AI for radiographic COVID-19 detection selects shortcuts over signal

Alex J. DeGrave[1,2,*], Joseph D. Janizek[1,2,*], and Su-In Lee[1,**]

[1]Paul G. Allen School of Computer Science and Engineering, University of Washington
[2]Medical Scientist Training Program, University of Washington
[*]These authors contributed equally and are ordered alphabetically
[**]Corresponding: suinlee@cs.washington.edu

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.

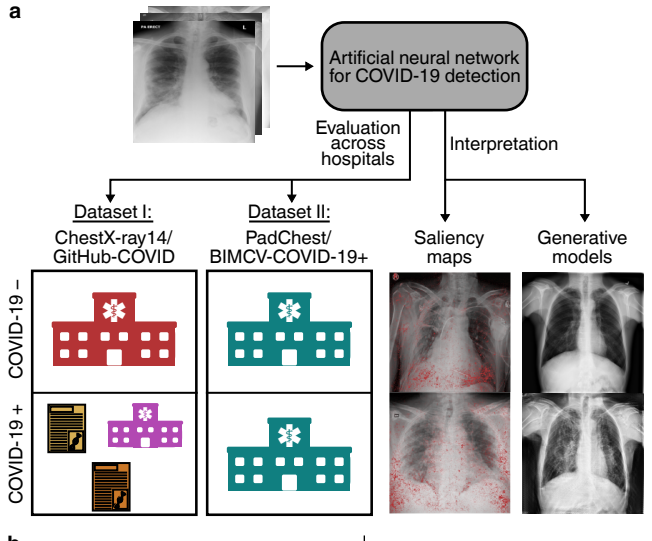
1 Main

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 [1, 2]. While numerous recent publications and preprints report machine learning models with high performance at this task [3–8], the trustworthiness of these models needs to be rigorously evaluated before deployment in a clinical setting [9]. 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 correlations between labels of COVID-19 positivity and confounding variables, i.e., "shortcuts" [10]. While such spurious correlations 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 [3–8]. To understand the extent of this issue, we trained deep convolutional neural networks [11, 12] (Methods Section 2.2, Supplementary Fig. 1) in the manner of these previous publications [3–8] and interrogated their reliance on confounds by evaluating their performance on radiographs from external sources and elucidating their inner workings via widely-used, state-of-the-art interpretability techniques [13–16].

To train and evaluate our models, we created two datasets (Fig. 1a, Supplementary Table 1). Dataset I consisted of COVID-19 positive radiographs from the GitHub-COVID dataset [17] supplemented with COVID-19 negative radiographs from the National Institutes of Health's ChestX-ray14 dataset [18]. Dataset I is similar to the datasets used for training in recent publications on AI for COVID-19 detection [3–8]. Unlike the datasets used in recent publications, which collected COVID-19 positive and negative images from different sources, Dataset II corresponds to a seemingly more ideal case where both COVID-19 positive and negative images were drawn from the same source. This dataset consisted of radiographs from a single regional healthcare system, originally published as the PadChest and BIMCV-COVID-19+ datasets (Fig. 1a-b) [19, 20].

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. 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 [10].

While we initially expected that a dataset built from radiographs drawn from a single hospital system would be less likely to contain spurious correlations that enable ML models to take shortcuts, we found that models trained on Dataset



b

	Dataset I			Dataset II		
	Combined	Chest-X-ray14	GitHub-COVID	Combined	PadChest	BIMCV-COVID-19+
# radiographs	112,528	112,120	408	97,866	96,270	1,596
# patients	31,067	30,805	262	64,954	63,939	1,105
% COVID-19+	0.2	0	76.5	1.6	0	100
% AP images	39.9	40	26	5.6	4.7	58.1

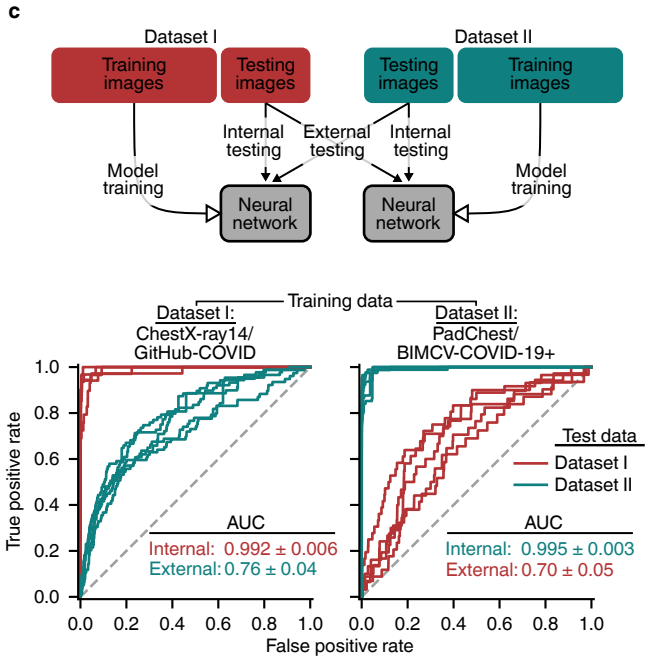


Fig. 1 | Overview of the study design. a, A neural network model is trained to detect COVID-19 using radiographs from either of two datasets, and then evaluated on both datasets to learn how performance may drop in deployment (i.e., a generalization gap). Intepretability methods are then applied to infer 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 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.

II also exhibit high performance on internal test data and low performance on external test data (Fig. 1c, right). Thus, dataset-level confounding may pose a severe issue even in datasets derived from a single hospital system, contrary to the conclusions of contemporary work [21]. 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 predict 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 [22–24]. 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 [25].

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 [26], 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 (Extended Data Fig. 1). 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 [27]. Likewise, we found that replacing the multi-label classification scheme of our original models with a simpler single-label classification scheme (see Methods Section 2.2) 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 (Extended Data Fig. 2), 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 [28].

We further interrogated the trained AI models using saliency maps [13, 29, 30], which highlight the regions of each radiograph that contribute most to the model's prediction (Supplementary Note 1), 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. 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 (Fig. 2a and Extended Data Fig. 3), 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 GitHub-COVID data source [17] (Supplementary Fig. 2), which aligns with a previous study finding that ML models can learn to detect pneumonia based on spurious differences in text on radiographs [31]. 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 [32] and instead likely reflect dataset-level differences in patient positioning and radiographic projection (i.e., AP vs. PA view) [24]. 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 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. 2b, blue boxes, and Supplementary Fig. 3 and 4). However, the generative networks frequently add or remove laterality markers and annotations (Fig. 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. 2b, dashed red boxes), supporting our previous assertion that systematic, dataset-level differences in patient positioning and radiographic projection provide ML models an undesirable shortcut to detection of 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

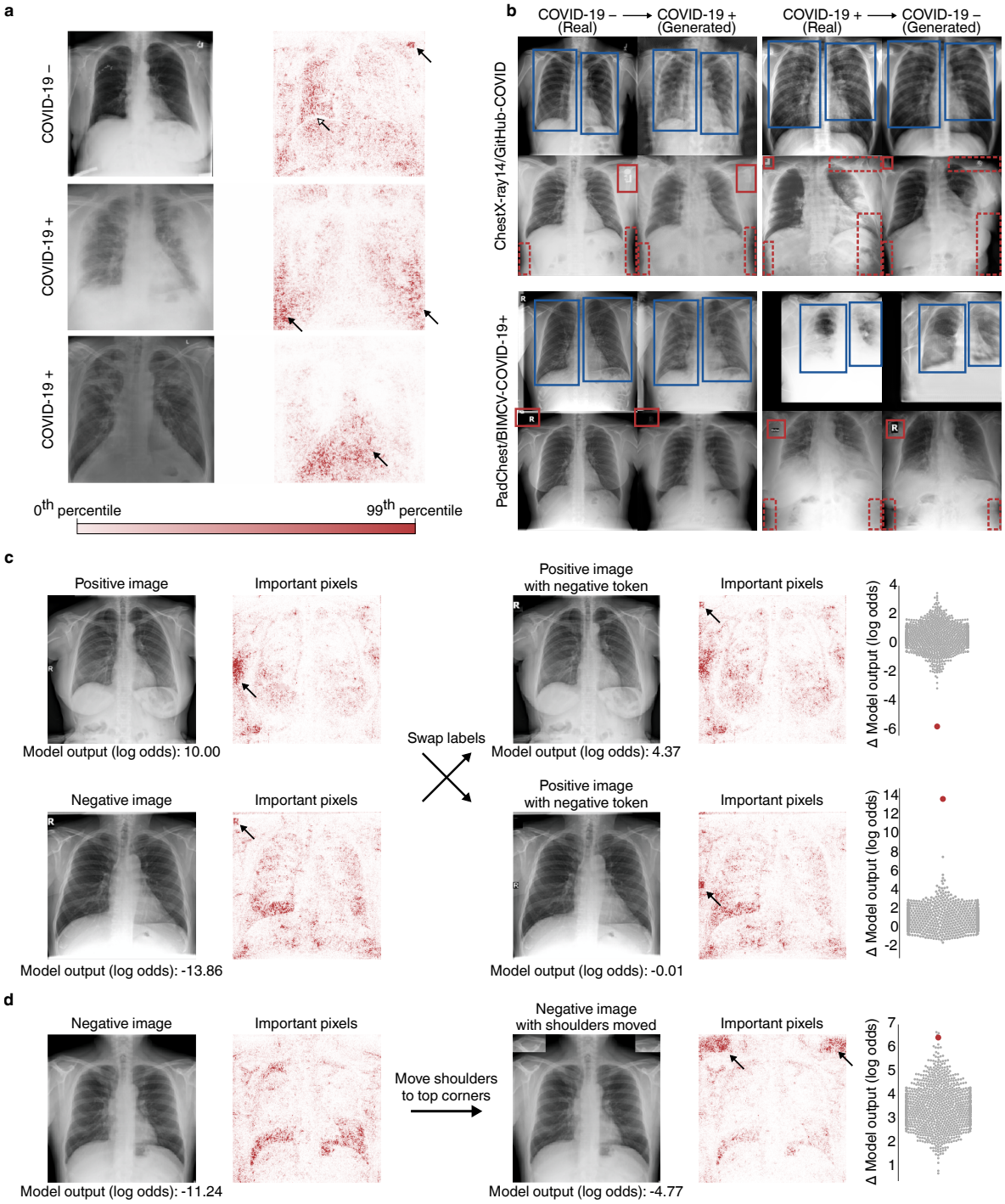


Fig. 2 | Explainable AI visualizes image factors important for deep neural networks trained to detect COVID-19 in radiographs. a, Saliency maps for our neural network models indicating the regions of each radiograph with the greatest influence on the model's prediction. Top, in a COVID-19 negative radiograph, in addition to the highlighting in the lung fields (open arrow), the saliency maps also emphasize laterality tokens (closed 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). Colorbar indicates saliency map pixel importances by percentile. (Caption continued on next page.)

Fig. 2 | (Previous page.) **b**, Radiographs and their corresponding transformations by a generative adversarial network (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. **c**, (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 $< 1 \times 10^{-3}$ based on Monte Carlo substitution of random image patches, n=1000); Δ negative image (log odds) $= 13.85$ ($p < 1 \times 10^{-3}$, n=1000) (Methods Section 2.5). 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. **d**, 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 $= 4 \times 10^{-3}$, 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.

class rather than the original class (Extended Data Fig. 4), demonstrating that the majority of features used by our classifiers were altered by the GAN. 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.

We next aimed to experimentally validate the importance of spurious confounds to our models by manually modifying key features (see also Methods Section 2.5, Supplementary Note 1, and Supplementary Fig. 7). We first swapped laterality markers from a COVID-19 positive and COVID-19 negative 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 (Fig. 2c). 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, 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 (Supplementary Fig. 2, 3, 4 and Extended Data Fig. 5), 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 and 5), 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, 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.

Importantly, we also observe that some potential confounders may generalize across datasets, meaning that some of the apparent external test set performance may *still* be due to spurious correlations rather than genuine radiographic evidence of COVID-19. For example, if there is a consistent difference in the proportion of men and women who are COVID-19 positive, and a model can predict patient sex with high accuracy in both internal and external test data, this indicates that the model's external test performance could be due to this clinical, but non-radiographic information. We find that the ability of deep convolutional neural networks to detect patient sex, and radiographic projection generalize well (Extended Data Fig. 6), which indicates that such concepts can be exploited as shortcuts by COVID-19 classifiers. Furthermore, we have already noted our models rely on patient positioning information, and we find that this confound indeed generalizes between Dataset I and Dataset II (Extended Data Fig. 5). 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.

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. Our findings support common-sense solutions to alleviate this problem, 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 [33, 34]. 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 [9].

2 Methods

2.1 Datasets and preprocessing

To train and evaluate our models, we combined images from four large open-access repositories of chest radiographs into two datasets (1). The first, which we refer to as Dataset I (multi-source dataset), was designed to replicate the datasets used to develop and evaluate the most popular COVID-19 diagnostic models [4]. 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 [18]. We collected COVID-19 positive images from the GitHub COVID-Chestxray dataset [17], 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. Following the recommendations in [35], we filtered lateral chest radiographs and AP supine frontal radiographs from the data, so that we only used frontal PA and frontal AP upright radiographs in the data. We also filtered out any chest CT images from the COVID-Chestxray data.

The second dataset, which we refer to as Dataset II (single-source dataset), was designed to represent a more ideal case in terms of domain confounding – both COVID-19 positive and COVID-19 negative images were taken from the same hospital system. We collected COVID-19 negative images from the PadChest repository, representing 96,270 radiographs from 63,939 patients from a single hospital system in Valencia, Spain (from 11 individual hospitals) [19]. The COVID-19 positive images in our dataset were taken from the BIMCV-COVID-19+ dataset, which represents 1,596 images from 1,015 patients, also from the same regional hospital system in Valencia, Spain [20].

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 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.

2.2 Model architecture and training procedure

In our investigation, we aimed to faithfully replicate the modeling choices employed in recent high-performance models for COVID-19 classification, while also following established best practices for the classification of pathologies from chest radiographs using deep learning. 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 [11]. This architecture has not only been used in a variety of recent models for COVID-19 classification [4, 5], but has also been used for the diagnosis of non-COVID pneumonia [24, 29], as well as for more general radiographic classification [36].

Following the approach in recent COVID-19 models [4, 5], we first pre-trained the model on ImageNet, a large database of natural images [37]. Forcing models to first learn general image features should also serve as an inductive bias to prevent overfitting on domain-specific features [24]. 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 [29].

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 [38], version 1.4, on NVIDIA RTX 2080 graphics processing units and required approximately 5 hours of training time per replicate.

To test the hypothesis that lower-capacity models may not learn spurious correlations [26], we also trained two lower-capacity models. The first, an AlexNet model [12], 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 [26], 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).

2.3 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* [16]. 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:

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

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 [16] 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 [39]. 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 1) [13].

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 compared 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. 6).

2.4 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 [14, 15]. 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:

$$\mathcal{L}_{\text{GAN}}(G, D_{\mathcal{Y}}, \mathcal{X}, \mathcal{Y}) = \mathbb{E}_{Y \sim p_{\text{data}}(Y)}[\log D_{\mathcal{Y}}(Y)] + \mathbb{E}_{X \sim p_{\text{data}}(X)}[\log(1 - D_{\mathcal{Y}}(G(X))], \tag{2}$$

$$\mathcal{L}_{\text{GAN}}(F, D_{\mathcal{X}}, \mathcal{Y}, \mathcal{X}) = \mathbb{E}_{X \sim p_{\text{data}}(X)}[\log D_{\mathcal{X}}(X)] + \mathbb{E}_{Y \sim p_{\text{data}}(Y)}[\log(1 - D_{\mathcal{X}}(F(Y))], \tag{3}$$

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:

$$\mathcal{L}_{\text{cyc}}(G, F) = \mathbb{E}_{X \sim p_{\text{data}}(X)}[\|F(G(X)) - X\|_1] + \mathbb{E}_{Y \sim p_{\text{data}}(Y)}[\|G(F(Y)) - Y\|_1]. \tag{4}$$

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

$$\mathcal{L} = \mathcal{L}_{\text{GAN}}(G, D_{\mathcal{Y}}, \mathcal{X}, \mathcal{Y}) + \mathcal{L}_{\text{GAN}}(F, D_{\mathcal{X}}, \mathcal{Y}, \mathcal{X}) + \mathcal{L}_{\text{cyc}}(G, F) \tag{5}$$

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 [40]. 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.

2.5 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 [41]. 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, we generate empirical p-values by comparing to a null distribution generated by swapping 1000 random patches of each image of the same dimensions as the text markers (Fig. 2c, Supplementary Fig. 8). 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. 2d). We tested for statistical significance 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.

3 Data availability

All radiographs are compiled from publicly-available data repositories and links for download are provided at https://github.com/suinleelab/cxr_covid.

4 Code availability

All of the code necessary to reproduce our experimental findings can be found at https://github.com/suinleelab/cxr_covid.

References

1. Mossa-Basha, M. *et al.* Policies and guidelines for COVID-19 preparedness: Experiences from the University of Washington. *Radiology,* 201326 (2020).
2. 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).
3. Ghoshal, B. & Tucker, A. Estimating uncertainty and interpretability in deep learning for coronavirus (COVID-19) detection. *arXiv:2003.10769* (2020).
4. Wang, L. & Wong, A. COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. *arXiv:2003.09871* (2020).
5. 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).
6. 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).

7. 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).
8. Karim, M. *et al.* DeepCOVIDeExplainer: Explainable COVID-19 predictions based on chest X-ray images. *arXiv:2004.04582* (2020).
9. Laghi, A. Cautions about radiologic diagnosis of COVID-19 infection driven by artificial intelligence. *The Lancet Digital Health* **2,** e225 (2020).
10. Geirhos, R. *et al.* Shortcut learning in deep neural networks. *arXiv:2004.07780* (2020).
11. 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.
12. Krizhevsky, A., Sutskever, I. & Hinton, G. E. *ImageNet classification with deep convolutional neural networks* in *2012 Conference on Neural Information Processing Systems* (2012).
13. 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.
14. Zhu, J.-Y., Park, T., Isola, P. & Efros, A. A. *Unpaired image-to-image translation using cycle-consistent adversarial networks* in *Proceedings of the IEEE international conference on computer vision* (2017), 2223–2232.
15. Singla, S., Pollack, B., Chen, J. & Batmanghelich, K. *Explanation by progressive exaggeration* in *International Conference on Learning Representations* (2019).
16. 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).
17. Cohen, J. P., Morrison, P. & Dao, L. COVID-19 image data collection. *arXiv 2003.11597.* `https://github.com/ieee8023/covid-chestxray-dataset` (2020).
18. 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.
19. 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. *arXiv:1901.07441* (2019).
20. 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).
21. Maguolo, G. & Nanni, L. A critic evaluation of methods for COVID-19 automatic detection from X-ray images. *arXiv:2004.12823* (2020).
22. Castro, D. C., Walker, I. & Glocker, B. Causality matters in medical imaging. *Nature Communications* **11,** 1–10 (2020).
23. Richens, J. G., Lee, C. M. & Johri, S. Improving the accuracy of medical diagnosis with causal machine learning. *Nature Communications* **11,** 1–9 (2020).
24. 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.
25. Ganin, Y. *et al.* