Auditing the Reasoning Processes of Medical-Image AI

Alex DeGrave

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington
2024

Reading Committee:
Su-In Lee, Chair
Marshall Horwitz
Roxana Daneshjou

Program Authorized to Offer Degree:
Paul G. Allen School of Computer Science & Engineering

University of Washington

Abstract

Auditing the Reasoning Processes of Medical-Image AI

Alex DeGrave

Chair of the Supervisory Committee:
Su-In Lee
Paul G. Allen School of Computer Science & Engineering

While medical artificial intelligence (AI) systems are achieving regulatory approval and clinical deployment across the world, the reasoning processes of these systems remain opaque to all stakeholders, including physicians, patients, regulators, and even the developers of these systems. Since the modern wave of medical AI relies on automatic learning of statistical patterns from large datasets—via 'machine–learning' techniques such as neural networks—they are prone to learning unexpected and potentially undesirable patterns, which may lead to pathological behavior in deployment. Here, we investigate the 'reasoning processes' of medical-image AI systems, that is, by forming a human-understandable, medically grounded conception of that mechanisms by which they generate predictions. Along the way, we develop new tools and frameworks as necessary to do so. Via these investigations, we uncover severe flaws in the reasoning of medical AI systems, and we build the first thorough, medically grounded picture of machine-learning–based medical-image AI reasoning processes.

To my family, including:

My partner Meghan, who has provided tremendous love and support throughout my education;

My parents Pam and John, who instilled in me the sense of curiosity that blossomed into a passion for research;

My sister Kara, who has always encouraged me and helped me believe I could succeed in research and medicine.

Contents

Chapter 1

Introduction

Artificial intelligence (AI) has gained an increasing hold in clinical medicine, owing largely to the performance of machine learning (ML) techniques such as neural networks. These ML systems contrast with medicine's earlier forays with AI, based on manually programmed 'expert systems', in that ML automatically learns patterns from large datasets, without humans programming those patterns or even necessarily understanding them. Hundreds of such ML-based AI systems have been approved by the United States Food and Drug Administration (FDA)[1] or have gained CE approval, countless more medical AI systems are under development by major players in industry and academia, and landmark funding of AI systems by the Centers for Medicare and Medicaid services[2] appears likely to attract additional investment. In principle, medical AI systems could offer numerous benefits to patients and their providers, including improved diagnostic accuracy or reduced cost, and a few systems have indeed shown evidence for meaningful benefit in prospective clinical trials.[3–6]

However, since ML-based AI systems automatically learn patterns from data, they do not necessarily follow the same reasoning processes as human experts. This creates risks for unexpected AI behavior. At its worst, such unexpected behavior could cause misdiagnoses or other patient harm, along with increased costs to the medical system. In principle, unexpected AI behavior might also offer benefits, for instance by discovering new ways to detect a disease. A medically-informed understanding of the mechanisms by which medical AI systems generate predictions—that is their 'reasoning processes'—could help guide (i) regulators in evaluating the adequacy of approval applications, (ii) patients and providers in choosing whether to trust an AI system, and (iii) developers in designing and testing new systems. In other words, we contend that a thorough understanding of the reasoning processes of ML-based medical AI systems should help maximize the potential benefit of AI in medicine.

In conjunction with the rising popularity of machine-learning systems in medicine and elsewhere, investigators have developed 'explainable AI' (XAI) approaches to understand these systems' complex, automatically-learned relationships between inputs and outputs. These XAI approaches are numerous and span a wide breadth in their intended applications: auditing AI systems, providing additional feedback to human decision-makers as part of a human-AI team, or aiding in scientific analysis of their underlying training data. They also range widely in their intended data type, with approaches tailored toward tabular data, natural language, time-series data, images, *etc.*

While technical and theoretical developments to these XAI approaches abound, their principled application to clinical AI systems has been comparatively scarce. In particular, despite that the majority of FDA-approved AI systems focus on medical images,[1,7] application of XAI to medical-image AI systems has struggled. Multiple reasons explain this observation: First, current XAI approaches to understand image-based models frequently explain the AI's prediction for a single image at a time, which provides limited information on the AI's behavior across a dataset.[8–10] Second, the *de facto* standard XAI techniques for image-based models only highlight the regions of the image that contribute most to the AI's predictions, providing limited information on the higher-level representations learned by the ML algorithm.[8,9] Third, application of XAI in the medical domain likely demands simultaneous expertise in both artificial intelligence and medicine.[11] These barriers have conspired to prevent pragmatic understanding of the reasoning processes of medical-image AI.

1.1 Background

Broadly, current understandings of how medical-image AI systems function lack sufficient detail and medical interpretability to enable key stakeholders, such as AI system developers, regulators, clinicians, and patients, to make actionable inferences on their trustworthiness. In this section, we describe how medical-image AI decisions are currently understood, the available tools to improve this understanding, and existing attempts to apply these tools.

The 'classic' understanding of medical-image AI reasoning processes

medical-image AI has enjoyed widespread deployment prior to the current wave of ML-powered AI, and the reasoning processes underlying these *expert systems* were well-understood by their developers. AI systems such as 'computer-aided detection' (CAD) tools for screening mammography relied on manually programmed routines to extract information, which was designed by experts to align with the reasoning of radiologists or other specialists. For instance, the M1000 ImageChecker, a CAD system approved by the FDA in 1998 for screening mammography, was programmed using classical computer vision techniques to detect two patterns: clusters of at least three bright spots, suggestive of microcalcifications, and groups of line segments radiating from a common origin, suggestive of a mass.[12] Developers sometimes explicitly described the programmed 'reasoning' of their system in premarket approval applications to the FDA.[12,13] In dermatology, AI systems were explicitly programmed to mimic clinical checklists,[14] such as the ABCD-E system,[15,16] 7-point checklist[17] or Menzies method.[18] In some cases, human-defined features were extracted and then applied as inputs to ML algorithms,[14] somewhat reducing the system's interpretability, but the use of human-defined features arguably facilitated greater transparency in the system's 'reasoning' than systems to come.

Following the 2012 success of a convolutional neural network named 'AlexNet'[19] in a popular, general image recognition challenge (*ImageNet*),[20] interest grew in applying similar systems toward medical tasks.[21–23] In contrast to earlier *expert systems*, these systems relied on *machine learning*. Critically, rather than attempting to recapitulate expert reasoning, ML instead automatically learns statistical patterns from large volumes of data. These ML systems in particular relied on the *representation learning* capabilities of deep neural networks: from an input of numeric pixel values, which have little meaning on their own, the ML systems automatically learn to extract useful visual patterns, such as edges, colors, or complex shapes. Most medical-image AI systems now rely on neural networks; unless otherwise specified, the term 'medical-image AI system' will hereinafter refer to neural-network–based ML systems.

Perhaps the most common conception of how neural networks 'reason' relates closely to their mathematical form. Typical neural networks applied to vision tasks consist of a sequence of *layers*, connected by mathematical functions. The first layer receives a numeric representation of the pixels of the input image, and each subsequent layer builds on the previous layer's output to detect increasingly complex patterns: perhaps edges in the first layer, basic shapes in the subsequent layer, more complex shapes in the third, *etc.*[24–26] Ultimately, representations of complex visual patterns are used to determine an output (probability of disease). Each layer consists of multiple nodes, which can be conceptualized as detecting different patterns (diagonal edges, horizontal edges, *etc.*), and multiple works have attempted to experimentally analyze individual nodes to learn what they represent.[24–26] These works are responsible for the above view that nodes early in the network extract simple features such as edges or colors, while nodes later in the network detect complex compositions of those simple features, such as faces or lettering.[24–26] However, each node may be individually challenging to understand in human terms. Moreover, any network may consist of perhaps millions of nodes, precluding such studies from providing an in-depth, systematic understanding.

This basic conceptualization—that neural networks automatically extract features of progressively increasing complexity—forms a core part of the current understanding of ML-based medical-image AI systems, but the practical implications are limited. Especially in recent years, numerous techniques have been developed to better understand the basis for AI decisions, and these techniques have provided some more practical glimpses at medical-image AI reasoning processes, including examples of errors in reasoning.

Tools to better understand medical-image AI systems

Numerous researchers have designed tools to better understanding the basis of the predictions of complex AI systems. For AI systems that analyze images, these tools broadly encompass (i) techniques that attribute predictions to specific input pixels or regions, (ii) techniques that attribute to higher-level concepts, and (iii) 'counterfactual' generation techniques, which show how the input image could differ in order to elicit a different prediction. This subsection provides a high-level technical overview of these methodologies, while the next subsection describes implications of their real-world use.

Attribution to specific input pixels or regions, known as 'feature attribution,' has become the *de-facto* standard XAI approach applied to image-based AI systems. Investigators calculate these attributions on an image-by-image basis

(*i.e.*, as 'instance-level' explanations), and display and analyze each result as a 'saliency-map,' *i.e.*, an image that encodes the feature attributions on a color scale while maintaining the same two-dimensional spatial arrangement as the corresponding original input pixels (or image regions). The set of particular methodologies to calculate these attributions ranges widely, and can be further subdivided to include categories such as gradient-based attributions and removal-based attributions, amongst others. Gradient-based approaches capture information about the AI system's reaction to small changes in the input pixels or higher-level visual features; most simply, an investigator may calculate the gradient of the model's prediction with respect to the input features, and a number of more complex variations such as Integrated Gradients,[9] Expected Gradients,[27] SmoothGrad,[28] Grad-CAM,[8] and others have been developed. Constrasting with gradient-based explanations, removal-based explanations instead capture information about the AI system's reaction to 'removing' parts of the input, *e.g.*, by setting pixels to a constant color or blurring them. Popular removal-based approaches used with image models include occlusion[25,29] and Shapley Additive Explanations,[30] alongside other Shapley-value–based techniques.[31,32] Our work described in this document uses previously developed feature attribution techniques, focusing on their practical application.

A few XAI techniques also attempt to explain image-based AI systems by attributing to higher-level concepts, which in principle could offer advantages when individual pixels lack semantic meaning. The technique *Testing with Concept Activation Vectors* (TCAV)[33] assumes that each of a set of high-level, human-understandable concepts is well-represented by a vector in the *latent space* of a neural network classifier (*i.e.*, a layer of nodes near to the output of the network). The derivative of the AI system's output along each vector then provides information about how the AI system uses the corresponding concept. In contrast to saliency map techniques, this approach provides global information about the AI system's behavior across a dataset. The technique Concept Bottleneck Models[34] also enables understanding in terms of higher-level concepts, but whereas other XAI techniques typically enable investigation of pre-existing AI systems, this technique requires that the AI system be specifically designed as a concept bottleneck model. Similar to how a physician may first observe high-level concepts ('this lesion is *asymmetrical* and has *multiple colors of pigment*') and combine that information to infer a diagnosis ('melanoma'), concept bottleneck models similarly break a task into two parts: first, a neural network predicts the presence of a series of pre-specified, human-understandable concepts (*e.g.*, asymmetry, lesion color, *etc.*), and second, a simple model (*e.g.*, logistic regression) then generates the final prediction (*e.g.*, presence of melanoma) from those concepts. We have not observed concept bottleneck models used in practice in clinical applications, and while TCAV would in principle be applicable to the questions examined in this study, numerous concerns about its validity, including that of the key assumption described above, precluded its use in the work described in this document.

Whereas these two groups of XAI approaches ask the question 'what most affected the model's prediction?' a third group of XAI approaches for image-based AI systems instead asks a slightly different question: 'how could the input have differed to elicit a different output?' An alternate version of the input image, which answers this question, is called a 'counterfactual,' and may be generated via a variety of techniques. Most simply, counterfactuals may be generated by manual image editing,[35] but approaches based on generative models have also been developed.[36,37] For instance, the approach 'Explanation by Progressive Exaggeration'[38] trains a generative adversarial network to create counterfactuals that change the predictions of a image-based AI system of interest. CycleGANs,[39] which learn to transform between distributions of images without respect to a particular classifier, can also be viewed as generating counterfactuals. Our work described in this document employs multiple counterfactual generation techniques, and it contributes important technical improvements necessary for rigorous analysis of image-based AI systems in medicine.

Glimpses at medical-image AI reasoning enabled by XAI

Investigators have applied both these XAI techniques as well as more classical techniques such as subgroup analyses to interrogate image-based AI systems in medicine.

Some studies have uncovered pathological behavior in AI systems for medical images, including that these systems often leverage spurious correlations from the training data, in a phenomenon termed 'shortcut learning'.[40] For instance, subgroup analyses combined with saliency map techniques uncovered that a deep learning model for pneumonia detection relied on the identity of a patient's hospital, which was inferred via laterality markers. Subgroup analyses analyses likewise revealed that AI systems designed to detect a pneumothorax relied on the presence of chest drains,[41,42] which physicians place in response to having already identified the condition. In the domain of dermatology, investigators found that a CE-approved AI system for detection of skin cancer relied in part on gentian violet surgical skin markings. This discovery was made by comparing AI outputs before and after *in vivo* application of the marking.[43] Experiments with manual image editing revealed that dermatology AI systems can learn to leverage the background skin, perhaps spuriously.[44] These studies provide a basic outline of errors in reasoning displayed by medical-image AI systems.

In general, where XAI has been applied toward model auditing, unsystematic application has limited its potential.

For instance, groups who developed AI systems for detection of COVID-19 in chest radiographs reported only one to three saliency maps as part of their evaluation,[45–48] and such a cursory analysis appears unlikely to uncover errors in the AI systems' reasoning processes. Moreover, these studies lack involvement of human domain experts, likely limiting their ability to interpret the XAI outputs. Existing attempts at systematic evaluation are limited. One study systematically assessed agreement betweeen saliency maps and regions delineated as important by physicians, but did not provide a fine-grained analysis of the specific image attributes.[10] Another assessed numerous saliency maps but limited the analysis to only three coarse-grained attributes.[49]

Evidence generated by multiple research groups also points toward that medical-image AI may rely in part on peculiar, 'AI-specific' image attributes that are difficult for humans to recognize, and which may enable medical-image AI to complete tasks challenging for human physicians, such as ophthalmologists. AI can predict multiple demographic factors from retinal fundus photographs,[49] including patient sex, a challenging task for human physicians. In an attempt to decipher this AI's reasoning process, a follow-up study hand-crafted higher-level visual features and then applied a logistic regression to infer which features provided useful signal.[50] While this study identified a few factors that can assist with differentiation between sexes in fundus photographs, it remains unclear whether these same features are used by the neural-network–based systems, and these features could only explain part of the AI system's high performance. AI is also able to predict patient race from various forms of radiological imaging, at task which to our knowledge has not be performed by physicians.[51] Investigations into potential confounders were unable to uncover a clear basis for the high performance at this task.[51] Difficulty explaining the reasoning processes that enables these prediction tasks supports the notion of 'AI-specific' image attributes. The nature of these attributes remains unclear; in principle these attributes could be learnable by humans, or they could be more akin to 'adversarial' perturbations to images, which substantially affect classifiers while remaining imperceptible to humans.[52]

In summary, the reasoning processes of modern, medical-image AI systems remain poorly understood and thus scarcely actionable. In contrast to earlier AI systems, current AI relies on machine learning, which automatically learns statistical relationships that may differ from human reasoning. The mathematical form of neural networks (which account for almost all recent medical-image AI systems) provides a basic level of understanding—the AI system progressively transforms an input image to an output prediction via mathematical functions that recognize increasingly complex patterns—but is too complex to enable systematic understanding. Numerous XAI tools have been developed to analyze image-based AI systems, but these have only been applied minimally toward medical images. In a few cases, these XAI tools, coupled with subgroup analyses, have revealed troubling dependencies of AI systems on spurious correlations, but the extent of this behavior is not known. Studies have also identified peculiar AI abilities that puzzle human physicians, and these too remain poorly explained. Overall, existing efforts only provide a sparse view of the mechanisms by which medical-image AI generates predictions, which is scarcely sufficient for actionability.

In this thesis, we aim to develop an actionable understanding of medical-image AI. We analyze a range of medical-image AI systems, ranging from academic to commercial. Using powerful tools from explainable AI, we are able to paint a detailed picture of how medical-image AI functions, from a mechanistic perspective. To guide our analyses, we also review the state of medical AI, with a view toward what medical professionals need to know.

1.2 Contributions

Through this work, we offer important contributions toward the fields of medicine and artificial intelligence:

1. Revelation of a widespread issue in medical AI in which AI relies on source-specific artifacts rather than bona fide medical signals
2. Demonstration that testing on external data may be inadequate to reveal 'shortcuts', as these may persist across data sources
3. Identification that a few (but only a few) FDA-approved medical AI devices show evidence of meaningful clinical benefits
4. Identification that medical AI under development frequently disregards systemic impacts on physicians
5. Elucidation that a range of dermatology AI systems rely (in part) on detailed, at times high-specific, medical features that are also used by dermatologists
6. Elucidation that dermatology AI systems frequently also rely on likely undesirable, medically irrelevant features, including image acquisition artifacts

7. Development of improved tools for analysis of medical-image AI
8. Improvements in rigor of explainable AI analyses, including analyses of large numbers of images, involvement of experts, and randomization.

Chapter 2

AI for radiographic COVID-19 detection selects shortcuts over signal

This section is adapted from the preprint 'AI for radiographic COVID-19 detection selects shortcuts over signal' by Alex J. DeGrave*, Joseph D. Janizek*, and Su-In Lee published in medRxiv (doi: 10.1101/2020.09.13.20193565 CC-BY 4.0 license). The final form[53] was published under the same title at *Nature Machine Intelligence* (doi: 10.1038/s42256-021-00338-7).

2.1 Abstract

Artificial intelligence (AI) researchers and radiologists have recently reported AI systems that accurately detect COVID-19 in chest radiographs. However, the robustness of these systems remains unclear. Using state-of-the-art techniques in explainable AI, we demonstrate that recent deep learning systems to detect COVID-19 from chest radiographs rely on confounding factors rather than medical pathology, creating an alarming situation in which the systems appear accurate, but fail when tested in new hospitals. We observe that the approach to obtain training data for these AI systems introduces a nearly ideal scenario for AI to learn these spurious 'shortcuts.' Because this approach to data collection has also been used to obtain training data for the detection of COVID-19 in computed tomography scans and for medical imaging tasks related to other diseases, our study reveals a far-reaching problem in medical imaging AI. In addition, we show that evaluation of a model on external data is insufficient to ensure AI systems rely on medically relevant pathology, since the undesired 'shortcuts' learned by AI systems may not impair performance in new hospitals. These findings demonstrate that explainable AI should be seen as a prerequisite to clinical deployment of ML healthcare models.

2.2 Introduction

The prospect of applying artificial neural networks to the detection of COVID-19 in chest radiographs has generated interest from machine learning (ML) researchers and radiologists alike, given its potential to (i) help guide management in resource-limited settings that lack sufficient numbers of the gold-standard reverse-transcription polymerase chain reaction (RT-PCR) assay, and (ii) clarify cases of suspected false negatives from the RT-PCR assay.[54,55] While numerous recent publications and preprints report machine learning models with high performance at this task,[45–48,56,57] the trustworthiness of these models needs to be rigorously evaluated before deployment in a clinical setting.[58]

Our findings in this study support the troubling possibility that these models fail to learn the true underlying pathology reflecting the presence of COVID-19 and instead leverage spurious associations between presence or absence of COVID-19 and radiographic features that reflect variations in image acquisition, *i.e.*, 'shortcuts'.[40] While such spurious

*equal contribution

associations may arise in any dataset, we observed that many recent ML models for radiographic detection of COVID-19 were trained using data with the potential for near *worst-case* confounding: these datasets are composed of an exclusively COVID-19 negative source and a COVID-19 positive source, such that any systematic differences between the sources correlate perfectly with COVID-19 status.[45–48,56,57] Similar combinations of data sources, where the source label correlates with disease status, have also been used to train AI systems for detection of COVID-19 in computed tomography scans[59] (though the non-public nature of the data precludes experimental verification of the extent of shortcut learning in this setting) and for other medical imaging tasks,[23,60] implying that our findings have broad implications to the field of medical machine learning.

In this study, we evaluate the trustworthiness of recent deep learning models for COVID-19 detection from chest radiographs. After training deep convolutional neural networks[19,61] (Supplementary Fig. 2.7) in the manner of these previous publications,[45–48,56,57] we evaluate their performance in new hospital systems. Then, we interrogate the extent to which these models rely on confounds by identifying the most important image features using state-of-the-art explainable AI techniques, including both saliency maps and generative adversarial networks (GANs).[9,27,36,39] These inquiries reveal how seemingly high-performance AI systems may derive the majority of their performance from the exploitation of undesired shortcuts, highlighting the need to verify that AI systems rely on the desired signals. Finally, we evaluate several methods to alleviate the problem of shortcut learning in this setting, demonstrating the importance of improved data quality for the creation of robust and useful models.

2.3 Results

Overview of the experimental approach

Before examining our main results, we first outline our experimental approach (Fig. 2.1a). To begin, we reviewed the literature to examine the datasets and models used for detection of COVID-19 from chest radiographs, with attention toward studies with the potential for 'worst-case confounding.' After choosing representative networks, we build two datasets: one that reproduces the data used in previous studies, and a second that enables external validation on new hospitals. In a first experiment, we evaluate models that were trained on one dataset using test images from the other dataset, under the expectation that a model that relies on valid medical pathology–which should not change between datasets–should maintain high performance. We then probe deeper into specific shortcuts that these models leverage, using techniques from explainable AI.

In a 'model-centric' approach, which focuses on the specific portions of the radiographs that contribute most to the predictions of our models in particular, we build saliency maps using Expected Gradients.[27] In essence, this approach attributes importance to each pixel of a radiograph based on the gradients of our models, while avoiding issues associated with other gradient-based approaches.[9,27] We complement this model-centric approach with a data-centric approach, which focuses on the key aspects of the data that could be used to distinguish COVID-19 positive and COVID-19 negative cases. Specifically, we apply generative adversarial networks (CycleGANs[39]) to transform COVID-19 positive radiographs to appear COVID-19 negative and vice versa, in the sense that key image features are transformed, such that a network can no longer discriminate between the real images of a given pathology label and the transformed images from the opposite class.[36] Rather than use our classifier networks to perform this discrimination task, we instead train new discriminator networks simultaneously with generator networks that transform the images, such that this experiment focuses on key aspects of our data, rather than our classifiers in particular.

To further validate these findings, we go on to perform 'region-swapping' experiments, in which we swap out portions of radiographs that our explainable AI approaches identify as important, with the expectation that changes to truly important regions will have a large impact on our classifiers' outputs. We conclude by evaluating approaches to mitigate shortcut learning from the perspectives of both generalization performance and model explainability.

Literature review of model and dataset construction

In our investigation, we aimed to determine the extent to which shortcut learning affects AI systems for COVID-19 detection in chest radiographs, which is complicated by the diversity of these systems. We therefore trained a series of ten models with varied architectures, including state-of-the-art networks that were tailor-made for detection of COVID-19 in chest radiographs[46,56,62] and multiple 'off-the-shelf,' general-purpose architectures.[19,61,63,64] For our primary models, we chose a network based on the DenseNet-121 architecture,[61] which we judged faithfully replicated the modeling choices of recent high-performance models for COVID-19 classification, while also following established best practices for classification of pathologies from chest radiographs using deep learning. Alongside these primary models, we also investigate multiple secondary models, to help probe the generality of our findings and the extent to

which they apply to AI systems found in the wild. These secondary models include: the COVID-Net network, which was custom designed for detection of COVID-19 via a machine-based architecture search:[56] the DarkCovidNet model, which was modified from a standard Darknet-19 model for the purpose of COVID-19 detection;[46] and the CV19-Net model,[62] which was built by ensembling twenty DenseNet-121 networks and motivates our primary model, which uses the same architecture without ensembling, given that ensembling did not provide performance gains but substantially increases computational complexity (see Results, 'Evaluation of models on new hospital systems').

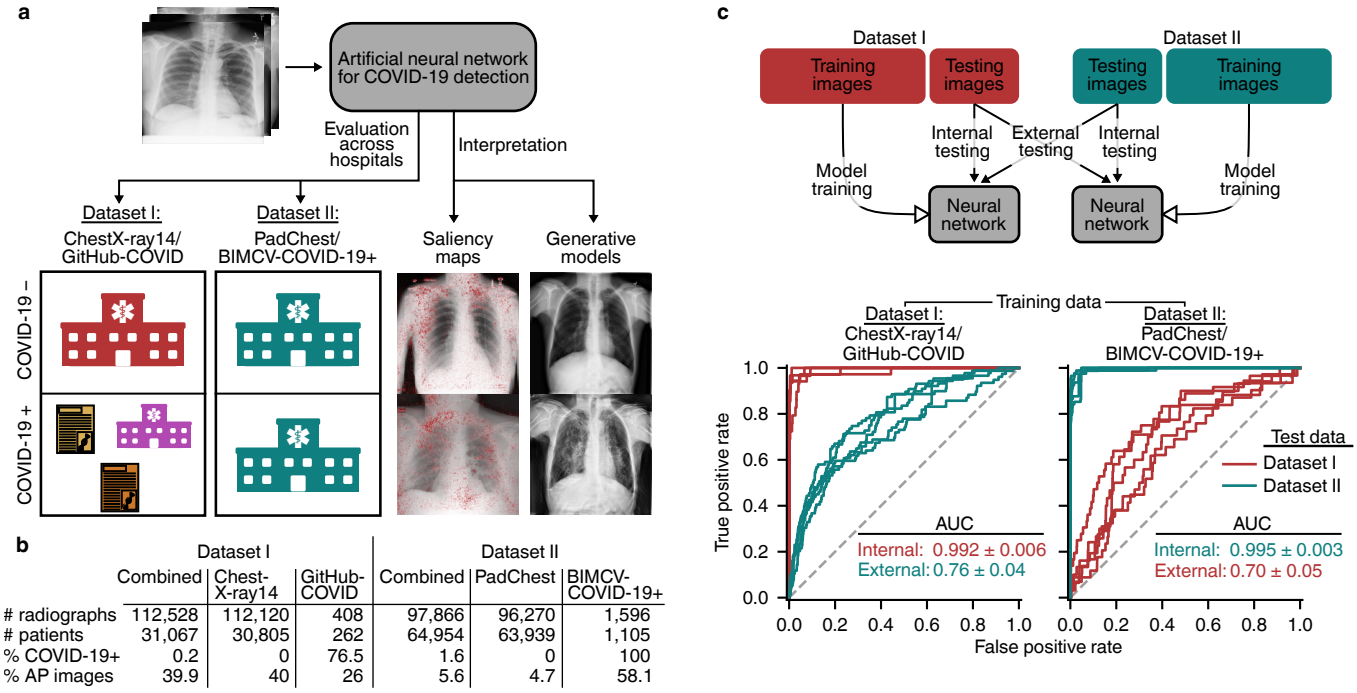


Fig. 2.1 | Overview of the study design. **a**, A neural network model is trained to detect COVID-19 using radiographs from either of two datasets, and then evaluated on both datasets to learn how performance may drop in deployment (i.e., a generalization gap). Intepretability methods are then applied to infer what the model learned and which features were important for its decisions. Whereas Dataset I draws radiographs from multiple hospital systems as well as cropped images from publication figures, Dataset II draws radiographs from multiple hospitals from a single regional hospital system. **b**, Characteristics of the datasets used in this study. **c**, Model evaluation scheme (top) and corresponding receiver operating characteristic (ROC) curves (bottom), which indicate the performance of our neural network models evaluated on both an *internal* test set (new, held-out examples from the same data source as the training radiographs) and an *external* test set (radiographs from a new hospital system). Inset numbers indicate area under the ROC curves, where larger area corresponds to higher performance (AUC, mean ± standard deviation). The difference between internal and external test set performance is the generalization gap.

To train and evaluate these models, we created two datasets (Fig. 2.1a, Supplementary Table 2.1). Dataset I consisted of COVID-19 positive radiographs from the GitHub-COVID repository,[65] which aggregates radiographs from publication figures and other online sources with varied geographic origin. We supplemented these with COVID-19 negative radiographs from the National Institutes of Health's (NIH) ChestX-ray14 repository,[66] which originates from a single hospital in the United States.

Dataset I is similar to the datasets used for training in recent publications on AI for COVID-19 detection.[45–48,56,57] Specifically, four of these publications[45–47,57] combine the GitHub-COVID repository with either the NIH repository[66] or the similar Radiological Society of North America pneumonia dataset,[67] which was derived from the NIH repository; two others[48,56] similarly combine these repositories and then supplement with additional COVID-19+ images from other online repositories, many of which have since been added to the GitHub-COVID repository. Given the continually evolving nature of many of these repositories, the precise set of images used in each study remains unclear, and additional uncertainty is introduced by the dearth of documentation on the source of some images or the validity of their labels (*e.g.*, in the ActualMed and Figure 1 databases at https://github.com/agchung/Actualmed-COVID-chestxray-dataset and https://github.com/agchung/Figure1-COVID-chestxray-539dataset). This uncertainty notwithstanding, our core observation is that numerous well-cited studies build their datasets by gathering COVID-19+ radiographs from varied sources, as exemplified most thoroughly by the GitHub-COVID repository (in which the image sources and labeling method are clearly documented), and then combining these with COVID-19 negative radiographs originating from the NIH repository, such that we judge our Dataset I fairly represents the key aspects of the data used in these

prior works. Other publications,[62,68–70] which generally use non-public data that precludes our ability to audit their models, do not share this issue of strong correlation between data source labels and COVID-19 status, but based on our review of the literature, we find this issue in an alarming proportion of the publications, including many of the most high profile studies[46,56,57].

Unlike the datasets used in recent publications, which collected COVID-19 positive and negative images from disparate sources, Dataset II corresponds to a seemingly more ideal case where both COVID-19 positive and negative images were drawn from similar sources. This dataset, which comprises the PadChest and BIMCV-COVID-19+ repositories (Fig. 2.1a-b), consisted of radiographs from a single region and published by a shared research team, though BIMCV-COVID-19+ represents a greater diversity of hospitals than PadChest, and the repositories were acquired over different time periods.[71,72]

Evaluation of models on new hospital systems

After training on Dataset I, we evaluated our models for reliance on confounding factors by comparing the predictive performance on an internal test set (new, held-out radiographs from Dataset I) to performance on external radiographs from Dataset II. While our models attain high performance on internal test data, *half of the model's predictive performance is lost* when testing on Dataset II (Fig. 2.1c, left). This performance drop (i.e., generalization gap) suggests these models rely on source-specific confounds in the radiographs, as we would expect models that use genuine markers of pathology to generalize well.[40] This finding held true for all nine additional architectures we examined, including those that were custom tailored in recent studies for detection of COVID-19 in radiographs (Supplementary Fig. 2.8-2.9).

While we initially expected that a dataset built from radiographs drawn from a single region would be less likely to contain spurious correlations that enable ML models to take shortcuts, we found that models trained on Dataset II also exhibit high performance on internal test data and low performance on external test data (Fig. 2.1c, right, and Supplementary Fig. 2.8). Thus, dataset-level confounding may pose a severe issue even in datasets derived from more similar sources, such as hospitals from a single region, contrary to the conclusions of contemporary work.[73] These findings argue for routine reporting of metadata on potential patient, hospital system, and preprocessing confounds. By illuminating the construction of radiographic datasets in greater detail, these data will make it easier for domain experts to identify likely sources of confounding. Additionally, these metadata enable the construction of models that explicitly control for confounds, providing a route to AI systems that generalize well even in the context of confounded training data.[74–76] In contrast, we note that a popular set of approaches to improve generalization performance, known as 'unsupervised domain adaptation,' are precluded by the presence of worst-case confounding because these methods rely on learning models invariant to data-source labels, which will be perfectly correlated with the pathology labels.[77]

Alternate hypotheses do not explain poor generalization

To verify the hypothesis that exploitation of dataset-specific confounding leads to poor generalization performance, we investigated alternative explanations for the generalization gap. Previous publications have suggested that more complex models, *i.e.*, those with higher *capacity*, may be particularly prone to learning confounds,[78] so we evaluated the generalization performance of simpler models, including a logistic regression and a simple convolutional neural network architecture, but found that the generalization gap did not improve (Supplementary Fig. 2.9). This result further supports the broad applicability of our findings, since the generalization gap was present regardless of network architecture, aligning with a previous study which showed that radiograph classification performance is robust to neural network architecture.[79] Likewise, we found that replacing the multi-label classification scheme of our original models with a simpler single-label classification scheme (see Methods Section 4.1) did not improve generalization performance.

In addition to the choice of model architecture, an alternative explanation for poor generalization performance is that, rather than the model learning a spurious correlation that does not generalize, the model learns a genuine relationship between a radiograph's appearance and its COVID-19 label that still does not generalize. One such scenario is that the COVID-19 detection task differs between training and test-time, which may occur in our datasets given that most of the images in the GitHub-COVID dataset were cropped from scientific publications and thus are perhaps more likely to show radiographic evidence of COVID-19, while labels in the BIMCV dataset are derived solely from RT-PCR or serology, and therefore may or may not feature radiographic evidence of COVID-19. However, when we modified the label scheme of BIMCV-COVID-19+ such that radiographs are only labelled positive if a radiologist noted evidence of COVID-19, the generalization gap persisted (Supplementary Fig. 2.10), suggesting that such *concept shift* between training and test time does not explain the performance difference and leaving the use of spurious correlations as the best explanation.[80]

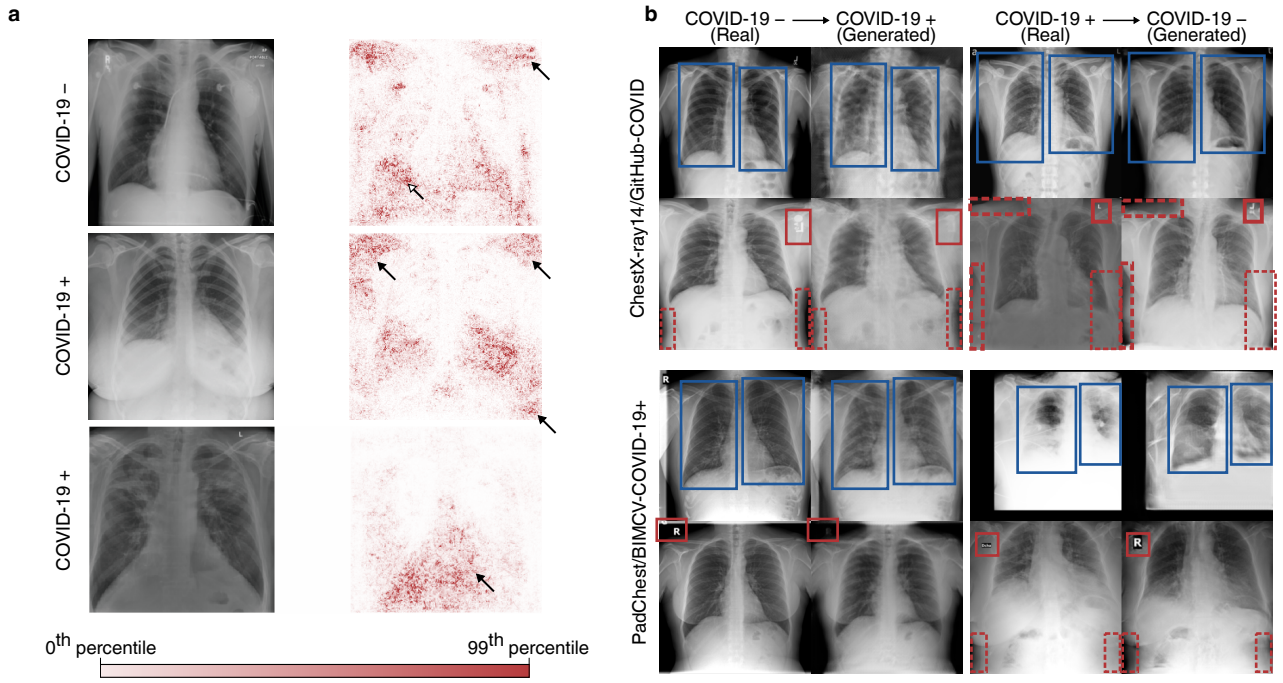


Fig. 2.2 | Explainable AI visualizes image factors important for deep neural networks trained to detect COVID-19 in radiographs. a, Saliency maps for our neural network models indicating the regions of each radiograph with the greatest influence on the model's prediction. Top: in a COVID-19 negative radiograph, in addition to the highlighting in the lung fields (open arrow), the saliency maps also emphasize laterality tokens (filled arrow). Middle: in a COVID-19 positive radiograph, the most intensely highlighted regions of the image are the bottom corners (arrows) outside of the lung fields. Bottom: in a COVID-19 positive radiograph, the only highlighted region is the diaphragm (arrow). The colour bar indicates saliency map pixel importances by percentile. **b**, Radiographs and their corresponding transformations by a GAN, illustrating systematic differences that enable neural networks to differentiate between COVID-19-positive and -negative radiographs. COVID-19-negative images are transformed by the GAN to appear as if they were COVID-19-positive, and vice versa. Comparison of images before and after transformation with a GAN visualizes important image features for COVID-19 prediction. Blue boxes indicate alterations to the opacity of the lung fields, which may represent the network's attention to genuine COVID-19 pathology. Red solid boxes indicate altered laterality markers, and red dashed boxes indicate altered radiopacity at the image borders, both of which may spuriously correlate with a patient's COVID-19 status in the training data. Figure adapted with permission from ref.,[81] H. Winther et al. (a, bottom; b, bottom row); and ref.,[82] Springer Nature Ltd (b, top row)

Explainable AI identifies spurious confounders

We further interrogated the trained AI models using saliency maps,[9,83,84] which highlight the regions of each radiograph that contribute most to the model's prediction (Supplementary Note and Supplementary Fig. 2.11), to determine specific confounds that deep convolutional networks for COVID-19 detection exploit. While our saliency maps sometimes highlight the lung fields as important (Fig. 2.2a), which suggests that our model may take into account genuine COVID-19 pathology, the saliency maps concerningly also highlight regions outside the lung fields that may represent confounds. The saliency maps frequently highlight laterality markers that originate during the radiograph acquisition process (Fig. 2.2a and Supplementary Fig. 2.12), which differ in style between the COVID-19-negative and COVID-19-positive datasets, and similarly highlight arrows and other annotations that are uniquely found in the publication-sourced radiographs of the GitHub-COVID data source[65] (Supplementary Fig. 2.13), which aligns with a previous study finding that ML models can learn to detect pneumonia based on spurious differences in text on radiographs.[85] Our saliency maps also indicate that the image edges, the diaphragm, and the cardiac silhouette are important for our models' predictions of a patient's COVID-19 status, though these regions are *not* among those routinely used by radiologists to assess for COVID-19[86] and instead likely reflect dataset-level differences in patient positioning and radiographic projection, *i.e.*, anterior-posterior (AP) vs. posterior-anterior (PA) view.[76] Reliance on such confounds, which do not consistently correlate with COVID-19 status in outside datasets, helps explain the previously observed poor generalization performance.

