Discovering mechanisms underlying medical AI prediction of protected attributes

Soham Gadgil[1,*], Alex J. DeGrave[1,2,*], Roxana Daneshjou[3,4], and Su-In Lee[1]

[1]Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, WA, USA
[2]Medical Scientist Training Program, University of Washington, Seattle, UW, USA
[3]Department of Dermatology, Stanford School of Medicine, Stanford, CA, USA
[4]Department of Biomedical Data Science, Stanford School of Medicine, Stanford, CA, USA
[*]Equal Contribution

Abstract

Recent advances in Artificial Intelligence (AI) have started disrupting the healthcare industry, especially medical imaging, and AI devices are increasingly being deployed into clinical practice. Such classifiers have previously demonstrated the ability to discern a range of protected demographic attributes (like race, age, sex) from medical images with unexpectedly high performance, a sensitive task which is difficult even for trained physicians. Focusing on the task of predicting sex from dermoscopic images of skin lesions, we are successfully able to train high-performing classifiers achieving a ROC-AUC score of ∼0.78. We highlight how incorrect use of these demographic shortcuts can have a detrimental effect on the performance of a clinically relevant downstream task like disease diagnosis under a domain shift. Further, we employ various explainable AI (XAI) techniques to identify specific signals which can be leveraged to predict sex. Finally, we introduce a technique to quantify how much a signal contributes to the classification performance. Using this technique and the signals identified, we are able to explain ∼44% of the total performance. This analysis not only underscores the importance of cautious AI application in healthcare but also opens avenues for improving the transparency and reliability of AI-driven diagnostic tools.

1. Introduction

Machine learning (ML) technologies have been widely applied to image classification problems in medicine, and this has led to the approval of hundreds of ML-based devices by the United States Food and Drug Administration (FDA) [1, 4, 34]. To generate their predictions, these classifiers rely on several image attributes, including medically important details also relied on by physicians, and image-acquisition artifacts or other features that are medically irrelevant and likely undesirable [12, 13]. Considered in combination with prior studies that identified disparate performance across protected demographic groups (for instance, based on race and/or sex) [6], including in the medical context [11], this raises the spectre that ML-based medical image classifiers may inappropriately leverage protected demographic information present in the data to generate predictions. In principle, inappropriate use of such attributes could lead to undesirable behavior, such as fragile performance or discrimination due to domain shift in a downstream, clinically relevant task. If these demographic variables correlate with a diagnosis or prediction target in the classifier's training data—either due to societal inequities or random chance—an ML system that uses all available information is likely to incorporate these variables in its prediction.

In further support of this possibility, ML classifiers have unexpectedly displayed an ability to predict a range of demographic variables directly from medical images [17, 28], that is, even when the demographic information is not supplied as metadata. Given that physicians in the relevant imaging field are unable to explain how ML classifiers predict certain demographic features from medical images, this raises another question: what image attributes do ML classifiers use to detect protected demographic attributes? The unexplained performance has led to the speculation of the existence of unique image attributes that are detectable only by machines, a hypothetical category of image attributes we refer to as 'AI-specific signals.'

Investigation of these AI-specific signals thus lies at the intersection of two fundamental questions in medical ML:

(i) What is the 'reasoning process' of medical-image classifiers? That is, on what attributes do they rely to generate predictions and to what extent? (ii) How might medical-image classifiers produce disparate outcomes among protected classes, particular in the presence of domain shifts? By interrogating the existence and nature of AI-specific signals that may be used to predict a protected demographic variable, we address a potential gap in our understanding of how medical-image classifiers generate predictions, which currently includes only image attributes readily recognized by humans [12]. Simultaneously, we detail a mechanism by which classifiers may display undesirable characteristics across protected classes.

In this study, we answer both the questions above by examining medical-image classifiers trained to predict patients' sex from dermoscopic images, which offer a magnified view of a patient's skin lesions. To do so, we apply a range of methodologies from the field of explainable AI, namely clustering analysis and counterfactual image generation. We further introduce a technique which we term 'removal via balancing' to quantify how much of the classifier's performance is explained by the signals identified.

2. Related Work

There has been growing academic and clinical interest in the prediction of protected attributes by training classifiers for different clinical modalities. Examples include the prediction of a patient's sex from retinal fundus images [28] and the prediction of a patient's race from different forms of radiological imaging [17]. A motivation for this type of work is provided by [37], showcasing the extent to which medical AI leverages demographic encodings in datasets for clinically relevant disease classification by focusing on potential fairness discrepancies within both in-distribution training sets and external test sets. Another study also confirms downstream performance disparities across race and sex in deep learning models using publicly available chest X-ray datasets [18]. However, none of these works probe into the specific signals that these classifiers leverage to achieve unexpectedly high performance in predicting these protected attributes.

There have been previous works attempting to identify specific signals, but have only pointed toward factors like statistical correlations between these protected demographic variables and diagnoses, or other variables that are more visible in the images (*e.g.*, age). [17, 36], but they haven't directly related it to the prediction task or quantified the contribution of these signals to the prediction performance.

