data* **5** (2018).
22. Combalia, M. *et al.* BCN20000: Dermoscopic Lesions in the Wild. *arXiv:1908.02288* (2019).
23. Rotemberg, V. *et al.* A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific Data* **8,** 34 (2021).
24. Dosovitskiy, A. *et al.* An image is worth 16x16 words: transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)* (2021).
25. Krizhevsky, A., Sutskever, I. & Hinton, G. E. *ImageNet classification with deep convolutional neural networks* in *2012 Conference on Neural Information Processing Systems* (2012).
26. MacQueen, J. *et al. Some methods for classification and analysis of multivariate observations* in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* **1** (1967), 281–297.

27. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 770–778 (2016).
28. Deng, J. *et al. Imagenet: A large-scale hierarchical image database* in *2009 IEEE conference on computer vision and pattern recognition* (2009), 248–255.
29. Sauer, A. & Geiger, A. Counterfactual generative networks. *arXiv preprint arXiv:2101.06046* (2021).
30. Chang, C.-H., Creager, E., Goldenberg, A. & Duvenaud, D. Explaining image classifiers by counterfactual generation. *arXiv preprint arXiv:1807.08024* (2018).
31. Yu, Y. *et al. Towards counterfactual image manipulation via clip* in *Proceedings of the 30th ACM International Conference on Multimedia* (2022), 3637–3645.
32. Joshi, S., Koyejo, O., Kim, B. & Ghosh, J. xGEMs: generating examplars to explain black-box models. *preprint at arXiv:1805.08867v1* (2018).
33. Balasubramanian, R., Sharpe, S., Barr, B. & Bruss, C. B. Latent-CF: a simple baseline for reverse counterfactual explanations. *preprint at arXiv:2012.09301v2* (2021).
34. Cohen, J. P., Blankemeier, L. & Chaudhari, A. Identifying spurious correlations using counterfactual alignment. *preprint at arXiv:2312.02186v1* (2023).
35. Karras, T. *et al.* Training generative adversarial entworks with limited data. *preprint at arXiv:2006.06676* (2020).
36. Singla, S., Pollack, B., Chen, J. & Batmanghelich, K. *Explanation by progressive exaggeration* in *International Conference on Learning Representations* (2019).
37. Zhu, J.-Y., Park, T., Isola, P. & Efros, A. A. *Unpaired image-to-image translation using cycle-consistent adversarial networks* in *Proceedings of the IEEE international conference on computer vision* (2017), 2223–2232.
38. Rosenbaum, P. R. & Rubin, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika* **70,** 41–55 (1 1983).
39. Rosenbaum, P. R. Model-based direct adjustment. *Journal of the American Statistical Association* **82,** 387–394 (398 1987).
40. Austin, P. C. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research* **46,** 399–424 (3 2011).
41. Brookhart, M. A., Wyss, R., Layton, J. B. & Stümer, T. Propensity Score Methods for Confounding Control in Non-Experimental Research. *Circulation: Cardiovascular Quality and Outcomes* **6,** 604–611 (2013).
42. Kirillov, A. *et al. Segment anything* in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), 4015–4026.
43. Karras, T. *et al. Analyzing and improving the image quality of StyleGAN* in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), 8107–8116.
44. Bommasani, R. *et al.* On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).
45. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T. J. & Zou, J. A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine* **29,** 2307–2316 (2023).
46. Li, C. *et al.* Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems* **36** (2024).
47. Tiu, E. *et al.* Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. *Nature Biomedical Engineering* **6,** 1399–1406 (2022).
48. Eslami, S., Meinel, C. & De Melo, G. *Pubmedclip: How much does clip benefit visual question answering in the medical domain?* in *Findings of the Association for Computational Linguistics: EACL 2023* (2023), 1181–1193.
49. Zhang, S. *et al.* BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915* (2023).
50. Xu, H. *et al.* A whole-slide foundation model for digital pathology from real-world data. *Nature,* 1–8 (2024).
51. Kim, C. *et al.* Transparent medical image AI via an image–text foundation model grounded in medical literature. *Nature Medicine,* 1–12 (2024).

Supplementary Tables

Lesion Diagnosis	OR, train	OR, test	N, train	N, test
solar lentigo	1.74	2.07	255	162
lentigo simplex	0.60	1.57	90	155
atypical melanocytic proliferation	1.46	1.07	13	86
squamous cell carcinoma	0.70	1.10	699	32
dermatofibroma	1.40	2.35	264	17
basal cell carcinoma	0.81	1.59	3278	21
seborrheic keratosis	0.87	0.55	1316	214
actinic keratosis	1.13	0.85	850	60
nevus	1.10	1.46	26005	1979
melanoma	0.99	0.48	5122	649

Supplementary Table. 1 | Odds ratios (ORs) for predicting a female sex based on the diagnosis. Diagnoses are sorted by the OR in the training data, with those lacking at least 10 corresponding images in both the training and test data excluded. Images lacking a diagnosis are excluded from the analysis (6575 of 45924 images in the training data, and 19926 of 23461 images in the test data). N indicates the number of images.

CXR Diagnosis	OR, train	OR, test	N, train	N, test
cardiomegaly	0.52	0.87	13333	5435
fracture	1.83	1.75	4681	837
pneumothorax	1.50	2.25	9734	767
pleural effusion	1.21	1.85	42502	5203
lung lesion	1.20	1.21	4237	996
edema	0.84	1.23	25782	2695
pneumonia	0.84	1.12	2964	2274
Atelectasis	1.14	1.24	16234	5637
consolidation	1.09	1.20	7165	1048
support devices	1.06	1.52	56838	5931
enlarged cardiomediastinum	0.96	1.23	5402	0878
lung opacity	1.01	1.15	52684	6558

Supplementary Table. 2 | Odds ratios (ORs) for predicting a white race from chest x-rays based on the diagnosis. Diagnoses are sorted by the OR in the training data, with those lacking at least 10 corresponding images in both the training and test data excluded. N indicates the number of images.

Dermoscopy method	OR, train	OR, test	N, train	N, test
contact polarized	0.99	1.11	5277	459
contact non-polarized	1.03	0.92	2047	7963
non-contact polarized	0.77	0.82	102	46

Supplementary Table. 3 | Odds ratios (OR) for prediction of female sex based on method of dermoscopy employed in image acquisition. Images lacking information on acquisition method are excluded (38498 of 45924 images in the training data, and 14993 of 23461 images in the test data). N indicates the number of images.

View Position	OR, train	OR, test	N, train	N, test
frontal	1.12	1.51	93990	28330
lateral	0.889	0.661	15972	8942

Supplementary Table. 4 | Odds ratios (OR) for prediction of white race based on the view position of the chest x-ray employed in image acquisition. N indicates the number of images.

Signal Identified	Description	Counterfactual Technique	Association
dermoscopic gel	Visible dermoscopic gel, also known as immersion fluid, in the image used to enhance visualization of the skin lesion.	Latent Space Optimization	Male
ruler	The dermoscopic image has visible ruler markings that serve as a scale reference.	Latent Space Optimization	Female
pigmentation	Presence of color or pigment within the lesion, ranging from amelanotic to black in color. Darker pigmentation was found to be associated with males.	EBPE	Male
redness	The skin lesion or the background shows signs of reddish dots, indicating inflamed skin or erythema.	EBPE	Male

Supplementary Table. 5 | AI-specific signals identified by analyzing the female and male counterfactual images obtained from the generative model. The description specifies what the signal corresponds to visually, and the association indicates the sex in which the signal was more prevalent.

Supplementary Methods

1 Additional experiments to verify identified signals

To further validate the identified signals in addition to the quantification, we manually manipulated the signals to add or remove them from the images and test the effect on the predictions. We focused on "redness" and "stickers" since those are signals which are easier to edit in the images. The other signals are more diffused and, therefore, difficult to edit in the images by direct manipulation.

1.1 Effect of redness

The presence of red dots on the dermoscopic images was identified as one of the potential signals associated with the male sex by the trained classifier. To replicate the effect of having red dots, we manually added red dots to the existing images from the test set. We first started with an blank image of the same width and height as the lesion and added 100 red dots of a fixed radius to random locations in the image. Then, we blended the original dermoscopic image with it to simulate the presence of red dots (Supplementary Figure 1a). We then observed the change in prediction accuracy and the male predicted probability before and after adding the red dots for the dermoscopic lesions in the male and female subgroups.

Considering the subgroup of ground truth male lesions, both the accuracy and the male predicted probability increased after adding the red dots (Supplementary Figure 1b). The accuracy increased from $\sim$0.718 to $\sim$0.871 (21%) while the male predicted probabilty increased from $\sim$0.716 to $\sim$0.0.861 (20%). Considering the subgroup of ground truth female lesions, the accuracy decreased and the male predicted probability increased after adding the red dots (Supplementary Figure 1c). The accuracy decreased from $\sim$0.702 to $\sim$0.325 (54%) while the male predicted probabilty increased from $\sim$0.312 to $\sim$0.659 (111%). Both of these results indicates that images with red dots are associated with male lesions by the classifier and further validates the signals identified through counterfactual image generation.

1.2 Effect of stickers

From the clustering based analysis, stickers were identified as a potential signal that the classifier could use to predict sex from dermoscopic lesions, with male being associated with the presence of stickers. Since stickers are artificats which are not part of the skin lesion, they are easier to manually manipulate. We used different techniques to add and remove stickers from the images in the dataset.