To further investigate what features could be used by an ML model to differentiate between the COVID-19 positive and COVID-19 negative datasets, we trained generative adversarial networks (GANs) to transform COVID-19 negative radiographs to resemble COVID-19 positive radiographs and vice versa. This technique should capture a broader range of features than saliency maps, as the GANs are optimized to identify all possible features that differentiate the datasets. Consistent with our knowledge of how radiologists detect evidence of COVID-19 in chest radiographs, the

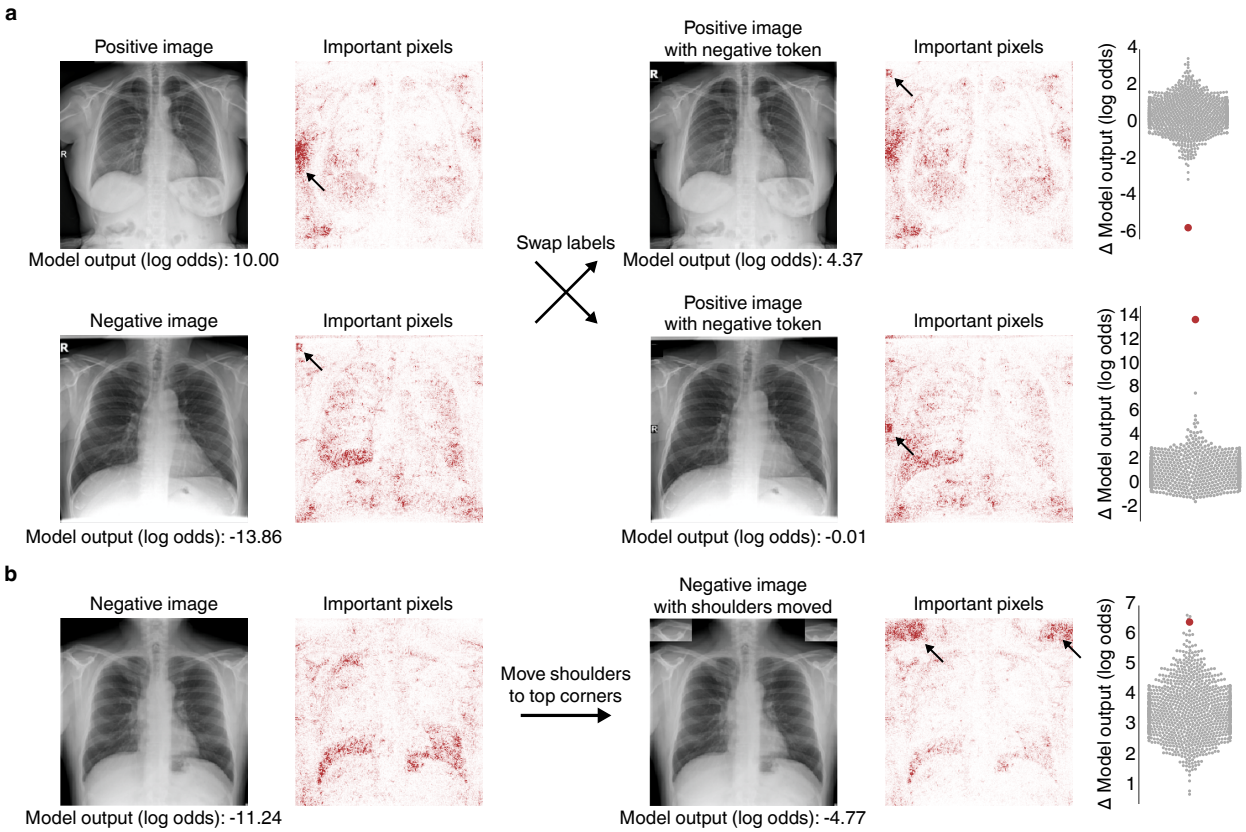


Fig. 2.3 | Experimental confirmation of insights from saliency maps and CycleGANs via radiograph modification. a, (Left) Text markers on radiographs are highlighted by saliency maps as important for COVID-19 prediction. The exchange of laterality markers between a pair of COVID-19 + and COVID-19 - images significantly shifts the output when compared to swapping random patches of the same size: Δ positive image (log odds) $= -5.63$ (empirical p-value $= 9.99 \times 10^{-4}$ based on Monte Carlo substitution of random image patches, n=1000); Δ negative image (log odds) $= 13.85$ ($p = 5.00 \times 10^{-3}$, n=1000) (Methods Sections 4.5 and 4.6). Gray dots in the distribution plots (right) correspond to the change in model output after swapping random image patches, which were used as a negative control, while the red dots correspond to the change in model output for the radiographs with swapped laterality markers. **b**, Positioning of patient shoulders may impact COVID-19 prediction. Saliency maps highlight the shoulder region as important predictors of COVID-19 positivity after (but not before) this region is moved to the top of the image (left). This patch increased model output significantly more than random patches of the same size moved to the same corners ($\Delta = 6.57$, empirical p-value $= 5.00 \times 10^{-3}$, n=1000). Gray dots in the distribution plot (Right) correspond to radiographs with randomly selected patches, while the red dot corresponds to the radiograph with the shoulder regions moved.

GAN increases the radiopacity or radiolucency of the lung fields bilaterally to respectively add or remove evidence of COVID-19, indicating that neural network models are capable of learning genuine markers of COVID-19 (Fig. 2.2b, blue boxes, and Supplementary Fig. 2.14 and 2.15). However, the generative networks frequently add or remove laterality markers and annotations (Fig. 2.2b, solid red boxes), reinforcing our observation from saliency maps that these spurious confounds also enable ML models to differentiate the COVID-19 positive and COVID-19 negative radiographs. The generative networks additionally alter the radiopacity of image borders (Fig. 2.2b, dashed red boxes), supporting our previous assertion that systematic, dataset-level differences in patient positioning and radiographic projection provide an undesirable shortcut for ML models to detect COVID-19. Given this strong evidence that ML models can leverage spurious confounds to detect COVID-19, we also investigated the extent to which our classifiers, in particular, relied upon the features altered by the GAN. We found that images transformed by the GANs were reliably predicted by the classifiers to be the transformed class rather than the original class (Supplementary Fig. 2.16), demonstrating that the majority of features used by our classifiers were altered by the GAN, *i.e.*, the features identified by the GAN are approximately a superset of those used by the classifiers. Thus, the image transformations from the GANs enable us to see hypothetical versions of the same radiographs that would have caused our classifiers to predict the opposite COVID-19 status.

Experimental validation of factors identified by interpretability methods

We next aimed to experimentally validate the importance of spurious confounds to our models by manually modifying key features (Fig. 2.3a-b). We first swapped laterality markers from a COVID-19 positive and COVID-19 negative

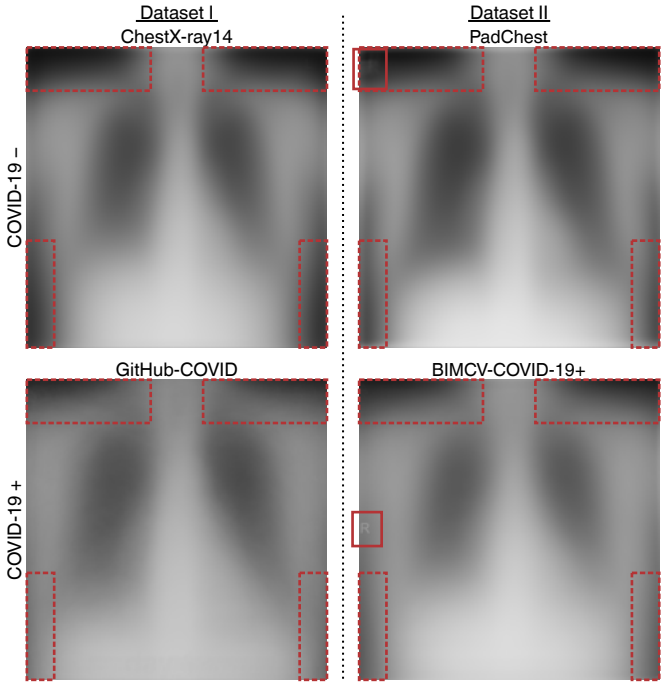


Fig. 2.4 | Average images from the four repositories used to construct datasets in this study, demonstrating systematic differences between the radiograph repositories that could be exploited by AI systems. Solid red boxes indicate systematic differences in laterality markers that are visible in the average images, and dashed red boxes indicate systematic differences in radiopacity of the image borders, which could arise from variations in patient position, radiographic projection, or image processing.

image, and found that introduction of a laterality marker more common in COVID-19 positive images increased the models' predicted odds that the patient had COVID-19, while the converse also held. As a control, we compared to randomly swapped image patches of the same size and found that the change in model output from swapping laterality markers is significantly greater than expected by random (Fig. 2.3a), indicating that laterality markers are key features leveraged by our models to determine a patient's COVID-19 status. While these markers vary consistently between the datasets (Fig. 2.4 and Supplementary Fig. 2.13, 2.14, and 2.15), these markers would not reliably indicate COVID-19 status in more general settings. We similarly investigated the shoulder region of radiographs, which was frequently highlighted as an important feature in our saliency maps (Supplementary Fig. 2.13), and found that moving the clavicle region of a radiograph to the top border of the radiograph increased the model's predicted odds that the patient has COVID-19 (Fig. 2.3b and Supplementary Fig. 2.17), suggesting that the models leverage the consistent but medically irrelevant difference in patient positioning between the COVID-19 negative and COVID-19 positive data sources. To verify whether these findings held on a population basis, we sampled a random subset of the radiographs and repeated our experiments involving the swapping of laterality markers and movement of the shoulder region (Supplementary Fig. 2.18), which confirmed that our models indeed leverage these shortcuts throughout the dataset.

Shortcuts have a variable effect on generalization

Importantly, some shortcuts will impair generalization performance, while other shortcuts will not; while the large generalization gap is explained well by shortcut learning, a portion of the remaining external test set performance may still be due to shortcuts that happen to generalize for our datasets. Both types of shortcut are undesirable, since even those that generalize between our datasets may not consistently generalize to other settings, and the use of clinical rather than strictly radiological information extracted from these radiographs may be redundant, depending on the clinical workflow.

To analyze which shortcuts may contribute to poor generalization, we considered clinical metadata (Supplementary Table 2.1) and average images from each repository (Fig. 2.4). Among the shortcuts that do not generalize are the textual markers, which were clearly identified by our explainability approaches as important for prediction of COVID-19 but appear differently in the COVID-19 negative and COVID-19 positive images from each repository (Fig. 2.4). In addition, the radiographic projection, which may contribute to (but does not completely explain) the importance of the image edges and shoulder position, does not generalize between the datasets (Fig. 2.1b, '% AP images' row) and therefore may contribute to poor generalization performance.

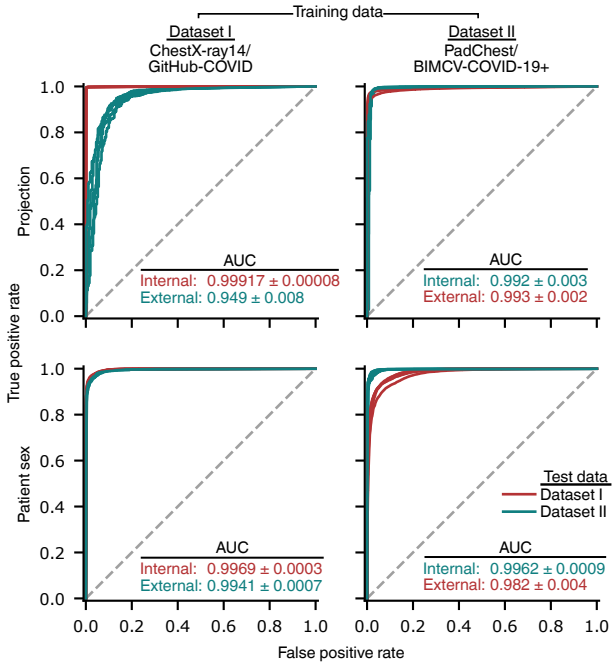


Fig. 2.5 | Evaluation of the extent to which the prediction of image factors that could be leveraged as shortcuts to detection of COVID-19 generalizes to new hospitals. Models were trained to predict radiographic projection (AP vs. PA view) and then evaluated on internal and external test radiographs. Inset values indicate area under the ROC curve (AUC, mean ± standard deviation, n=5).

Among the shortcuts that do generalize (at least between our datasets) are aspects of patient positioning that do not result from the radiographic projection. These aspects of patient positioning also likely contribute to the previously observed importance of image edges and shoulder position, and they maintain a consistent relationship with COVID-19 negative and COVID-19 positive radiographs in each dataset (Fig. 2.4), despite the inconsistent relationship of the radiographic projection with COVID-19 status. An additional factor that may generalize well is patient sex, since within both datasets, a higher proportion of males were COVID-19 positive (Supplementary Table 2.1). Taken together with our observation that half of our models' performance is attributable to confounds that do not generalize well, we conclude that only a minority of our models' performance is attributable to monitoring for genuine COVID-19 pathology.

Given that radiographic projection and patient sex are diffusely represented in radiographs and therefore less clearly pointed out by our explainability approaches, we additionally validated whether our models could leverage these factors as shortcuts. We reasoned that for a model to be able to leverage these concepts as shortcuts, the same model (when retrained) must be able to predict these concepts well. Indeed, our models accurately predict both the radiographic projection and patient sex for both internal and external test data (Fig. 2.5), which supports that these concepts are easily learned and available to be leveraged as shortcuts. Considering that these concepts are easily learned and are also predictive of COVID-19 status (*i.e.*, they are correlated with COVID-19 in our datasets), we judge that our networks likely incorporate this information to predict COVID-19 status.

Improved data mitigates shortcut learning

Given this strong evidence that neural networks leverage dataset-level differences as shortcuts for COVID-19 status, we inquired to what extent this issue might be mitigated. While an initial hypothesis may be that the choice of neural network architecture determines the propensity for shortcut learning, all architectures that we examined displayed similar evidence for shortcut learning, as quantified by the generalization performance (Supplementary Fig. 2.8). While our tests hinted that data augmentation may help alleviate shortcut learning, the effect was small and not statistically significant (Supplementary Fig. 2.8b; external test set ROC-AUC of 0.76 ± 0.04 vs. 0.79 ± 0.03 before and after data augmentation, respectively, when trained on dataset I, p = 0.22, U = 6 based on a Mann-Whitney U-test; external test set ROC-AUC of 0.70 ± 0.05 vs 0.69 ± 0.05 before and after data augmentation, respectively, when trained on dataset II, p = 1.00, U = 13 using Mann-Whitney U-test).

In principle, an attractive solution to mitigate shortcut learning is to remove the image factors that the models leverage as shortcuts. However, in practice, it is difficult to remove all such image factors. As a simple test case, we inquired

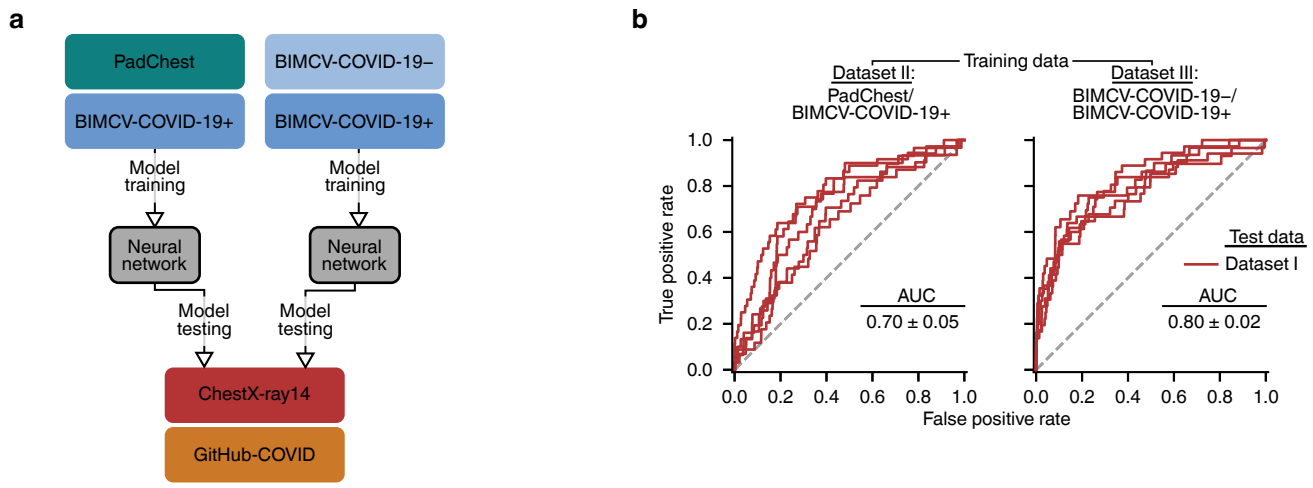


Fig. 2.6 | Mitigation of shortcut learning via collection of improved data. **a**, To evaluate whether improved data collection mitigates shortcut learning, we train classifiers on dataset II and dataset III, then test both on the same external data (dataset I). **b**, Evaluation of generalization performance as measured by receiver operating characteristic (ROC) curves. Inset values indicate area under the ROC curve (AUC, mean ± standard deviation, n=5). *$p = 0.016$ based on a two-tailed Mann-Whitney U-test (corresponding U=-2.4).

whether removing textual markers by cropping to the center 75% of each radiograph would reduce shortcut learning and thus improve generalization performance. After retraining our models on these cropped radiographs, we found that such cropping does not improve generalization performance (Supplementary Fig. 2.19), which naïvely may suggest that these textual markers do not contribute to shortcut learning. However, considering the consistent identification of this factor by saliency maps, the CycleGANs, and manual image modifications (Fig. 2.2a-b, Fig. 2.3a), a more likely explanation is that a multitude of redundant shortcuts exist, such that a model may shift its attention toward other shortcuts in absence of a particular shortcut; conjecturally, such image attributes could include the size of the lung fields relative to the image, the positioning of the scapular shadows, the size of the cardiac silhouette, image intensities, or textural features that enable inference of the data source.

Perhaps a more reliable solution to remove the image factors that enable shortcut learning is to simply collect data that is less confounded. To test this hypothesis, we created a third dataset (Dataset III) to represent a nearly optimal case; the COVID-19 positive and negative cases were taken from the BIMCV-COVID-19+ repository and its paired BIMCV-COVID-19− repository (https://bimcv.cipf.es/bimcv-projects/bimcv-covid19/), respectively, which were collected from the same hospitals over the same time period (Supplementary Fig. 2.20). If this near-optimal dataset solved the 'shortcut problem,' then we would expect that models trained on these data may (i) attain higher performance on an external test set, since bonafide pathology should transfer between datasets while shortcuts may or may not, and (ii) exhibit a lower generalization gap, in the sense that performance on an internal test set would not as drastically misrepresent the true performance, as measured on external data. We trained models to detect COVID-19 in Dataset III and then tested these models on external data from Dataset I, and compared these results to models that were trained on Dataset II and tested on Dataset I. Despite that Dataset III contains approximately $1/20^{th}$ the images of Dataset II, it attains significantly higher performance on external data (Fig. 2.6), and exhibits little generalization gap (Supplementary Fig. 2.21), suggesting that collection of less confounded data indeed alleviates the issue of shortcut learning. Furthermore, saliency maps for the model trained on Dataset III tend to attribute more importance to the lung fields, where COVID-19 pathology would be expected, than to potentially confounding regions, as compared to the equivalent saliency maps generated for the model trained on Dataset II (Supplementary Fig. 2.22), though the saliency maps still show some attention toward shortcuts. Taken together, these findings argue for careful collection of data so as to minimize potential for shortcut learning, with continued caution that improved data collection may only partially solve the problem.

2.4 Discussion

ML models that were built and trained in the manner of recent studies generalize poorly and owe the majority of their performance to the learning of shortcuts. This undesired behavior owes partially to the synthesis of training data from separate datasets of COVID-19 negative and COVID-19 positive images, which introduces near worst-case confounding

and thus abundant opportunity for models to learn these shortcuts. Importantly, since undesirable 'shortcuts' may be consistently detected in both internal and external domains, our results warn that external test set validation alone may be insufficient to detect poorly behaved models.

Previous studies also audited AI systems for detection of COVID-19 in radiographs, with mixed success at identification of shortcuts. In a simple yet clever approach, one study found that models retain high performance when examining only the borders of radiographs, such that genuine COVID-19 pathology was removed from the images.[73] This study concurs with our findings but comments primarily on the possibility of this issue rather than its occurrence in the wild, though it is nonetheless alarming. The study that introduces the COVID-Net model also audits its model, using a saliency map approach known as 'GSInquire,' but in contrast does not identify evidence of shortcut learning in a set of three published images.[56] Given the similarity of that study's training data to our own Dataset I and the large generalization gap that we observe with the same architecture, we suspect shortcut learning likely indeed occurred, and it remains unclear whether auditing decisions about additional radiographs beyond the three presented would have revealed evidence of shortcut learning, or if the GSInquire approach, which is not available through a public-facing repository, fails to identify the shortcuts. A number of other studies that involve datasets with severe confounding between pathology and image source[45–48,57] similarly audit their models using saliency map approaches (most prominently, the Grad-CAM approach[8]) and report findings on one to three radiographs, without noting evidence of shortcut learning. Based on this pattern, we recommend that researchers examine and report results from explainable AI or saliency map approaches on a population level, employing a sampling-based approach as necessary, and to remain skeptical of high performances in the absence of external validation. Moreover, we find that population-level audits using saliency maps are highly labor intensive to perform in a rigorous manner and may depend on domain knowledge, which motivates future approaches for explainable AI in medical imaging that simplify population-level analysis.

Our findings support common-sense solutions to alleviate shortcut learning in AI systems for radiographic COVID-19 detection, including (i) improved collection of training data, *i.e.*, data in which radiographs are collected *and processed* in a way matching the target population of a future AI system and (ii) improved choice of the prediction task to involve more clinically relevant labels, such as a numeric quantification of the radiographic evidence for COVID-19.[69,87] However, we demonstrate that shortcut learning may occur even in a more ideal data collection scenario, highlighting the importance of explainable AI and principled external validation. While AI promises eventual benefits to radiologists and their patients, our findings demonstrate the need for continued caution in the development and adoption of these algorithms.[58]

2.5 Methods

Model architecture and training procedure

For our primary neural network, we used a convolutional neural network with the DenseNet-121 architecture to predict the presence versus absence of COVID-19.[61] This architecture has not only been used in a variety of recent models for COVID-19 classification,[56,57] but has also been used for the diagnosis of non-COVID pneumonia,[76,83] as well as for more general radiographic classification.[88]

Following the approach in recent COVID-19 models,[56,57] we first pre-trained the model on ImageNet, a large database of natural images.[20] Forcing models to first learn general image features should also serve as an inductive bias to prevent overfitting on domain-specific features.[76] After ImageNet pre-training, the final 1000-node classification layer of the trained ImageNet model was removed and replaced by a 15-node layer, corresponding to the 14 pathologies recorded in the ChestX-ray14 dataset plus an additional node corresponding to COVID-19 pathology; while only the prediction for COVID-19 was used for evaluating the model, we followed previous works that showed simultaneous learning of multiple tasks was useful for achieving highest predictive performance.[83] To obtain a consistent label scheme, labels in the GitHub-COVID, PadChest, and BIMCV-COVID-19+ repositories were mapped to the 14 ChestX-ray14 categories.

The model was optimized end-to-end using mini-batch stochastic gradient descent with a batch size of 16, momentum parameter of 0.9, weight decay of 10^{-4}, and learning rate of 0.01, which was decreased by a factor of 10 every 5 epochs. We chose a binary cross entropy loss as the optimization criterion. To prevent overfitting, we monitored the area under the ROC curve (AUROC) for COVID-19 classification on a held-out validation set, and chose the epoch with the highest validation AUROC as the final model. All models were trained for 30 epochs, which was long enough for all models to reach a maximum in the validation AUROC. All models were trained using the PyTorch software library,[89] version 1.4, on NVIDIA RTX 2080 TI graphics processing units and required approximately 5 hours of

training time per replicate.

We additionally examined three architectures that were designed in previous publications specifically for the task of COVID-19 detection, with the hypothesis that these specialized architectures may better learn genuine COVID-19 pathology and generalize better to external data. These architectures include CV19-Net,[62] DarkCovidNet,[46] and COVID-Net.[56] We trained these models on datasets I and II, following the image pre-processing procedures, data augmentation pipelines, and optimization schemes used in the original publications (we note that while dataset I is analogous to the original datasets used to train DarkCovidNet and COVID-Net, CV19-Net was trained on data that is not publicly available). For both CV19-Net and DarkCovidNet, the base architectures were dowloaded from the torchvision library,[89] then modified to match the descriptions in each respective paper. The COVID-Net network was adapted from an open-source, PyTorch implementation (by Ilias Papastratis; https://github.com/iliasprc/COVIDNet). For the CV19-Net paper, the data augmentation pipeline was altered to match the pipeline in the original paper: when loading images, each radiograph is additionally randomly flipped with probability 0.5 then rotated between -30 and 30 degrees. To disentangle performance differences due to the ensembling present in the CV19-Net architecture from performances differences due to the change in data augmentation, we also trained a single DenseNet-121 model with the same data augmentation steps as the CV19-Net. In the case of the CV19-Net and DarkCovidNet, we maintained the same multilabel classification task (*i.e.*, the 14 ChestX-ray14 labels plus a label for COVID-19) to facilitate optimal comparison between architectures. In the case of the COVID-Net architecture, due to problems with vanishing and exploding gradients when using the full multilabel classification task, we reduced our full label set to only the 3 labels used in the COVID-Net paper (COVID-19 Pneumonia, Non-COVID Pneumonia, No Pneumonia). We also trained additional, popular architectures that were not tailored specifically for COVID-19 detection, including MobileNetv2[63] and ResNeXt-50.[64] These networks were again modified from the ImageNet-pretrained base models in the torchvision library.[89] We trained these architectures using the same pre-processing scheme and optimization parameters as our DenseNet-121 models, again replacing the standard, 1000-label classification layers with an analogous layer for our 15 labels.

To test the hypothesis that lower-capacity models may not learn spurious correlations,[78] we also trained two lower-capacity models. The first, an AlexNet model,[19] was trained in the same manner as the DenseNet-121, with the weights randomly initialized rather than pretrained on ImageNet. The second was a logistic regression with 'deep features': since individual pixels do not have stable semantic meaning over different samples in the dataset, we first extract a set of 1024 higher-level features using the feature embedding (*i.e.*, the activations of the penultimate layer) of a DenseNet-121 trained on ImageNet and then fit a logistic regression to these fixed features. This procedure is accomplished by training the DenseNet-121 architecture with the weights of its feature embedding subnetwork frozen. The AlexNet and logistic regression were optimized using the same training parameters as the full DenseNet-121 model specified above. The fact that lower-capacity models did not generalize better in our setting may be due to the fact that Sagawa et al. focus on a reweighted training scheme,[78] while our models were trained to minimize empirical risk in order to replicate the training schemes used by recent COVID-19 detection models (see above).

Datasets and preprocessing

To train and evaluate our models, we combined images from five large open-access repositories of chest radiographs into three datasets (Fig. 2.1a, Supplementary Table 2.1). The first, which we refer to as Dataset I, was designed to replicate the datasets used to develop and evaluate the most popular COVID-19 diagnostic models.[56] In this dataset, we collected COVID-19 negative images from the NIH ChestX-ray14 repository, representing 112,120 radiographs from 30,805 patients from the NIH Clinical Center.[66] We collected COVID-19 positive images from the GitHub-COVID repository[65] (commit ID 9b9c2d5), representing 408 radiographs from 262 patients, where this data was originally collected from figures in scientific publications and assorted web sources of COVID-19 positive cases.

The second dataset, which we refer to as Dataset II, was designed to represent a more ideal case in terms of domain confounding – both COVID-19 positive and COVID-19 negative images were acquired from hospitals from a common region and were published by a shared research team. We collected COVID-19 negative images from the PadChest repository, representing 96,270 radiographs from 63,939 patients from a hospital in Valencia, Spain.[71] The COVID-19 positive images in our dataset were taken from the BIMCV-COVID-19+ dataset (version 1), which represents 1,596 images from 1,015 patients (after exclusions), from the same regional hospital system in Valencia, Spain.[72] We note that while PadChest and BIMCV-COVID-19+ originate from the same region, potential for confounding remains since (i) PadChest was collected from a single hospital whereas BIMCV-COVID-19+ was collected from multiple hospitals, and (ii) the repositories were collected over different time periods, over which image acquisition techniques may have changed.

The third dataset, referred to as Dataset III, was designed to represent the most ideal case in terms of domain confounding. Unlike dataset II, the COVID-19 positive and COVID-19 negative images were collected from not only the same region, but from the same hospitals and over the same time period. Like dataset II, the COVID-19 positive images were collected from the BIMCV-COVID-19+ repository. The COVID-19 negative images were taken from the corresponding BIMCV-COVID-19− repository, which includes 3086 images from 2327 patients (after exclusions).

Following the recommendations by Cohen et al.,[90] we filtered radiographs from the online repositories to include only PA and upright AP radiographs. Lateral radiographs, AP supine radiographs, radiographs with unknown projections, and computed tomography scans were excluded from the datasets. Images with absent radiographic windowing information, which was necessary to display radiographs from the BIMCV-COVID-19+ and BIMCV-COVID-19− repositories, were also excluded.

We partitioned each repository into training, validation, and test folds, ensuring that all radiographs of any given patient belong to a single fold. Since the ChestX-ray14 dataset specifies a 'test' partition, we used these radiographs as part of our dataset I test fold. Of the remaining portion, 5% were reserved as a validation fold, while the rest were used directly for training. In the PadChest, BIMCV-COVID-19+, and BIMCV-COVID-19− repositories, we reserved 5% of the radiographs for testing, and 5% of the remaining radiographs for validation. Due to the smaller size of the GitHub-COVID repository, we reserved 10% of the radiographs for testing, and 10% of the remaining radiographs for validation. With the exception of the ChestX-ray14 test fold, which was held fixed as explained above, the folds were drawn at random for each model replicate.

Model interpretability using saliency maps

To generate saliency maps, which enable interpretation of machine learning models by assigning importance values to each pixel of an input image, we apply a state-of-the-art approach known as *Expected Gradients*.[27] Broadly, this approach captures the notion of "importance" by tracking how each pixel of an image impacts the output of the model when contrasted with a set of noninformative baseline examples, where the impact is measured by accumulating the model's gradients (a mathematical measure of a model's sensitivity to small changes in a feature) as the image is interpolated from the baseline example to the image of interest. Formally, the Expected Gradients attribution ϕ for an input sample x and input feature i is defined:

$$\phi_i(x) := \mathop{\mathbb{E}}_{x' \sim D, \alpha \sim U(0,1)} \left[(x_i - x'_i) \times \frac{\delta f(x' + \alpha \times (x - x'))}{\delta x_i} \right], \tag{2.1}$$

where D represents a *background distribution* from which reference samples x' are drawn. This method is an extension of the popular saliency map approach Integrated Gradients, which is the special case of Expected Gradients in which there is only a single reference sample.

For our application, Expected Gradients improves over Integrated Gradients in terms of the accuracy of its saliency maps[27] and the inclusion of multiple reference samples, which avoids the choice of a single reference that may be arbitrary but nonetheless impactful upon the resultant saliency maps.[91] Finally, path-based approaches like Expected Gradients and Integrated Gradients are preferable to other methods for generating saliency maps because they are theoretically principled: these methods are provably guaranteed to attribute importance to important pixels and guaranteed not to attribute importance to unimportant pixels (also see Supplementary Note).[9]

As the background distribution D for Expected Gradients, we used the COVID-19-negative images from the training dataset for each model we explain. Intuitively, we are explaining how the output of our model for our input image x differs on average from the output of the model for images in the training data D. We demonstrate that Expected Gradients is not overly sensitive to choice of D by comparing the saliency maps for several radiographs with a background distribution of images from the training data to attributions for those same radiographs with a background distribution of images from the external dataset, and found the resultant attributions are similar (Supplementary Fig. 2.23).

Data interpretability using CycleGAN

To attain visual explanations of the differences between COVID-19 positive and COVID-19 negative images in each dataset, we aimed to understand which characteristics of the chest radiograph would have to change to make a COVID-19 negative image appear to be a COVID-19 positive image, and vice versa. Formally, let $\mathcal{X}$ be a domain of COVID-19

negative images, and let $\mathcal{Y}$ be a domain of COVID-19 positive images. Our goal is to learn a mapping $G : \mathcal{X} \mapsto \mathcal{Y}$ that takes a COVID-19 negative chest radiograph, $X \in \mathcal{X}$, and transforms it so that it is indistinguishable from COVID-19 positive chest radiographs. We also aim to learn the inverse transformation, $F : \mathcal{Y} \mapsto \mathcal{X}$.

Since generative adversarial networks have previously been shown to be effective for the interpretation of neural networks, we learn these two transformations using the CycleGAN approach.[36,39] The mappings G and F are learned by two neural networks, which are optimized in conjunction with two discriminator networks $D_{\mathcal{Y}}$ and $D_{\mathcal{X}}$. These networks are optimized to minimize a series of losses. The first, referred to as the *adversarial loss*, encourages the mapping functions G and F to match the distribution of generated images from each source domain to the true data distribution of each target domain:

$$\mathcal{L}_{\text{GAN}}(G, D_{\mathcal{Y}}, \mathcal{X}, \mathcal{Y}) = \mathbb{E}_{Y \sim p_{\text{data}}(Y)}[\log D_{\mathcal{Y}}(Y)] + \mathbb{E}_{X \sim p_{\text{data}}(X)}[\log(1 - D_{\mathcal{Y}}(G(X))], \tag{2.2}$$

$$\mathcal{L}_{\text{GAN}}(F, D_{\mathcal{X}}, \mathcal{Y}, \mathcal{X}) = \mathbb{E}_{X \sim p_{\text{data}}(X)}[\log D_{\mathcal{X}}(X)] + \mathbb{E}_{Y \sim p_{\text{data}}(Y)}[\log(1 - D_{\mathcal{X}}(F(Y))], \tag{2.3}$$

where $p_{\text{data}}(X)$ and $p_{\text{data}}(Y)$ represent the data distributions for each domain. In addition to the adversarial loss, the networks are also trained to enforce *cycle consistency*, meaning that $F(G(X)) = X$. This is desirable, since it enforces a similarity between the original and transformed images. The loss here is:

$$\mathcal{L}_{\text{cyc}}(G, F) = \mathbb{E}_{X \sim p_{\text{data}}(X)}[\|F(G(X)) - X\|_1] + \mathbb{E}_{Y \sim p_{\text{data}}(Y)}[\|G(F(Y)) - Y\|_1]. \tag{2.4}$$

The full loss that is optimized then is simply the sum of these three losses:

$$\mathcal{L} = \mathcal{L}_{\text{GAN}}(G, D_{\mathcal{Y}}, \mathcal{X}, \mathcal{Y}) + \mathcal{L}_{\text{GAN}}(F, D_{\mathcal{X}}, \mathcal{Y}, \mathcal{X}) + \mathcal{L}_{\text{cyc}}(G, F) \tag{2.5}$$

To understand which image features are important in distinguishing the domains $\mathcal{X}$ and $\mathcal{Y}$, we transform a COVID-19 negative radiograph $X \in \mathcal{X}$ or a COVID-19 positive radiograph $Y \in \mathcal{Y}$ using the learned generator networks G or F to map the image to the opposite domain. We then compare which image features are changed in the transformation.

Our CycleGAN networks were implemented in Python 3.7 using the PyTorch software library and an open-source implementation of the CycleGAN approach (by Aitor Ruano; https://github.com/aitorzip/PyTorch-CycleGAN). To attain comparable training time, the networks for trained for 3000 epochs (Dataset I) or 1000 epochs (Dataset II). Each network required approximately one week of training time on an NVIDIA RTX 2080 graphics processing unit.

Experimental validation of feature attributions

We experimentally validated our findings from saliency maps and GANs by modifying important radiographic features. To detect whether the higher-level features that our saliency maps highlight are major contributors to the model's classification, we used methods inspired by a behavioral testing approach.[92] For example, saliency maps highlight dataset-specific laterality markers and text within the images. If these text markers are indeed important, then moving a marker from a COVID-19 positive image to a COVID-19 negative image should increase the predicted log odds of COVID-19. For a pair of COVID-19 positive and COVID-19 negative images, we swap the text markers and measure the change in the output for each image. To assess the significance of the change in the model's output at the level of each individual image, we generate empirical p-values by comparing to a null distribution generated by swapping 1,000 random patches of each image of the same dimensions as the text markers (Fig. 2.3a). We conduct a similar experiment to validate whether the shoulder regions frequently highlighted in the saliency maps have a significant impact on the model's decisions. We observe that the shoulder region of COVID-19 positive images tends to appear at the upper image border, while the shoulder region of COVID-19 negative images appears slightly lower. Furthermore, the saliency maps highlight the clavicles and shoulders of the COVID-19 positive images, but not in the COVID-19 negative images. We hypothesized that the model was looking for the presence of shoulders in the upper corners of the image. To test our hypothesis, we moved the clavicles and shoulders of a COVID-19 negative image to the top corners of the radiograph and measured the change in model output (Fig. 2.3b). We tested for statistical significance at the level of individual images by generating empirical p-values. Our distribution was generated by randomly sampling and replacing 1000 patches of the same size as the shoulder region, following the same procedure described for the laterality markers.

In order to verify the significance of these regions for our models at a *population level*, we repeated the procedure described in the paragraph above for a sample of randomly selected radiographs from the datasets (see Supplementary Fig. 2.18). For the dataset-specific laterality markers (Supplementary Fig. 2.18, left), we randomly sampled 10 COVID-19 negative images with laterality or other text markers and 10 COVID-19 positive images with laterality or other text markers. To test for the significance of the text markers across the datasets, we used a Wilcoxon signed rank test to compare the distribution of the magnitudes of changes in model output after swapping the text markers to the distribution of the magnitudes of the average changes in model output after swapping 1000 random patches of the same size ($p = 8.86 \times 10^{-5}$, Siegel's T statistic $= 0.0$). For the positioning of the shoulder regions (Supplementary Fig. 2.18, right), we randomly sampled 20 COVID-19 negative images. We then used a Wilcoxon signed rank test to compare the distribution of changes in model output after moving the clavicles and shoulder regions to the top of the image with the distribution of the average changes in model output after moving 1000 random patches of the same size ($p = 8.86 \times 10^{-5}$, Siegel's T statistic $= 0.0$).

Statistics

In our experiments involving manual modification of radiographs (Fig. 2.3a-b, Supplementary Fig. 2.17), we computed empirical p-values by first generating the distribution of the change in the model output (in log odds space) for a set of random, non-specific modifications as described in each caption. The p-value was then calculated as $(r + 1)/(n + 1)$ where r is the number of non-specific modifications that produced a greater increase in model output (greater magnitude decrease in Fig. 2.3a, top row) and n is the total number of non-specific modifications.[93]

To compare the generalization performance of models (*e.g.*, Fig. 2.6), we performed a two-tailed Mann-Whitney U-test, given that the ROC-AUC values are bounded by 0 and 1 and therefore unlikely to be normally distributed.

Data availability

All radiographs are compiled from publicly-available data repositories. The ChestX-ray14 repository is available at https://nihcc.app.box.com/v/ChestXray-NIHCC. The GitHub-COVID dataset is available at https://github.com/ieee8023/covid-chestxray-dataset. The PadChest repository is available at https://bimcv.cipf.es/bimcv-projects/padchest/. The BIMCV-COVID19 repositories are available at https://bimcv.cipf.es/bimcv-projects/bimcv-covid19/.

Code availability

All of the code necessary to reproduce our experimental findings can be found at https://github.com/suinleelab/cxr_covid (archived at https://doi.org/10.5281/zenodo.4623792).

Materials & Correspondence

Correspondence to Su-In Lee.

Acknowledgments

This work was funded by the National Science Foundation [CAREER DBI-1552309 to S.-I.L.] and the National Institutes of Health [R35 GM 128638 and R01 AG061132 to S.-I.L.]. We thank Hugh Chen and Gabriel Erion for providing feedback while writing the manuscript. We thank Dr. Aurelia Bustos for clarifying characteristics of the PadChest and BIMCV-COVID-19+ datasets. We also thank Dr. David Janizek for insight into the interpretation of COVID-19 on chest radiographs.

Author contributions

J.D.J. conceived the study. A.J.D. and J.D.J. prepared datasets, designed experiments, and wrote software. S.-I.L. supervised the study. A.J.D. and J.D.J. and S.-I.L. wrote the manuscript.

Competing interests

The authors declare no competing interests.

2.6 Supplementary Information

Supplementary Note

While saliency maps are widely used to interpret image-based artificial intelligence systems [83, 84, 94], the reliability of these approaches has been disputed by contemporary work, which observes that saliency maps explaining medical imaging classifiers fail to localize medically relevant pathology [95]. However, this prior work did not disentangle whether (i) the saliency maps fail to identify the features that are important for the classification models, or (ii) the saliency maps faithfully identify the features that are important for the classification models, but the models do not depend on medically relevant pathology. We hypothesised the latter, that attribution maps fail to localize relevant pathology because the models they explain do not rely on relevant pathology [96].

To validate that the pixels selected by our saliency maps are truly important for the models they explain, we chose 100 images that our model predicted are COVID-19 negative, then masked and mean-imputed a subset of pixels. If we selected these pixels at random, we would expect the models output to regress to the mean output (become more positive) since the negative images become more like the mean image (which is predicted to be more positive than the COVID-19 negative images). If the pixels identified by Expected Gradients are important for the model's prediction, we would anticipate that masking these pixels should make the model's output *more positive* than masking randomly selected pixels. When we mask the top 10% of pixels identified by EG as contributing to the negative prediction of the model, we see that the model's output is shifted to be significantly more negative than when we mask pixels selected at random (Supplementary Fig. 2.11).

Supplementary Figures

	Dataset I			**Dataset II**		
	Combined	CXR14	Cohen et al.	Combined	PadChest	BIMCV-COVID
CXR #s	112,528	112,120	408	97,866	96,270	1,596
Patients, #s	31,067	30,805	262	64,954	63,939	1,015
Age, mean (std)	46.9 (16.8)	46.9 (16.8)	57.0 (16.4)	65.4 (20.1)	65.5 (20.1)	61.2 (16.0)
Sex, N women (%)	48,926 (43.5)	48,780 (43.5)	146 (35.8)	49,700 (50.8)	49,010 (50.9)	690 (43.2)
AP Images (%)	44,916 (39.9)	44,810 (40.0)	106 (26.0)	5,485 (5.6)	4,557 (4.7)	928 (58.1)
COVID + (%)	312 (0.2)	0 (0.0)	312 (76.5)	1,596 (1.6)	0 (0.0)	1,596 (100.0)
Non-COVID Pneumonia (%)	1,494 (1.3)	1,413 (1.3)	81 (19.9)	4,145 (4.2)	4,145 (4.3)	0 (0.0)

Table 2.1: Summary characteristics of our two main datasets (multi-source and single-source), as well as the summary characteristics of the data sources that are combined to yield these datasets.

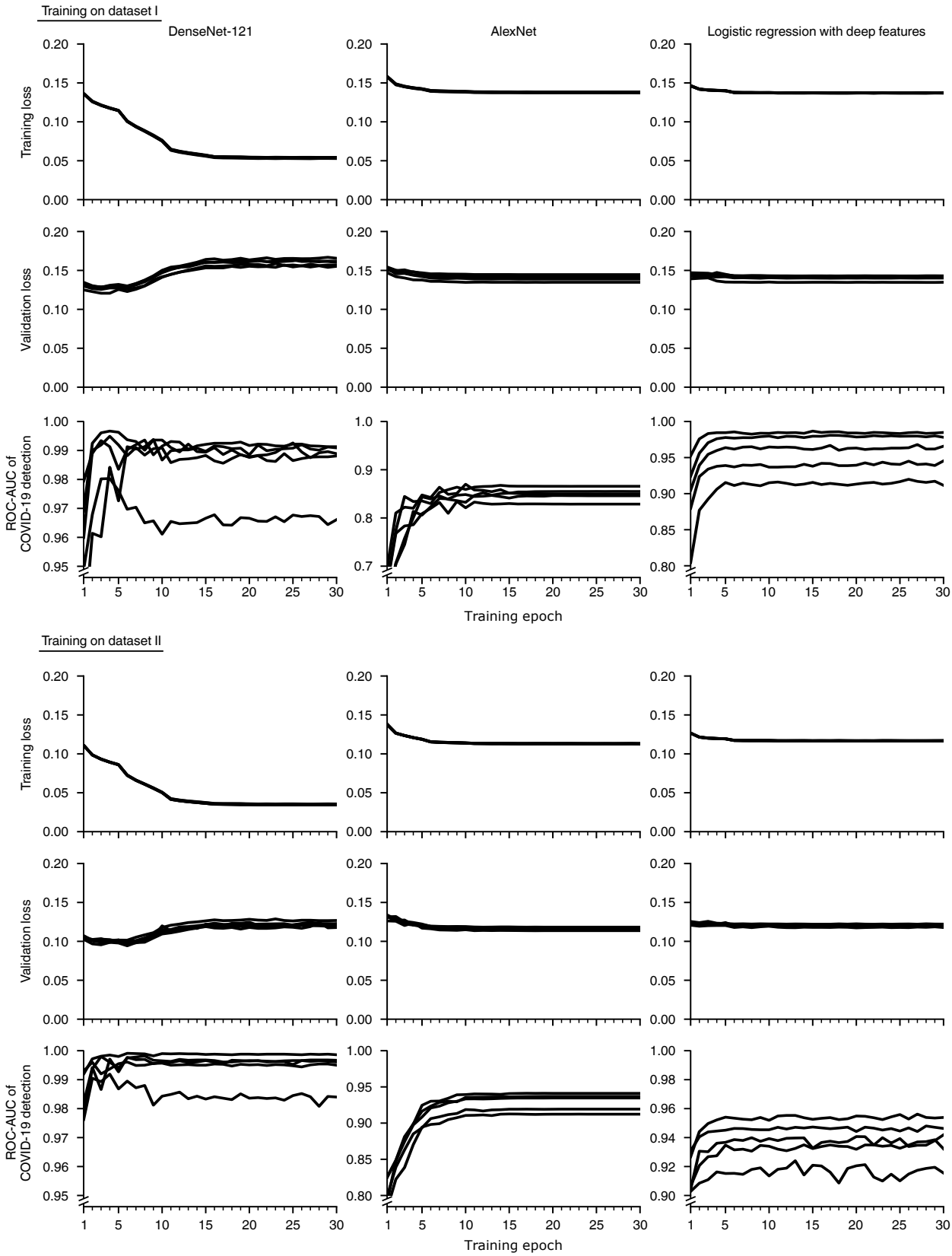


Fig. 2.7 | (Caption next page.)

Fig. 2.7 | (Previous page.) Evolution of metrics that monitor the artificial neural network training process. Training curves are shown for each of 5 random train/validation/test splits of the datasets. During the training procedure, the model is progressively optimized to decrease the training loss, for which we chose the *binary cross entropy*. The validation loss monitors the same metric on a subset of the training radiographs that is held-out from the optimization process (and that is also entirely separate from testing data). Increases in the validation loss may indicate that the model has *overfit* the training data, *i.e.*, the model has memorized the training data rather than learning general principles that apply to new radiographs, such as those in the validation set. To prevent overfitting, we save models when they achieve a maximum in the area under the receiver operating characteristic curve (ROC-AUC) for COVID-19 classification in the held-out validation set, and we use these models for all subsequent analysis. All models were trained for a total of 30 epochs, which was sufficient to attain a maximum in the ROC-AUC of COVID-19 classification. Note that to permit visualization of the maximum in the ROC-AUC of COVID-19 detection, the plots that visualize this quantity feature variable y-axis scales.

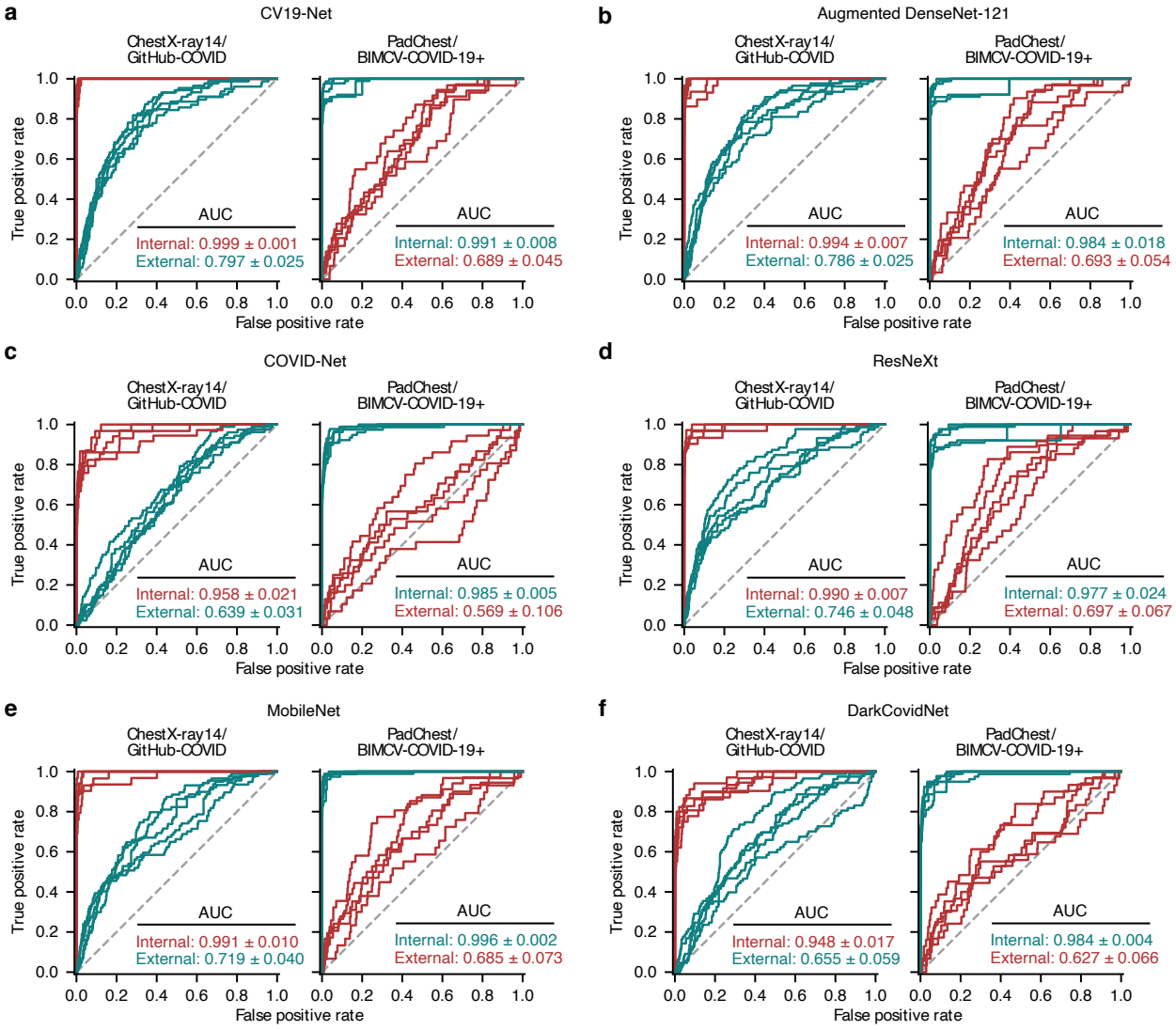


Fig. 2.8 | Generalization performance of models that were specifically designed in previous studies for detection of COVID-19 in chest radiographs as well as additional 'off-the-shelf' architectures. Generalization performance is examined by comparing the performance of each model on held out test data from the same source as the training data (internal) to its performance on test data from new hospitals (external), where we use receiver-operating characteristic (ROC) curves to quantify performance. The architectures designed specifically for detection of COVID-19 in radiographs include CV19-Net,[62] COVID-Net,[56] and DarkCovidNet.[46] The additional 'off-the-shelf' models include ResNeXT[64] and MobileNet.[63] The 'augmented DenseNet-121' is the same as our primary DenseNet-121 model with the addition of the data augmentation scheme from CV19-Net; it therefore represents an intermediate between our primary model and CV19-Net, which is an ensemble of twenty of the 'augmented DenseNet-121' models, and it is provided to disentangle the effects of the CV19-Net data augmentation scheme from the effects of ensembling. For example, while the data-augmented DenseNet-121 provides a small but insignificant improvement in external test set performance over the same network without data augmentation for one of the two datasets (panel b, external test set AUC of 0.76 ± 0.04 vs. 0.79 ± 0.03 before and after data augmentation, respectively, when trained on dataset I, p = 0.22, U = 6 using two-tailed Mann-Whitney U-test; external test set AUC of 0.70 ± 0.05 vs 0.69 ± 0.05 before and after data augmentation, respectively, when trained on dataset II, p = 1.0, U = 13 using two-tailed Mann-Whitney U-test), we find no evidence of significant improvement between the ensembled and single DenseNet-121 models for either dataset (panels a and b, external test set AUC of 0.79 ± 0.04 vs. 0.80 ± 0.02 before and after ensembling, respectively, when trained on dataset I, p = 0.5476, U = 16 using two-tailed Mann-Whitney U-test; external test set AUC 0.69 ± 0.05 vs. 0.69 ± 0.04 before and after ensembling, respectively, when trained on dataset II, p = 0.84, U = 11 using two-tailed Mann-Whitney U-test). Inset values indicate area under the ROC curve (AUC, mean ± standard deviation, n=5).

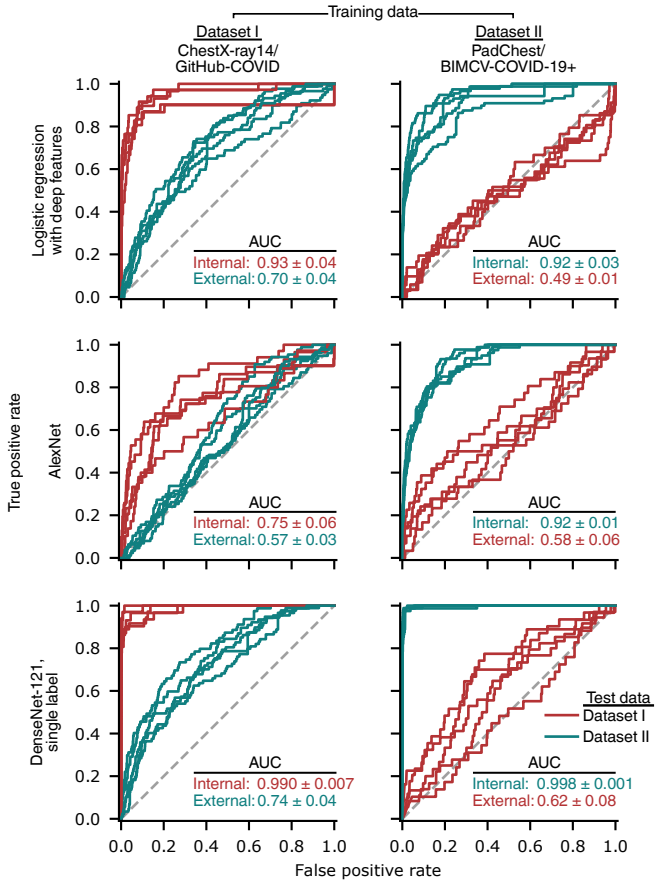


Fig. 2.9 | Generalization performance of models with lower capacity or reduced label information, as measured by receiver-operating characteristic (ROC) curves. The first two rows correspond to models in which the capacity to overfit, which has been implicated in learning of spurious associations [78], has been reduced. The logistic regression with deep features comprises a neural network with the DenseNet-121 architecture that was trained on the ImageNet dataset to derive a set of of 1024 general image features, *i.e.* those output by the penultimate layer of the network, which were used as inputs for a logistic regression; the weights of the neural network were held fixed during training of the logistic regression. The AlexNet models follow the original AlexNet model architecture [19] but with the final 1000-class classification head replaced by a 15-class classification head, corresponding to the 14 ChestX-ray14 labels plus an additional label for COVID-19. The final row represents models with an identical architecture and training scheme to those in the main text, except with only a single output corresponding to presence/absence of COVID-19. Red and teal numbers indicate area under the ROC curves (AUC, mean ± standard deviation, n=5).

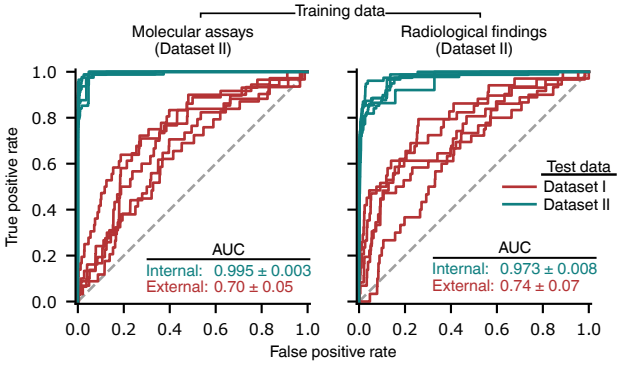


Fig. 2.10 | Evaluation of the impact on generalization performance of *concept shift*, a change in the classification task between the training and testing datasets. In addition to the learning of spurious correlations that do not remain constant between datasets, generalization performance may also drop due to changes in non-spurious correlations between datasets, including a shift in how the labels are generated. In particular, the GitHub-COVID dataset [65], which consists largely of radiographs published in academic articles, may predominantly feature COVID-19+ images with radiological evidence of COVID-19, while COVID-19 labels for the BIMCV-COVID-19+ dataset [72] may be derived from molecular assays (left panel), including reverse-transcription polymerase chain reaction and serology, or from a radiologist's assessment for radiological evidence of COVID-19 (right panel) in addition to confirmation by molecular assays. Specifically, we defined 'radiological evidence of COVID-19' as presence of *COVID-19* or *COVID-19 uncertain* in the radiologist-derived labels of BIMCV-COVID-19+. In the event that poor generalization performance is due to a shift from predicting presence of COVID-19, with or without radiological evidence, in the training data, to predicting radiological evidence of COVID-19 in the test data, generalization performance would be expected to increase substantially. Red and teal numbers indicate area under the ROC curves (AUC, mean ± standard deviation, n=5).

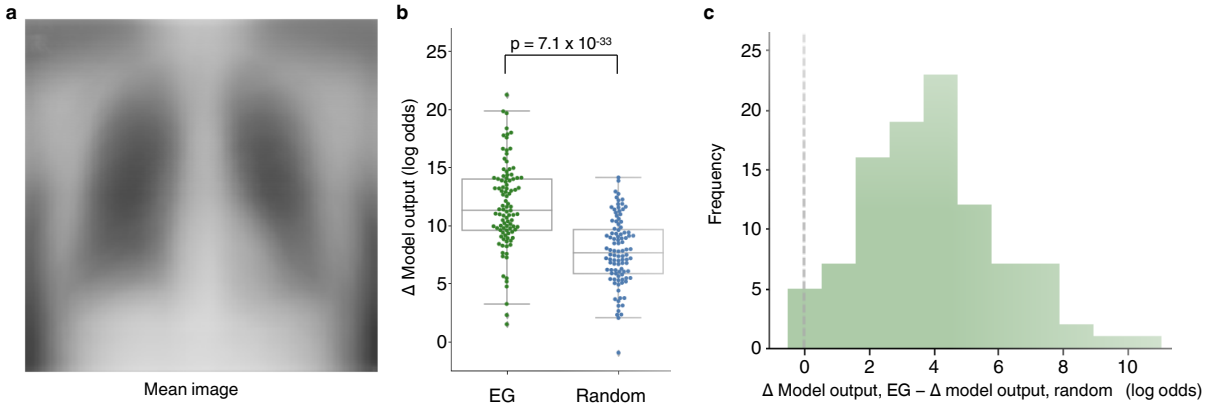


Fig. 2.11 | Ablation tests to assess the importance of pixels that are highlighted by saliency maps. **a**, Average image of COVID-19+ radiographs from dataset I, from which pixels are drawn to 'ablate', *i.e.*, hide, putatively important parts of individual radiographs in our experiment. **b**, Comparison of the change in an AI-based COVID-19 classification model's predictions when pixels are ablated based on their saliency map importance scores or by random. For a randomly chosen subset of radiographs, the 10% of pixels with the highest magnitude expected gradients (EG) scores were ablated by replacing those pixels with the corresponding pixels from the average COVID-19+ image, and as a control, an equivalent number of pixels were replaced at random. Note that in both cases, the model's predicted log odds that the radiograph represents a COVID-19+ patient is expected to increase, since pixels are replaced with pixels from the mean COVID-19+ image. The boxes mark the quartiles (25th, 50th, and 75th percentiles) of the distribution, while the whiskers extend to show the minimum and maximum of the distribution (excluding outliers). Each boxplot marks the 25th, 50th, The p-value is calculated by a two-sided Wilcoxon signed-rank test, n=100 (Siegel's T statistic = 7.69, $p = 1.48 \times 10^{-14}$. **c**, Pairwise comparison of the change in the model's predictions, to assess the superiority of EG relative to random choice at determining important pixels. Since the potential for ablation to change the model's prediction varies from image to image, overlap in the distributions of 'EG' and 'random' in **b** does *not* imply that for any given image random choice is superior to EG. If for any image a random choice of pixels were superior to EG at determining important pixels, we would expect to observe values less than zero in the histogram, which shows image-level, pairwise differences between EG and random choice.

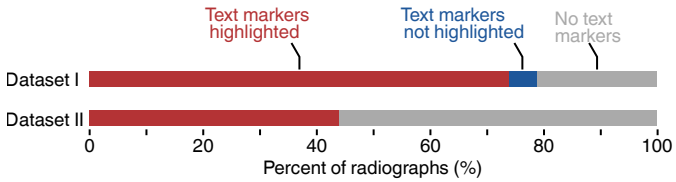


Fig. 2.12 | Analysis of the frequency at which saliency maps highlight laterality markers as important features. To assess the frequency, a random sample of 100 radiographs and their corresponding saliency maps was chosen from each dataset, and each radiograph was manually categorized as (i) contains a laterality marker that is highlighted by the saliency map, (ii) contains a laterality marker that is not highlighted by the saliency map, or (iii) does not contain a laterality marker.

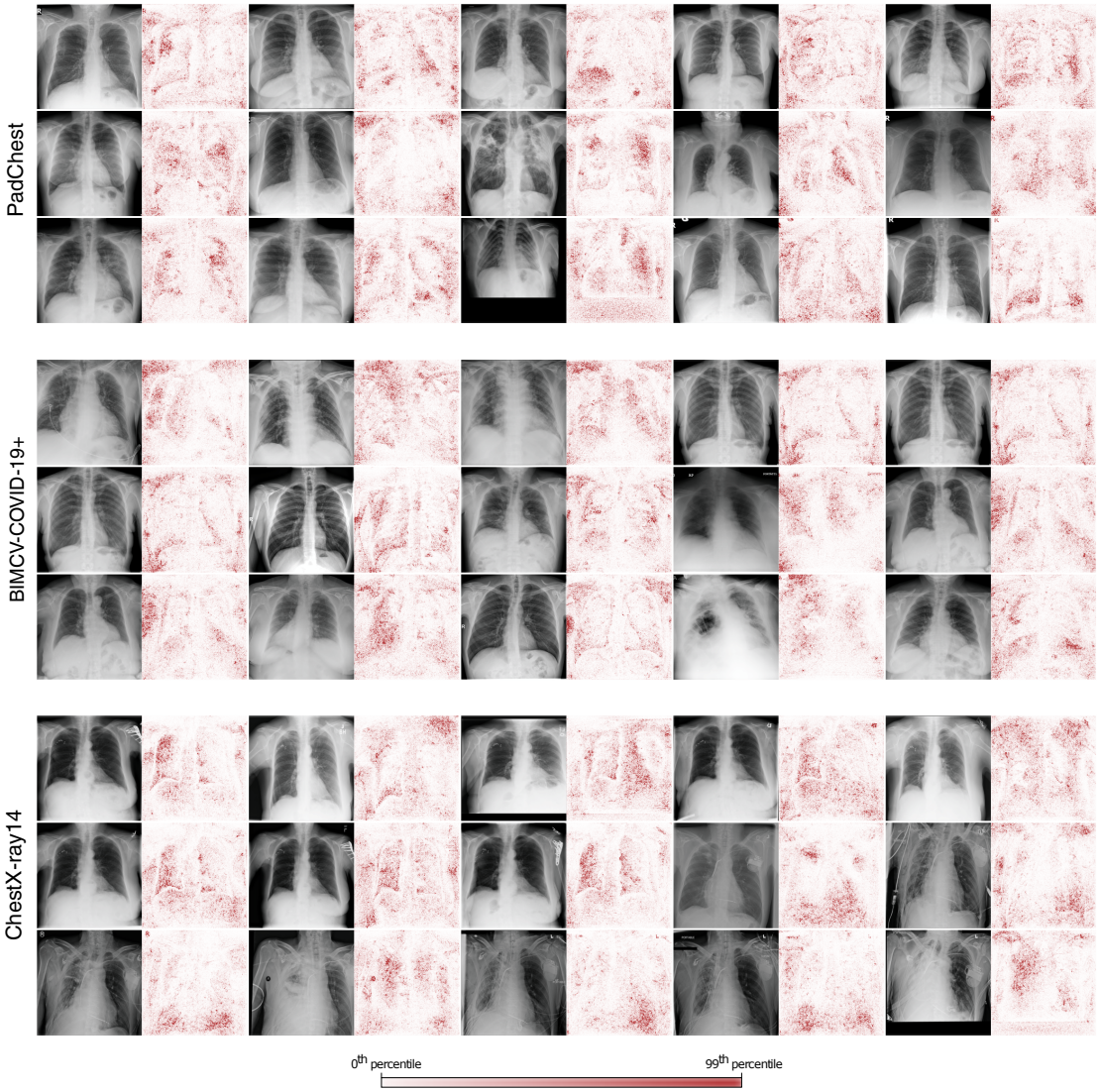


Fig. 2.13 | Saliency maps for 15 radiographs from the PadChest, BIMCV-COVID-19+, and ChestX-ray14 repositories. Across the data sources, saliency maps highlight text tokens and laterality markers (e.g., the first radiograph-saliency map pair in the first row of the PadChest examples, the second-to-last and last radiograph-saliency map pairs in the third row of the PadChest examples, the first four radiograph-saliency map pairs in the second row of the BIMCV examples, and all five radiograph-saliency map pairs in the third row of the ChestX-ray14 examples). For a version of this figure that includes example attributions for the GitHub-COVID repository, see our GitHub repository at `https://github.com/suinleelab/cxr_covid`.

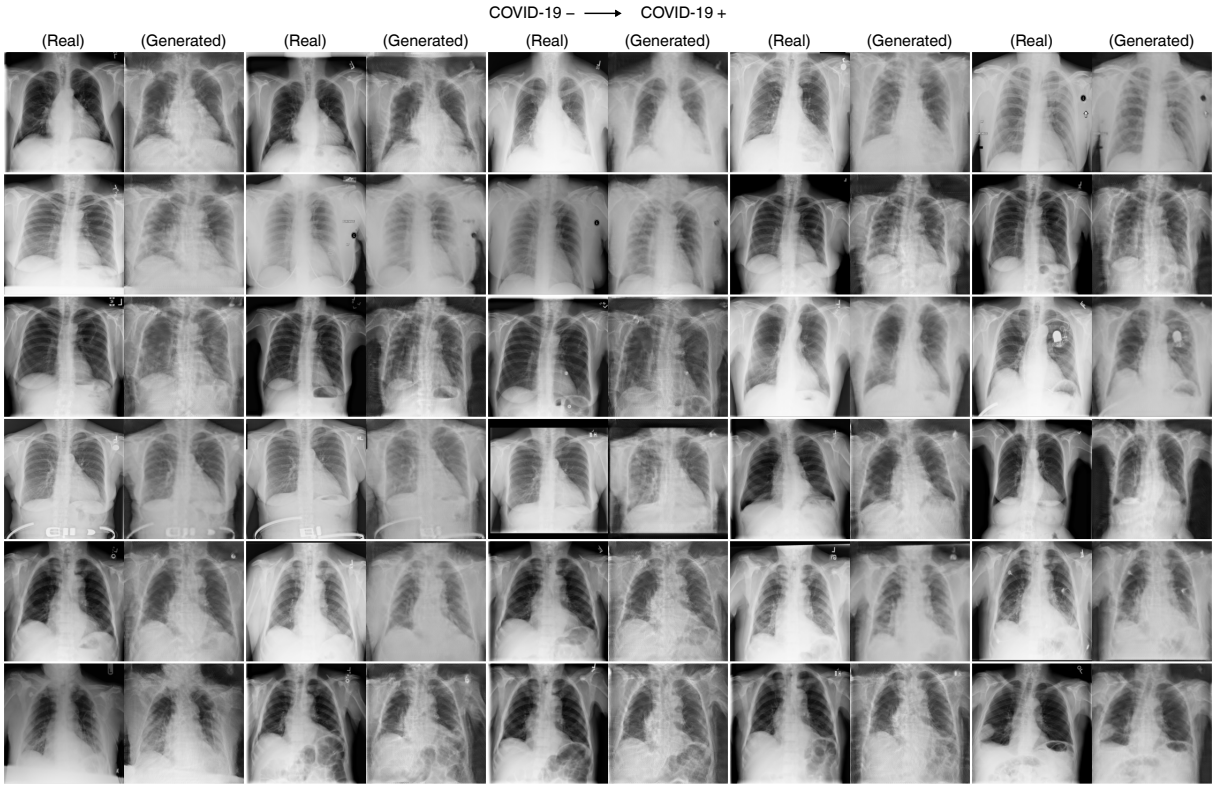


Fig. 2.14 | Examples images generated by a CycleGAN that was trained to alter COVID-19 negative images from the ChestX-ray14 dataset to appear like COVID-19 positive images from the GitHub-COVID dataset and vice versa. See our GitHub repository at `https://github.com/suinleelab/cxr_covid` for a version of this figure that includes images from the GitHub-COVID repository.

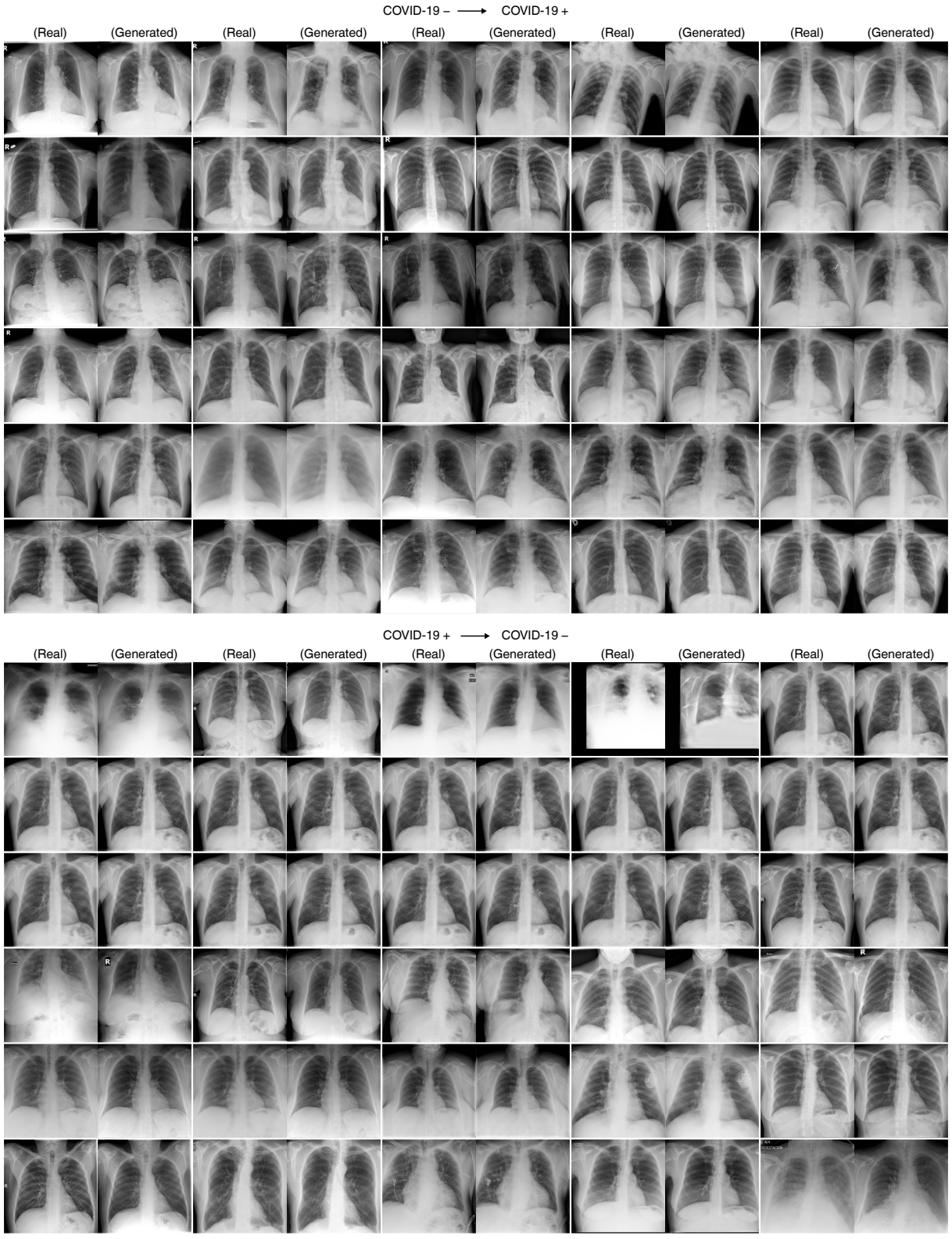


Fig. 2.15 | Examples images generated by a CycleGAN that was trained to alter COVID-19 negative images from the PadCheset dataset to appear like COVID-19 positive images from the BIMCV-COVID-19+ dataset and vice versa.

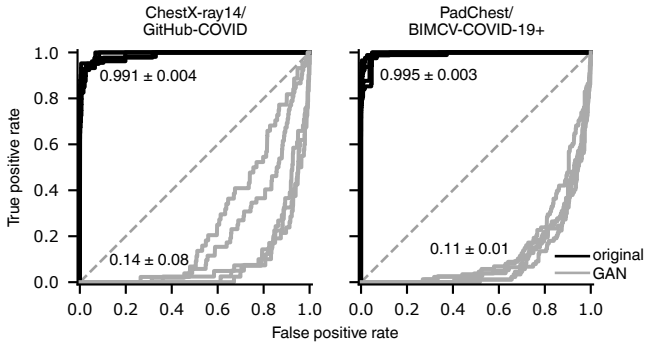


Fig. 2.16 | Evaluation of the extent to which features relied upon by the COVID-19 detection models are altered by the CycleGAN, as measured by the drop in classification performance following transformation by the CycleGAN. A CycleGAN that more reliably alters images such that they appear to the classifier to be of the COVID-19 label opposite their original will achieve an area under the ROC curve (AUC) closer to zero. Inset values indicate AUC (mean ± standard deviation, n=5).

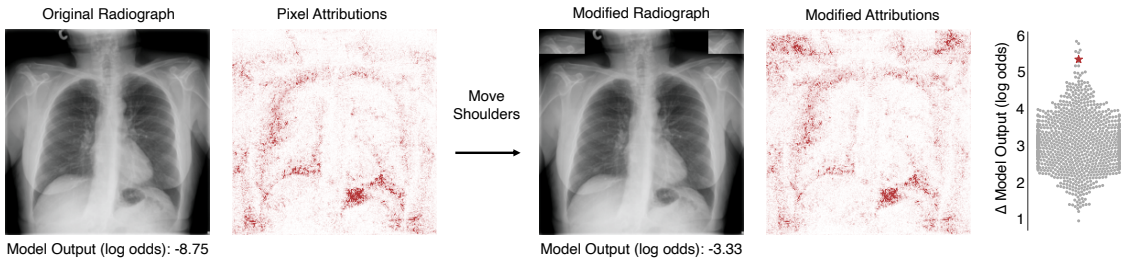


Fig. 2.17 | Additional assessment of the importance of shoulder positioning to an AI model for radiographic COVID-19 detection. The procedure to generate Figure 2d was replicated with a new radiograph; *i.e.*, a patch of the radiograph containing the patient's clavicles was copied to the top corners of the image, and the increase in the model's predicted log odds of COVID-19 was compared to that produced by copying random image patches of the same size ($\Delta = 5.42$, empircal p-value $= 7 \times 10^{-3}$ based on Monte Carlo substitution of random image patches, n=1000) (see Methods Section 2.5).

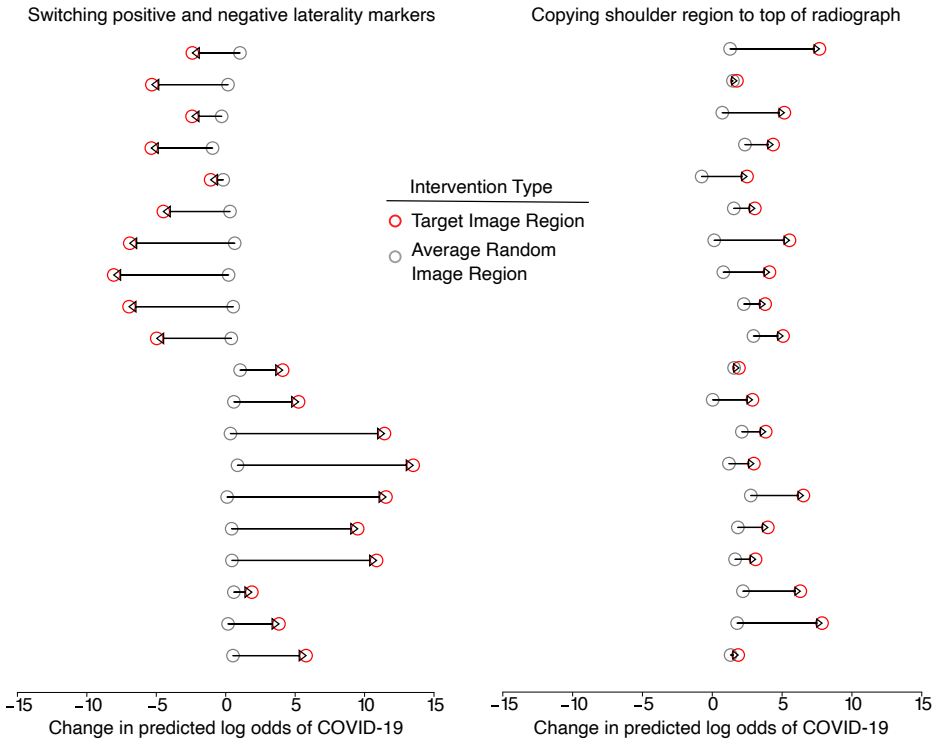


Fig. 2.18 | Population-level analysis of importance of laterality markers and shoulder positioning. Each pair of dots corresponds to an radiograph sampled at random from the larger population, which enables inference of our findings to the population level, despite the infeasibility of completing these experiments for the complete dataset (Dataset II). In each pair, the red dot indicates the difference between the model's predicted log odds of COVID-19 following a targeted intervention on the region of interest and the model's predicted log odds of COVID-19 for the original, unaltered image. The gray dot provides a negative control by repeating the intervention with 1000 random, rather than targeted, image patches of the same size, and then taking the average over the resulting set of changes in the model output. In the left panel, the targeted intervention is to replace the laterality marker on a radiograph from the COVID-19+ repository with a laterality marker on a radiograph from the COVID-19− repository (top 10 radiographs) or vice versa (bottom 10 radiographs), while the untargeted intervention is to swap random image patches of the same size. In the experiments in the left panel, radiographs were sampled at random from the subset with laterality markers. In the right panel, the targeted intervention is to copy the shoulder region of the radiograph and move it to the top of the image, while the untargeted intervention is to copy a random region of the same dimensions as the targeted intervention and move it to a random position. In the experiments in the right panel, radiographs were sampled at random from the full set of images. Swapping of laterality markers between COVID-19+ and COVID-19− radiographs produces a significantly greater change in model output than swapping random image patches ($p=8.9 \times 10^{-5}$, Siegel's T statistic = 0.0, by two-tailed Wilcoxon signed rank test, n=20 random radiographs), and similarly, movement of the shoulder regions to the top of the radiograph produces a significantly greater change in model output than moving random image patches of the same size ($p=8.9 \times 10^{-5}$, Siegel's T statistic = 0.0 by two-tailed Wilcoxon signed rank test, n=20 random radiographs).

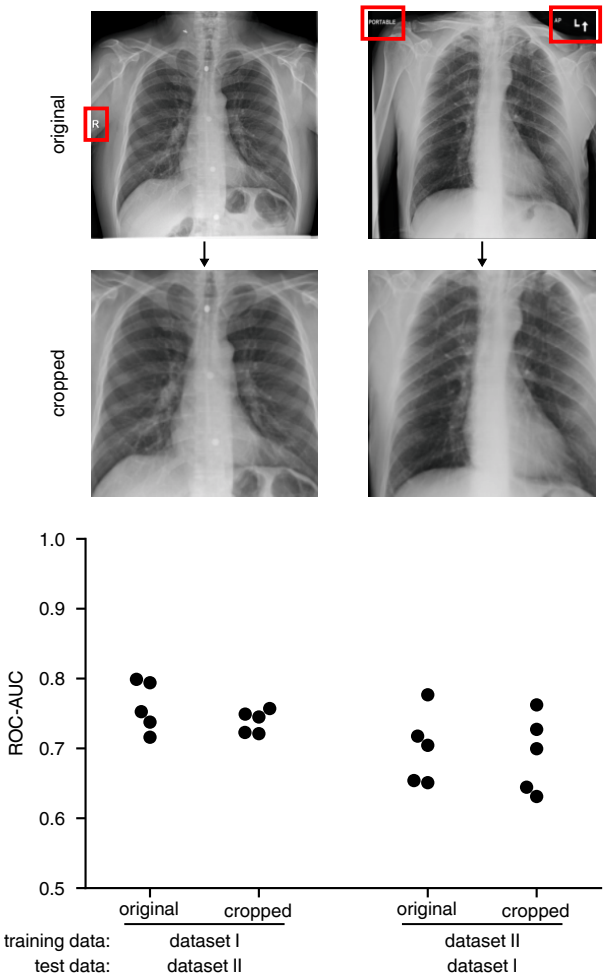


Fig. 2.19 | Evaluation of the extent to which image cropping mitigates shortcut learning. For each dataset, models were trained before and after cropping to the center 75% of the radiograph, which removes from the edge of radiographs the textual markers (red boxes) that may contribute to shortcut learning. Models were then evaluated on an external test set, consisting of radiographs from a different hospital than the training data, to evaluate the generalization performance. Cropping of images did not significantly improve generalization performance based on a one-tailed signed-rank test, where the alternative hypothesis is that the median ROC-AUC of the model trained on cropped images is greater than that trained on the original images (p=0.46 and p=0.60 for models trained on datasets I and II, respectively, based on the Mann-Whitney U-test; corresponding test statistics are U=0.73 and U=0.52, respectively ; n=5 independently trained models).

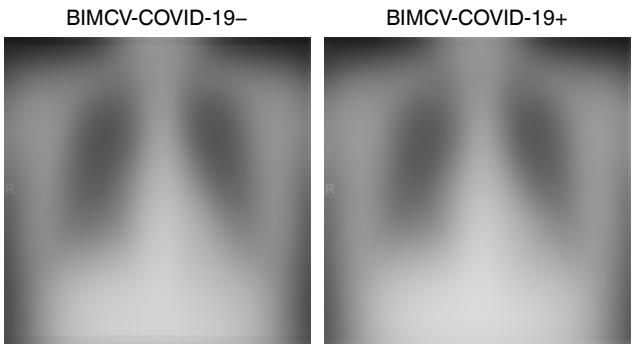


Fig. 2.20 | Average images of the BIMCV-COVID-19– and BIMCV-COVID-19+ repositories. Note consistency in the laterality markers, shoulder positioning, and radiopacity of image borders.

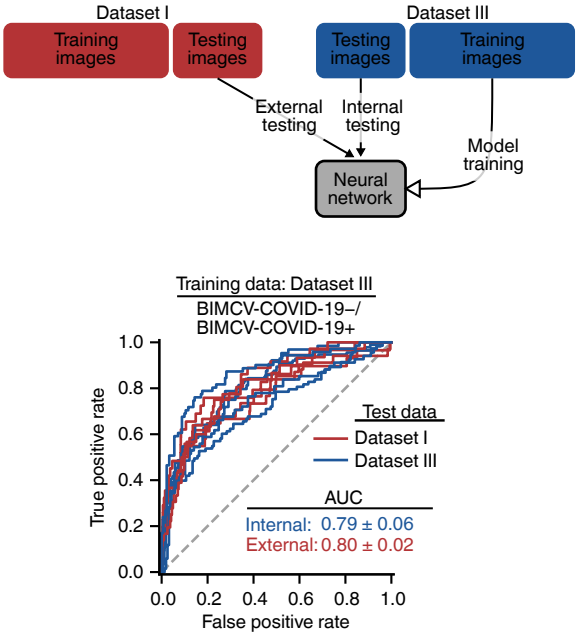


Fig. 2.21 | Evaluation of the generalization performance of models trained on dataset III, via ROC curves. Models are evaluated on both an internal test set (new, held-out examples from the same data source as the training radiographs), and an external test set (radiographs from a new hospital system). Inset numbers indicate the area under the ROC curves (AUC, mean ± standard deviation), where larger area corresponds to higher performance. The difference between internal and external test set performance is the generalization gap.

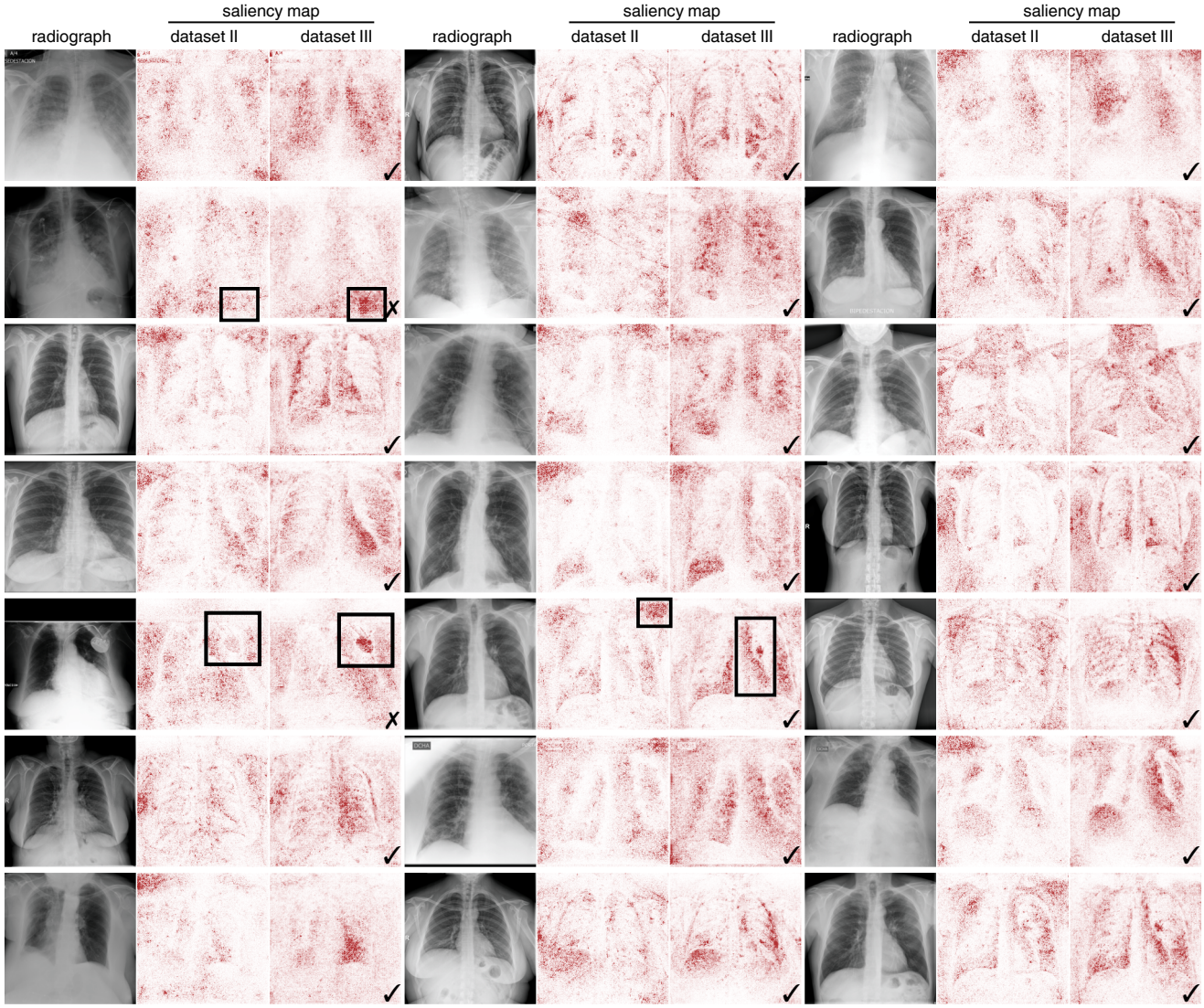


Fig. 2.22 | Evaluation of the extent to which improved training data mitigates shortcut learning, evaluated by comparison of saliency maps for models trained on dataset II and dataset III. For a set of images randomly chosen from the BIMCV-COVID-19+ repository, saliency maps were generated for models trained on Dataset II and models trained on Dataset III, which we expect to contain fewer image factors that spuriously enable COVID-19 positive and COVID-19 negative radiographs to be distinguished. As a basic validation, a model that focuses less on shortcuts would be expected to exhibit saliency maps with increased emphasis on the lung fields and decreased emphasis on the image edges; radiographs for which we judged, on this basis, that the model exhibits less dependence on shortcuts when trained on dataset III than dataset II are marked with a check mark, while radiographs that exhibit greater dependence are marked with an "x". The saliency maps of the two radiographs (out of 21) that did not show improvement exhibit increased attention toward a gastric bubble (black boxes, row two) and a medical device (black boxes; row 5, column 1). While gastrointestinal symptoms are sometimes associated with COVID-19,[97] we were unable to identify reports of an association between gastric bubbles and COVID-19, and therefore judged that this factor likely represents a spurious confound. We additionally annotate an example in which the model exhibits increased attention toward relevant factors (black boxes; row 5, column 2), namely a decrease in attention toward the region above the patient's left shoulder, and an increase in attention toward the left perihilar region.

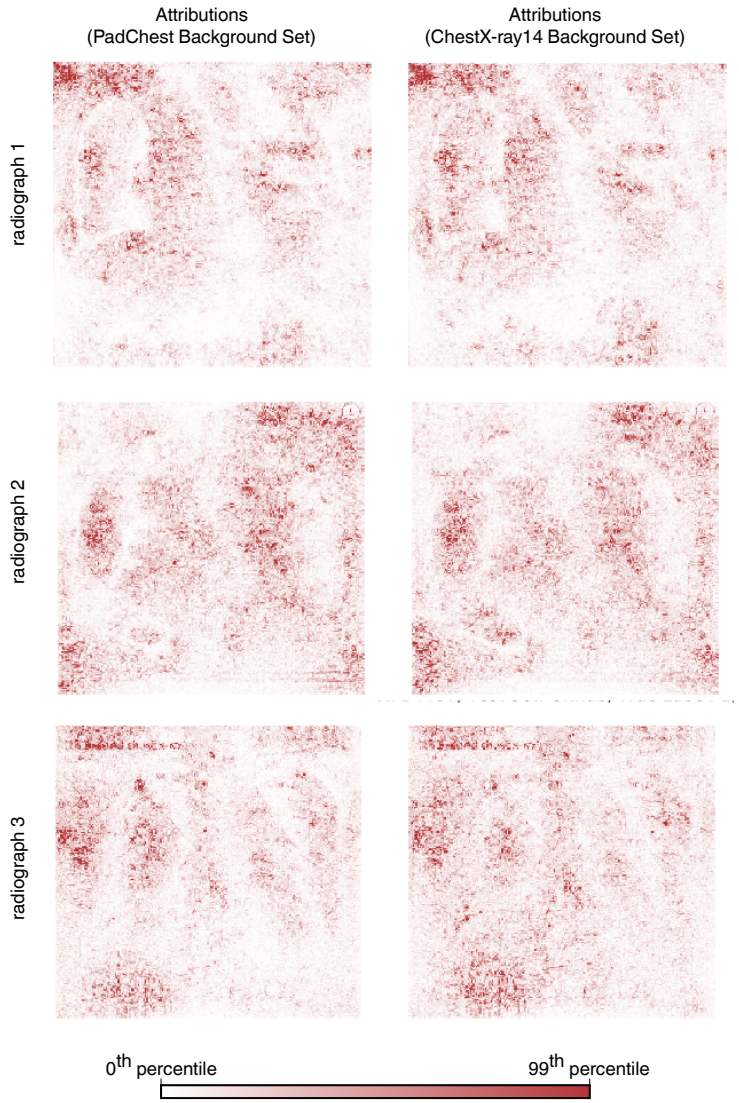


Fig. 2.23 | Comparison of expected gradients saliency maps generated from varied reference distributions, which provide the baseline radiographs from which the expected gradients algorithm integrates.

Chapter 3

Course corrections for clinical AI

Following our work auditing AI systems for detecting COVID-19 in chest radiographs, we encountered reports of similar systems that were being either tested or fully deployed in clinical settings. Considering also that our work in that study had been restricted to preclinical, academic AI systems, we became interested in assessing the contemporary state of the field with respect to how medical AI devices are deployed and regulated. This assessment would then help guide our future investigations to maximize their practical relevance.

These assessments culminated in our perspective piece, 'Course corrections for clinical AI' by Alex J. DeGrave*, Joseph D. Janizek*, and Su-In Lee published in the journal Kidney360.[11] Please refer to the original publication for full details; we provide a brief overview of the key observations here:

Key observations:

1. Most medical AI devices approved in the United States are approved through the '510k' pathway, which provides an easier path to approval for medical devices with a substantially equivalent 'predicate device' previously approved. In many cases, a device marketed as using machine learning or artificial intelligence relies on a predicate device that *did not* use AI, raising the question of how the new AI-based device could provide increased benefit when it is 'substantially equivalent'.
2. Most medical AI devices are tested on retrospective data, and in some cases appear to only use 'internal' test data held out at training time. Given the poor generalization performance of many AI systems, such devices may perform unexpectedly poorly in deployment. This raises a need for new, more rigorous, testing scenarios.
3. We identified a few medical AI devices that we believe are likely to indeed benefit patients, that is, in terms of meaningful clinical endpoints and on the basis of clinical testing. For instance, one AI system triages computed tomography scans of the head to help ensure that patients with emergent conditions (e.g., bleeding in the brain) have their scans read by a radiologist sooner, thus decreasing time to diagnosis. Most clinical evidence of AI devices benefiting patients in terms of meaningful clinical endpoints is limited by the small quantity of patients examined or other aspects of the study design, such lack of randomization and potential for bias.
4. Based on a historical assessment of AI devices, including early computer-aided detection systems for screening mammography, we believe there is an important risk of harm from medical AI devices when they are insufficiently tested. AI systems can have unexpected performance in deployment, and could contribute to issues such as alert fatigue.
5. We believe that AI systems are likely to substantially impact how physicians work. Early studies point toward issues in integrating AI systems into the clinical workflow, and they suggest that depending on their implementation, medical AI systems could generate additional work for physicians, potentially decreasing satisfaction.
6. Overall, we recommend increased involvement of medical practitioners in the AI development workflow, that is, by collaborating with AI researchers to help guide technology toward the most important applications, and by helping test promising new systems.

*equal contribution

Chapter 4

Dissection of medical AI reasoning processes via physician and generative-AI collaboration

This section is adapted from the preprint 'Dissection of medical AI reasoning processes via physician and generative-AI collaboration' by Alex DeGrave, Zhuo Ran Cai, Joseph D. Janizek, Roxana Daneshjou*, and Su-In Lee*, published in medRxiv (doi 10.1101/2023.05.12.23289878, CC-BY-NC 4.0 license). The final form[98] was published under the title 'Auditing the inference processes of medical-image classifiers by leveraging generative AI and the expertise of physicians' in the journal *Nature Biomedical Engineering* (doi 10.1101/2023.05.12.23289878).

4.1 Abstract

Despite the proliferation and clinical deployment of artificial intelligence (AI)-based medical software devices, most remain black boxes that are uninterpretable to key stakeholders including patients, physicians, and even the developers of the devices. Here, we present a general model auditing framework that combines insights from medical experts with a highly expressive form of explainable AI that leverages generative models, to understand the reasoning processes of AI devices. We then apply this framework to generate the first thorough, medically interpretable picture of the reasoning processes of machine-learning–based medical image AI. In our synergistic framework, a generative model first renders 'counterfactual' medical images, which in essence visually represent the reasoning process of a medical AI device, and then physicians translate these counterfactual images to medically meaningful features. As our use case, we audit five high-profile AI devices in dermatology, an area of particular interest since dermatology AI devices are beginning to achieve deployment globally. We reveal how dermatology AI devices rely both on features used by human dermatologists, such as lesional pigmentation patterns, as well as multiple potentially undesirable features, such as background skin texture and image color balance. Our study also sets a precedent for the rigorous application of explainable AI to understand AI in any specialized domain and provides a means for practitioners, clinicians, and regulators to uncloak AI's powerful but previously enigmatic reasoning processes in a medically understandable way.

4.2 Introduction

Medical artificial intelligence (AI) systems have proliferated in recent years,[99] but currently, the scientific and medical community poorly understands what factors influence AI outputs and whether these factors could lead to failures and harm to patients when AI is deployed in practice. The reasoning processes of these high-stakes systems—namely those that rely on neural networks and other complex 'machine-learning' techniques, which automatically learn statistical patterns in large datasets—remain opaque to all stakeholders, including patients, medical providers, regulators, and

*indicates co-senior authorship

even the developers of these AI systems. In principle, a detailed understanding of the reasoning processes of these AI systems could help us predict and prevent AI failures, help us improve AI models, and offer scientific value by contributing to the community's knowledge of AI reasoning processes or their underlying training data. However, to our knowledge, no thorough medically interpretable picture of the reasoning process of a machine-learning–based medical image AI system yet exists. Prior efforts provide extremely limited *peeks* at medical AI reasoning processes,[35,100] typically via techniques that 'sanity check' whether a model is looking in the correct place,[10,43,44,53] and both these and more expressive techniques[37,38] typically suffer from lack of principled, medically informed analysis, precluding a thorough understanding. Indeed, despite technical developments in these explainable AI (XAI) tools, the gap between XAI tool output and pragmatic understanding of an AI system, particularly for image analysis and other 'representation learning' AI systems, remains so large that efforts to apply XAI often miss severe faults in an AI system's logic,[45–48] such as strong dependence on spurious 'shortcut' features.[40,53]

In exploring the reasoning processes of medical image AI, dermatology AI systems serve as a particularly impactful use case, for several reasons: numerous academic papers report high performance;[21,101,102] the first handful of companies have received CE approval to deploy their AI systems on patients in the European Economic Area;[103,104] and multiple developers are working on approval from the United States Food and Drug Administration.[105] Dermatology AI systems, often targeted directly at consumers, may pose particular risks due to the lack of involvement from healthcare providers, potential for bias on skin tone[106] and other sensitive attributes, and heterogeneity of user-acquired images, resulting from variability in lighting conditions, image acquisition devices, and digital processing procedures, none of which are standardized. Simultaneously, the *de facto* standard[10] XAI modality to analyze image models—saliency maps, which highlight the regions of an image that most influence a model's prediction—appear poorly suited to understand dermatology AI systems, which may be best explained in terms of dermatological concepts (e.g., 'multiple colors of pigment', 'atypical pigment networks') that spatially overlap or manifest diffusely throughout an image (Supplementary Fig. 4.5). Explanation of even a single prediction involves simultaneously high levels of technical AI knowledge and dermatology expertise, impeding a global understanding of the AI system's behavior.

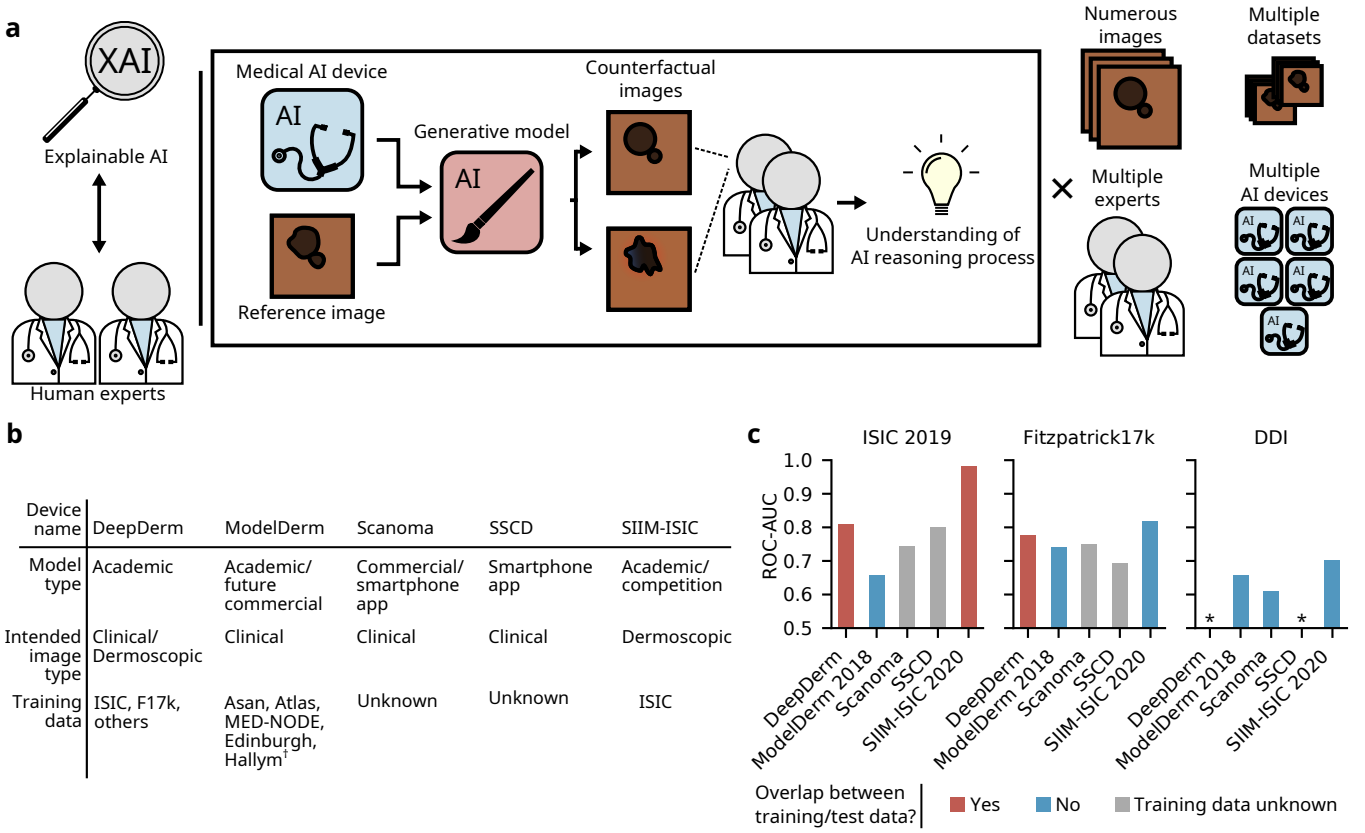
Here, we scrutinize numerous high-profile dermatology AI models to obtain the first thorough, medically interpretable picture of medical image AI reasoning processes. In the process, we showcase our workflow that combines explainable AI with human domain expertise (Fig. 4.1a). We demonstrate solutions to severe practical issues with explainable AI in the imaging domain, including (i) conceptualizing AI behavior in medically meaningful terms, (ii) addressing sampling challenges to form robust conclusions, and (iii) scaling from explanations of individual predictions to a global understanding of an AI system's reasoning processes. At a high level, our workflow involves generative-AI–based synthesis of *counterfactual* images, which circumvent limitations of the *de facto* standard XAI modality (saliency maps) in medical image analysis. Here, we define counterfactuals as images that answer the question 'what realistic alterations elicit a different prediction from the AI?' We constrain the alterations to appear realistic, such that the differences between counterfactuals may be interpreted by medical experts (see also Methods). Our workflow continues with the analysis of thousands of such counterfactual images by dermatology experts, to characterize AI systems in human-understandable medical terms. Throughout the process, we emphasize rigor by mitigating problems of sampling and bias, via examination of numerous images, consideration of multiple datasets, and solicitation of insights independently from two dermatologists via a randomized and blinded analysis.

4.3 Results

Overview of dermatology AI system selection and reproduction

Aiming to best represent the current state-of-the-art in dermatology AI systems, we explored the scientific literature and commercial market, ultimately choosing five AI systems to audit (Fig. 4.1b). These systems span the spectrum from academic to commercial, and include systems already distributed for use by consumers. The five AI systems are: (i) DeepDerm, a previously developed reproduction[106]—using the original training data—of the classifier from a seminal academic publication,[21] which hailed the classifier for its 'dermatologist-level' performance; (ii) ModelDerm 2018,[107] an academic classifier for which a later version (which we were unable to obtain) was CE approved for use in the European economic zone; (iii and iv) Scanoma and Smart Skin Cancer Detection (SSCD), two consumer-facing, smartphone apps; and (v) a 'competition-style' classifier, designed to mimic the key design decisions of the winning model[109] from the 2020 SIIM-ISIC Melanoma Classification Kaggle challenge[110] while circumventing that model's prohibitive computational burden. Authors of additional AI systems declined to make available their full models (*i.e.*, model weights), preventing us from analyzing other high-profile systems.[101,102]

Since these diverse AI systems were trained on highly varied training data, we hypothesize they may exhibit a wide



Device name	DeepDerm	ModelDerm	Scanoma	SSCD	SIIM-ISIC
Model type	Academic	Academic/ future commercial	Commercial/ smartphone app	Smartphone app	Academic/ competition
Intended image type	Clinical/ Dermoscopic	Clinical	Clinical	Clinical	Dermoscopic
Training data	ISIC, F17k, others	Asan, Atlas, MED-NODE, Edinburgh, Hallym†	Unknown	Unknown	ISIC

Fig. 4.1 | Overview of joint expert, XAI auditing procedure and audited AI systems. **a,** Our auditing procedure unites explainable AI with analysis by human experts to understand medical AI systems. Specifically, we leverage generative models to create *counterfactual* images that alter the prediction a medical AI system; analysis of the counterfactuals by human experts (dermatologists) reveals the medical AI system's reasoning processes. We perform the analysis on numerous images from each of multiple datasets, gathering insights from two experts, for each of five different dermatology AI systems. **b,** Key details of dermatology AI systems audited in this study. **c,** Performance of the dermatology AI systems on three datasets, including a dataset (DDI) external to the training data of every system. We examine the area under the receiver operating characteristic curve (ROC-AUC) to focus on the model's internal reasoning processes rather than emphasize the authors' original choices of model calibration. † Asan, Atlas, and Hallym datasets described in ref.;[107] MED-NODE is described in ref.;[108] Edinburgh is available at `https://licensing.edinburgh-innovations.ed.ac.uk/product/dermofit-image-library` *ROC-AUC<0.5 (i.e., worse than random performance).

range of internal reasoning processes, for instance focusing on varied dermatological features or spurious signals. The training data include both dermoscopic images (taken through a specialized dermatological tool that magnifies and enables visualization of deeper layers of the skin) and clinical images (acquired with a digital camera, without the use of a dermatoscope). Dermoscopic and clinical images feature unique profiles of potential signals for AI systems to learn: for instance, dermoscopic images better reveal a lesion's fine details, such as pigmentation patterns, and exhibit unique artifacts, such as ruler markings and dark corner artifacts; clinical images likewise may provide more information on a lesion's context (location, surrounding lesions), in addition to their own characteristic artifacts, such as presence of markings or patient clothing. Dermoscopic images from the ISIC database[110–112] were used to train both DeepDerm and SIIM-ISIC, though the particular subsets of data used for each model differed. DeepDerm also included clinical images in its training set, gathered from numerous online sources. ModelDerm trained on only clinical images, including publicly available images as well as images that were never made publicly available. The training procedures for the smartphone app AI systems have not been published, but based on the wide public availability of dermatology image datasets, we speculate they could have trained at least in part on images from ISIC, Fitzpatrick17k,[113] or other sources. Beyond the variability introduced by differences in training data, additional variation between the models may also arise from their diverse architectures, preprocessing schemes, ensembling, and other computational differences.

We frame our analysis around the clinical task of differentiating melanomas from melanoma look-alikes (*e.g.*, benign nevi, seborrheic keratoses, dermatofibromas), which has received historically recieved great attention within the AI community, and which aligns with the intended use cases of the AI systems. Four of the five AI systems explicitly predict melanoma, while the remaining AI system (DeepDerm) provides a more general prediction of 'benign' or

'malignant.' To model this clinical task, we construct our test data to contain only melanomas and melanoma look-alikes; in this setting, DeepDerm effectively functions as a melanoma classifier, though the DeepDerm's training for a more general task could still impart variation relative to the other systems. We frame our analysis through this narrower problem, which has historically received great attention within the AI community, and which models a well-defined clinical task. Since some classifiers were designed to function on dermoscopic images, others on clinical images, and at least one (DeepDerm) both, we examine all classifiers in each context, using ISIC as our source of dermoscopic images, and Fitzpatrick17k for clinical images (note that, since we are most interested in what alterations cause images to appear more benign or malignant and not benchmarking AI performance, we do not expect our XAI analysis to be sensitive to overlap between the training and test data).[38]

We carefully adapted each AI system for use with our XAI tools, such that all analyses could be performed in a uniform software environment, thus eliminating a potential source of variation. Wherever feasible (*i.e.*, with the exception of SIIM-ISIC), we used the original model weights, to ensure that the original reasoning processes for that AI system could not change. While we suspect that the reasoning process of SIIM-ISIC should closely match the original 2020 SIIM-ISIC Kaggle competition winning model—we use the same training data, training procedure, and test-time image augmentations/ensembling—we intend our audit of SIIM-ISIC to shed light on the influence of these common, performance-boosting techniques rather than to definitively comment on the reasoning process of that original model. We verified our adaptations against the original implementations and achieved close reproduction of the original results; only slight differences arose due to platform-dependent implementation differences in preprocessing or arithmetic (Supplementary Fig. 4.6).

Dermatology AI systems vary in melanoma-detection performance

As a first step toward understanding dermatology AI systems, we evaluated the performance of each system for differentiation between melanoma and melanoma look-alikes. While most AI systems detected melanomas in most datasets with at least limited success, performance was variable and often low. All failed to achieve satisfactory performance in DDI, the only of our three datasets known not to overlap with the training data of any AI system. This performance gap could come from DDI's inclusion of diverse skin tones and rare diseases, but may also arise from other out-of-distribution features.[106] Despite training on no clinical images, SIIM-ISIC—which utilizes ensembling in conjunction with more modern neural network architectures—outperforms all other models on clinical images. Overall, our performance evaluation provides a sanity check that the dermatology AI systems likely rely in part on medically relevant attributes, given that most generalize, albeit to a limited extent, to external datasets. In addition, our evaluation suggests that the five dermatology AI systems likely differ in their internal reasoning processes, since the pattern of performance gains or losses across the three datasets does not hold consistent among the AI systems. The findings from this retrospective analysis (which we do not intend as estimates of real-world performance as might be observed in deployment) motivate further analysis via XAI.

Counterfactual images reveal basis for AI decisions

To understand the reasoning processes of the AI systems, we examined each AI system via an XAI tool: generation of counterfactual images. Counterfactual images reveal the basis of an AI system's decisions by altering attributes of a reference image so as to produce a similar image that elicits a different prediction from the AI system. For instance, consider the case that an AI system predicts a lesion is malignant, while a counterfactual predicted by the AI system to be benign differs in that it features lighter, more uniform pigmentation, and fewer brown spots on the background skin; provided that we ensure all differences in the counterfactual push the AI system's predictions in the desired direction (more benign), we may infer that the classifier uses darker pigmentation of the lesion and brown spots on the background skin as part of its reasoning process (Fig. 4.2a).

To this end, we improved and applied a previously developed[38] technique for generation of counterfactual images, Explanation by Progressive Exaggeration, with updates to enable more rigorous conclusions. In the context of our dermatology AI systems, this techniques enables generation of both 'benign' and 'malignant' counterfactuals from a reference image (Fig. 4.2a). We can then learn from comparing *two opposing counterfactuals*, which guards against potential misinterpretations, should the technique introduce any systematic changes to the counterfactuals. Explanation by Progressive Exaggeration trains a generative AI model in conjunction with an AI system, such that the generative model learns how to alter images to change the AI system's predictions. We train the generative model to create counterfactuals that are similar to the reference image and appear realistic, but differ from the reference image in order to elicit the desired prediction from the AI system. Importantly, since the generated counterfactuals may alter more than one attribute, we updated the technique to ensure that we train the generative model to only change attributes when those changes elicit the desired effect on the AI system's output, whereas the previously published version of this technique may also alter attributes irrelevant to the classifier's output (Supplementary Fig.

4.7). Additional updates enabled generation of higher quality images that retain fine details, such as hair, that might be important for dermatology AI systems (Supplementary Fig. 4.8). We separately trained such generative models for each AI system, for each of the ISIC and Fitzpatrick17k datasets, for a total of ten generative models (Methods, Supplementary Fig. 4.9-4.10); a uniform set of training parameters facilitates comparison between the AI systems (Supplementary Fig. 4.11).

While examination of a single counterfactual pair provides some information about an AI system's reasoning process, to obtain a more complete and rigorous understanding of the AI systems and enable direct comparisons between systems, we systematically interrogated thousands of counterfactual images, in a randomized and blinded fashion (Fig. 4.2b). We began our analysis by pre-screening the counterfactuals, to ensure we only examined high-quality counterfactuals and to facilitate comparisons between AI systems. We excluded counterfactuals that failed to produce the desired output from our AI systems (*i.e.*, we ensured the 'malignant' and 'benign' counterfactuals lie on the correct sides of the decision threshold), or that contained visual artifacts (*e.g.*, 'water-droplet–like' artifacts[122]), as judged by dermatologists. Two dermatologists then independently annotated each counterfactual pair, which was randomized and blinded to reduce bias. To learn whether the dermatologists' general impressions of the counterfactuals agreed with each AI system regarding what appears more or less malignant, we first inquired, 'Which image appears most likely to represent a melanoma?' We then asked the dermatologists to record individual image attributes that differ between the 'benign' and 'malignant' counterfactuals, such that we could learn which attributes each AI system uses, and how it uses them (Supplementary Fig. 4.12).

We aggregated the dermatologists' insights over thousands of counterfactuals to determine the reasoning process of each dermatology AI system. We conceptualize the reasoning process as swayed toward a benign or malignant prediction by key attributes identified as differing in counterfactual pairs; our analysis provides the typical direction of an attribute's effect, based on whether that attribute was predominant in the benign or malignant counterfactuals, as well as an approximate idea of the extent of the effect, based on the frequency with which dermatologists observed that attribute differing in counterfactuals. Note that we expect this frequency to depend on multiple factors, including the fraction of the dataset to which that attribute is relevant, inductive biases of our generative models, and perhaps a combination of a dermatology AI system's sensitivity to an attribute and the sensitivity of our evaluators in detecting that attribute (which may be at odds, in the case of a visually subtle change that sizeably affects a prediction). Our analysis reveals that the AI systems focus on both medically relevant and putatively spurious attributes, and exhibit considerable heterogeneity in how they interpret those attributes (Fig. 4.2c).

A detailed view of medical AI reasoning

Our counterfactual analysis highlights the pigmentation of lesions as a key attribute in determining the predictions of all dermatology AI systems examined, for both dermoscopic and clinical images. In all cases, 'darker pigmentation' surpassed all other attributes in frequency, with dermatologists noting this change in the majority of counterfactual pairs. Consistent with dermatologists' interpretation of more darkly pigmented lesions, dermatology AI systems typically associate darker pigmentation of lesions with increased likelihood of melanoma; the only exception is ModelDerm when evaluated on dermoscopic images—an image type upon which this model was never trained. Dermoscopic counterfactuals from a subset of the dermatology AI systems (DeepDerm, Scanoma, and SSCD) also displayed atypical pigment networks, featuring these in the more 'malignant' images, in agreement with dermatologists' use of this attribute during pattern analysis of melanocytic lesions.[114,115]

Our counterfactual analysis suggested that dermatology AI systems also depend on a variety of other attributes of the lesion, many of which dermatologists also consider when analyzing melanocytic lesions. In both dermoscopic and clinical images, counterfactuals from all AI systems varied the number of colors in a lesion, typically associating a greater number of colors with predictions of malignancy.[116] Some AI systems, most prominently SIIM-ISIC, also elicited counterfactuals with blue/white veils, which has previously been reported as a specific finding for melanoma.[117,118] Other attributes of the lesion that may factor into the AI systems' decisions include presence of structureless areas or regression in dermoscopic images, and uneven pigmentation or erythema in clinical images. Aside from erythema, which varies between a benign or malignant signal depending on the AI system, these attributes typically associate with the malignant counterfactuals. Their frequency, however, varies considerably between AI systems, pointing out heterogeneity in the systems' reasoning processes.

Analysis of each AI systems's top attributes (Supplementary Fig. 4.13-4.14) revealed additional lesional attributes highlighted by counterfactuals from only a subset of the AI systems In dermoscopic images, these attributes included patchiness (DeepDerm and SSCD), strawberry pattern (ModelDerm), white spots (SSCD), prominence of follicles or pores (SSCD), white striae (SIIM-ISIC), and scale (SIIM-ISIC). In clinical images, these attributes included erosion or ulceration (DeepDerm and Scanoma), nodular or papular appearance (ModelDerm), uneven borders (ModelDerm),

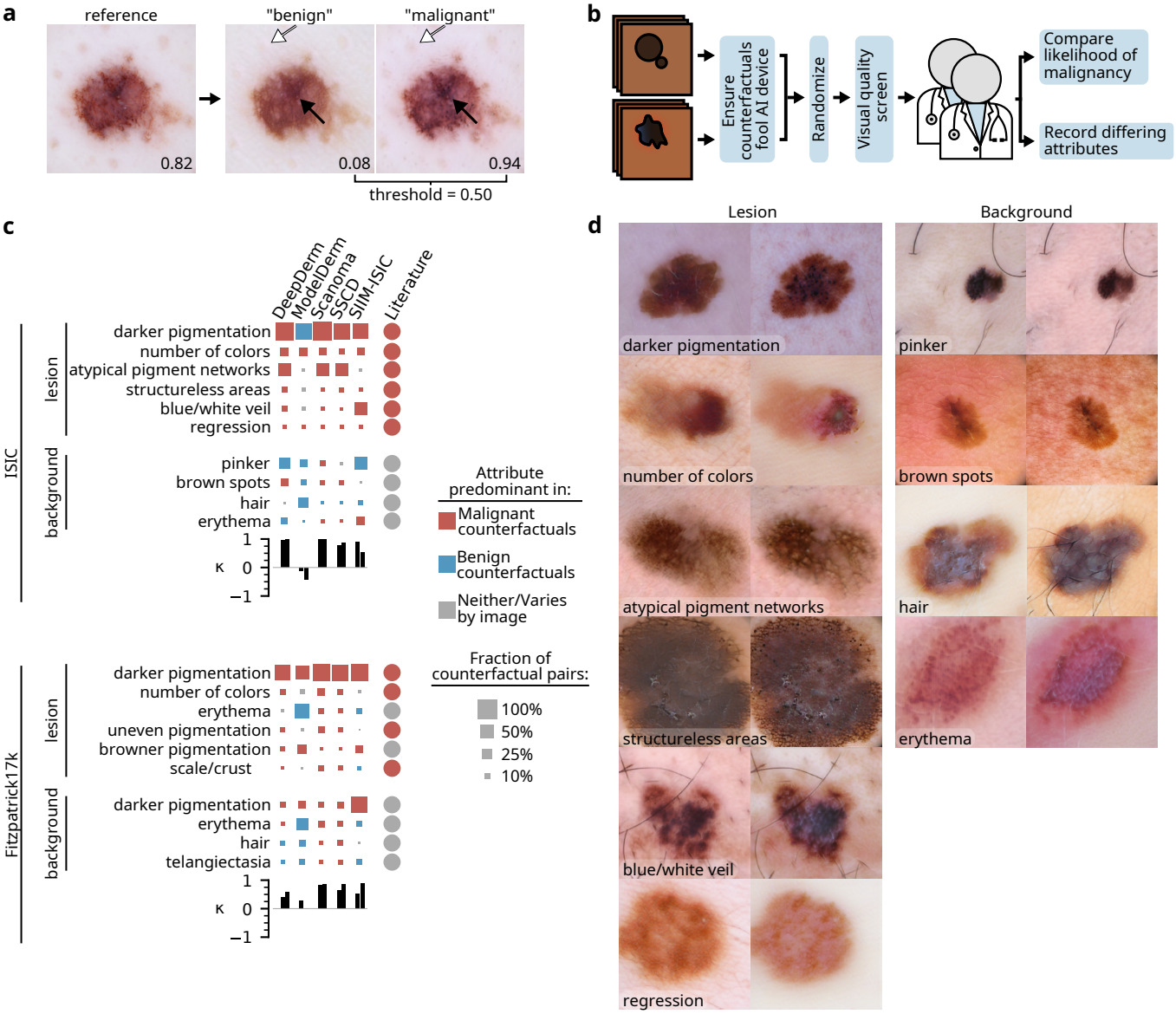


Fig. 4.2 | Joint expert, XAI auditing procedure reveals reasoning processes of dermatology AI systems. **a,** Given a reference image and an AI system to investigate, our generative model produces 'benign' and 'malignant' counterfactuals, which resemble the reference image but differ in one or more attributes (e.g., pigmentation, solid arrows, and dots on the background skin, open arrows). When evaluated by the AI system, the counterfactuals' outputs lie on opposite sides of the decision threshold. Higher values indicate greater likelihood of malignancy, as predicted by an AI system (Scanoma). **b,** To obtain robust conclusions, dermatology experts evaluate numerous counterfactuals after pre-screening and randomization of the images. **c,** Attributes identified by our joint expert-XAI auditing procedure as key influences on the output of dermatology AI systems. For each attribute/system pair, we count the proportion of counterfactual pairs in which experts noted that attribute differs; we display the global top-10 attributes as determined by lowest rank-sum over all AI systems, then group by attribute category ('Lesion' or 'Background'). Based on expert evaluation of whether the attribute was present to a greater extent in the malignant or benign counterfactual of each pair, we determine whether that attribute was 'predominant' in benign or malignant counterfactuals, *i.e.*, present to a greater extent in benign (malignant) counterfactuals in at least twice as many images as malignant (benign) counterfactuals. The size of each square (the 'fraction of counterfactual pairs') is then determined as the proportion of counterfactual pairs with a difference noted in the predominant direction, averaged over both readers. For comparison, we specify how human dermatologists use each attribute ('Literature'), based on our review of the literature[114–120] combined with expert opinion from two board-certified dermatologists; see Discussion for additional information. Bar charts indicate Cohen's κ values for agreement between each expert and the AI system, where each is asked which image in each counterfactual pair appeared more likely to be malignant. **d,** Examples of counterfactuals that differ in each of the top ten attributes identified in the ISIC data; the attribute is present to a greater extent in the right image of each pair. For conciseness, some attribute names were shortened; refer to Supplementary Table 4.1 for full names. Images adapted with permission from ref.[112] Combalia et al., ref.[111] Tschandl et al., and ref.[121] Codella et al.

and the shininess of a lesion (SIIM-ISIC).

Typically, inter-reader variability did not result in conflicting conclusions about the presence or direction of an attribute's effect (Supplementary Fig. 4.15).

Our counterfactuals indicate attributes of the background skin also influence the dermatology AI systems, and in comparison to attributes of the lesion, often elicit more diverse responses among the systemss: Counterfactuals for multiple AI systems display brown spots on the background skin, and these variably associate with either malignant or benign predictions, depending on the system. Hair typically associates with benign counterfactuals in dermoscopic images, but can also associate with malignant counterfactuals in clinical images. Reticulation of the background skin associates with the benign counterfactuals of Scanoma and ModelDerm (Supplementary Fig. 4.13), but is rarely highlighted by the counterfactuals of other systems. Erythema or telangiectasias of the background skin also feature prominently in the results of our counterfactual analysis, and the effects of these attributes vary both between AI systems and within an AI system, depending on whether an image is clinical or dermoscopic. Finally, counterfactuals highlighted the 'pinkness' of background skin as influencing AI systems' decisions, particularly in dermoscopic images. In contrast to erythema, this attribute often applies uniformly across an image (Fig. 4.2d), consistent with effects of lighting or an image's color balance. Similarly, we recorded overall darker images and cooler color temperatures as influential for one classifier (SIIM-ISIC). Similar to other background skin attributes, lighting or color balance changes may sway an AI system toward a more benign or more malignant prediction depending on the system. Aside from brown spots on the background skin, which could be interpreted as sun damage,[123] we were unable to identify dermatological literature that establishes these attributes of the background skin as signals commonly used by dermatologists.

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

In general, AI systems and human dermatologists agreed on which image in the counterfactual pair most likely depicted a malignancy. The exception, ModelDerm, exhibited negative Cohen Kappa values compared to dermatologists on dermoscopic images, in alignment with the unique profile of attributes highlighted in our analysis. This system also agreed poorly on clinical images, again coinciding with its focus on a unique profile of attributes. Curiously, Scanoma achieved the best agreement with dermatologists on both datasets, despite other AI systems achieving higher predictive performance (even when that performance was on external data and therefore not inflated by train-test overlap, e.g., SIIM-ISIC with Fitzpatrick17k; Fig. 4.1c).

Validation of insights from counterfactuals

While we engineered our counterfactual generation procedure to ensure that detected attributes indeed influence AI systems' predictions, we performed additional analyses to verify these conclusions. Ideally, we may confirm our findings by performing a targeted intervention to experimentally modify a single attribute of an image, in a well-defined fashion, then monitor the intervention's effect on each AI system's prediction. While existing techniques such as cycleGANs[39] or manual image editing do not enable reliable modification of most attributes detected in our analysis (*e.g.*, addition or removal of atypical pigment networks without altering other attributes), transformation to a suitable color space (CIELUV)[124] enables programmatic modification of the color of an image, permitting us to experimentally produce images that are more or less 'pink', an attribute detected as influential to most classifiers (Fig. 4.2c and Fig. 4.3a). We shifted the color (i.e., the u' and v' chromaticity coordinates in the CIELUV color space[124]) of each image in the ISIC dataset, then monitored how each AI system's prediction changed for a range of colors (Fig. 4.3b).

These experimental modifications of image color and their impact on the predictions of the AI systems recapitulates the trend observed in our previous analysis of counterfactual images (Fig. 4.3c; compare to 4.3a): e.g., pinker images elicit more benign predictions from DeepDerm and more malignant predictions from Scanoma. Multiple factors including the 'sensitivity' of an AI system to changes in an attribute determine the relative frequency of an attribute among counterfactuals (Fig. 4.3a); thus, magnitudes are not directly comparable (see Results: 'Counterfactual images reveal basis for AI decisions'). This experiment validates that the attributes identified in our previous analysis of counterfactual images indeed influence the output of the AI systems in the direction described by the counterfactual analysis. In addition, this experiment validates our interpretation of 'pinker background skin' as a global change in lighting or color balance. Indeed, our experimental procedure mirrors computational techniques used to perform white balancing (correction for chromatic adaptation) in digital cameras[125] and highlights how changes to lighting or camera

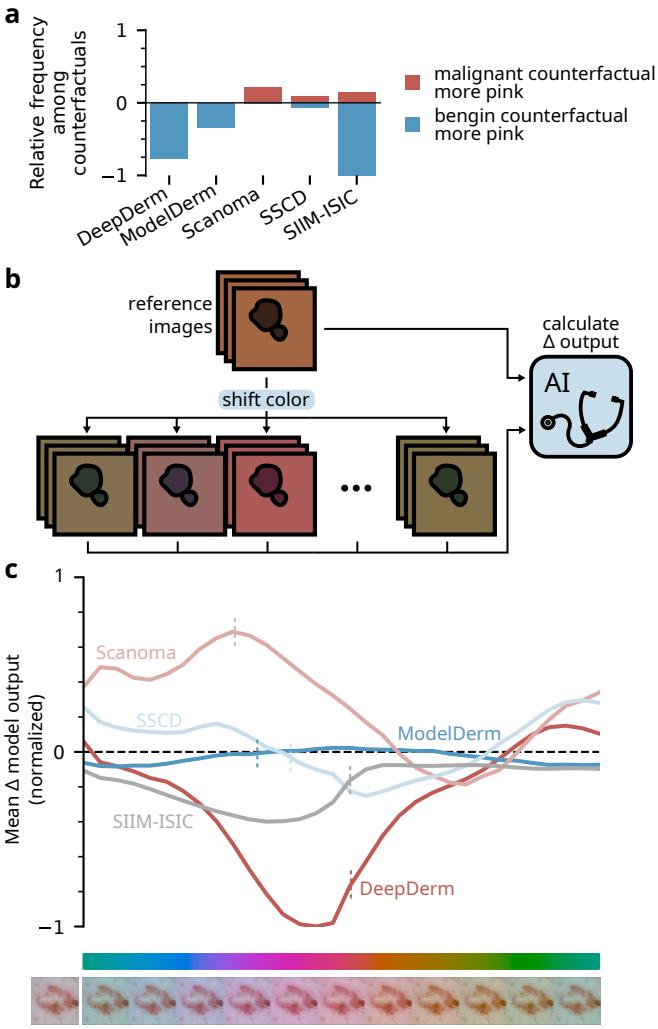


Fig. 4.3 | Experimental validation of findings from expert analysis of counterfactual images. **a,** Frequency with which experts noted that either the benign or malignant image in a pair of counterfactuals displayed a pinker background; this view details our observations from the ISIC dataset summarized in Fig. 2c, in the row 'B: pinker'. The vertical axis is normalized relative to the maximum observed frequency, that is, 42% of counterfactual pairs from SIIM-ISIC. **b,** Experimental setup used to verify the importance of a pink tint to the AI systems' predictions. We programmatically color-shifted each image in the ISIC dataset ($n = 20260$) by modifying its chromaticity coordinates in the CIELUV color space (see Methods), then compared each AI system's predictions between the original and color-shifted images. **c,** Sensitivity of each AI system to programmatic color shifts, mirroring observations from our counterfactual experiments regarding the effect of pinker tints on the AI systems' predictions. The vertical axis is normalized relative to the maximum change in AI system output, *i.e.*, a decrease of 0.17 with DeepDerm. Vertical dashed lines indicate the mean change in chromaticity (color) among counterfactual pairs annotated as differing in their pink tone. Example color-shifted images (below color bar) display the extent of the color shift; the reference image, adapted with permission from the ISIC archive,[121] appears at far left.

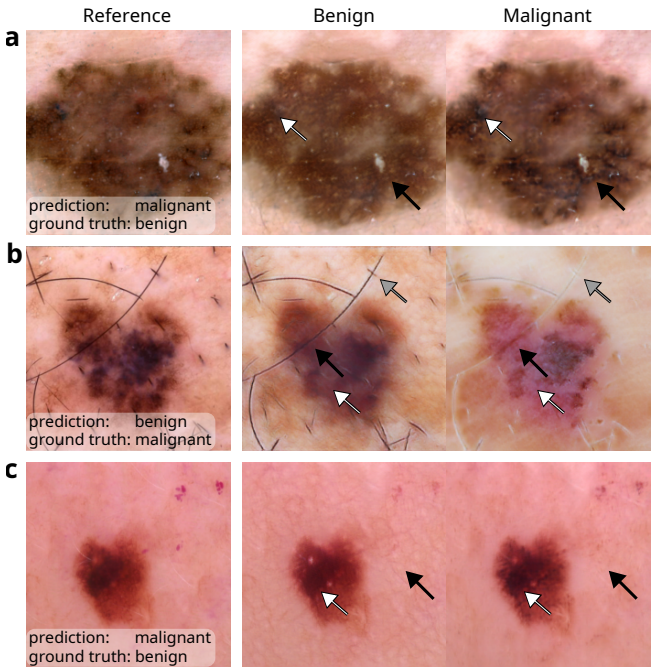


Fig. 4.4 | Explanations of failure cases of dermatology AI systems, illustrating key findings from our systematic analysis. **a,** Presence of atypical pigment networks (black arrows) and darker pigmentation (white arrows) contributed to a false positive prediction from Scanoma. **b,** Lack of more colors of pigment may have contributed to a false-negative prediction from ModelDerm. Curiously, ModelDerm may have also required lighter pigmentation (black arrows), increased erythema (white arrows), and less hair on background skin (gray arrows) to correctly predict this image pictures a melanoma. **c**, Lack of prominent reticulation on the background skin (black arrows), alongside darker pigmentation of the lesion (white arrows), contributed to another false positive prediction from Scanoma. Images adapted, with permission, from the ISIC archive.[111,112,121]

settings might affect AI dermatology systems' predictions in undesirable ways.

Counterfactuals explain failure cases

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

The reliance of dermatology AI models on the pigmentation of a lesion can lead to failures that are 'reasonable', in that they might also be expected from human dermatologists (Fig. 4.4a): for instance, while presence of atypical pigment networks and darker pigmentation lead one AI system to predict a lesion was malignant, it turned out to be benign; indeed, authors of this present study who practice dermatology find this lesion concerning for the same reason, and would have opted to biopsy the lesion.

In other cases, dermatology AI models rely on potentially relevant attributes of an image, but use these attributes incorrectly. ModelDerm misclassified a malignant lesion as benign, and examination of the corresponding counterfactuals revealed attributes such as darker pigmentation of the lesion and absence of erythema as influential for this decision (Fig. 4.4b). However, dermatologists would not typically associate darker pigmentation with decreased likelihood of melanoma, and the distribution of erythema does not match the 'pink rim' sometimes associated with melanoma.[119]

Dermatology AI systems also utilize likely irrelevant attributes in their reasoning process, including associating hair on background skin with benign lesions (Fig. 4.4b). In another example (Fig. 4.4c), a classifier misclassifies a benign lesion as melanoma in part due to an attribute of the background skin, namely lack of prominent reticulation.

4.4 Discussion

Relative to previous techniques to analyze medical image AI classifiers, our framework provides numerous advantages, which together enable us to present a detailed view of the reasoning processes of AI systems for medical images. Whereas the *de facto* standard XAI technique for image models, saliency maps, best reveals the importance of

localizable attributes, our discovery of dependencies on numerous overlapping, textural, and tonal changes to an image showcases the importance of our use of XAI based on counterfactual images, and highlights limitations of previous work that relied only on saliency maps.[10] In fact, we surmise that most attributes identified by our framework, such as darker pigmentation of lesions, number of colors in a lesion, presence of erythema, pigmentation patterns, etc., would be unlikely to be identified by saliency maps. Our framework also improves upon previous efforts[37,38] to analyze medical image AI systems via counterfactual images. In contrast to other generative techniques[37,38] for counterfactual generation (including the original Explanation by Progressive Exaggeration) or simply comparing real images predicted as benign and malignant, our method enables the inference that each attribute that differs in a benign/malignant pair is indeed important for the AI classifier's predictions (Supplementary Fig. 4.7). Our method also offers more detailed reproduction of fine-grained features such as hair (Supplementary Fig. 4.8), which we discovered to influence some AI classifiers. Perhaps more importantly, our framework introduces a means to translate XAI outputs to a human-understandable, medically meaningful form, namely via systematic, randomized, blinded analysis by medical experts. Particularly for a high-stakes application such as medical decision-making, we contend that such a medically-grounded understanding offers greatest potential for actionability.

We find that dermatology AI classifiers leverage a number of medically meaningful attributes found within lesions—including attributes related to a lesion's pigmentation—in a manner consistent with human experts. Dermatology AI classifiers also rely on numerous attributes with debatable medical relevance and unclear desirability. Brown spots on the background skin may signify a patient's age or history of sun exposure (a risk factor for melanoma[123]) but are not in any established melanoma diagnosis guidelines. Our observation of this attribute is consistent with prior works that suggested AI classifiers may rely in part on perilesional sun damage when examining actinic keratoses or Bowen disease;[126,127] a later study further corroborated that perilesional sun damage also enhances human diagnosis of actinic keratoses and, more directly relevant to the prediction task in our study, also provided evidence that melanomas too display perilesional sun damage more frequently than benign nevi.[128] Erythema, particularly in a 'pink rim' distribution around a lesion,[119] has been associated with melanoma, but also with benign melanoma look-alikes such as irritated seborrheic keratoses.[120] Hair may suggest a lesion's location on the body while skin grooves may provide clues on a lesion's location (e.g., acral), the patient's age, or history of sun exposure. Lighting conditions or color balance also influence many dermatology AI classifiers, and we surmise these almost certainly undesirable dependencies arise from spurious differences in image acquisition or preprocessing. The examined AI classifiers display considerable variability in their reasoning processes, especially with respect to their use of background attributes. While such variability might be partially explained by one model (DeepDerm) differing in its intended task (differentiation between malignant and benign lesions in general, as opposed to melanomas and benign melanoma look-alikes), the remaining models differ in reasoning processes despite sharing a common task. Beyond the fundamental scientific interest of this detailed characterization of AI reasoning processes, our approach could be used by AI developers to improve their models and to inform stakeholders on the trustworthiness of medical AI classifiers.

This methodology can help uncover idiosyncratic failure modes of AI, with implications for its regulation and medical use. We expect distributional shifts in medical AI to be common–especially in dermatology AI, given the diversity of image acquisition devices, lighting conditions, skin appearances across demographics, and lack of implemented image standards. Our findings suggest that common distributional shifts, such as changes in lighting or color balance, will alter AI performance. Thus, we caution potential users of such classifiers that a classifier's advertised performance, which is often estimated in a well-circumscribed setting, may not be achieved in real-world use.[106] Our findings also imply that regulators should scrutinize the distribution of data on which a classifiers is evaluated, with particular attention toward (i) ensuring it well reflects the intended deployment distribution, and (ii) considering differential performance across subgroups (e.g., varied acquisition devices or regions, or key potential dependencies such as lighting and skin tone). For AI developers, we envision that our methodology may enable more tractable debugging of AI classifiers prior to more expensive and time-consuming multi-site performance evaluations.[129] Finally, our framework might directly assist physicians by revealing new attributes that they could subsequently use to improve their diagnostic skills, as was previously exemplified with perilesional sun damage as a diagnostic clue.[128] In contrast, while use of XAI outputs to support the case-by-case decision-making of physicians as part of a human-AI team has received attention within the XAI community, our framework is not directly applicable to this task but focuses more on large-scale auditing, and additional studies would be required to ascertain the utility of the underlying counterfactuals for verifying AI decisions.[130,131]

In light of a recent study that highlighted how dermatology AI classifiers perform worse on darker skin tones,[106] we considered how our analysis might detail the underpinnings of this behavior, *i.e.*, which aspects of the classifiers' reasoning processes might lead to inequitable performance across skin tones. In some malignant counterfactuals, particularly those of SIIM-ISIC, annotators noted diffusely darker background skin as compared to the benign counterfactual. Since multiple real-world variations, including differences in skin tone or lighting conditions, might recapitulate this effect,

the precise explanation remains unclear, but either case may be concerning. To the extent that real-world variations in skin tone may mirror this difference between the counterfactuals, a dermatology AI model may depend directly on skin tone. To the extent that real-world variation in lighting conditions or camera settings might mirror this difference, there is also potential for an indirect dependence of dermatology AI models on skin tone: camera designs are often biased toward ensuring appropriate exposure and color in light skin tones, but not dark skin tones,[132] implying that an AI classifiers that depends on lighting and color balance may as a result perform inequitably across skin tones. While we performed additional experiments modifying image brightness in hopes of better disentangling effects of lighting and skin tone, conclusions (Supplementary Fig. 4.16) varied considerably with the methodology employed (in contrast to our experiments with image chromaticity, Supplementary Fig. 4.17). Finally, our counterfactuals occasionally highlighted reflections as influential, which could systematically bias predictions in images of dark skin acquired with suboptimal lighting (*e.g.*, use of camera flash).[133] Thus, our study suggests multiple potential avenues by which inequitable performance of dermatology AI classifiers may arise from a mechanistic point of view, though future studies would be required to alleviate ambiguity and verify potential links between skin tone and variations in image acquisition on a dataset-by-dataset basis.

While our framework provides a detailed picture of the reasoning processes of medical AI classifiers, limitations remain. First, we aimed to characterize the classifiers in medically meaningful, human-derived terms, but AI reasoning processes *a priori* need not coincide with human concepts. For instance, AI classifiers can predict sex from fundoscopic images,[49] a challenging task for ophthalmologists, and the struggle to conceptualize these decisions in terms simple to humans[50] suggests the existence of peculiar, AI-specific abilities to detect certain attributes. While our use of an expressive XAI technique in combination with free-text annotations may improve our chances of capturing such AI-specific attributes, human biases may nonetheless prevent their detection or description. Second, while our counterfactual generation technique is highly expressive (Supplementary Fig. 4.5), inductive biases may still limit detection of some attributes. For instance, considering similarities between our generative models and the CycleGAN, which struggles to produce large-scale geometric changes,[39] our models may similarly be less likely to produce certain alterations in the counterfactuals, such as changes to the size of a lesion. Third, our approach does not provide information on the relationships between multiple attributes, *e.g.*, on the extent of any 'interactions' between attributes. Fourth, while we examine multiple modalities of dermatological images (clinical and dermoscopic), our analysis provides limited information on out-of-distribution features, or features that rarely appear in the examined images (*e.g.*, sutures). Fifth, use of human annotators introduces variability, both due to stochastic effects of whether an annotator notices an attribute in a given image, and due to variation in the background and training of experts. We found that our annotators typically agreed on the presence and direction of an attribute's effect, but the frequency with which they noted that attribute was not quantitatively consistent (Supplementary Fig. 4.15). Thus, while the 'fraction of counterfactual pairs' in which an attribute was noted may help gauge our confidence in the attribute's effect or enable approximate comparisons, granular comparisons of the 'extent' of an attribute's effect are likely not meaningful. Moreover, domain experts from varied backgrounds may tend to focus on different attributes (*e.g.*, a dermoscopy expert may focus on traditional features of pattern analysis). Finally, while our use of free-text entry likely improves the expressiveness of our framework, there is no uniquely correct way to distill these responses into a uniform taxonomy, implying that another set of domain experts may have chosen different levels of granularity. Despite these limitations, we believe our use of an expressive XAI technique, expert annotators, and free text entry together enable detailed, medically meaningful inferences on the AI classifiers' reasoning processes and how they could lead to desirable or undesirable behavior in deployment.

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

4.5 Methods

Image selection and preprocessing

To interrogate the performance of AI-based dermatological classifiers, we collected images of melanomas and melanoma look-alike lesions from multiple sources. We focus on this specific task for multiple reasons, including (i) the substantial attention it has received within the machine learning community,[110,121] (ii) alignment between this task and the intended use cases of the five AI classifiers, so as to enable comparison between classifiers, and (iii) the improved likelihood of generating interesting information on the reasoning processes of dermatology AI classifiers, as compared to simpler tasks that feature more visually salient signals.

Our first source, Fitzpatrick17k,[113] consists of clinical (rather than dermoscopic) images previously aggregated from online dermatology atlases. We filtered Fitzpatrick17k to include only melanomas, benign melanocytic lesions, seborrheic keratoses, and dermatofibromas. We additionally excluded diagramatic and histopathological images, and images that could be clearly identified as pediatric; after exclusions, the dataset consisted of 889 images. Advantages of Fitzpatrick17k include closer approximation of the expected inputs to consumer-facing dermatology AI tools (as compared to dermoscopic images, which require specialized tools) and inclusion of a variety of skin tones. Disadvantages include its relatively small size after filtering and noise in the diagnosis labels, which may not have been acquired via histopathological analysis or other gold-standard means.

Our second source, the ISIC 2019 challenge dataset,[111,112,121] consists of dermoscopic images from a variety of primary sources, including HAM10000[111] and BCN20000.[112] Like Fitzpatrick17k, we filtered the dataset to include melanomas, as well as melanoma look-alikes: benign melanocytic lesions, seborrheic keratoses, and dermatofibromas. After filtering, the ISIC dataset consisted of 20260 images. Most lesions were confirmed via histopathology (n=13072) or serial imaging showing no change (n=3704), while a smaller number were confirmed by single image expert consensus (n=1207), confocal microscopy with consensus dermoscopy (n=712), or unspecified means (n=1565). Compared to Fitzpatrick17k, ISIC thus offers more reliable diagnoses, but it lacks diversity in skin tones, featuring predominately light skin.