3. Methods

3.1. 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 [8, 10, 31, 33]. The ISIC archive comprises primarily dermoscopic images collected by medical professionals across the world, 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, which is intended for crediting images under Creative Commons Attribution licenses but often also provides information on the image acquisition site (Table 1). This partitioning scheme minimizes chance of overlap in patients between the training and test data and additionally provides a more robust test scenario for domain shift analysis, since spurious associations present in the 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
Anonymous	Memorial Sloan Kettering Cancer Center
Department of Dermatology, Hospital Clinic de Barcelona	Syndney Melanoma Diagnostic Center at Royal Prince Alfred Hospital
Department of Dermatology, Medical University of Vienna	Syndney 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
Dermoscopedia	
For educational purpose only	
Hospital Clinic de Barcelona	
Hospital Italiano de Beunos Aires	
Konstantinos Liopyris	

Table 1. Hospital sites used in the train and test sets.

3.2. Model training

We trained image transformers (ViT-Base architectures) [15] to predict a patient's sex based on a dermoscopic image. To train these classifiers, we started with classifiers pre-trained on ImageNet [25], 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 (that is, $lr_{new} = 0.2 \times lr_{old}$) if the model's loss did not improve for 5 epochs. Finally, we used the epoch with the highest ROC-AUC on the validation data for all subsequent experiments. We repeated this procedure for 5 replicates.

3.3. Clustering-based analysis

To identify the signals that the classifier relies on to predict sex, we started by grouping together images in the test dataset based on their visual similarity by using the K-means clustering algorithm [27] (K=20). To obtain the clustering features, we used the first 50 principal components derived by running Principal Component Analysis (PCA) on the embeddings of the penultimate layer of the Resnet50 model [20] (pretrained on ImageNet [14]).

Once we had the clusters, we calculated the ratio $\frac{\#(\text{predicted males})}{\#(\text{predicted females})}$ for each cluster and retrieved the two with the maximum and minimum ratios corresponding to the male-dominant and female-dominant clusters respectively. Then, we analyzed a subset of 100 images selected uniformly at random from each of the two clusters to identify the signals which differed. Visual clustering of the images before comparison allows us to hypothesize that the signals identified would be significant to the model prediction and are not just occurring by chance. In addition to visual inspection, we also used a Large Language Model (LLM) identify these signals. LLMs like ChatGPT have shown great success in modeling and generating natural language [35]. Specifically, GPT-4 with vision allows the model to take in images and answer questions about them. We leveraged the `gpt-4-vision-preview` model to automatically identify visual signals differing between the image clusters. To achieve this, we passed in both the image clusters at the same time to the model along with the prompt: *"Tell me the differences between the two image clusters in terms of visual artifacts"*.

3.4. Generation of counterfactuals

Another XAI tool we used to understand the signals that can potentially guide the prediction of AI classifiers is called counterfactual image generation [7, 32, 38]. 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 from the classifier. For instance, consider the case that an AI classifier predicts a lesion as female, while a counterfactual predicted by the AI classifier to be male differs in some visual signals; 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.

To generate counterfactual images, we followed prior work [3, 9, 22] in which an image is optimized to elicit a desired output from a classifier. At a high level, we optimize an image using gradient descent, and in order 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 optimizing the image directly in pixel-space (Figure 1). This ensures that the counterfactual images remain within the data manifold of plausible images [9]. Since we are interested in broadly understanding the predictions of classifier rather than explaining the predictions for a specific output, we utilized randomly generated images as our references, eliminating the need for an encoder network as was used in prior efforts [22].

We generated a pair of counterfactual 'female' and 'male' 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 $C : \mathcal{R}^{224\times224} \rightarrow [0, 1]$ that quantifies the probability of the image representing a female patient, we performed gradient descent on z to optimize $C(G(z))$. Based on any given z, we generated a female counterfactual by minimizing $-C(G(z))$ until $C(G(z')) > 0.95$ (where z' represents the updated latent vector) and a male counterfactual by minimizing $C(G(z))$ until $C(G(z')) < 0.05$. Since the optimization is deterministic and the optimization problem is not convex, a portion ($\sim$60%) of initial vectors z failed to produce either a female or male counterfactual by the above criteria; we stopped optimization after a maximum of 10 steps and excluded these from further analysis. During optimization, we used a learning rate of 0.02.

As the generator G, we chose a styleGAN2 (a generative adversarial network, GAN) with adaptive discriminator augmentation [23]. We trained the GAN using 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 learn-

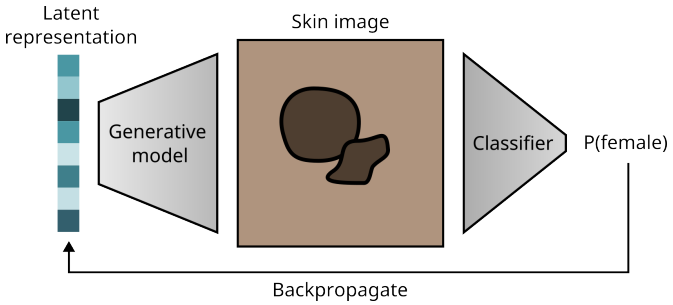


Figure 1. Latent space optimization for generating counterfactuals using a StyleGAN.

ing rate of 0.0025, $\beta_1 = 0$, $\beta_2 = 0.99$, and batch size of 64. For adaptative discriminator augmentation, we set the target to 0.6 [24]. We performed optimization for a total of 25000 kilo-images, requiring approximately four days of training on four NVIDIA RTX 2080TI GPUs.

3.5. Quantification of explained performance with 'Removal via balancing'

Once we identified the signals that can potentially exist, we quantified the amount of the ML classifier's performance that is explained by a putatively important signal (or signals), which we term the 'query signal', 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 updated the test data to form a new pseudopopulation 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 may be represented by a scalar or vector in $\mathbb{R}^n$ (where $n \in \mathbb{N}$), for instance as a scalar quantification of that signal or a one-hot 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 [2, 5, 29, 30] 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 [2] 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; for instance, 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 drop in performance after balancing to the maximum possible drop in performance (to random performance of ROC-AUC=0.5):

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

4. Results

4.1. ML classifiers successfully predict patient sex from dermoscopic images

To investigate how ML classifiers predict protected attributes from medical images, we trained image transformers as described in section 3.2 for the specific task of identifying a patient's sex based on a dermoscopic image of a skin lesion (Figure 2a). Dermoscopic images offer a magnified view of the skin and typically lack anatomical landmarks (eyes, nose, fingers, *etc.*), which could offer a route to the identification of a patient's sex. We trained our classifiers on images from the International Skin Imaging Collaboration (ISIC) archive. All images were resized to 224×224 pixels. To mitigate the possibility that the networks rely on source-specific confounding rather than signals that generalize across data sources, we partitioned the ISIC archive based on the image collection site, using a disjoint group of collection sites for training and testing. In this external test scenario, our classifiers could predict patient sex with substantial performance (area under the receiver operating characteristic curve, ROC-AUC, of 0.782 ± 0.012, mean $\pm$ standard deviation; Figure 2b).

4.2. Prediction of protected features enables undesirable outcomes

While we anticipate that medical-image classifiers would not be typically used for the prediction of protected attributes, we hypothesized that their ability to predict these attributes may lead to undesirable behavior in medically related prediction tasks. If true, this hypothesis motivates the need for a better understanding of the mechanisms that underlie the classifier's prediction of protected attributes. 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' [16] for identification of a disease, due to an association between the attribute and disease in the training data. While some associations may reflect genuine medical differences (for instance, rates of breast cancer among females and males) [26], they may also reflect soci-

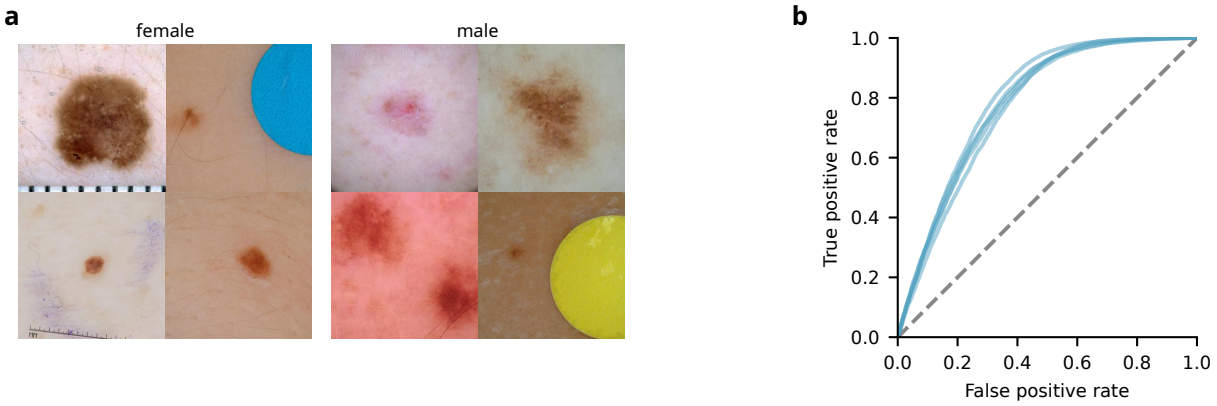


Figure 2. **AI prediction of patient sex from dermoscopic images. a,** A randomly selected set of dermoscopic images from female and male patients from the ISIC archive. **b,** Performance of trained ML classifiers at prediction of sex from dermoscopic images. Receiver operating characteristic curves are generated based on external data (image acquisition sites held out from the training data) and represent each of five training replicates, trained on the same data but with different random seeds. In this figure and throughout our study, the positive label indicates 'female'.

etal 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 exhibit a drop in performance.

To test our hypothesis, we focused on the task of differentiating melanoma from look-alike lesions (benign nevi, seborrheic keratoses, solar lentigo, lentigo NOS, pigmented benign keratosis, dermatofibroma, lentigo simplex, cafe-au-lait macule) and engineered datasets to exhibit an association between a patient's sex and melanoma (Figure 3a). We carried out our tests for a variety of odds ratios, in each case setting the odds ratio in the external test data to the inverse of the odds ratio in the training data.

We observed that while performance on internal validation data remained high across scenarios (indeed, even improving with stronger associations between sex and melanoma), external test set performance dropped as the odds ratios varied from unity (Figure 3b). Performance dropped precipitously for extreme changes in odds ratios (in particular when the protected attribute correlated perfectly with the prediction target). Performance changed more modestly for moderate differences in odds ratio (e.g., a drop of 6% in external test set ROC-AUC from an odds ratio of 1 to an odds ratio of 0.5/2 in the train/test data, respectively). We observed that among the data sources that comprise the ISIC archive, the odds ratio for 'female' as a predictor for 'melanoma' varied 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.

4.3. Statistical associations with patient sex

As a first step toward understanding how the classifiers may identify sex from dermoscopic images, we examined statistical associations, using the odds ratio, with the available metadata characteristics, like diagnosis, dermoscopy method, and age, finding a few characteristics that could potentially associate with sex (Appendix A).

Multiple diagnoses weakly associated with sex, and a few of these associations persisted from the training data, where an association must be present for the classifier to learn it, and the testing data, where the association must persist to benefit performance (Supplementary Table 3). These included 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 [19], 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 (Supplementary 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 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 4). 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.

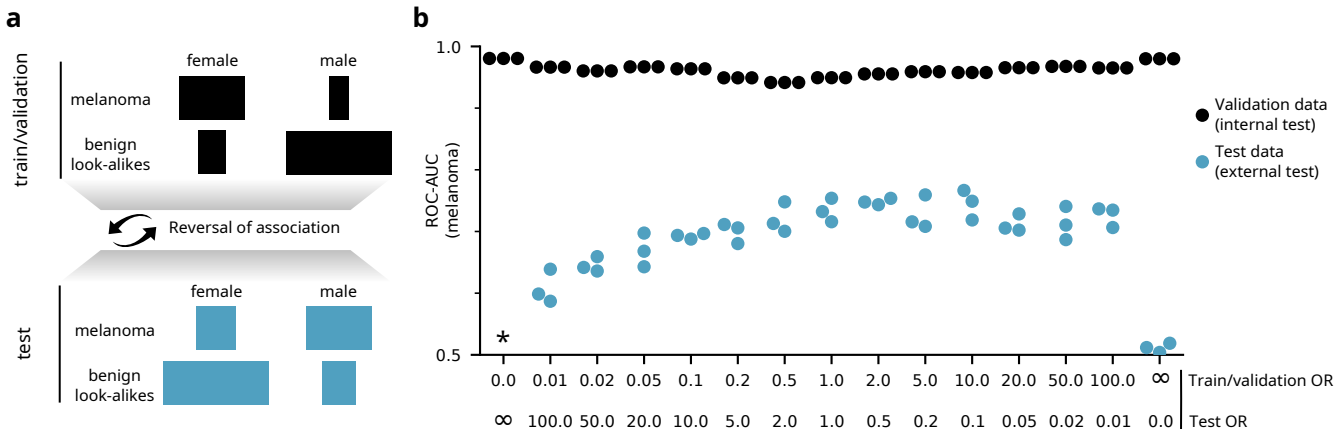


Figure 3. **Spurious correlation with a protected attribute may result in performance deficit. a,** Overview of the experimental setup. We engineer datasets (by subsampling the original data) such that the prediction target (melanoma) correlates with a protected attribute (patient sex). The correlation is inverted in the test data. **b,** Performance of the trained classifiers on the melanoma-prediction task using 3 replicates. Odds ratios measure the association between the demographic 'female' and the prediction target 'melanoma', that is, an odds ratio greater than one indicates that a higher proportion of female than male images depict melanoma. * indicates ROC-AUC<0.5 (that is, worse than random performance). OR, odds ratio.

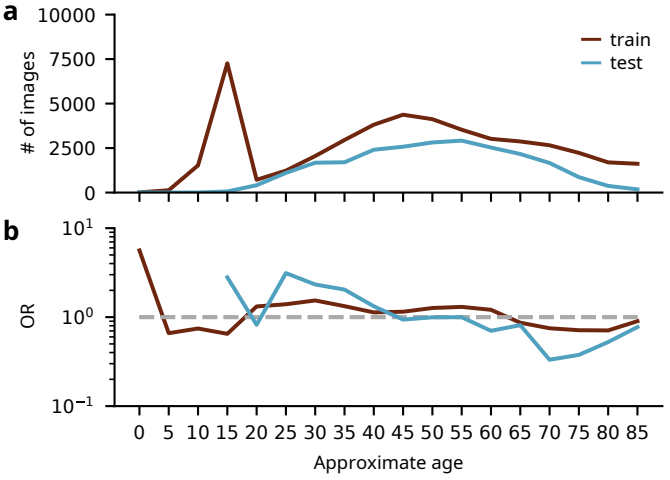


Figure 4. **Association of patients' sex with their age. a,** Histogram of patients' ages in the training and test data. **b,** Odds ratio (OR) for prediction of female sex on the basis of a patient's age.

4.4. Clustering-based analysis helps identify global artifacts

Using the clustering analysis described in section 3.3, we examined visually similar clusters of images in the training data which differed the most in terms of ratio of males to females predicted by the trained classifier. Hair differed strikingly between the two clusters, being highly prevalent in the cluster with more predicted males. This was also confirmed by querying GPT4-Vision (Figure 5a).

Since hair was a strong signal, to identify other signals, we equalized the training set by sub-sampling images so that there was no statistical correlation between hair and sex. To label the images (which lack detailed annotation) for presence of hair, 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 train-set split). We then used that classifier to label the rest of the dataset. After sub-sampling, the new training set had 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 and stickers were identified as being more prevalent in the cluster with the most number of predicted males, indicating that stickers could be a potential signal associated with males by the sex classifier (Figure 5b). This signal was also recovered by GPT4-Vision as before. Stickers are small adhesive markers that may be placed on the skin to indicate the location of lesions or areas of interest. This helps in guiding biopsies, surgical excisions, or other treatments. The exact type and color of the stickers used can vary across hospital sites, making it an easy 'shortcut'[16] to learn for predicting protected attributes (like sex) that may vary between sites.

4.5. Generative image AI reveals local prediction mechanism

Our generative technique (section 3.4) was able to produce realistic counterfactual images of dermoscopic lesions; the distribution of images produced by the final network dif-

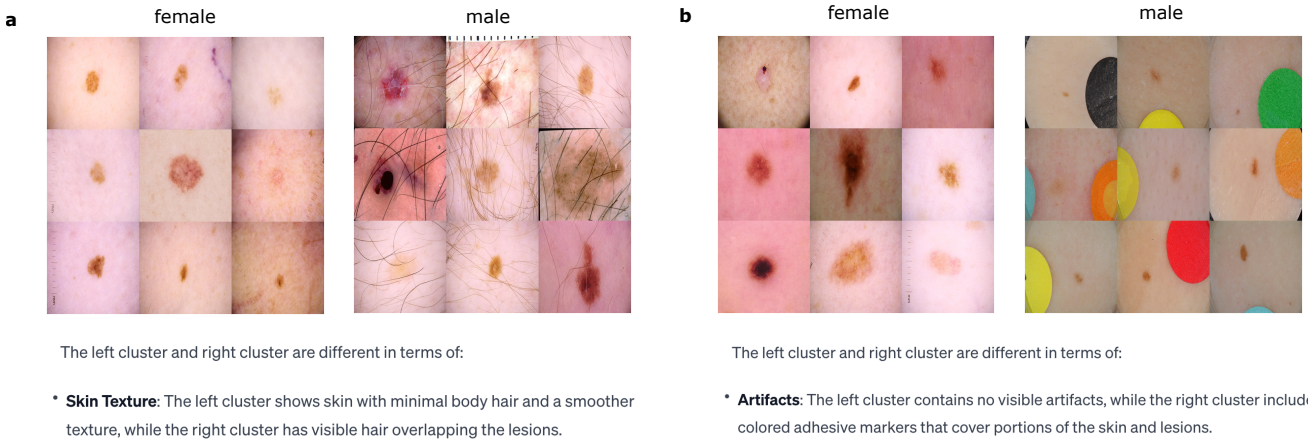
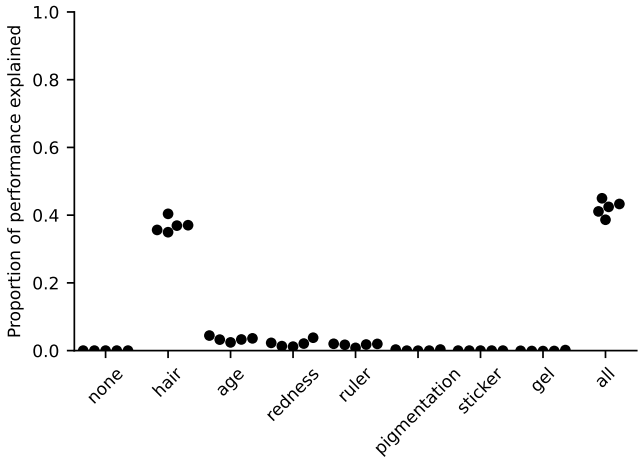


Figure 5. **Clustering analysis.** A sample of images from the visually similar clusters with the lowest and highest predicted male to female ratios along with the signals identified by GPT4-vision for **a,** The unequalized sex classifier, in which hair is identified as a potential signal. **b,** The sex classifier equalized for hair, in which sticker is identified as a potential signal.



Signal Identified	Quantification
hair	36.983 ± 1.64
age	3.405 ± 0.57
redness	2.133 ± 0.82
ruler	1.663 ± 0.39
All	42.093 ± 1.87

Figure 6. Quantification of the proportion of performance explained by signals identified in our prior experiments. **Top:** Visualization of the proportion of ROC-AUC explained across five replicates. **Bottom:** Quantification with 95% confidence intervals.

fered from the distribution of training images by a Fréchet Inception Distance (FID score) [21] of 6.29. Some samples of the generated counterfactuals are shown in Appendix B. We generated 1000 pairs of counterfactuals which elicited a desired sex prediction from the classifier, that is, a pair classified as 'female' and 'male'. The image pairs were then manually analysed to identify signals in addition to hair and sticker that were prevalent in either of the sexes and could potentially be used by the classifier for predictions.

Table 2 lists the signals identified along with the prevalence in either the male or female sex. These signals were identified in multiple pairs of images in the same direction (always in male or always in female), indicating a possible correlation with sex. We also identified differences based on other signals but they were only observable in one or two image pairs, so we discarded them as being insignificant for the sex prediction task. The signals identified were also confirmed by a board-certified dermatologist. A qualitative visualization of the different signals is provided in Figure 7.

4.6. 'Removal-via balancing' quantifies classifier performance

After applying a range of techniques like statistical association, clustering analyses, and generative modeling to identify putatively important signals, we confirmed and quantified their importance using 'removal via balancing' (section 3.5). This technique quantifies the importance of a putatively important 'query signal' for a classifier's predictions in a particular test set. To do so, 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 expect that classifier's performance to drop when there is no difference in the query signal between target classes (female and male), and the degree by which the performance drops quantifies the importance of that signal. The 'removal via balancing' technique required a numerical representation for each signal in each

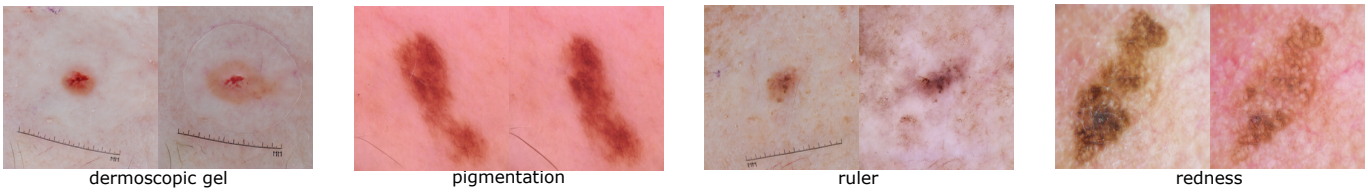


Figure 7. Visualizing the signals obtained from generating counterfactual pairs using the generative model. For each pair, the left lesion is the female counterfactural and the right one is the male counterfactual.

Signal Identified	Description	Association
redness	The skin lesion or the background shows signs of reddish, inflamed skin, or erythema	Male
ruler	The dermoscopic image has visible ruler markings which serve as a scale reference	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.	Male
gel	Visible dermoscopic gel, also known as immersion fluid, in the image used to enhance visualization of the skin lesion	Male

Table 2. AI-specific signals identified through analysis of 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.

image, and since these annotations were absent from the original data, we produced annotations using a hybrid approach, similar approach to the one used for equalizing hair in the clustering analysis; we manually labeled 500 female and 500 male images from the test set for each of the signals identified. We then trained separate classifiers (ViT-Base architecture, ROC-AUC: 0.95) for each signal to label the rest of the test set.

Quantifying the importance 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, with the largest single contributor being hair, which alone explained about 37% of the classifier's performance. Other attributes such as age, redness, and presence of rulers explained a smaller but non-negligible proportion of the classifier's performance (Figure 6). A few 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, there are only two images with stickers in the test set, indicating that the classifier cannot rely on this signal for prediction in the test set even though it might have learned this signal during training. Some of the signals identified, like hair, can be explained by the physiological differences based on sex, since males typically grow more hair on their bodies than females. However, the other signals identified, like age, redness, and ruler, do not conform with known biological insights and can be specific to the training data.

5. Conclusion

This study illuminates the intricate dynamics of AI classifiers in predicting protected attributes within medical imaging, focusing on patient sex prediction from dermoscopic images. Through a domain shift analysis, we motivated the need for rigorous data curation to ensure that trained medical AI classifiers do not encode demographic shortcuts leading to degraded performance or discrimination in downstream, clinically relevant tasks. By employing a suite of explainable AI (XAI) techniques, we revealed specific visual signals that the AI could potentially leverage for prediction. We further introduced a technique to quantify the signals in terms of their contribution to the AI classifier's performance. Remarkably, it was revealed that the hair signal accounted for a substantial portion ($\sim$ 37%) of the prediction performance, challenging conventional understandings of AI's diagnostic capabilities. We hope this study highlights the critical importance of scrutinizing AI's decision-making processes in healthcare, paving the way for future research aimed at enhancing the transparency and accountability of AI systems in clinical settings.

6. Acknowledgements

We thank Chanwoo Kim and other members of the Lee Lab for their helpful discussions. We also thank Jesutofunmi Omiye for helping with manuscript writing.

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. 1

[2] Peter C. Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46:399–424, 2011. 4

[3] Rachana Balasubramanian, Sam Sharpe, Brian Barr, and C. Bayan Bruss. Latent-cf: a simple baseline for reverse counterfactual explanations. *preprint at arXiv:2012.09301v2*, 2021. 3

[4] Stan Benjamens, Pranavsingh Dhunnoo, and Bertalan Meskó. The state of artificial intelligence-based fda-approved medical devices and algorithms: an online database. *npj Digital Medicine*, 3(118), 2020. 1

[5] M. Alan Brookhart, Richard Wyss, J. Bradley Layton, and Til Stümer. Propensity score methods for confounding control in non-experimental research. *Circulation: Cardiovascular Quality and Outcomes*, 6:604–611, 2013. 4

[6] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of Machine Learning Research Conference on Fairness, Accountability and Transparency*, pages 1–15, 2018. 1

[7] Chun-Hao Chang, Elliot Creager, Anna Goldenberg, and David Duvenaud. Explaining image classifiers by counterfactual generation. *arXiv preprint arXiv:1807.08024*, 2018. 3

[8] Noel C. F. Codella, David Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, and Allan Halpern. 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. 2

[9] Joseph Paul Cohen, Louis Blankemeier, and Akshay Chaudhari. Identifying spurious correlations using counterfactual alignment. *preprint at arXiv:2312.02186v1*, 2023. 3

[10] Marc Combalia, Noel C. F. Codella, Veronica Rotemberg, Brian Helba, Veronica Vilaplana, Ofer Reiter, Allan C. Halpern, Susana Puig, and Josep Malvehy. Bcn20000: Dermoscopic lesions in the wild. *arXiv:1908.02288*, 2019. 2

[11] Roxana Daneshjou, Kailas Vodrahalli, Roberto A. Novoa, Melissa Jenkins, Weixin Liang, Veronica Rotemberg, Justin Ko, Susan M. Swetter, Elizabeth E. Bailey, Olivier Gevaert, Pritam Mukherjee, Michelle Phung, Kiana Yekrang, Bradley Fong, Rachna Sahasrabudhe, Johan A. C. Allerup, Utako Okata-Karigane, James Zou, and Albert S. Chiou. Disparities in dermatology ai performance on a diverse, curated clinical image set. *Science Advances*, 8(32):eabq6147, 2022. 1

[12] Alex DeGrave, Zhuo Ran Cai, Joseph D. Janizek, Roxana Daneshjou, and Su-In Lee. 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. 1, 2

[13] Alex J. DeGrave, Joseph D. Janizek, and Su-In Lee. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3:610–619, 2021. 1

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: transformers for image recognition at scale. *International Conference on Learning Representations (ICLR) 2021*, 2021. 3

[16] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *arXiv:2004.07780*, 2020. 4, 6

[17] Judy Wawira Gichoya, Imon Banerjee, Ananth Reddy Bhimireddy, John L Burns, Leo Anthony Celi, Li-Ching Chen, Ramon Correa, Natalie Dullerud, Marzyeh Ghassemi, Shih-Cheng Huang, Po-Chih Kuo, Matthew P Lungren, Lyle J Palmer, Brandon J Price, Saptarshi Purkayastha, Ayis T Pyrros, Lauren Oakden-Rayner, Chima Okechukwu, Lalah Seyyed-Kalantari, Hari Trivedi, Ryan Wang, Zachary Zaiman, and Haoran Zhang. Ai recognition of patient race in medical imaging: a modelling study. *The Lancet Digital Health*, 4:E406–E414, 2022. 1, 2

[18] Ben Glocker, Charles Jones, Mélanie Bernhardt, and Stefan Winzeck. Algorithmic encoding of protected characteristics in chest x-ray disease detection models. *Ebiomedicine*, 89, 2023. 2

[19] Quishen Ha, Bo Liu, and Fuxu Liu. Identifying melanoma images using efficientnet ensemble: winning solution to the siim-isic melanoma classification challenge. *Preprint at arXiv:2010.05351*, 2020. 5

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770 – 778, 2016. 3

[21] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 7

[22] Shalmali Joshi, Oluwasanmi Koyejo, Been Kim, and Joydeep Ghosh. xgems: generating examplars to explain black-box models. *preprint at arXiv:1805.08867v1*, 2018. 3

[23] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial entworks with limited data. *preprint at arXiv:2006.06676*, 2020. 3

[24] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *2020 IEEE/CVF Confer-*

ence on Computer Vision and Pattern Recognition (CVPR), pages 8107–8116, 2020. 4

[25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *2012 Conference on Neural Information Processing Systems*, 2012. 3

[26] Diana Ly, David Forman, Jacques Ferlay, Louise A. Brinton, and Michael B. Cook. An international comparison of male and female breast cancer incidence rates. *International Journal of Cancer*, 132:1918–1926, 2013. 4

[27] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, pages 281–297. Oakland, CA, USA, 1967. 3

[28] Ryan Poplin, Avinash V. Varadarajan, Katy Blumer, Yun Liu, Michael V. McConnell, Greg S. Corrado, Lily Peng, and Dale R. Webster. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering*, 2:158–164, 2018. 1, 2

[29] Paul R. Rosenbaum. Model-based direct adjustment. *Journal of the American Statistical Association*, 82:387–394, 1987. 4

[30] Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 1983. 4

[31] Veronica Rotemberg, Nicholas Kurtansky, Brigid Betz-Stablein, Liam Caffery, Emmanouil Chousakos, Noel Codella, Marc Combalia, Stephen Dusza, Pascale Guitera, David Gutman, Allan Halpern, Brian Helba, Harald Kittler, Kivanc Kose, Steve Langer, Konstantinos Lioprys, Josep Malvehy, Shenara Musthaq, Jabpani Nanda, Ofer Reiter, George Shih, Alexander Stratigos, Philipp Tschandl, Jochen Weber, and H. Peter Soyer. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific Data*, 8:34, 2021. 2

[32] Axel Sauer and Andreas Geiger. Counterfactual generative networks. *arXiv preprint arXiv:2101.06046*, 2021. 3

[33] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5(180161), 2018. 2

[34] Eric Wu, Kevin Wu, Roxana Daneshjou, David Ouyang, Daniel E. Ho, and James Zou. How medical ai devices are evaluated: limitations and recommendations from an analysis of fda approvals. *Nature Medicine*, 27:582–584, 2021. 1

[35] Tianyu Wu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, Qing-Long Han, and Yang Tang. A brief overview of chatgpt: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, 10(5):1122–1136, 2023. 3

[36] Takehiro Yamashita, Ryo Asaoka, Hiroto Terasaki, Hiroshi Murata, Minoru Tanaka, Kumiko Nakao, and Taiji Sakamoto. Factors in color fundus photographs that can be used by humans to determine sex of individuals. *Transitional Vision Science and Technology*, 9, 2020. 2

[37] Yuzhe Yang, Haoran Zhang, Judy W Gichoya, Dina Katabi, and Marzyeh Ghassemi. The limits of fair medical imaging ai in the wild. *arXiv preprint arXiv:2312.10083*, 2023. 2

[38] Yingchen Yu, Fangneng Zhan, Rongliang Wu, Jiahui Zhang, Shijian Lu, Miaomiao Cui, Xuansong Xie, Xian-Sheng Hua, and Chunyan Miao. Towards counterfactual image manipulation via clip. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3637–3645, 2022. 3

Discovering mechanisms underlying medical AI prediction of protected attributes

Supplementary Material

A. Association between metadata and sex

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

Table 3. Odds ratios (OR) for prediction of female sex based on diagnosis. Diagnoses are sorted by the OR in the training data, with diagnoses 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, 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

Table 4. 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, number of images.

B. Counterfactuals generated by generative technique

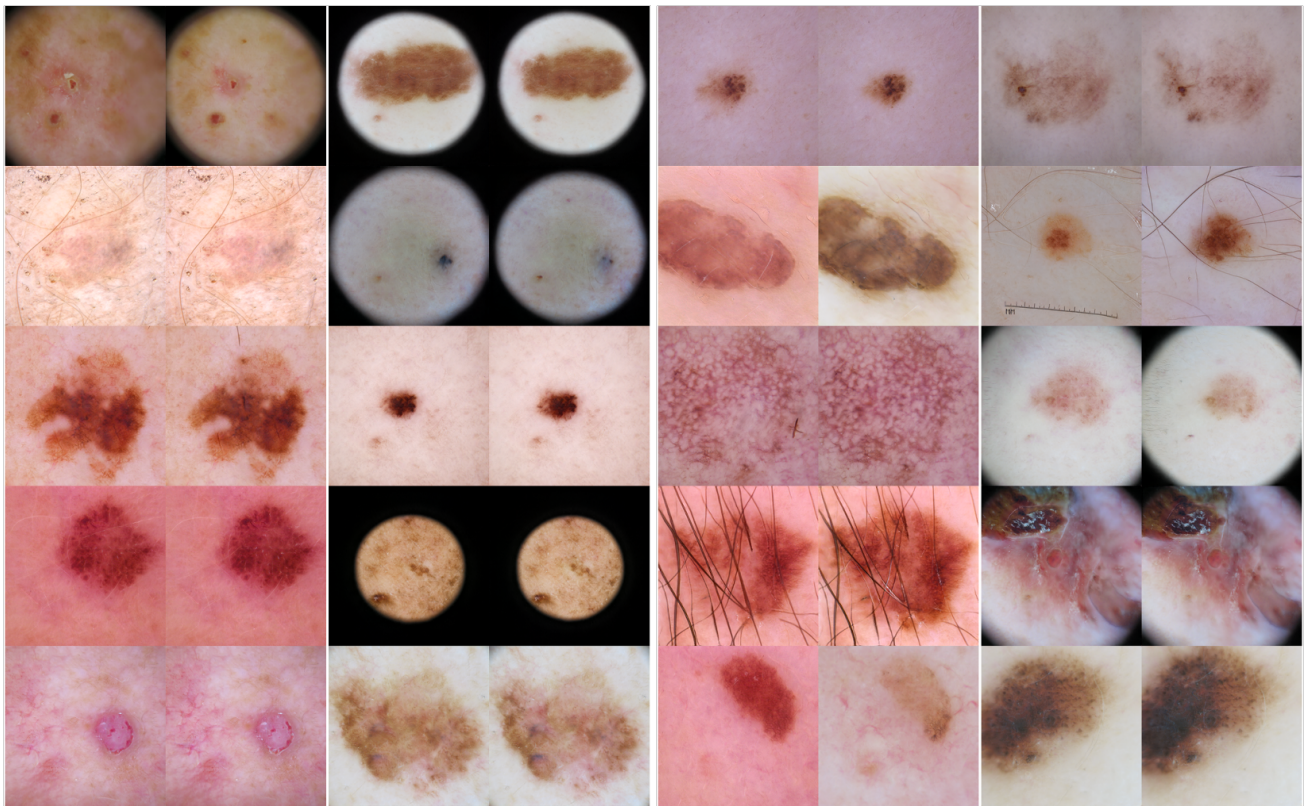


Figure 8. Visualization of the counterfactuals generated by our latent space optimization techinique. In each pair of images, the left one is the female counterfactual and the right one is the male counterfactual.