To remove the stickers from the images which had stickers, we first had to segment them within the image. We used the recently developed Segment Anything Model (SAM)[42] to segment out the stickers in the images and it worked well for our purposes (accuracy of 90% from 500 manually tested images). Once we had the masks for the stickers, we inpainted them with the color of the skin to emulate the removal of stickers. However, the skin had different coloration in different regions, so using a simple mean color of the unmasked pixels did not look realistic. Instead, to do the inpainting, for each mask, we considered the closest 5 pixels that were bordering the sticker mask and took the mean color of those to inpaint the masked region (Supplementary Figure 2a). To emulate the addition of stickers to the images without stickers, we selected a set of three stickers from existing dermoscopic images with stickers representing

different sticker patterns that were observed in the daatasts. We then overlaid one of them selected at random on each of the images either at the top, bottom, left, or right sides (Supplementary Figure 2b). Once we had a set of images in which the stickers were either inpainted or overlaid, we observed the change in the accuracy and the male predicted probability of the classifier before and after the manipulation for the male and female subgroups.

Considering the case in which we start with dermoscopic images which have stickers and then inpaint them using the technique mentioned above, for the subgroup of male lesions, both accuracy and the male predicted probability decrease, while for the subgroup of female lesions, both of them remain almost the same (Supplementary Figure 2c). This indicates that removing stickers from images makes them more "female-like" to the classifier, which corroborates the findings from the clustering analysis. In the case where we start with dermoscopic images which don't have stickers and manually add stickers to them, for the subgroup with male lesions, both the accuracy and male predicted proabiblity remains the same before and after the addition. For the subgroup with female lesions, the accuracy decreases and the male predicted probability increases. This again confirms the observation that adding stickers to images make them more "male-like" to the classifier.

2 Frequency Spectrum Analysis

AI classifiers can be sensitive to features that humans can't directly visualize in the original images and they can use those differences to make predictions. To this end, we followed prior work[9] and performed a frequency spectrum analysis on the dermoscopic images to break them into their component frequencies and evaluated the image frequencies for which the classifier had a higher sensitivity for the task of sex prediction. Spatial frequency describes how quickly pixel values change in space, with the higher frequency features representing sharp transitions between different regions, such as edges, lines, textures, and the lower frequency features representing gradual changes in color or intensity, such as those found in backgrounds or large, uniformly colored areas.

To obtain the component frequencies from an image, we first applied a fast fourier transform (fft) to the image which converts it to frequency spectrum. Then we shifted the zero-frequency component to the center of the spectrum to make filtering easier. In the shifted fourier transform, the center of the image represents the highest frequencies and they decrease as we move away from the center. Following this, to perform the frequency filtering, we created two circular masks, one with a radius of 5 pixels and the other with a radius of 50 pixels. For for the smaller circular mask, we applied a high-pass filter, causing only the highest frequencies present inside the circle to pass through while filtering the rest out. For the bigger circular mask, we applied a low-pass filter, causing the highest frequencies within the circle to get filtered out and only allowing the lower frequencies outside the circle to pass through. After applying these filters, we then used the inverse fast fourier tranform (ifft) to convert the images back from the frequency domain to the pixel domain (Supplementary Figure 3a). The high-pass filter results in images in which only high frequency features (like hair) are retained, while the low-pass filter results in images in which only low frequency features are retained, which creates a softening or blurring effect.

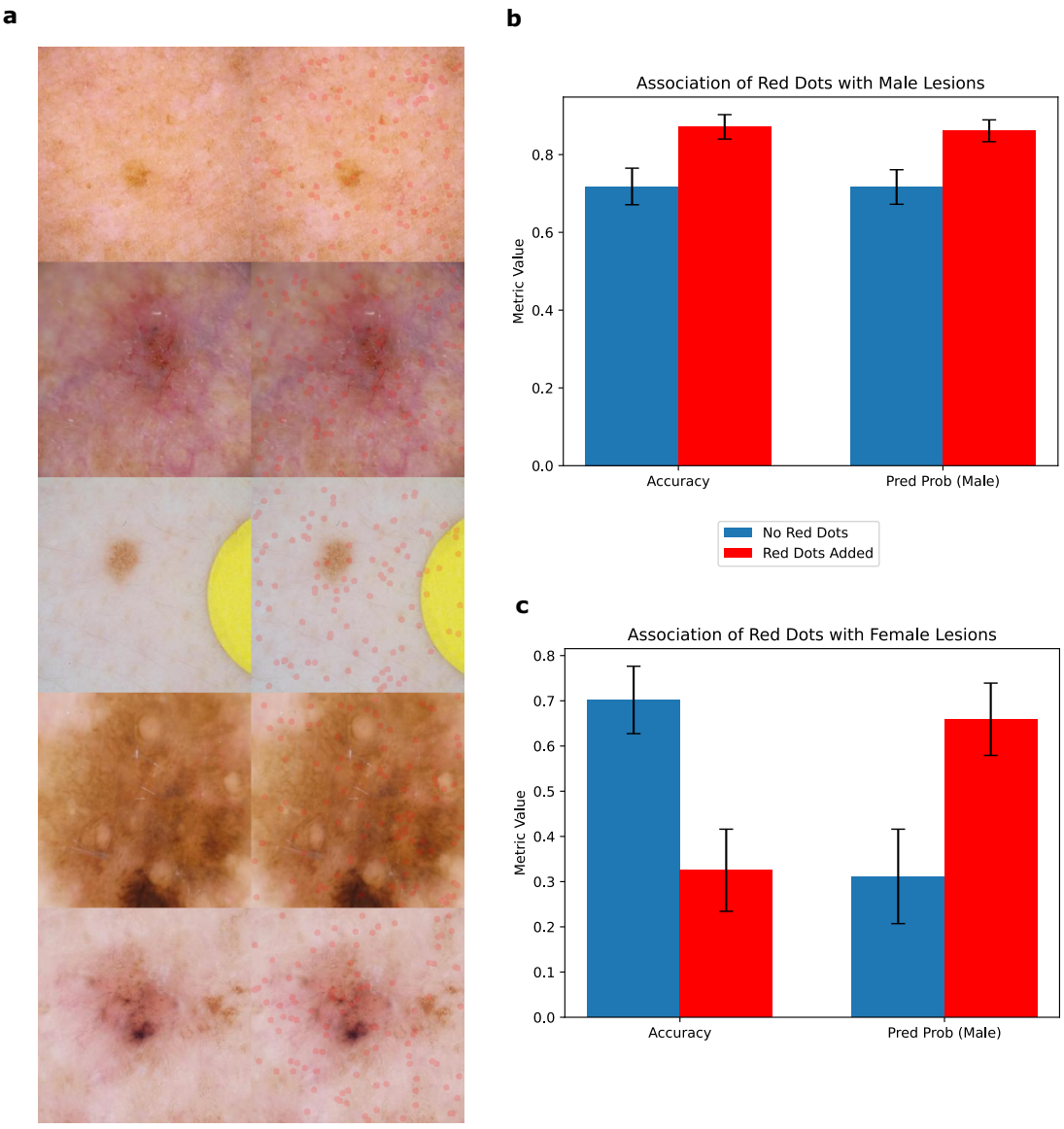
Once we had the filtered images, we used the evaluated them using the trained classifier (Supplementary Figure 3b). Unsurprisingly, the original images without any filtering perform the best. However, contrary to expectations, the images obtained from the high-pass filter (TODO: Include values) outperform the images obtained from the low-pass filter even though most of the image is not visible after the high-pass filtering. Low-pass filtering only causes the image to get blurred while returning most of the structural and color-based variations but this still causes the performance to drop significantly. On the other hand, we lose most of the visual information using the high-pass filter but it still retains enough information to have high performance. These results further validate that hair is an important signal that the classifier relies on to make predictions since the high-pass frequency filtering retains features like hair.

3 Training details for the generative models to generate counterfactual

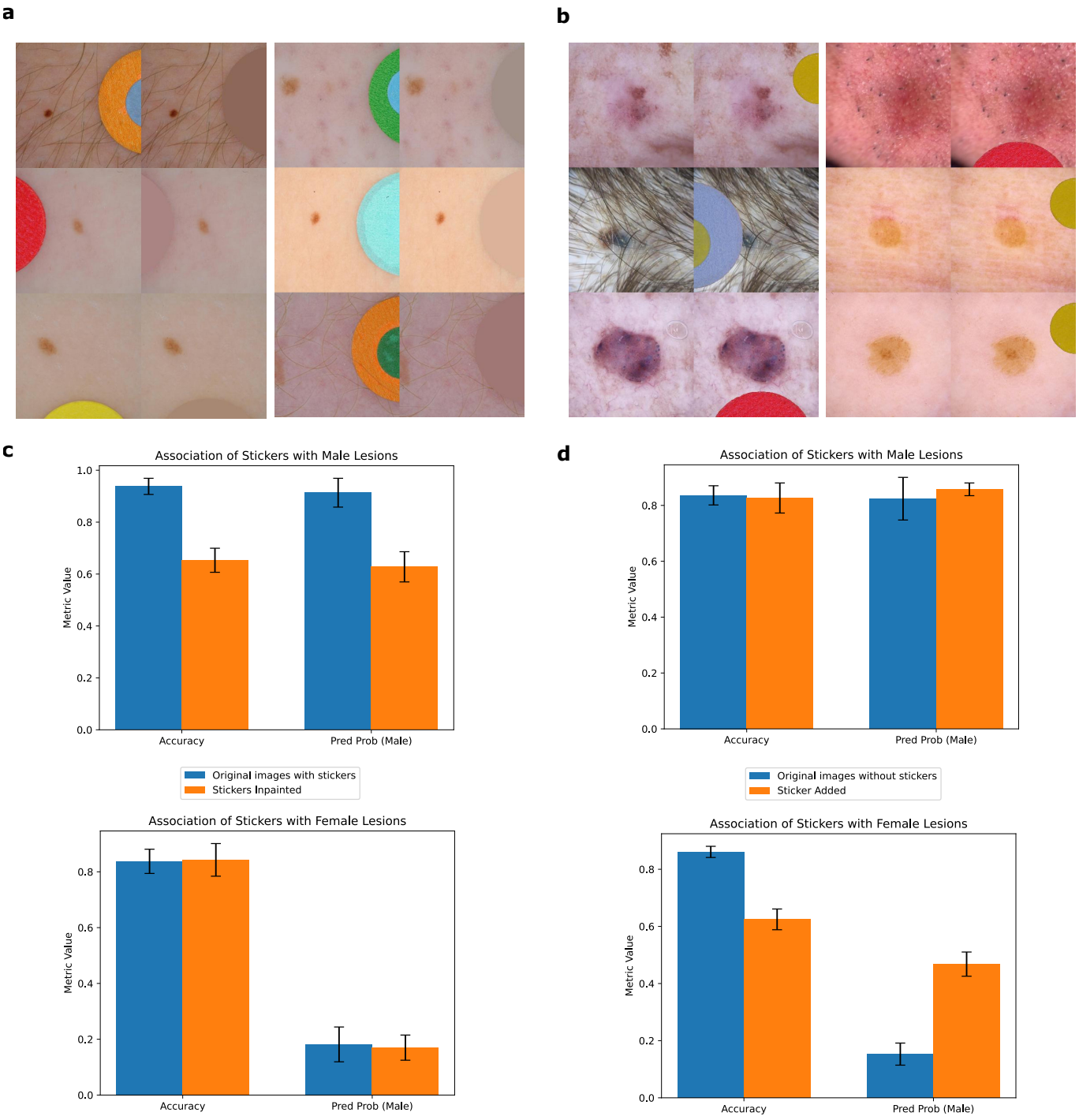
Here we provide additional training details on the two methods we employed to generate counterfactuals, latent space optimization and explanation by progressive exaggeration (EBPE).

3.1 Latent Space Optimization

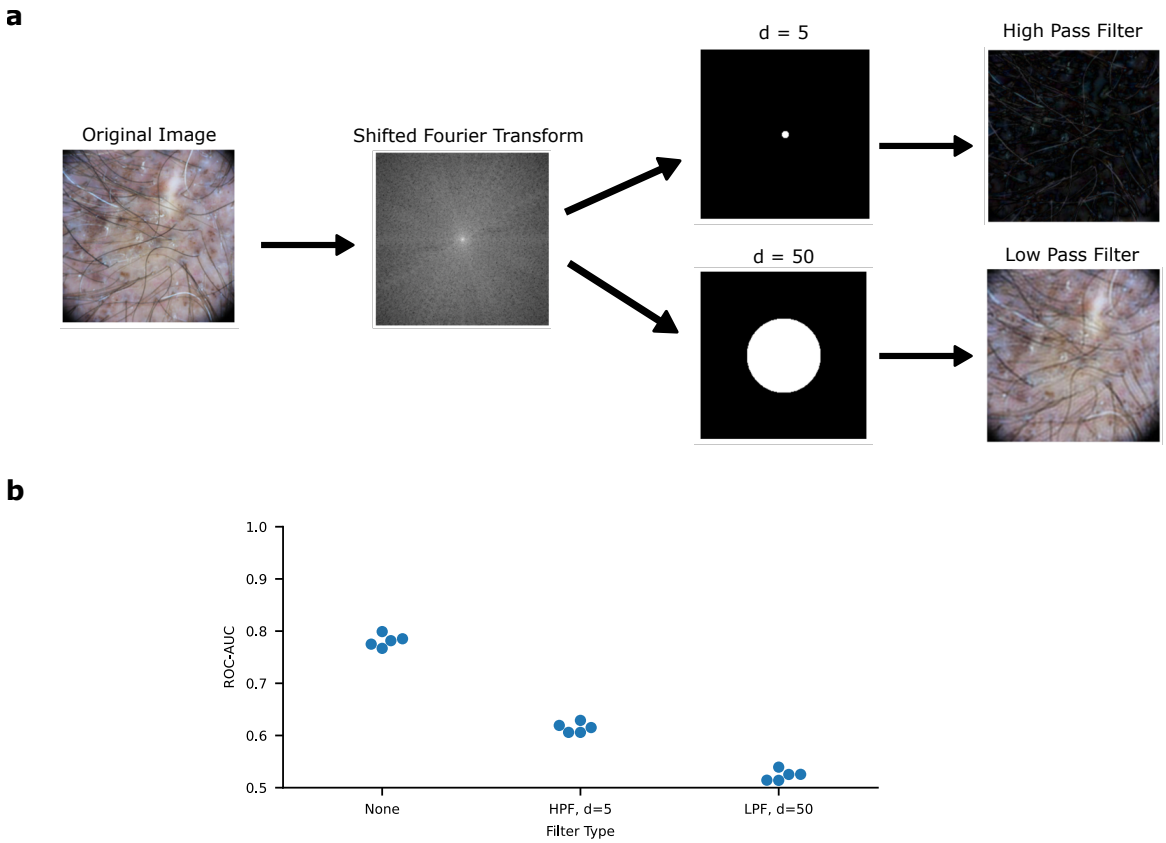
To train the styleGAN2 that was used as the generator, we used all images from the ISIC 2019 dataset, which were resized such that the short edge measured 256 pixels and then center-cropped to 256×256 pixels. We fine-tuned the model starting from a checkpoint pre-trained on Flickr Faces High Quality 256 (FFHQ256). During training, we augmented the training data by randomly, horizontally flipping images. We optimized the networks using the Adam optimizer with a learning rate of 0.0025, $\beta_1 = 0$, $\beta_2 = 0.99$, and batch size of 64. For adaptative discriminator



Supplementary Fig. 1 | Additional experiments to verify impact of redness on sex prediction a, Sample images with redness manually added using red dots. The left image in each pair is the original real image and the right image is the manipulated one with red dots added. **b,** Impact of adding red dots on the accuracy and predicted probability of the male lesions. When red dots are added, the accuracy and predicted probability both increase. **c,** Impact of adding red dots on the accuracy and predicted probability of the female lesions. When red dots are added, the accuracy decreases and the predicted probability increases.



Supplementary Fig. 2 | Additional experiments to verify impact of stickers on sex prediction a, Sample images with stickers manually inpainted with background color. The left image in each pair is the original real image and the right image is the manipulated one with the stickers removed using inpainting. **b,** Sample images with stickers manually added. The left image in each pair is the original real image and the right image is the manipulated one with the stickers added. **c,** Impact of removing the stickers by inpainting the original images (with stickers) on the accuracy and predicted probability of the male (top) and female (bottom) lesions. When the stickers are removed, the accuracy and predicted probability of the male lesions decreases while that of the female lesions relatively stays the same. **d,** Impact of removing the stickers by adding stickers to the original images (without stickers) on the accuracy and predicted probability of the male (top) and female (bottom) lesions. When the stickers are added, the accuracy and predicted probability of the male lesions remains the same while for female lesions the accuracy decreases and predicted probability increases.



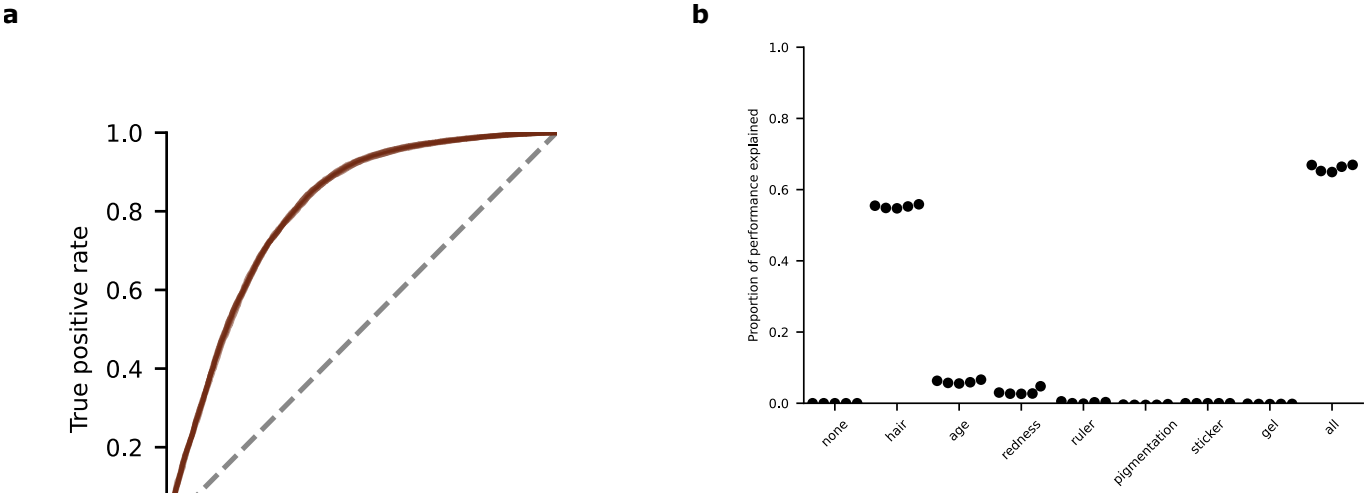
Supplementary Fig. 3 | Frequency filtering experiment a, We start with a real dermoscopic image and retrieve the shifted fourier trasnform to break it down into its component frequencies. withe the frequencies decreasing from the centre. Then, a circle mask is applied to the transform of varying pixel diameters to filter out different frequencies. For d=5, we use a high pass filter to only the allow the frequencies within the circle to pass through. For d=50, we use a low pass filter to only allow for frequencies outside the circle to pass through. **b,** Performance of using different filter types and circle masks for the sex prediction task with 5 random seeds.

augmentation, we set the target to 0.6.[43] We performed optimization for a total of 25000 kilo-images, requiring approximately four days of training on four NVIDIA RTX 2080TI GPUs. The distribution of images produced by the final network differed from the distribution of training images by a Fréchet Inception Distance of 6.29.

3.2 Explanation by Progressive Exaggeration

To optimize the discriminator and the CycleGAN-based generator, we followed the PyTorch reference implementation[5] and used an Adam optimizer with a learning rate of 2×10^{-4}, $\beta_1 = 0$, and $\beta_2 = 0.9$, with a mini-batch size of 32. To prevent the discriminator from outpacing the generator, we trained the discriminator for 5 mini-batches for each mini-batch that the generator was trained, and we applied spectral normalization to the discriminator's parameters. To avoid overfitting, we also applied data augmentation, including random cropping and random brightness modifications. To choose the hyperparameters λ, we followed the reference implementation and chose $\lambda_{GAN} = 1$ and $\lambda_f = 1$. To balance the magnitude of the generator's alterations such that the counterfactuals were similar to original images but still contained perceptible differences (based on manual visual analysis of images), we chose $\lambda_{rec} = 3$ after gradually relaxing the λ_{rec} term from the value $\lambda_{rec} = 100$ suggested in the original publication. The generative models for each classifier and for each dataset were all trained using identical parameters. Comparison of counterfactuals generated by independent re-trainings of a generative model preserved which attributes varied between the male and female counterfactuals.

The generative models were trained for either 500 epochs (ISIC dataset) or 10^4 epochs (Fitzpatrick17k dataset), to achieve approximately equal total training time for each dataset (∼10,000 kilo images); training time for a single generative model amounted to between one week and one month on an NVIDIA RTX 2080 TI graphics processing unit, depending on the complexity of the classifier. To generate counterfactuals, we choose the extreme values for c ($c = 0$ and $c = 9$, corresponding to classifier predicted probability outputs of 0.05 and 0.95 respectively) in order to



Signal Identified	Quantification
hair	55.248 ± 0.415
age	6.016 ± 0.347
redness	3.167 ± 0.711
All	65.089 ± 0.744

Supplementary Fig. 4 | CLIP-based foundation model encodes demographic information a, Performance of training a linear classifier using a CLIP-based foundation model (MONET) embeddings for predicting sex. Each of the curves represent one of five training replicates using different random seeds. **b,** Visualization of the proportion of ROC-AUC explained across five replicates along with quantification using 95% confidence intervals.

obtain counterfactuals most likely to elicit the prediction of the protected attributes classes.

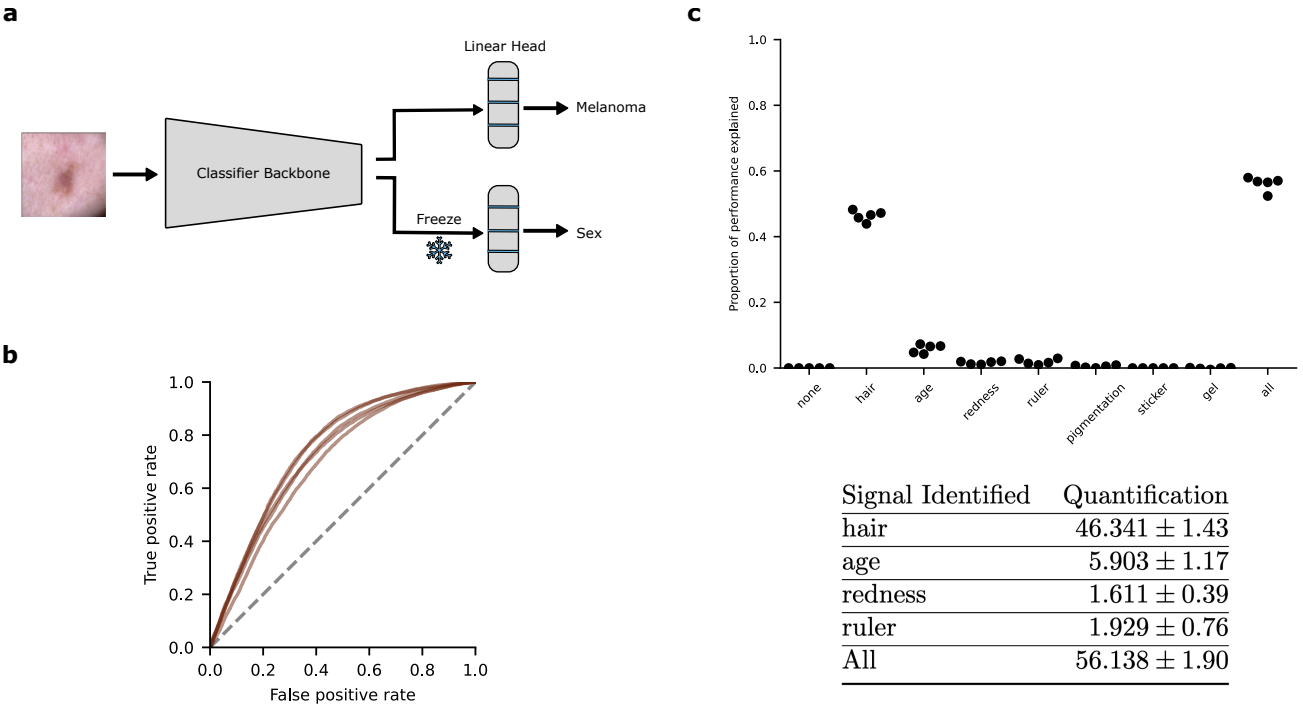
3.3 Melanoma classifier embeddings encode demographic information

In a practical scenario, deep learning classifiers would not be trained specifically for the sex prediction task, but would rather be trained for disease prediction. However, even such classifier embeddings can encode some information for protected attribute classification. To test this out, we first trained a ViT-based classifier to predict melanoma versus melanoma look alikes. Then, we used transfer learning to retrain just the linear head to predict sex instead of melanoma, while keeping the rest of the model weights frozen (Supplementary Figure 5a). We used the same data splits as the ones used for training the ViT-based sex classifier. If the embeddings obtained from the melanoma classifier rely on demographic information, the transfer-learned classifier should also perform well on the sex classification task.

We observed that the transfer-learned classifier is still able to predict sex with a high performance that is comparable to the performance using the original ViT-based classifier (ROC-AUC of 0.742 $\pm$ 0.015; mean $\pm$ standard deviation; Supplementary Figure 5b). Next, using the 'removing via balancing' technique, we quantified the same signals that were identified using the original ViT-based classifier. All the signals that were identified earlir were also important for the transfer-learned classifier, with hair explaining a bulk of the performance (about 45%). Overall, we were actually able to explain more of the performance of the transfer-learned classifier compared to the original classifier (56% versus 42%). This indicates that the transfer-learned classifier can encode demographic information to a larger extent, and it can be used for predicting clinically relevant downstream tasks like melanoma.

3.4 Foundation Models Encode Demographic Information

More recently, foundation models[44] are gaining popularity over specialized models trained for specific downstream tasks due to their ability to generalize across a wide range of tasks, leveraging extensive pretraining on diverse, large-scale datasets. One such family of foundation models is based on CLIP (Contrastive Language-Image Pre-Training), initially developed by OpenAI by training a self-supervised contrastive learning objective on a vast corpus of images sourced from the internet paired with textual descriptions to learn a joint embedding space. Such multi-modal models are especially effective in the clinical domain where there is a lack of structured, specialized datasets with ground



Signal Identified	Quantification
hair	46.341 ± 1.43
age	5.903 ± 1.17
redness	1.611 ± 0.39
ruler	1.929 ± 0.76
All	56.138 ± 1.90

Supplementary Fig. 5 | Melanoma classifier encodes demographic information a, Performance of training a linear classifier using a CLIP-based foundation model (MONET) embeddings for predicting sex. Each of the curves represent one of five training replicates using different random seeds. **b,** Visualization of the proportion of ROC-AUC explained across five replicates along with quantification using 95% confidence intervals.

truth. This has led to the development of numerous CLIP-like biomedical foundation models by finetuning using domain-specific datasets in the form of image caption pairs[45–50].

These foundation models can also encode biases similar to fully-supervised models and learn information about protected attributes like sex which can result in spurious correlations and be detrimental to downstream performance. To test this hypothesis, we consider one such dermatology foundation model called MONET[51], trained on dermatological images paired with natural language descriptions obtained from PubMed articles and medical textbooks. We used the trained image encoder to get the image embeddings and fit a linear classifier on top for the sex prediction task. We used the same training, validation, and testing sets as the ones used for the ViT-based sex classifier.

The classifier had high performance on the external test set (ROC-AUC of 0.798 ± 0.002; mean $\pm$ standard deviation; Supplementary Figure 4a). Using the 'removal via balancing' technique to quantify the importance of the same signals that were identified with the ViT-based classifier, we observed that the same signals were also important for this classifier. 'Hair' again accounted for the largest contribution, explaining about 55% of the classifiers performance, which was significantly higher than the performance explained of the ViT-based classifier. 'Age' and 'Redness' accounted for a smaller but non-negligible proportion of the performance while 'Ruler' was not important for the classifiers prediction, contrary to what was observed with the ViT-based classifier. Overall, we were able to explain about 65% of the classifier's performance, which is significantly higher than what we observed with the ViT-base classifier (Supplementary Figure 4b). This experiment shows that even foundation models do encode demographic information and their learning mechanisms seem to be similar to those of fully-supervised classifiers, with similar signals being important for prediction albeit with varying quantification.