Finally, our third source, DDI,[106] consists of clinical images gathered from Stanford Clinics. Like other datasets, we filtered DDI to include only melanomas and melanoma look-alikes. In the case of DDI, which contains more granular and varied diagnoses, we included the following labels in our 'melanoma' category: acral lentiginous melanoma, melanoma *in situ*, nodular melanoma, as well as the general tag 'melanoma'. As melanoma look-alikes, we included the following labels: acral melanotic macule, atypical spindle cell nevus of reed, benign keratosis, blue nevus, dermatofibroma, dysplastic nevus, epidermal nevus, hyperpigmentation, keloid, inverted follicular keratosis, melanocytic nevi, nevus lipomatosus superficialis, pigmented spindle cell nevus of reed, seborrheic keratosis, irritated seborrheic keratosis, and solar lentigo. After filtering, DDI included 282 images; due to the comparatively high volume of data required for training our generative models, DDI was used only for performance evaluation (Fig. 4.1), rather than for our in-depth analysis of medical AI reasoning processes. However, DDI offers a number of desirable characteristics for evaluation purposes: (i) its images were not publicly available until after we obtained the five audited dermatology AI classifiers, precluding train-test overlap; (ii) DDI images have diverse skin tones, including enrichment for Fitzpatrick skin types V and VI; (iii) DDI contains a wide variety of skin conditions, including uncommon conditions; and (iv) the lesions are histopathologically proven, guaranteeing label accuracy. We note also that DDI is likely enriched for challenging lesions, since these are the lesions likely to require a biopsy.

For all evaluations, we preprocess the images to match the native input resolution of the AI classifier, which is 299×299 pixels for DeepDerm, and 224×224 pixels for all other classifiers. When evaluating AI classifier performance or generating counterfactuals (after generator training is complete), we resize the image via bilinear interpolation such that its shorter edge matches the input size of the AI classifier, then center-crop to obtain a square image. When training our generative models, which benefit from image augmentation, we instead resize the image such that its shorter edge is 120% of the input size of the corresponding AI classifier, then perform a random square crop matching the input size.

Classifier reproduction

We reproduced five AI-based dermatological classifiers, including prominent academically designed classifiers proposed for clinical use and classifiers currently in use by the public. Two of the classifiers, *Scanoma* and *Smart Skin Cancer Detection* (SSCD) are designed for use on mobile devices by the general public. The DeepDerm classifier is a previously published reproduction[106] of a prominent academic model,[21] sharing its training data and architecture. The

ModelDerm 2018 classifier is a publicly distributed academic model,[107] of which a later iteration (for which model weights are not publicly available) has been CE marked for use by the general public in Europe. The SIIM-ISIC Kaggle competition classifier is a reproduction of the first-place classifier[109] in the 2020 SIIM-ISIC Kaggle competition.[110] These models cover a broad range of architectures, pre-processing techniques, and training data sources; as such we believe these models offer a thorough view of both current practices and the state-of-the-art in dermatology AI.

Scanoma is commercial software available for mobile platforms including iOS and Android; at the time of writing, the app's AI classifier is free to use, while follow-up human evaluation is available for a fee. Architecturally, it is a custom convolutional neural network consistent with a MnasNet,[134] that is further optimized for use on mobile devices via quantization.[135] We obtained and unzipped the Scanoma APK file (normally installed on Android devices) to examine its TensorFlow Lite (TFlite) file, which contains the model specification and weights. Since our analysis tools are based on the PyTorch software library, we converted the network to the cross-library Open Neural Network Exchange (ONNX) format, which we then parsed in PyTorch. To maintain consistency with the original, quantized network while maintaining useful gradients, we implement the network using 'fake quantization'.[135] We verified that our PyTorch re-implementation matches the TensorFlow Lite implementation by comparing a series of 1000 test images, and we achieved nearly identical outputs (r=0.99, Supplementary Fig. 4.6a). To account for the small discrepancy between the classifiers, we analyzed the processing pipeline step-by-step and found slight differences in the bilinear rescaling preprocessing step, which may differ due to different antialiasing constants; the remaining differences were explained by sporadic single-bit differences in the quantized feature maps, likely resulting from numerical differences between TensorFlow Lite's native integer arithmetic routines and the equivalent operations performed in floating point arithmetic followed by fake quantization.

Like Scanoma, SSCD is a publicly available app intended for use on mobile devices. The architecture is a MobileNetV1, evaluated using floating-point (non-quantized) arithmetic. We followed a similar process to re-implement the SSCD classifier in PyTorch: a TFLite file was obtained from the app's APK package, then converted to ONNX before loading in PyTorch. We again verified our reproduction using a series of 1000 images and found that our PyTorch re-implementation of the neural network exactly matched the original Tensorflow Lite network. However, to ease comparison between classifiers, we update the input image resizing routine (a pre-processing step, prior to the neural network) in our implementation relative to the original app. Whereas the original app asks a user to specify a bounding box and then scales this box to the 224×224-pixel input image (warping the aspect ratio), we use the same preprocessing routine as for all other networks, in which we first center-crop the image and then resize the image using a bilinear filter. To assess the impact of this change in image preprocessing, we compared our PyTorch implementation against (i) the original TFLite model accompanied by preprocessing with square center-cropping and nearest-neighbor resizing and (ii) the original TFLite model with variable aspect-ratio resizing using nearest-neighbor rescaling (matching the original Android implementation, under the assumption that the uncropped image represents a user-defined bounding box), and we observed Pearson correlation coefficients of 0.97 and 0.92, respectively (Supplementary Figs. 4.6b-c). While evaluation of the entire processing pipeline including user selection of bounding boxes and choice of resampling filters is important for clinical evaluation of an AI system, our study instead focuses on the decision-making processes of the neural networks.

ModelDerm[107] is an academic classifier that has undergone multiple iterations, some of which have been tested in clinical settings, and one version of which has been approved for use in Europe via CE marking. We analyze the latest version for which model weights are publicly available, which we term ModelDerm 2018 based on the date of the accompanying publication;[107] authors declined to provide weights for the latest version of the model due to commercialization plans. ModelDerm is a ResNet-152[136] that runs natively in PyCaffe, with preprocessing performed in OpenCV. We parse the model architecture and weights directly from Caffe Protocol Buffer files and reconstruct the model in PyTorch. While the majority of the processing pipeline is highly reproducible in PyTorch relative to the original implementation, the original implementation preprocesses images channel-by-channel using the histogram equalization function in OpenCV, which we could not exactly reproduce in PyTorch while maintaining meaningful gradients during backpropagation. Instead, we implemented a custom, differentiable analogue of histogram equalization, in which the empirical cumulative density function used in OpenCV's implementation is replaced with a piecewise-linear approximation. Our PyTorch reimplementation of ModelDerm 2018, including the differentiable histogram equalization preprocessing step, retains close correspondence to the original PyCaffe/OpenCV implemention (r=0.96, Supplementary Fig. 4.6d).

The SIIM-ISIC competition classifier is intended to represent key features responsible for the high performance of the first-place winning classifier from the 2020 SIIM-ISIC melanoma classification Kaggle challenge, while reducing the computational complexity to permit feasible analysis. The original classifier is an extremely large ensemble of 90 networks, comprising mostly EfficientNets,[137] but also a few SE-ResNext 101s[138] and ResNest101s,[139] all of which are

evaluated at test time on 8 flips and rotations of the test image, for a total of 720 model evaluations per prediction. We reduced the computational complexity by retraining an ensemble of 3 EfficientNets (an EfficientNet-B5, -B6, and -B7), which comprise 80 of the 90 classifiers in the original ensemble, using a lower resolution of 224×224 pixels. To encourage similarity to the original model, we use the same training data, augmentation scheme and hyperparameters as the original classifiers. Our classifier additionally retains 8-fold image augmentation at test time, which we suspected may reduce the classifier's sensitivity to subtle image variations. While not intended to be an exact reproduction of the original winning classifier, our classifiers attain only slightly lower classification performance in 5-fold cross validation as compared to the original classifier (area under the receiver operating characteristic curve of 0.966 vs. 0.985).

The DeepDerm classifier is a previously published reproduction[106] of an academically developed model that was acclaimed for performing similarly well to dermatologists.[21] DeepDerm shares the same architecture (Inception-V3[140]) and importantly, the same training data as the original model, which was not publicly released. Since DeepDerm is distributed natively in PyTorch, no conversion steps were necessary for this classifier.

Counterfactual generation

To identify specific image factors responsible for each classifier's predictions, we generated counterfactual images using a variant of the technique 'Explanation by Progressive Exaggeration'.[38] However, to improve image quality, stabilize training, and better restrict generated alterations to those that cause a classifier to output a different prediction, we introduce multiple updates. We begin with an overview of the technique, then explain our specific updates. Full details of our generative models, including a formal mathematical treatment and an explanation of training parameters, are described in the Supplementary Methods.

At a high level, a counterfactual considers a scenario that did not occur, typically for the purpose of comparison to a scenario that did occur or to another counterfactual scenario. Such comparison may enable inferences about how a different AI classifier output may have been achieved, or which factors lead to that outcome. To enable these inferences, a counterfactual must typically be sufficiently similar to allow comparison, while differing in a realistic manner and eliciting a different outcome. In our case, we consider counterfactual images, which are alternate versions of real images. To create these counterfactual images, we use a type of generative image AI based on *generative adversarial networks*. We train our generative AI models to produce counterfactuals by altering real, reference images, with the goal of eliciting different predictions from an AI classifier; we also constrain these differences to be realistic (Supplementary Table 4.2). Then, examination of the differences between counterfactuals thus enables inferences regarding which image attributes influence an AI classifier's predictions.

We updated an existing generative AI technique for counterfactual images, Explanation by Progressive Exaggeration[38] (Supplementary Fig. 4.18-4.19), to better suit our purposes: First, we found that the original formulation of this technique could alter attributes of an image upon which the classifier does not depend, but which correlate with attributes upon which it does depend (Supplementary Fig. 4.7). We found that this behavior, which could lead to misinterpretations about the reasoning processes of the AI classifiers, arose from the specification of the *discriminator*, a component of the generative model that helps ensure realism of the generated images, and thus we updated our discriminator to remove this behavior (see Supplementary Methods for full details). Second, we also update the generator component of our model to use an architecture similar to that used in CycleGANs.[39] This network is similar to the residual network-based autoencoder used in the original implementation of Explanation by Progressive Exaggeration, but we found it produced images of higher visual quality (Supplementary Fig. 4.8). Finally, we applied data augmentation, including random cropping and random brightness modifications, to improve training when only a modest number of images are available (*e.g.*, with Fitzpatrick17k).

Expert evaluation of counterfactuals

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

To facilitate interpretation of the dermatologists' responses and comparison of the classifiers, we prescreened the

counterfactual images before analysis of the alterations within counterfactual pairs. Our prescreening consisted of a 'classifier-consistency' criterion to ensure that the alterations between each pair of counterfactuals meaningfully changed the classifiers' predictions, and a 'visual quality' criterion to mitigate the presence of artifacts, which could impede our ability to infer the importance of non-artifactual alterations. Our classifier-consistency criterion required the 'benign' and 'malignant' images in a counterfactual pair lay on opposite sides of the decision threshold (*i.e.*, they were classified as benign and malignant). In the visual-quality prescreening step, two board-certified dermatologists independently evaluated for artifacts each image that passed the classifier-consistency criterion, and we excluded images rejected by either evaluator. To ease comparison between classifiers, we included the same set of counterfactual pairs (modulo counterfactual alterations) for all classifiers; more precisely, for each reference image x_r, we included the corresponding counterfactual images $\{G_i(x_r)\}_{i \in C}$, where C represents the set of classifiers, if and only if $G_i(x_r)$ passed the prescreen for each classifier i. For subsequent analysis, we included the 92 images from Fitzpatrick17k that passed our pre-screening criteria, and we included 100 images from ISIC to achieve a similar quantity of images.

To learn which attributes differ between benign and malignant counterfactuals—and thus influence an AI classifier's predictions—we developed a two-stage annotation approach. We designed the first stage of this approach to encourage discovery of a wide variety of attributes, which we then leverage in the second stage to more efficiently collect data. Both stages leverage a graphical interface that runs locally in a web browser; expert evaluators view a pair of benign and malignant counterfactuals, then answer questions regarding (i) which member of the pair appears *most* likely to be malignant, and (ii) what attributes differ, and how they differ, between the counterfactuals. In the first stage, evaluators enter attributes as free text (*e.g.*, 'skin lines more prominent'), accompanied by a 'direction' specifying how the images differ (see Supplementary Fig. 4.12). After the first 100 pairs were evaluated by each expert, we pooled and grouped the free text terms to determine 'preset' attributes (*e.g.*, 'skin lines more prominent' and 'more skin lines' map to the preset 'Prominence of skin grooves/dermatoglyphs') that could be selected during the second stage of annotation. This stage also retained the option for free text entry, in case a new attribute were discovered. To mitigate potential bias, we randomized and blinded evaluators to (i) the appearance order of a counterfactual pair (*i.e.* whether the benign or malignant counterfactual appeared on the left/right) and (ii) the overall order of the counterfactual pairs, including randomization of the corresponding reference images and shuffling counterfactual pairs from the various AI classifiers. Evaluators annotated the counterfactual pairs in sets of twenty, which required approximately 30 minutes to complete.

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

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

To visualize the resulting values (Fig. 4.2), we further aggregated the 'directions' d, which originally included five options: *benign only*, *benign < malignant*, *different*, *benign > malignant*, and *malignant only* (during data collection, which was blinded, these terms appeared as *A only*, $A < B$, etc., where images A and B were randomized to benign or malignant). We aggregated *benign only* and *benign > malignant* into a new category, *benign*, and likewise aggregated *benign < malignant* and *malignant only* into the new category *malignant*. Finally, for each pair of attribute and AI classifier, we determined the 'predominate direction' of that attribute, which we defined as *benign* if $\bar{s}_{c,a,\text{benign}} > 2 \cdot \bar{s}_{c,a,\text{malignant}}$, we defined as *malignant* if $\bar{s}_{c,a,\text{malignant}} > 2 \cdot \bar{s}_{c,a,\text{benign}}$, and we defined as *neither* otherwise, where the cutoff factor of 2 was chosen to prevent emphasis on small differences in frequency between the benign and malignant directions. In Fig. 4.2, the size of the square is then proportional to $\bar{s}$ for the predominate direction, or the average of the directions if neither was predominate.

Experimental validation of findings from counterfactuals via color shifts

To validate the attributes identified as important for dermatology AI classifier's predictions in our counterfactual experiments, we aimed to experimentally modify a single attribute and observe the effect on each AI classifier; we

chose image color as a test case, since existing mathematical tools[124] enable well-defined, unambiguous changes to this attribute. To alter the color of each image, we converted from the sRGB color space to the CIE 1976 L*, u*, v* color space (CIELUV),[124] added an offset to the chromaticity coordinates (u^*, v^*), then converted back to sRGB. Different chromaticity shifts were generated by varying the offset along a circle centered at $(u^*, v^*) = (0, 0)$ with radius 20, where the factor 20 was chosen heuristically to produce color changes that we deemed visible while remaining plausible.

Data availability

Images used in this study were obtained from publicly available repositories. ISIC images are available at `https://challenge.isic-archive.com/data/`. Fitzpatrick17k images are available at `https://github.com/mattgroh/fitzpatrick17k`. The DDI images are available at `https://stanfordaimi.azurewebsites.net/datasets/35866158-8196-48d8-87bf-50dca81df965`.

Model weights for the DeepDerm classifier are available at `https://zenodo.org/record/6784279#.ZFrDc9LMK-Z`. The weights and model specification for the ModelDerm classifier are available at `https://figshare.com/articles/Caffemodel_files_and_Python_Examples/5406223`. Model weights for our retrained variant of the SIIM-ISIC competition classifier are available at `https://zenodo.org/doi/10.5281/zenodo.10049216`. Scanoma and Smart Skin Cancer Detection are third party software for which we cannot redistribute model weights. At the time of writing, both are apps are available for download with no fee from the Google Play store and third-party APK package download sites.

Code availability

Our code, including a PyTorch implementation of explanation by progressive exaggeration and classes for loading datasets and classifiers are available at `https://github.com/suinleelab/derm_audit`. Weights for our trained generative models and the re-trained SIIM-ISIC classifier are available at `https://zenodo.org/doi/10.5281/zenodo.10049216`.

Author contributions

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

Funding

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

Ethics declarations

Competing interests

R.D. reports fees from L'Oreal, Frazier Healthcare Partners, Pfizer, DWA, and VisualDx for consulting; stock options from MDAcne and Revea for advisory board; and research funding from UCB.

4.6 Supplementary Information

Supplementary Methods

Additional details of counterfactual generation

Our choice of method to create counterfactuals, Explanation by Progressive Exaggeration, uses generative adversarial networks to create alternate versions of images that (i) appear 'realistic', in the sense that they lie on the manifold of training images, (ii) 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 (iii) are similar to the original image, in the sense that the original image may be approximately reconstructed by passing an altered, generated image back through the generator.

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 $f : [0,1]^{d^2} \to [0,1]$ be a classifier to be audited. Our goal is to obtain a generator $G : [0,1]^{d^2} \times \mathcal{C} \to [0,1]^{d^2}$ 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 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 explanation by progressive exaggeration, 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 translate to (i) the range of the generator $G(\mathcal{X}, \mathcal{C})$ is contained in the data manifold $\mathcal{M}_{\mathcal{X}}$, (ii) the prediction of the classifier 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 (iii) 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 $D : [0,1]^{d^2} \to \mathbb{R}$ that attempts to distinguish real from generated images. In contrast to the original implementation, we update the discriminator such that it does not depend on a condition c. The original implementation of the discriminator attempts to differentiate generated images from real images that elicit a particular prediction from the classifier, which may encourage generated images to appear similar to that subset of real images including potentially via changes that do not alter the output of the classifier. In contrast our implementation of the discriminator instead attempts to differentiate generated images from any real image, such that it only encourages that the generated images appear similar to real images (Supplementary Fig. 4.7). To reflect this update, we choose the following functions for the loss of the discriminator L_D and of the generator L_G (see also Supplementary Fig. 4.18). 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 as follows:

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

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

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

In addition our introduction of a non-conditional discriminator, we also update G to use an architecture similar to that used in CycleGANs.[39] This network is similar to the residual network-based autoencoder used in the original

implementation of explanation by progressive exaggeration, but we found it produced images of higher visual quality (Supplementary Fig. 4.8).

To optimize the networks, we followed the reference implementation 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 prevent overfitting, we also applied data augmentation including random cropping and random brightness modifications. To choose the hyperparameters λ, we followed the original publication 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 (Supplementary Fig. 4.11). 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 benign and malignant counterfactuals (Supplementary Fig. 4.19), so we focused on a single generative model for each combination of AI device and generative model (Supplementary Table 4.2).

To train our models, we reimplemented the original TensorFlow library for explanation by progressive exaggeration using PyTorch. 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 AI device outputs of 0.05 and 0.95 respectively) in order to obtain counterfactuals most likely to elicit benign and malignant predictions. As in all components of our study, we use only reference images of melanomas and melanoma look-alikes when generating counterfactuals.

Saliency map generation

In initial efforts to understand the reasoning processes of the dermatology AI devices, we generated saliency maps, which highlight the regions of an image that contribute most to the AI's prediction. To mitigate the possibility that a particular technique for saliency map generation may produce less useful results, we applied three popular techniques separately.

Following our previous work that analyzed radiology AI devices,[53] we first applied Expected Gradients.[27] This gradient-based feature attribution technique mitigates shortcomings of previous techniques,[9] including the tendency to fail to highlight darker regions of an image,[91] which would be problematic given that melanomas and melanoma look-alikes are typically darker than background skin. At a high level, this technique captures the importance of an input pixel by measuring the sensitivity of the AI devices's prediction to small changes in that pixel (in mathematical terms, calculating the gradient), and averaging this value as the image is interpolated from a number of baseline images to the image of interest. Formally, the Expected Gradients attribution ϕ for a sample x, input feature (pixel) i, baseline distribution D, and AI device f is given by:

$$\phi_i(x) := \mathop{\mathbb{E}}_{x' \sim D, \alpha \sim U(0,1)} \left[(x_i - x'_i) \times \frac{\partial f(x' + \alpha \times (x - x'))}{\partial x_i} \right] \tag{4.1}$$

As our background distribution, we chose the full ISIC 2019 dataset; attributions were estimated via Monte Carlo sampling, using 1000 samples.

As our second feature attribution technique, we next calculated saliency maps via KernelSHAP.[30] This technique characterizes the importance of an input pixel by measuring how the model's prediction changes when that feature is 'removed' (in our case, replaced by the mean color of that image). Importantly, feature interactions are properly accounted by removing multiple features at a time, then summarizing a feature's effects over these subsets via the *Shapley value*,[31] a well-established technique grounded theoretically in game theory. KernelSHAP estimates the Shapley value by casting it as the solution to a least squares problem, which can be solved by sampling random sets of features to remove, rather than requiring exhaustive enumeration of every possible set of features. To enable tractable calculation, we define 16×16 super-pixels as features, then upsample the final result via bilinear interpolation to match the original image size. For each image, we perform the KernelSHAP estimate using 10^5 samples, which required approximately one hour of computation on an NVIDIA RTX 2080Ti graphics processing unit, per image.

Finally, we calculated saliency maps via the highly popular GradCAM approach.[8] This technique characterizes the importance of a region of an image by monitoring the activation of individual neurons in a neural network, which may retain coarse spatial information even in layers far from the input. Specifically, for each channel of an activation map, the technique multiplies that activation by the derivative of the network's output with respect to that neuron, then sums over all channels to determine an aggregate value for a spatial location, before finally discarding negative values. Formally, let A denote the activations of a neural network f at the layer of interest, let k represent each channel of those activations, and let x denote the input. Then the GradCAM attributions ϕ are given by:

$$\phi(x) := min\left[\sum_k A_k \frac{\partial f}{\partial A_k}(x), 0\right] \tag{4.2}$$

We take derivatives of the model's prediction of the likelihood of melanoma such that intuitively, these attributions can be understood as identifying the regions of the image that contribute toward a prediction of melanoma. As the 'layer of interest,' we target the layer immediately prior to the final global pooling. To account for model ensembling in the AI device SIIM-ISIC, which includes three individual models, each of which is evaluated at test time on eight versions of the input image (the original, plus a series of flips and rotations), we treat the channels of that layer of the twenty-four resulting 'sub-models' as channels of one aggregated layer. In other words, in the above equation, k runs over the channels of each sub-model's final layer prior to global pooling, as well as over all sub-models; to preserve spatial relationships, we reverse the augmentations before averaging. The resulting saliency maps match the spatial dimensions of the layer of interest, which (as is typical with GradCAM) are lower resolution than the input image; we upsample the saliency map via a bilinear filter to match the dimensions of the original image.

In all cases, we display the final saliency map by taking its absolute value, then overlaying it on a desaturated version of the original image, with the saliency map blended at $\alpha = 80\%$. To mitigate overemphasis of the color scale on outlier values, we clip the maximum value of the saliency map at the 99^{th} percentile of each image.

Additional colorimetric experiments

To help clarify observations from our counterfactual experiments that some AI devices may rely in part on the pigmentation of background skin, or (less frequently) the brightness of images, we performed additional experiments manipulating image brightness or color.

To modify image brightness, we separately applied each of three methods: (i) Following the brightness routines implemented in Python Image Library (Pillow) and PyTorch, we multiply RGB values by a constant B, where $B < 1$ darkens the image and $B > 1$ brightness the image. We assume images are encoded as sRGB, and to make the transform not dependent on this encoding, we first convert to linear RGB (*i.e.*, gamma-expanded) values by applying the sRGB gamma function, before multiplying by the brightness factor B. (ii) We add an offset to the perceptual lightness, as defined by the L^* channel of the CIELUV color space.[124] (iii) We convert the image to the recently developed $J_z a_z b_z$ color space,[141] which compares favorably with other color spaces in perceptual uniformity, and multiply the lightness channel J_z by a brightness factor (as done in the linear RGB space). We manually tune the degree of brightness modification (*i.e.*, the multiplicative or additive factor) to produce images that noticeably vary the brightness of the image while avoiding excessive clipping. Since annotators noted variations in image brightness more frequently among counterfactuals derived from clinical images (Fitzpatrick17k), we carried out brightness modification experiments using the clinical images.

To modify image chromaticity, we again separately applied each of three methods, each of which can be conceptualized as a chromatic adaptation (white balance) transform: (i) As described in the main text methods, we convert the image to the CIELUV color space,[124] then add an offset to the u^*, v^* channels, *i.e.*, following CIELUV's native translational[142] chromatic adaptation transform. (ii) We convert the image to the CIELAB color space, shifting chromaticity by selecting colored reference illuminants, which the CIELAB conversion applies using a von Kries-like transform.[143,144] That is, we convert from CIEXYZ tristimulus values to CIELAB using a white point (X_n, Y_n, Z_n in the notation of the CIELAB specification[144]) chosen to be different from D65 (the sRGB white point), then convert back to CIEXYZ using the D65 white point. (iii) We apply the CAM16 chromatic adaptation transform (CAT16)[145] for colored reference illuminants. In all cases, we select the reference illuminants uniformly from a circle spanning the chromaticity axes of the corresponding color space, *i.e.*, u^* and v^* with CIELUV, a^* and b^* with CIELAB, or a and b with CAM16. In all cases, we also manually tune the magnitude of the chromaticity modification (radius of circle) to produce modifications that are noticeable while avoiding obvious clipping. Since annotators noted variations in

the ‘pinkness’ of an image predominantly in the dermoscopic images (ISIC), we performed chromaticity modification experiments using the dermoscopic images.

Supplementary Tables

	Category	Shortened	Original
ISIC			
	lesion	darker pigmentation	color of pigmentation - darker
	lesion	number of colors	number of colors in lesion
	background	pinker	color - pink
	lesion	structureless areas	structureless area(s)
	background	brown spots	number or prominence of brown spots
	background	hair	number or prominence of hairs
	lesion	regression	prominence of regression
Fitzpatrick17k			
	lesion	darker pigmentation	color of pigmentation - darker
	background	darker pigmentation	color of pigmentation - darker
	lesion	number of colors	number of colors in lesion
	lesion	erythema	redness/erythema
	background	hair	prominence of hair
	lesion	browner pigmentation	color of pigmentation - brown
	lesion	scale/crust	presence of scale/crust

Table 4.1: **Reference for attribute names from main text Fig. 4.2, which for conciseness shortens the original attribute names used during our annotation procedure.** Some attributes were not shortened: 'atypical pigment networks', 'blue/white veil', and 'erythema' (ISIC); 'erythema', 'telangiectasia', and 'uneven pigmentation' (Fitzpatrick17k).

	DeepDerm	ModelDerm	Scanoma	SSCD	SIIM-ISIC
ISIC	13.1	19.5	9.6	16.1	16.0
Fitzpatrick17k	22.7	19.6	23.0	8.8	23.6

Table 4.2: **Kernel inception distances ($\times 10^{-3}$) between generated images and the reference dataset.** The reference dataset contains all images from ISIC or Fitzpatrick17k, after exclusions (see Methods: Image selection and preprocessing), i.e., it was not limited to those images evaluated by experts. The generated dataset contains, for each image in the reference dataset, either the 'benign' or 'malignant' counterfactual chosen uniformly at random.

Lesion	Background	Other
area of pigmentation	color - pink	blurriness
atypical pigment networks	color - yellow	prominence of artifacts
atypical vessels	color of pigmentation - darker	
black striae	erythema	
blue/white veil	number or prominence of brown spots	
color of pigmentation - blue	number or prominence of hairs	
color of pigmentation - darker	number or prominence of red spots	
color of pigmentation - purple	number or prominence of white spots	
color of pigmentation - yellow	prominence of skin grooves/dermatoglyphs	
crust	reticulation	
erythema	scale	
number of colors in lesion	smoother	
patchiness	sun damage	
presence of blue globules	telangiectasia	
presence of filiform lesion		
prominence of follicles/pores		
prominence of radial streaks		
prominence of regression		
raised appearance/papular/nodular		
reticulated		
salt and pepper sign		
scale		
size of lesion		
strawberry pattern		
structureless area(s)		
uneven borders		
warty		
white spots		
white striae		

Table 4.3: List of attributes featured in analysis of counterfactuals generated from ISIC images. Each attribute was observed differing in at least one counterfactual pair.

Lesion	Background	Other
area of pigmentation	addition of pigmented peripheral lesion	blurriness
bullas/vesicles	changes in ulceration (outside primary lesion)	cooler color temperature
cerebriform appearance	color - more orange	darker image
color of pigmentation - black	color - more pink/red	darker shadow
color of pigmentation - blue	color - more yellow	prominence of artifacts
color of pigmentation - brown	color of pigmentation - darker	prominence of medical artifact
color of pigmentation - darker	color of pigmentation darker	prominence of reflection
color of pigmentation - pink	erythema	removal of anatomical features
color of pigmentation - purple	number of peripheral lesions	
color of pigmentation - yellow	presence of brown spots	
confluence	presence of crust	
crateriform	presence of peripheral papules	
darkness of hair	presence of pores	
friable	presence of red spots	
more textured appearance	presence of scale	
more translucency in the center	prominence of background skin texture	
number of colors in lesion	prominence of hair	
number of lesions	prominence of peripheral lesions	
number of red lesions/dots	prominence of skin grooves/dermatoglyphs	
number or prominence of pigment globules	prominence of veins	
presence of erosion or ulceration	prominence of wrinkles	
presence of fissures	reticulated	
presence of scale/crust	telangiectasia	
presence of white area or hypopigmentation		
presence of white/blue area		
prominence of follicles		
prominence of skin grooves/dermatoglyphs		
raised appearance/nodular/papular		
redness/erythema		
regression		
shininess		
size of lesion		
telangiectasia		
uneven borders		
uneven pigmentation		
white halo		

Table 4.4: List of attributes featured in analysis of counterfactuals generated from Fitzpatrick17k images. Each attribute was observed differing in at least one counterfactual pair.

Supplementary Figures

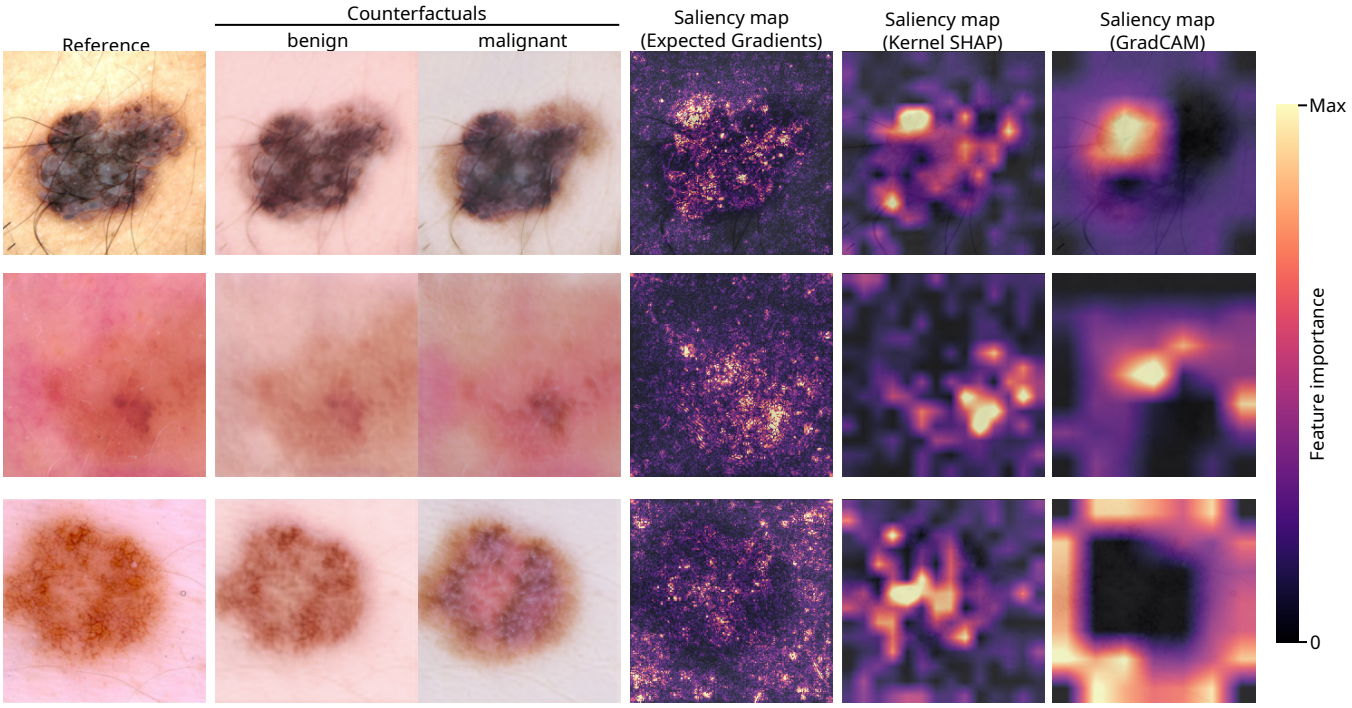


Fig. 4.5 | Comparison of insights from counterfactuals and saliency maps. We calculated feature attributions using three popular techniques, Expected Gradients,[27] Kernel SHAP,[30] and GradCAM[8] (see Supplementary Methods) and then produced our best-effort visualizations of the resulting saliency maps. We failed to gather insights from the saliency maps, except that the AI device may focus on the lesion (but perhaps not always, depending on the saliency technique). In contrast, the counterfactuals provided more granular and medically interpretable insights: for instance, based on the malignant counterfactuals we inferred that multiple colors of pigment (top + bottom), erythema (middle + bottom), darker pigmentation (all), and blue-white veil (bottom) tend to elicit more malignant predictions. In this figure, all saliency maps and counterfactuals were generated in reference to our AI device 'SIIM-ISIC'.

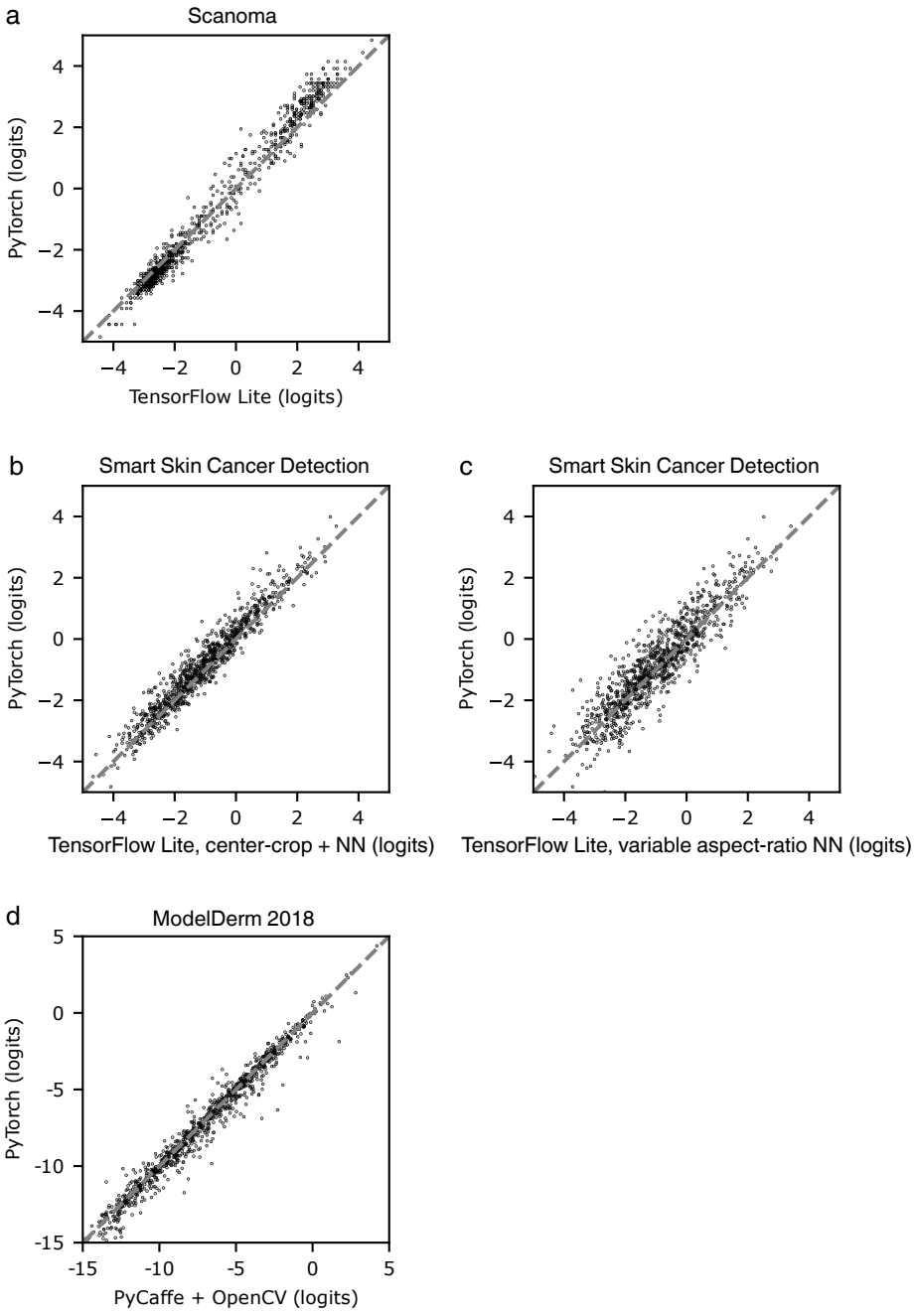


Fig. 4.6 | Similarity of predictions from original classifiers and our PyTorch re-implementations. To evaluate our PyTorch re-implementations' similarities to the original models, we compared the classifiers' predictions on a series of 1000 images from the ISIC dataset. **a**, Comparison of our PyTorch re-implementation of Scanoma with a TensorFlow Lite implementation of Scanoma, which differs from the original Android implementation only by antialiasing constants in the bilinear filtering preprocessing step. We compared our PyTorch re-implemention of Smart Skin Cancer Detection (SSCD), which uses square center-cropping and bilinear resizing to preprocess images, against the original TensorFlow Lite implementation with square center-cropping and nearest-neighbor resizing (**b**), and nearest-neighbor resampling to a square input (allowing changes to the aspect ratio, **c**). Aside from the input resizing routine, our PyTorch implementation achieves identical outputs to the original TensorFlow Lite classifier. **d**, Comparison of our PyTorch reimplemention of ModelDerm 2018, including our differentiable histogram equalization preprocessing step, with the original PyCaffe and OpenCV implementation. NN, nearest-neighbor interpolation.

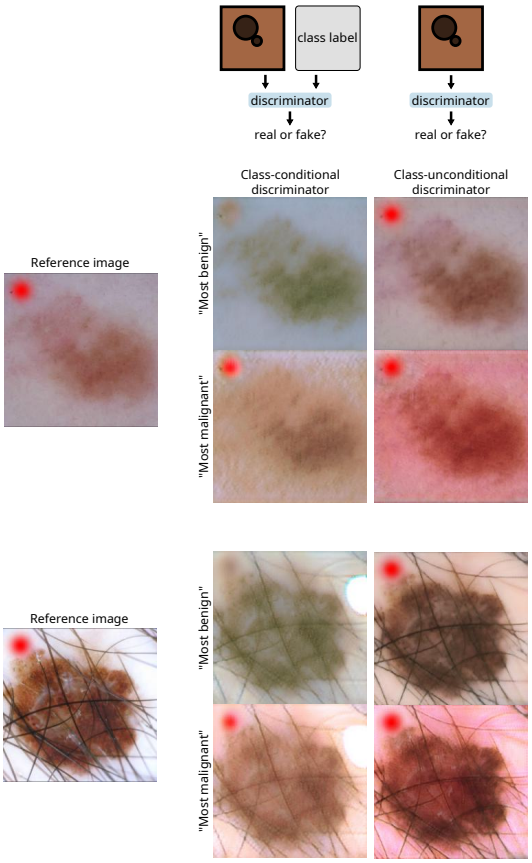


Fig. 4.7 | Comparison of a class-conditional discriminator with a discriminator not conditioned on class, with respect to their treatment of features correlated with the classifier's output. Hypothesizing that a class-conditional discriminator would alter features correlated with a classifier's predictions, even when not used by the classifier, we designed a scenario in which the classifier is unlikely to depend on the presence of a test artifact (a red dot in the corner of the image), but the test artifact correlates with melanoma status in the training data for the generative model. In particular, we trained an EfficientNet-B7 to detect benign versus malignant lesions among the melanomas and melanoma-lookalikes of the ISIC 2019 training data; since this training data lacked the test artifact in any image, the classifier is unlikely to depend strongly on the presence of the artifact. When training the generative models, we introduced the test artifact into every image of a melanoma, such that it correlates perfectly with melanoma status. While the test artifact is altered by the generator that was trained in conjunction with the class-conditional discriminator, which could mislead an investigator to conclude that the classifier's prediction is based in part on the presence of the test artifact, the generator trained with the discriminator that is not conditioned on class leaves the test artifact unaltered. In addition, we anecdotally noted that the generator trained with the unconditional generator produced images of higher visual quality (*e.g.*, with fewer 'water-droplet'–like artifacts; see lower set of images).

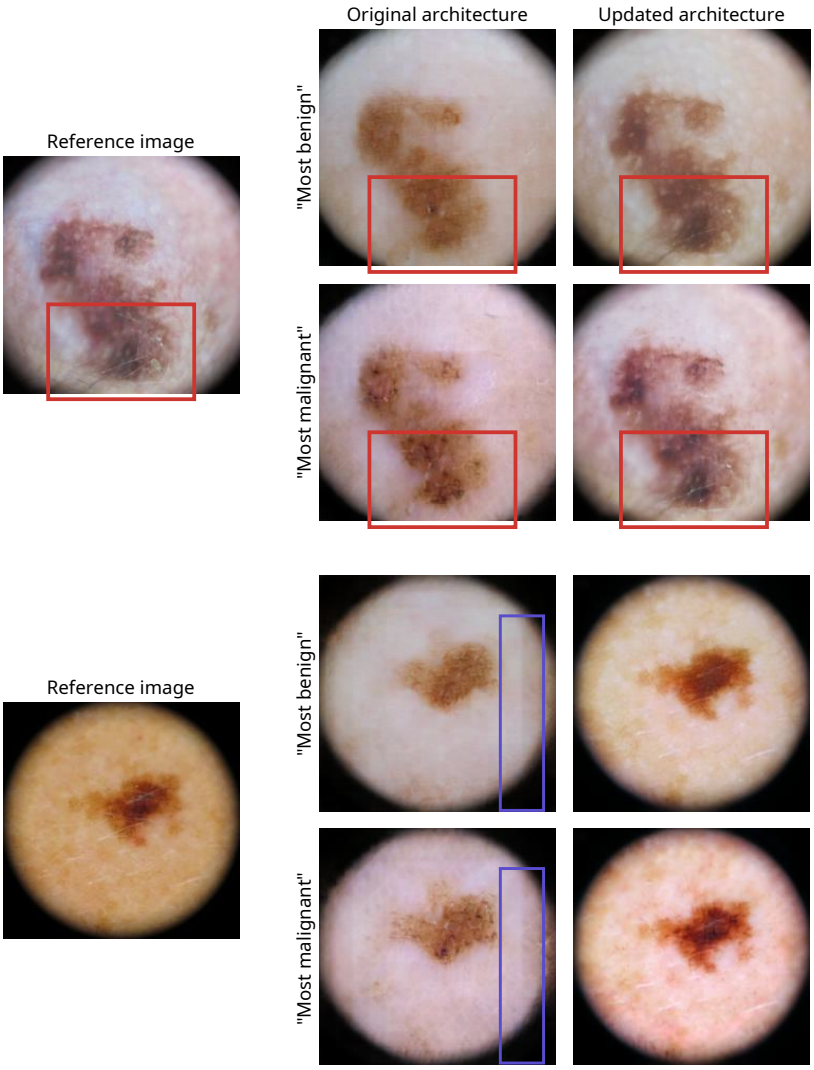


Fig. 4.8 | Comparison of the visual quality of images produced by the original generator architecture from ref.[38] with those produced by our updated architecture. Our updated architecture successfully reproduces details such as hairs, which the original architecture fails to capture (red boxes). The original architecture also introduces linear artifacts (blue boxes) not present in the original image, while we noted no such artifacts in images generated by the updated architecture.

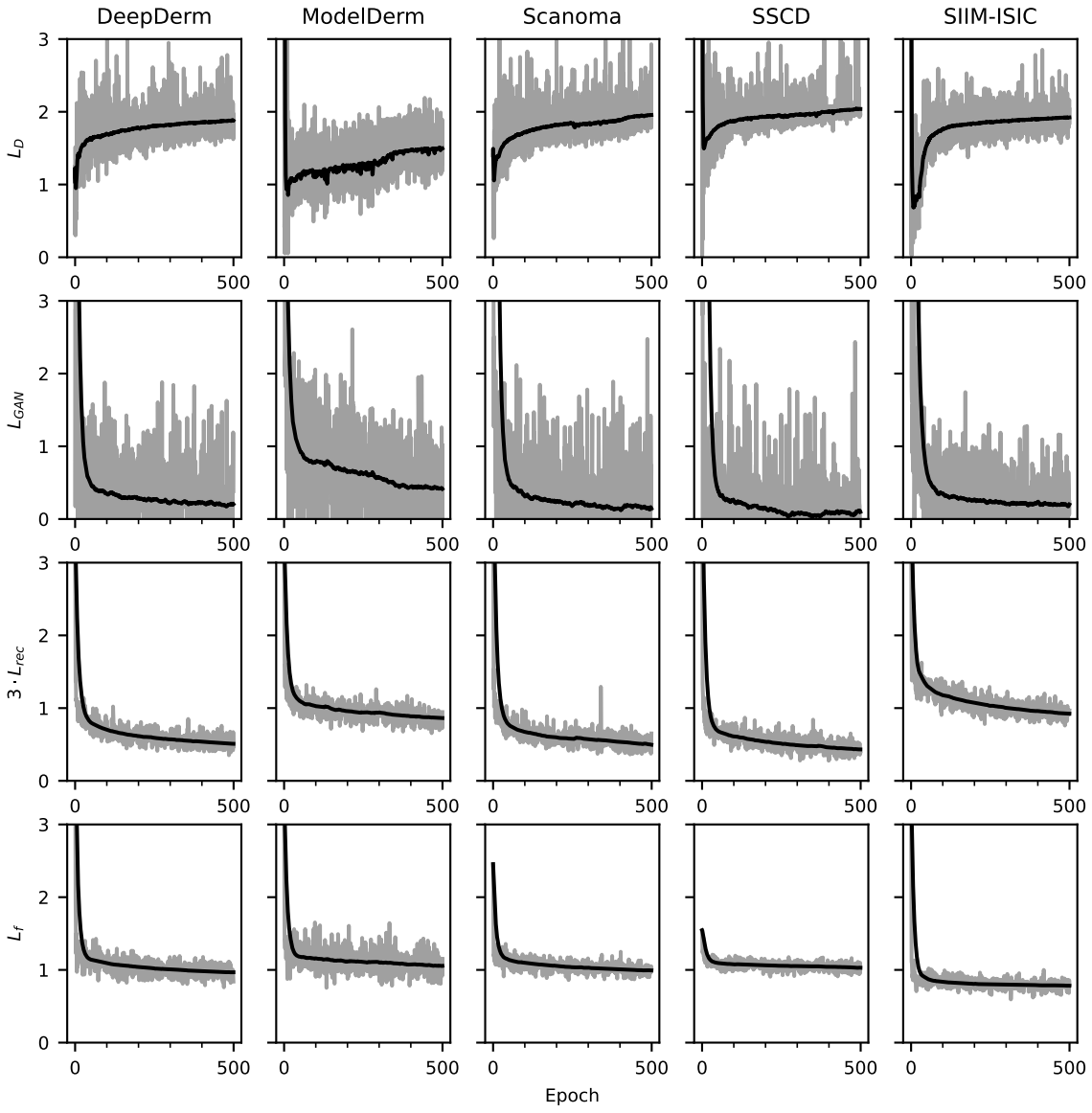


Fig. 4.9 | Evolution of loss terms during training of our generative models on the ISIC dataset. Loss terms are plotted after multiplication by their respective scaling factors ($\lambda_{rec} = 3$, $\lambda_D = \lambda_{GAN} = \lambda_f = 1$). Gray lines indicate the instantaneous loss, and black lines indicate the exponential moving average ($\alpha = 0.001$; loss terms were recorded at each gradient update of their respective model).

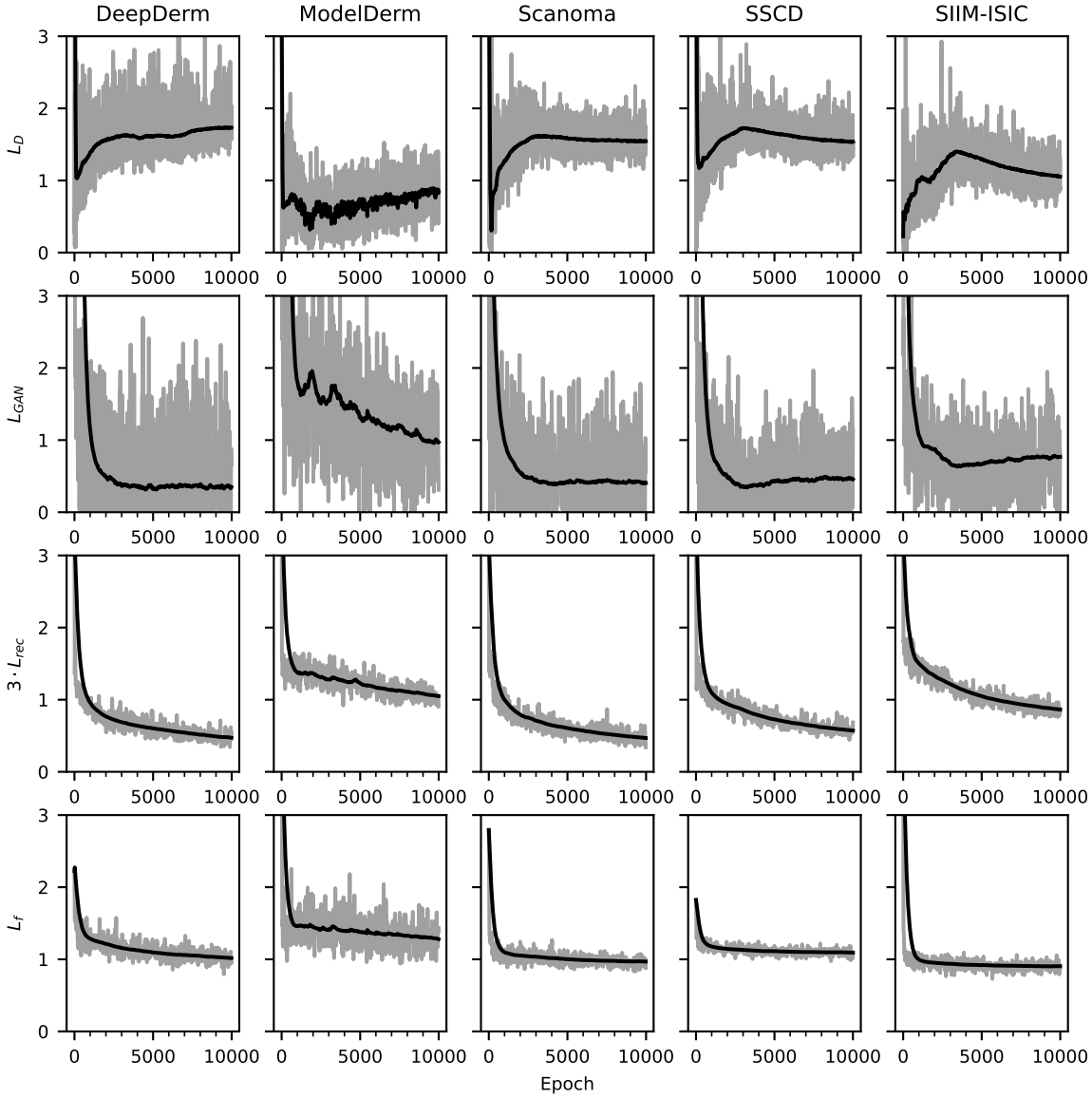


Fig. 4.10 | Evolution of loss terms during training of our generative models on the Fitzpatrick17k dataset. Loss terms are plotted after multiplication by their respective scaling factors ($\lambda_{rec} = 3$, $\lambda_D = \lambda_{GAN} = \lambda_f = 1$). Gray lines indicate the instantaneous loss, and black lines indicate the exponential moving average ($\alpha = 0.001$; loss terms were recorded at each gradient update of their respective model).

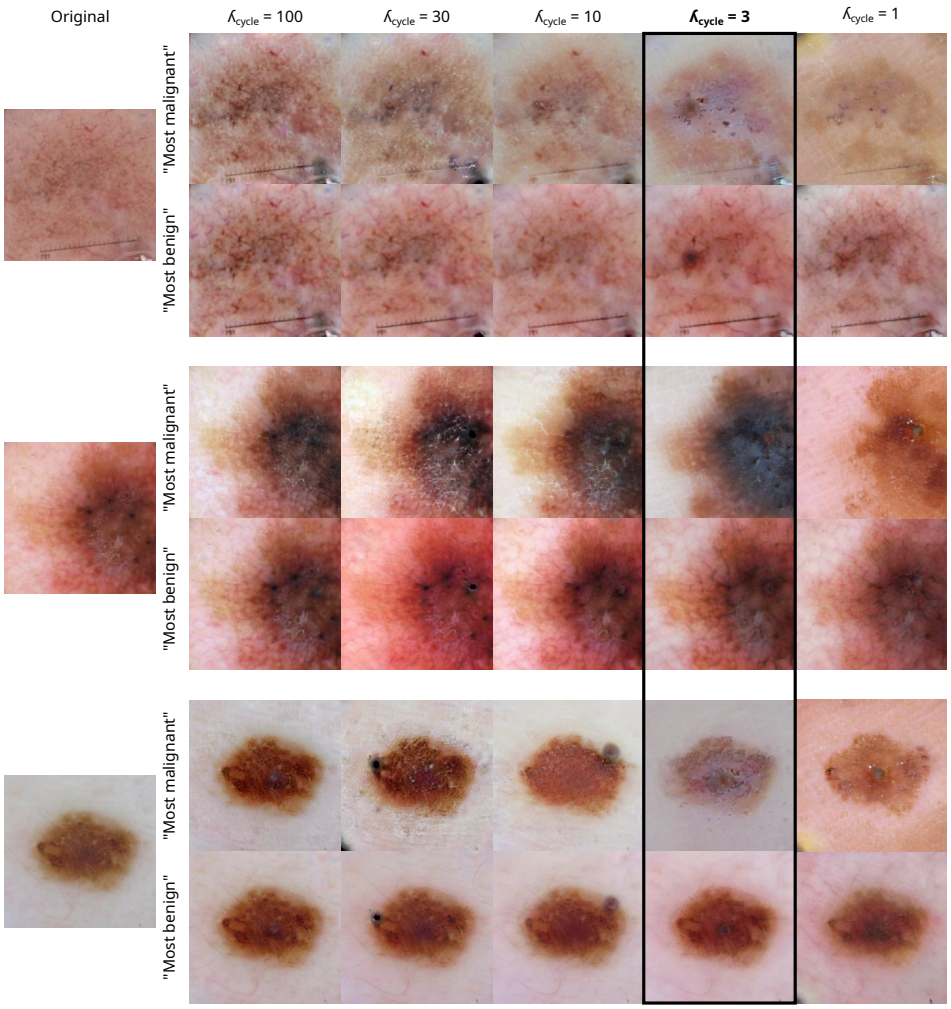


Fig. 4.11 | Tuning of the hyperparameter λ_{cyc} in the generative models. To tune the hyperparameter λ_{cyc}, we started with the value of 100 reported in the original publication of Explanation by Progressive Exaggeration,[38] then progressively decreased its value until the alterations between the 'most benign' and 'most malignant' images became apparent (based on manual, visual inspection), while ensuring that the generated images still appeared similar to the original, reference image. Counterfactuals in this figure were generated to analyze the AI device ModelDerm; images were chosen uniformly at random.

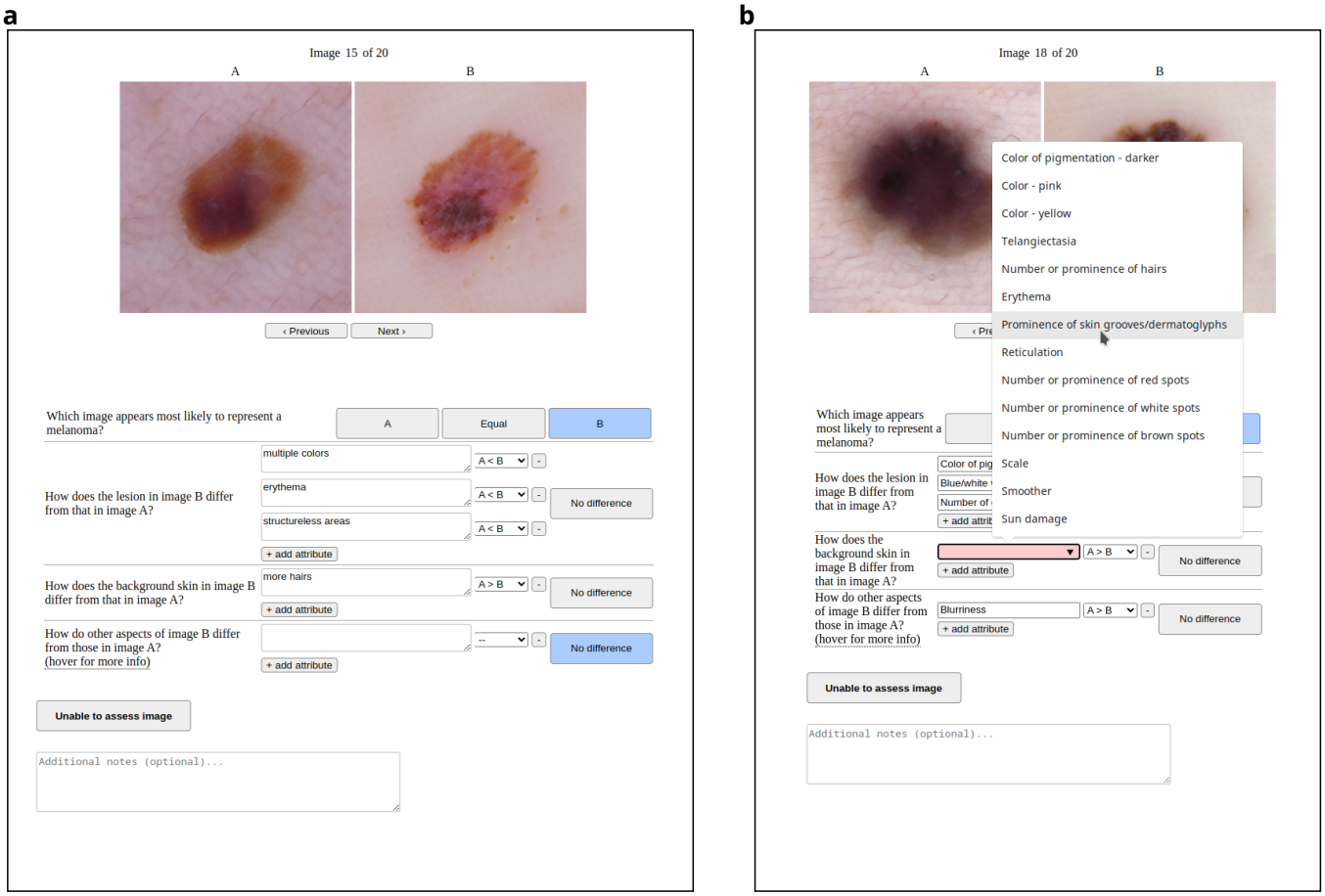


Fig. 4.12 | Screenshots of app for expert analysis of counterfactuals. a, 'Free text' version of the app, used during the initial phase of data collection to encourage collection of a broad, diverse set of attributes that differ between benign and malignant counterfactuals. The expert annotator enters an attribute (*e.g.*, 'structureless areas') and then specifies how that attribute differs between the two images by selecting a comparator ('A only', 'A > B', 'A < B', 'B only', or 'different') from the drop down menu. The app allows entry of an arbitrary number of attributes, and contains multiple categories of attributes ('lesion', 'background', and 'other') to remind annotators to pay attention to each part of the counterfactuals. **b**, After the initial phase of free-text data collection, attributes are pooled and grouped in collaboration with the expert annotators, to produce a list of 'preset' responses that enables faster, more uniform analysis. In the remaining modules, expert annotators may select a preset from a drop-down list, or continue to enter attributes as free text, accounting for the possibility that new attributes are discovered after the initial free-text phase.

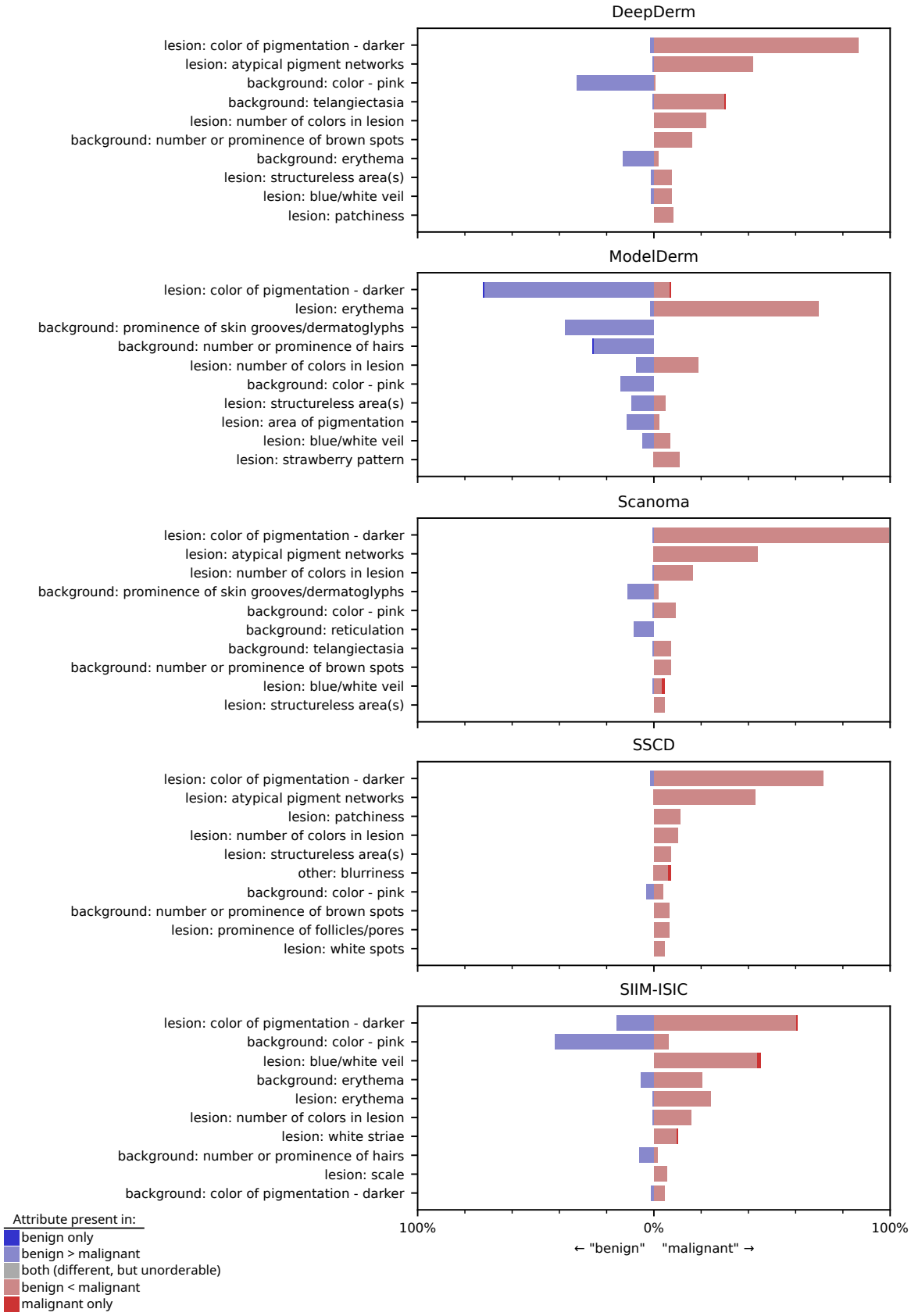


Fig. 4.13 | Attributes identified by our joint expert-XAI auditing procedure as key influences on the output of individual dermatology AI devices, when evaluated on the ISIC dataset. In contrast to main text Fig. 4.2, attributes are ordered by the proportion of counterfactual pairs *from the specified AI device* in which experts noted that attribute differs, enabling examination of attributes relevant to a particular AI device but not necessarily to most AI devices (*e.g.*, prominence of skin grooves or dermatoglyphs, which influences Scanoma and ModelDerm).

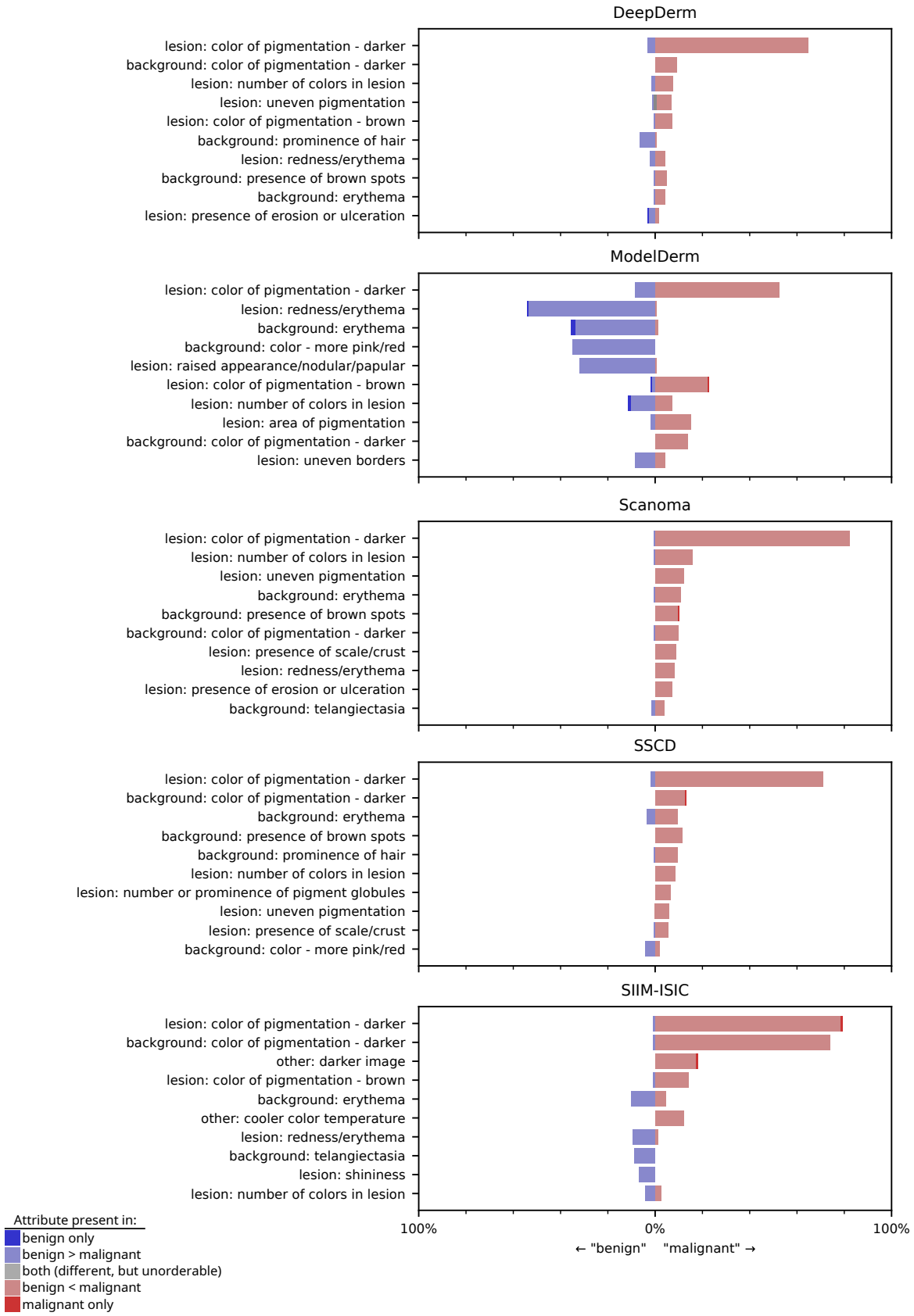


Fig. 4.14 | Attributes identified by our join expert-XAI auditing procedure as key influences on the output of individual dermatology AI devices, when evaluated on the Fitzpatrick17k dataset. In contrast to main text Fig. 4.2, attributes are ordered by the proportion of counterfactual pairs *from the specified AI device* in which experts noted that attribute differs, enabling examination of attributes relevant to a particular AI device but not necessarily to other AI devices.

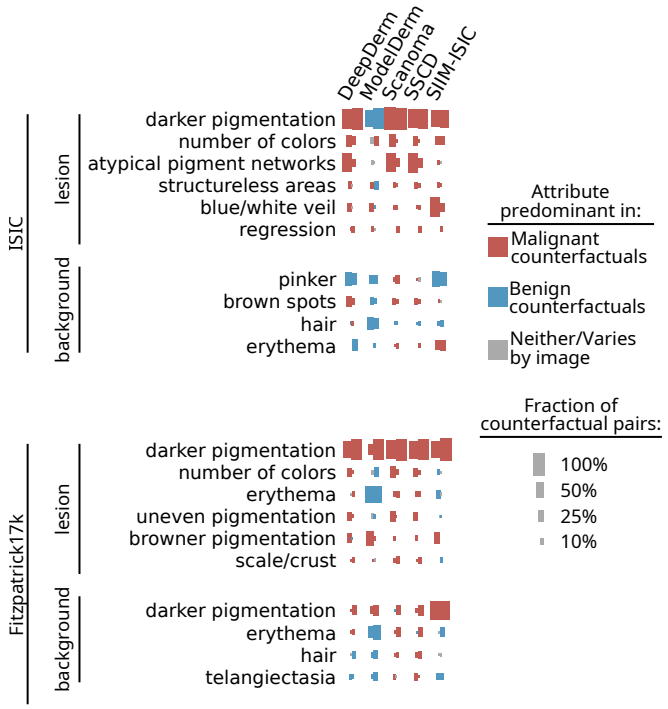


Fig. 4.15 | Analysis of inter-reader variability, displaying the two readers' individual conclusions side by side for each attribute. For each reader, we separately determine whether that attribute was 'predominant' in benign or malignant counterfactuals, *i.e.*, present to a greater extent in benign (malignant) counterfactuals in at least twice as many images as malignant (benign) counterfactuals. The size of each rectangle (the 'fraction of counterfactual pairs') is then determined as the proportion of counterfactual pairs with a difference noted in the predominant direction, for that reader alone. While readers typically do not attain quantitative agreement on the fraction of counterfactual pairs for a given attribute, the presence and direction of an attribute's effect typically remains consistent. For conciseness, attribute names are shortened as described in Supplementary Table 4.1.

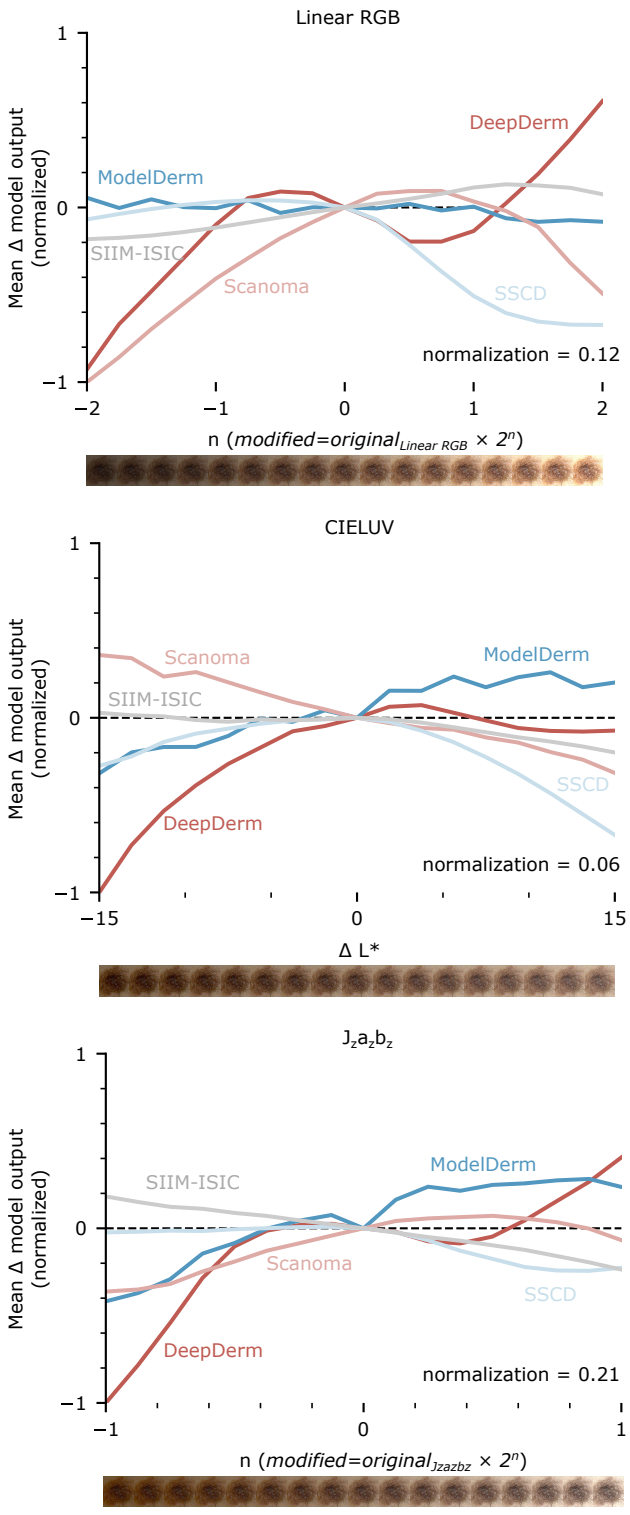


Fig. 4.16 | Effect of programmatic modification of image brightness on AI device predictions. We separately applied three methods of image brightness modification (see Supplementary Methods), then calculated the mean change in AI device output relative to the original, unaltered images. For modifications in linear RGB or $J_z a_z b_z$ space, we modified brightness by applying a multiplicative factor $B = 2^n$; we display AI device responses as a function of n. For modifications in CIELUV space, we add a constant δL^* to the perceptual lightness L^*, where the maximum value of L^* is 100. To facilitate visualization, the vertical axis is normalized to the maximum absolute change in AI device output observed for a given method; the normalization factors are displayed at bottom right. Images indicate the effect of each given brightness modification.

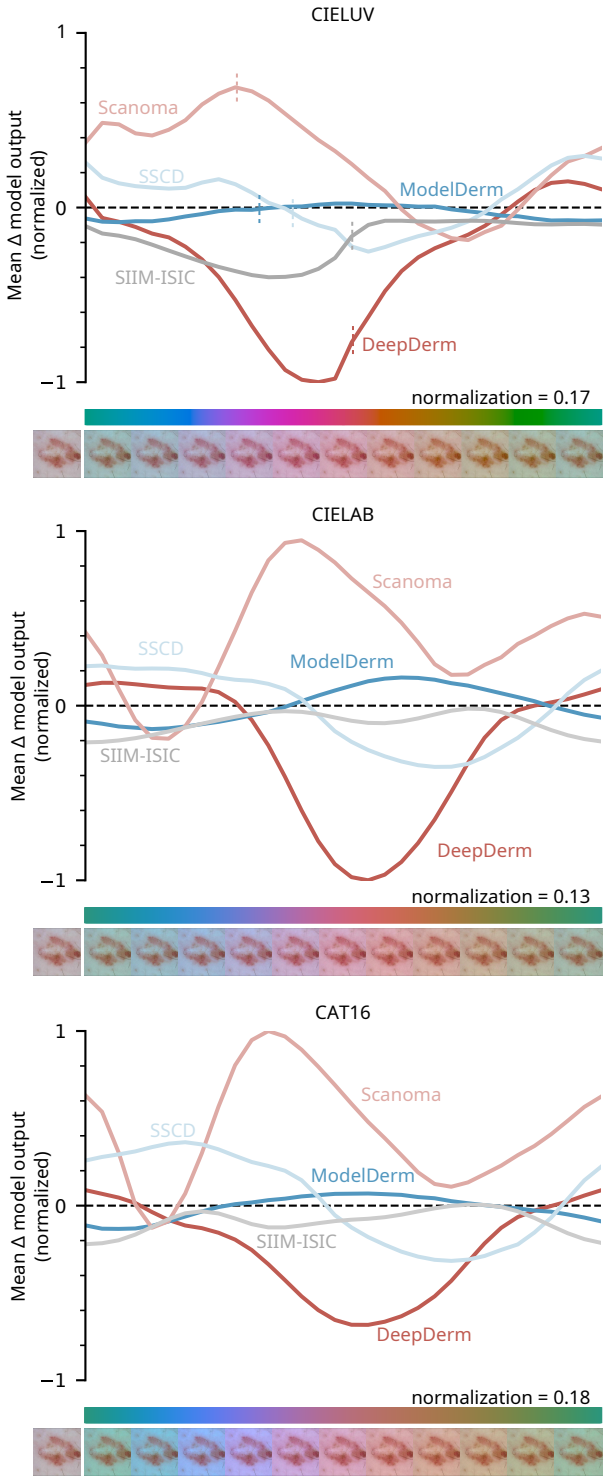


Fig. 4.17 | Effect of programmatic modification of image chromaticity on AI device predictions. We separately applied three methods of image chromaticity modification (see Supplementary Methods), then calculated the mean change in AI device output relative to the original, unaltered images. Each method of chromaticity modification reflects the chromatic adaptation transform (white balancing method) provided by the corresponding color appearance model (CIE 1976 L* u* v*, CIE 1976 L* a* b*, or CAM16). To facilitate visualization, the vertical axis is normalized to the maximum absolute change in AI device output observed for a given method; the normalization factors are displayed at bottom right. Images indicate the effect of each given chromaticity modification. Color bars indicate the hue to which a neutral color (white) is shifted by the chromaticity modification; colorfulness in the color bar (but not example images) is exaggerated for ease of viewing.

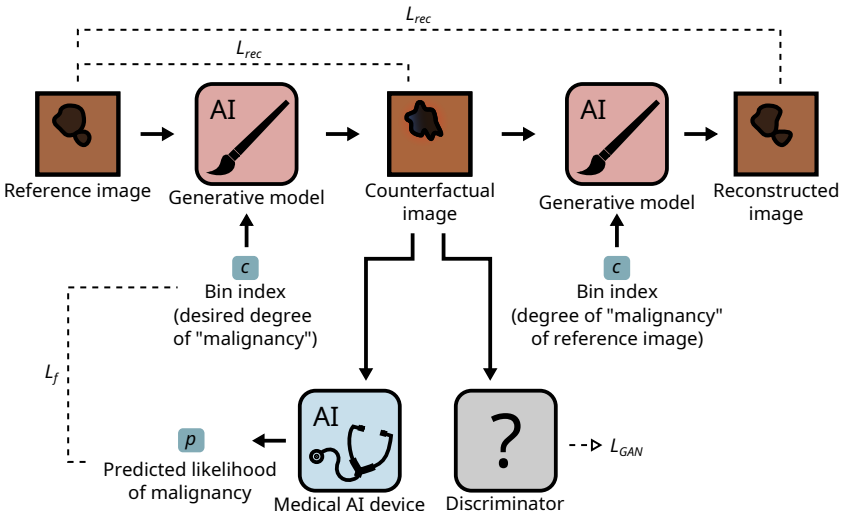


Fig. 4.18 | Schematic of the generative adversarial network (GAN) setup used during training. Given a reference image and a bin index representing the desired prediction from the AI device, the generative model creates a counterfactual image. The loss term L_f enforces that the counterfactual, when evaluated by the AI device (which is held fixed), elicits an output matching that 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}). Finally, we enforce that the counterfactual is similar to the reference image via the reconstruction loss L_{rec}, which is composed of two components: (i) The counterfactual is passed back to the generative model, along with the bin index that corresponds to the AI device's prediction on the reference image, in an attempt to reconstruct the reference image (L_{rec}, top). (ii) 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 the matches the AI device's output on the reference image. In both cases, we compare the reconstructed and reference image via an L1 loss.

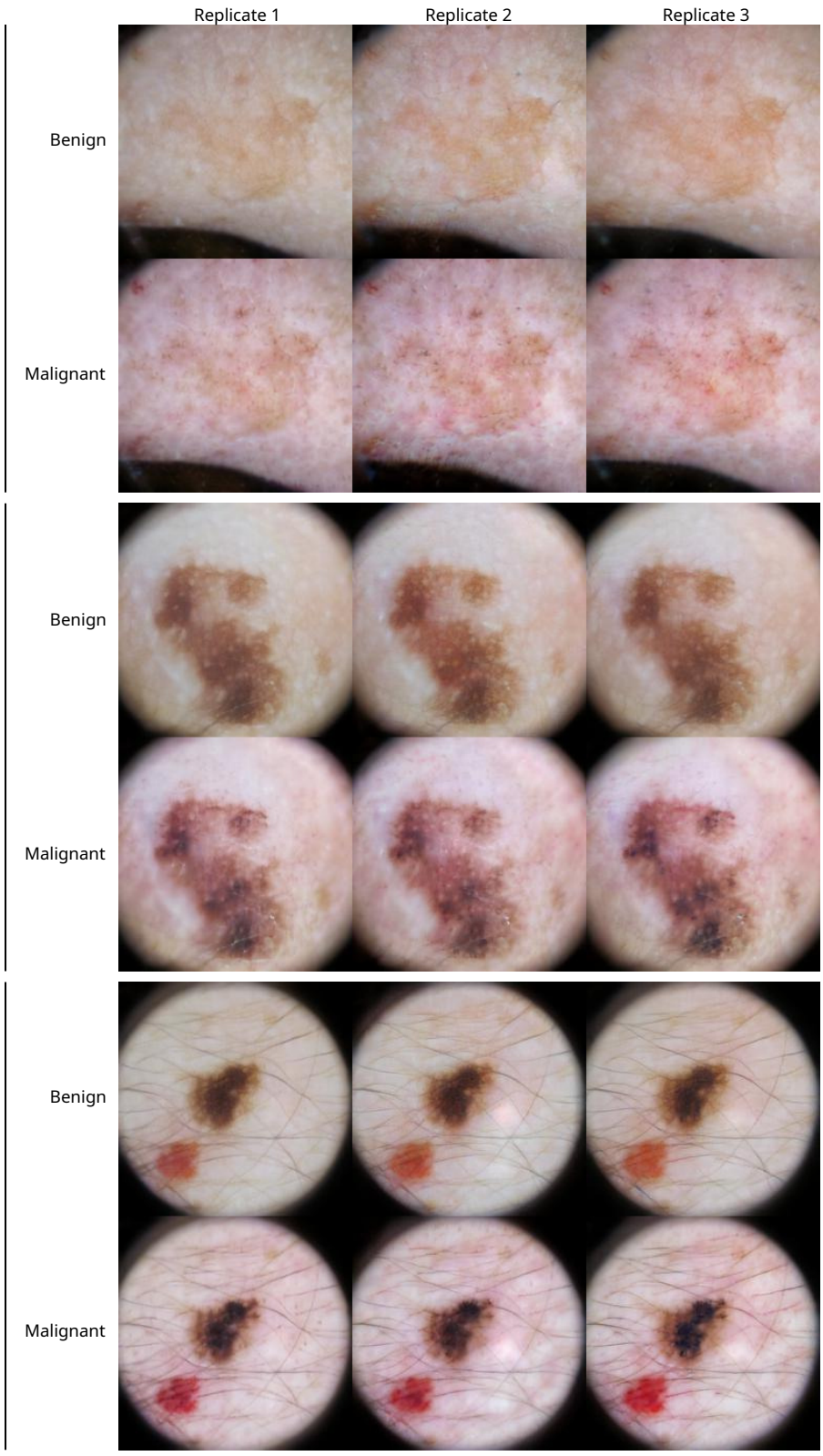


Fig. 4.19 | Independent re-trainings of a generative model using the same training data and AI device. Retrainings preserve key attributes that vary between benign and malignant counterfactuals, such as erythema of the background skin (top), darker pigmentation of the lesion (middle), and multiple colors of pigment in the lesion (bottom). The generative models were trained to evaluate the AI device Scanoma. Images are adapted with permission from the ISIC dataset.[111,112,121]

Chapter 5

Conclusion

The work described in this document has made substantial progress toward a functional, medically-informed understanding of the reasoning processes of machine-learning–based medical-image AI systems.

Major contributions include:

1. Detection of widespread flaws in AI systems for detecting COVID-19
2. Identification of spurious 'shortcuts' that still sometimes generalize to external data
3. Generation of the most complete picture to date of the reasoning processes of machine-learning–based medical-image AI systems
4. Identification of previously unreported signals used by dermatology AI systems, including potentially undesirable signals
5. Introduction of a framework for rigorous combination of expert insights with explainable AI
6. Technical improvements to enable cleaner interpretation of generative AI-based counterfactual images

We also note that this work has promoted important scientific values, including *rigor* and *reproducibility*. Prior to the works described herein, many XAI-based analyses of medical-image AI systems relied on anecdotal results, whereas we have analyzed randomized sets of hundreds to thousands of images, incorporating other important facets such as blinding and expert analysis. We intend that this pioneering level of rigor sets a new standard for the field. By examining previously published models, and by re-implementing previously developed XAI techniques, these works also provide an important service to the scientific community, as they help ensure reproducibility. Likewise, the software and data for each of our studies has been fully open-sourced under permissive licenses.

Based on the work described here, a more detailed picture of how ML-based medical-image AI systems 'reason' has emerged. These systems learn a complex array of signals, including highly detailed, medically important attributes used by physicians. These ML-based systems also often behave in unexpected ways, learning strong dependencies on spurious shortcuts, and sometimes learning to use medical attributes in the opposite manner as physicians. Stakeholders who rely on these systems should be aware of their fragile performance and pay particular attention that their use-case closely matches the scenarios in which the system was evaluated. Regulators should also be aware of the peculiar fragility of these systems, which differs from the failure modes of typical medical devices, including older medical-image analysis software that relied on 'expert systems': in evaluation of medical-image AI systems, retrospective analyses (particularly with 'internal' test data) may greatly overestimate real-world performance, and detailed sub-group analyses may be necessary to help ensure adequate performance across image acquisition devices, geographic locations, and demographic groups.

We also have shown that undesirable 'shortcuts' may persist across data sources, such that external testing (in principle, perhaps even prospective testing) may be inadequate to preclude the possibility of shortcut learning[40] or learning of other undesirable dependencies. This observation argues for dedicated, in-depth analyses of the mechanisms by which medical AI systems generate predictions—*e.g.*, using explainable AI—as an adjunct to robust clinical testing.

To conclude, we draw an analogy to the mechanistic understandings of biology and pharmaceuticals commonly sought in medicine: while dedicated clinical trials provide the strongest evidence for treatment decisions, mechanistic knowledge enables engineering of new treatments, informs design of clinical trials to properly evaluate risks, and may guide

physicians' decisions where better evidence is sparse.[146] We envision that such a mechanistic understanding of medical-image AI will offer similar benefits for AI-based disease detection tools: new abilities to engineer improved systems, better approval processes to mitigate common flaws, and improved abilities of patients and providers to decide whether to use or trust an AI system.

Chapter 6

Acknowledgements

I would like to acknowledge the following individuals and organizations:

- My graduate research mentor, Su-In Lee, for supporting me throughout my PhD and helping guide my projects to ensure a high impact on the field.
- The members of the AI for Medicine and Science (AIMS) lab for thoughtful feedback and research inspiration.
- The University of Washington Medical Scientist Training Program, for supporting me through medical school and research rotations, and for connecting me with a community of colleagues sharing my career path.
- The Seattle Chapter of the ARCS foundation for their generous fellowship.
- My undergraduate research mentor, Lillian Chong, for supporing me through my undergraduate research experiences and teaching me foundations of research.
- The members of the Chong lab for providing mentoring throughout my undergraduate research experience.

Bibliography

1. 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).
2. *Billing and Coding: Remote Imaging of the Retina to Screen for Retinal Diseases* `https://www.cms.gov/medicare-coverage-database/view/article.aspx?articleid=58914`. Accessed: 2023-09-29.
3. Arbabshirani, M. R. *et al.* Advanced machine learning in action: identification of intracranial hemorrhage on computed tomography scans of the head with clinical workflow integration. *npj Digital Medicine* **1** (2018).
4. Wijnberge, M. *et al.* Effect of a machine learning-derived early warning system for intraoperative hypotension vs standard care on depth and duration of intraoperative hypotension during elective noncardiac surgery. *JAMA* **323,** 1052–1060 (11 2020).
5. U.S. Food and Drug Administration Center for Devices and Radiological Health. *De novo classification request for Caption Guidance* `https://www.accessdata.fda.gov/cdrh_docs/reviews/DEN190040.pdf`.
6. Shimabukuro, D. W., Barton, C. W., Feldman, M. D., Mataraso, S. J. & Das, R. Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial. *BMJ Open Respiratory Research* **4.** eprint: `https://bmjopenrespres.bmj.com/content/4/1/e000234.full.pdf`. `https://bmjopenrespres.bmj.com/content/4/1/e000234` (2017).
7. *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.
8. Selvaraju, R. R. *et al.* Grad-CAM: visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision* **128,** 336–359 (2020).
9. Sundararajan, M., Taly, A. & Yan, Q. *Axiomatic attribution for deep networks* in *Proceedings of the 34th International Conference on Machine Learning-Volume 70* (2017), 3319–3328.
10. Singh, N. *et al.* Agreement between saliency maps and human-labeled regions of interest: applications to skin disease classification (2020).
11. DeGrave, A. J., Janizek, J. D. & Lee, S.-I. Course corrections for clinical AI. *Kidney360* **2,** 2019–2023 (12 2021).
12. Food, U. S. & Administration, D. Premarket approval of H1000 ImageChecker: summary of safety and effectiveness data. `https://www.accessdata.fda.gov/cdrh_docs/pdf/p970058.pdf` (1998).
13. Food, U. S. & Administration, D. Premarket approval of SecondLook: summary of safety and effectiveness data. `https://www.accessdata.fda.gov/cdrh_docs/pdf/P010038B.pdf` (2002).
14. Masood, A. & Al-Jumaily, A. A. Computer aided diagnostic support system for skin cancer: a review of techniques and algorithms. *Internation Journal of Biomedical Imaging* (2013).
15. Nachbar, F. *et al.* The ABCD rule of dermatoscopy: High prospective value in the diagnosis of doubtful melanocytic skin lesions. *Journal of the American Academy of Dermatology* **30,** 551–559 (4 1994).
16. Abbasi, N. R. *et al.* Early diagnosis of cutaneous melanoma: revisiting the ABCD criteria. *Journal of the American Medical Association* **8,** 2771–2776 (292 2004).
17. An evaluation of the revised seven-point checklist for the early diagnosis of cutaneous malignant melanoma. *British Journal of Dermatology* **130,** 48–50 (1994).
18. Menzies, S. W., Ingvar, C., Crotty, K. A. & McCarthy, W. H. Frequency and morpholigic characteristics of invasive melanomas lacking specific surface microscopic features. *Archives of Dermatology* **132,** 1178–1182 (1996).
19. Krizhevsky, A., Sutskever, I. & Hinton, G. E. *ImageNet classification with deep convolutional neural networks* in *2012 Conference on Neural Information Processing Systems* (2012).
20. Deng, J. *et al. Imagenet: A large-scale hierarchical image database* in *2009 IEEE conference on computer vision and pattern recognition* (2009), 248–255.
21. Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542,** 115–118 (2017).

22. Bhatia, K., Arora, S. & Tomar, R. Diagnosis of diabetic retinopathy using machine learning classification algorithm. *2016 2nd International Conference on Next Generation Computing Technologies (NGCT),* 347–351 (2016).
23. Lakhani, P. & Sundaram, B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology* **284,** 574–582 (2017).
24. Erhan, D., Bengio, Y., Courville, A. & Vincent, P. *Visualizing higher-layer features of a deep network* tech. rep. 1341 (University of Montreal, 2009).
25. Zeiler, M. D. & Fergus, R. Visualizing and understanding convolutional networks. *Europan Conference on Computer Vision* (2014).
26. Olah, C., Mordvintsev, A. & Schubert, L. Feature visualization: how neural networks build up their understanding of images. *Distill* (2017).
27. Erion, G., Janizek, J. D., Sturmfels, P., Lundberg, S. & Lee, S.-I. Learning explainable models using attribution priors. *arXiv:1906.10670.* `https://arxiv.org/abs/1906.10670` (2019).
28. Smilkov, D., Thorat, N., Kim, B., Viégas, F. & Wattenberg, M. SmoothGrad: removing noise by adding noise. *arXiv:1706.03825v1* (2017).
29. Petsiuk, v., Das, A. & Saenko, K. RISE: randomized input sampling for explanation of black-box models. *British Machine Vision Conference (BMCV) 2018* (2018).
30. Lundberg, S. M. & Lee, S.-I. *A unified approach to interpreting model predictions* in *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems* (2017), 4768–4777.
31. Shapley, L. S. in *Contributions to the Theory of Games* (Princeton University Press, 1953).
32. Ribeiro, M. T., Singh, S. & Guestrin, C. "Why should I trust you?": explanaing the predictions of any classifier. *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016).
33. Kim, B. *et al.* Interpretability beyond feature attribution: quantitative testing with concept activation vectors. *arXiv:1711.11279v5* (2018).
34. Koh, P. W. *et al.* Concept bottleneck models. *Proceedings of the 37th International Conference on Machine Learning* (2020).
35. Young, A. T. *et al.* Stress testing reveals gaps in clinic readiness of image-based diagnostic artificial intelligence models. *npj Digital Medicine* (4 2021).
36. Singla, S., Pollack, B., Chen, J. & Batmanghelich, K. *Explanation by progressive exaggeration* in *International Conference on Learning Representations* (2019).
37. Mertes, S., Huber, T., Weitz, K., Heimerl, A. & André, E. GANterfactual–counterfactual explanations for medical non-experts using generative adversarial learning. *Frontiers in Artificial Intelligence* **5** (2022).
38. Singla, S., Pollack, B., Chen, J. & Batmanghelich, K. Explanation by Progressive Exaggeration. *International Conference on Learning Representations* (2020).
39. Zhu, J.-Y., Park, T., Isola, P. & Efros, A. A. *Unpaired image-to-image translation using cycle-consistent adversarial networks* in *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)* (2017), 2223–2232.
40. Geirhos, R. *et al.* Shortcut learning in deep neural networks. *Nature Machine Intelligence* **2,** 665–673 (2020).
41. Oakden-Rayner, L. Exploring large-scale public medical image datasets. *Academic Radiology* **27,** 106–112 (1 2020).
42. Oakden-Rayner, L., Dunnmon, J., Carneiro, G. & Ré, C. Hidden stratification cuases clinically meaningful failures in machine learning for medical imaging. *ACM Conference on Health Inferenence and Learning,* 151–159 (2020).
43. Winkler, J. K. *et al.* Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA Dermatology* **155,** 1135–1141 (10 2019).
44. Bissoto, A., Fornaciali, M., Valle, E. & Avila, S. *(De) Constructing bias on skin lesion datasets* in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (2019), 2766–2774.
45. Ghoshal, B. & Tucker, A. Estimating uncertainty and interpretability in deep learning for coronavirus (COVID-19) detection. *arXiv:2003.10769* (2020).
46. Ozturk, T. *et al.* Automated detection of COVID-19 cases using deep neural networks with X-ray images. *Computers in Biology and Medicine,* 103792 (2020).
47. Brunese, L., Mercaldo, F., Reginelli, A. & Santone, A. Explainable deep learning for pulmonary disease and coronavirus COVID-19 detection from X-rays. *Computer Methods and Programs in Biomedicine* **196,** 105608 (2020).
48. Karim, M. *et al.* DeepCOVIDExplainer: Explainable COVID-19 predictions based on chest X-ray images. *arXiv:2004.04582* (2020).

49. Poplin, R. *et al.* Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering* **2,** 158–164 (2018).
50. 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).
51. 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).
52. Szegedy, C. *et al.* Intriguing properties of neural networks. *ICLR.* `https://arxiv.org/abs/1312.6199` (2014).
53. 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).
54. Mossa-Basha, M. *et al.* Policies and guidelines for COVID-19 preparedness: Experiences from the University of Washington. *Radiology,* 201326 (2020).
55. Kundu, S., Elhalawani, H., Gichoya, J. W. & Kahn Jr, C. E. How might AI and chest imaging help unravel COVID-19's mysteries? *Radiology. Artificial Intelligence* **2** (2020).
56. Wang, L., Lin, Z. Q. & Wong, A. COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. *Scientific Reports* **10,** 19549 (2020).
57. Hemdan, E. E.-D., Shouman, M. A. & Karar, M. E. COVIDX-Net: A framework of deep learning classifiers to diagnose COVID-19 in X-ray images. *arXiv:2003.11055* (2020).
58. Laghi, A. Cautions about radiologic diagnosis of COVID-19 infection driven by artificial intelligence. *The Lancet Digital Health* **2,** e225 (2020).
59. Harmon, S. A. *et al.* Artificial intelligence for the detection of COVID-19 pneumonia on chest CT using multinational datasets. *Nature Communications* **11,** 4080 (2020).
60. Al-Masni, M. A., Kim, D.-H. & Kim, T.-S. Multiple skin lesions diagnostics via integrated deep convolutional networks for segmentation and classification. *Computer Methods and Programs in Biomedicine* **190,** 105351 (2020).
61. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. *Densely connected convolutional networks* in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), 4700–4708.
62. Zhang, R. *et al.* Diagnosis of COVID-19 Pneumonia Using Chest Radiography: Value of Artificial Intelligence. *Radiology* **In press.** `https://doi.org/10.1148/radiol.2020202944`.
63. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. & Chen, L.-C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. *arXiv:1801.04381* (2019).
64. Xie, S., Girshick, R., Dollár, P., Tu, Z. & He, K. Aggregated Residual Transformations for Deep Neural Networks. *arXiv:1611.05431* (2016).
65. Cohen, J. P., Morrison, P. & Dao, L. COVID-19 image data collection. *arXiv 2003.11597.* `https://github.com/ieee8023/covid-chestxray-dataset` (2020).
66. Wang, X. *et al. ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases* in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), 2097–2106.
67. Of North Ameria, R. S. *RSNA pneumonia detection challenge* `https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/data`.
68. Wehbe, R. M. *et al.* DeepCOVID-XR: an artificial intelligence algorithm to detect COVID-19 on chest radiographs trained and tested on a large US clinical dataset. *Radiology,* 203511 (2020).
69. Li, M. D. *et al.* Automated assessment and tracking of COVID-19 pulmonary disease severity on chest radiographs using convolutional siamese neural networks. *Radiology: Artificial Intelligence* **2,** e200079 (2020).
70. Murphy, K. *et al.* COVID-19 on chest radiographs: a multireader evaluation of an artificial intelligence system. *Radiology* **296,** E166–E172 (2020).
71. Bustos, A., Pertusa, A., Salinas, J.-M. & de la Iglesia-Vayá, M. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical Image Analysis* **66,** 101797 (2020).
72. Vayá, M. d. l. I. *et al.* BIMCV COVID-19+: A large annotated dataset of RX and CT images from COVID-19 patients. *arXiv:2006.01174* (2020).
73. Maguolo, G. & Nanni, L. A critic evaluation of methods for COVID-19 automatic detection from X-ray images. *arXiv:2004.12823* (2020).
74. Castro, D. C., Walker, I. & Glocker, B. Causality matters in medical imaging. *Nature Communications* **11,** 1–10 (2020).
75. Richens, J. G., Lee, C. M. & Johri, S. Improving the accuracy of medical diagnosis with causal machine learning. *Nature Communications* **11,** 1–9 (2020).

76. Janizek, J. D., Erion, G., DeGrave, A. J. & Lee, S.-I. *An adversarial approach for the robust classification of pneumonia from chest radiographs* in *Proceedings of the ACM Conference on Health, Inference, and Learning* (2020), 69–79.
77. Ganin, Y. *et al.* Domain-adversarial training of neural networks. *The Journal of Machine Learning Research* **17,** 2096–2030 (2016).
78. Sagawa, S., Raghunathan, A., Koh, P. W. & Liang, P. *An investigation of why overparameterization exacerbates spurious correlations* in *International Conference on Machine Learning (ICML)* (2020).
79. Bressem, K. K. *et al.* Comparing different deep learning architectures for classification of chest radiographs. *arXiv:2002.08991* (2020).
80. Quionero-Candela, J., Sugiyama, M., Schwaighofer, A. & Lawrence, N. D. *Dataset shift in machine learning* (The MIT Press, 2009).
81. Winther, H. *et al.* COVID-19 image repository. *figshare.* `https://doi.org/10.6084/m9.figshare.12275009`.
82. Jin, Y.-H. *et al.* A rapid advice guideline for the diagnosis and treatment of 2019 novel coronavirus (2019-nCoV) infected pneumonia (standard version). *Military Medical Research* **7,** 4 (2020).
83. Rajpurkar, P. *et al.* CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. *arXiv:1711.05225* (2017).
84. Mitani, A. *et al.* Detection of anaemia from retinal fundus images via deep learning. *Nature Biomedical Engineering* **4,** 18–27 (2020).
85. Zech, J. R. *et al.* Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS medicine* **15,** e1002683 (2018).
86. Ng, M.-Y. *et al.* Imaging profile of the COVID-19 infection: Radiologic findings and literature review. *Radiology: Cardiothoracic Imaging* **2,** e200034 (2020).
87. Wong, H. Y. F. *et al.* Frequency and distribution of chest radiographic findings in COVID-19 positive patients. *Radiology,* 201160 (2020).
88. Gale, W., Oakden-Rayner, L., Carneiro, G., Bradley, A. P. & Palmer, L. J. Detecting hip fractures with radiologist-level performance using deep neural networks. *arXiv:1711.06504* (2017).
89. Paszke, A. *et al.* in *Advances in Neural Information Processing Systems 32* (eds Wallach, H. *et al.*) 8024–8035 (Curran Associates, Inc., 2019). `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf`.
90. Cohen, J. P. *et al.* COVID-19 image data collection: prospective predictions are the future. *arXiv 2006.11988.* `https://github.com/ieee8023/covid-chestxray-dataset` (2020).
91. Sturmfels, P., Lundberg, S. & Lee, S.-I. Visualizing the impact of feature attribution baselines. *Distill* **5,** e22. `https://distill.pub/2020/attribution-baselines/` (2020).
92. Ribeiro, M. T., Wu, T., Guestrin, C. & Singh, S. *Beyond accuracy: Behavioral testing of NLP models with CheckList* in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, Online, July 2020), 4902–4912. `https://www.aclweb.org/anthology/2020.acl-main.442`.
93. North, B. V., Curtis, D. & Sham, P. C. A note on the calculation of empirical p values from Monte Carlo procedures. *American Journal of Human Genetics* **71,** 439–441 (2002).
94. Bien, N. *et al.* Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet. *PLoS medicine* **15,** e1002699 (2018).
95. Arun, N. *et al.* Assessing the (un) trustworthiness of saliency maps for localizing abnormalities in medical imaging. *medRxiv* (2020).
96. Ghorbani, A., Abid, A. & Zou, J. *Interpretation of neural networks is fragile* in *Proceedings of the AAAI Conference on Artificial Intelligence* **33** (2019), 3681–3688.
97. Gu, J., Han, B. & Wang, J. COVID-19: gastrointestinal manifestations and potential fecal-oral transmission. *Gastroenterology* **158,** 1518–1519 (2020).
98. 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).
99. 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).
100. Reddy, S. Explainability and artificial intelligence in medicine. *The Lancet Digital Health* **4,** E214–E215 (4 2022).
101. Liu, Y. *et al.* A deep learning system for differential diagnosis of skin diseases. *Nature Medicine* **26,** 900–908 (2020).

102. Han, S. S. *et al.* Augmented intellignece dermatology: deep neural networks empower medical professionals in diagnosing skin cancer and predicting treatment options for 134 skin disorders. *Journal of Investigative Dermatology* **140,** 1753–1761 (9 2020).
103. Sun, M. *et al.* Accuracy of commercially available smartphone applications for the detection of melanoma. *British Journal of Dermatology* **186,** 744–746 (4 2022).
104. Freeman, K. *et al.* Algorithm based smortphone apps to assess risk of skin cancer in adults: systematic review of diagnostic accuracy studies. *British Medical Journal* **368** (2020).
105. Beltrami, E. J. *et al.* Artificial intelligence in the detection of skin cancer. *Journal of the American Academy of Dermatology* **87,** 1336–1342 (6 2022).
106. Daneshjou, R. *et al.* Disparities in dermatology AI performance on a diverse, curated clinical image set. *Science Advances* **8,** eabq6147 (2022).
107. Han, S. S. *et al.* Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. *Journal of Investigative Dermatology* **138,** 1529–1538 (2018).
108. Giotis, I. *et al.* MED-NODE: A computer-assisted melanoma diagnosis system using non-dermoscopic images. *Expert Systems with Applications* **42,** 6578–6585 (2015).
109. 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).
110. Rotemberg, V. *et al.* A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific Data* **8,** 34 (2021).
111. 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).
112. Combalia, M. *et al.* BCN20000: Dermoscopic Lesions in the Wild. *arXiv:1908.02288* (2019).
113. Groh, M. *et al.* Evaluating deep neurla networks trained on clinical images in dermatology with the Fitzpatrick 17k dataset. *Proceedings of the Computer Vision and Pattern Recognition (CVPR) Sixth ISIC Skin Image Analysis Workshop* (2021).
114. Shi, K. *et al. Journal of the American Academy of Dermatology* **83,** 1028–1034 (4 2020).
115. Yélamos, O. *et al. Journal of the American Academy of Dermatology* **80,** 365–377 (2 2019).
116. Halpern, A. C., Marghoob, A. A. & Reiter, O. Melanoma warning signs: what you need to know about early signs of skin cancer. `https://www.skincancer.org/skin-cancer-information/melanoma/melanoma-warning-signs-and-images/` (2023) (2021).
117. Massi, D., De Giorgi, V., Carli, P. & Santucci, M. Diagnostic significance of the blue hue in dermoscopy of melanocytic lesions: a dermoscopic-pathologic study. *The American Journal of Dermatopathology* **23,** 463–469 (2001).
118. Marghoob, N. G., Liopyris, K. & Jaimes, N. Dermoscopy: a review of the structures that facilitate melanoma detection. *Journal of Osteopathic Medicine* (2019).
119. Rader, R. K. *et al.* The pink rim sign: location of pink as an indicator of melanoma in dermoscopic images. *Journal of Skin Cancer* (2014).
120. Fitzpatrick, J. E., High, W. A. & Kyle, W. L. in, 477–488 (Elsevier, 2018).
121. 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).
122. 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.
123. Oliveria, S. A., Saraiya, M., Geller, A. C., Heneghan, M. K. & Jorgensen, C. Sun exposure and risk of melanoma. *Archives of Disease in Childhood* **91,** 131–138 (2 2006).
124. On Illumination, I. C. ISO/CIE 11664-5:2016(E) colorimetry - part 5: CIE 1976 L*u*v* colour space and u', v' uniform chromaticity scale diagram (2016).
125. Deng, Z., Gijsenij, A. & Zhang, J. Source camera identification using auto-white balance approximation. *2011 IEEE International Conference on Computer Vision,* 57–64 (2011).
126. Tschandl, P. *et al.* Human-computer collaboration for skin cancer recognition. *Nature Medicine* **26,** 1229–1234 (2020).
127. Tschandl, P. *et al.* Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based international, diagnostic study. *Lancet Oncology* **20,** 938–947 (7 2019).
128. Weber, P., Sinz, C., Rinner, C., Kittler, H. & Tschandl, P. Perilesional sun damage as a diagnostic clue for pigmented actinic keratosis and Bowen's disease. *Journal of the European Academy of Dermatology and Venereology* **35,** 2022–2026 (10 2021).

129. Wu, E. *et al.* Toward stronger FDA approval standards for AI medical devices. *Stanford University Human-centered Artificial Intelligence* (2022).
130. Bansal, G. *et al.* Does the whole exceed its parts? The effect of AI explanations on complementary team performance. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (2021).
131. Rok, R. & Weld, D. S. In Search of Verifiability: Explanations Rarely Enable Complementary Performance in AI-Advised Decision Making. *arXiv:2305.07722v3* (2023).
132. Roth, L. Looking at Shirley, the ultimate norm: colour balance, image technologies, and cognitive equity. *Canadian Journal of Communication* **34,** 111–136 (2009).
133. Lester, J., Clark, L., Linos, E. & Daneshjou, R. *British Journal of Dermatology* **184,** 1177–1179 (6 2021).
134. Tan, M. *et al.* MnasNet: platform-aware neural architecture search for mobile. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* 2820–2828 (2019).
135. Jacob, B. *et al.* Quantization and training of neural networks for efficient integer-arithmetic-only inference. *Preprint at arXiv:1712.05877* (2017).
136. 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).
137. Tan, M. & Le, Q. EfficientNet: rethinking model scaling for convolutional neural networks. *Proceedings of the 36th International Conference on Machine Learning (ICML 2019),* 6105–6114 (2019).
138. Hu, J., Shen, L. & Sun, G. Squeeze-and-excitation networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* 7132–7141 (2018).
139. Zhang, H. *et al.* ResNeSt: split-attention networks. *Preprint at arXiv:2004.08955* (2020).
140. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlesn, J. & Wojna, Z. Rethinking the inception architecture for computer vision. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 2818–2826 (2016).
141. Safdar, M., Cui, C., Kim, Y. J. & Luo, M. R. Perceptually uniform color space for image signals including high dynamic range and wide gamut. *Optics Express* **25,** 15131–15151 (13 2017).
142. Judd, D. B. Hue, saturation, and lightness of surface colors with chromatic illumination. *Journal of Research of the National Bureau of Standards* **24,** 293–333 (1940).
143. Von Kries, J. A. Die Gesichtsempfindungen. *Handbuch der Physiologie des Menschen* **3,** 109–282 (1905).
144. On Illumination, I. C. ISO/CIE 11664-4:2008(en) colorimetry - part 4: CIE 1976 L*a*b* colour space (2008).
145. Li, C. *et al.* Comprehensive color solutions: CAM16, CAT16, and CAM16-UCS. *Color Research & Application* **42,** 703–718 (6 2017).
146. Howick, J., Glasziou, P. & Aronson, J. K. Evidence-based mechanistic reasoning. *Journal of the Royal Society of Medicine* **103,** 433–441 (11 2010).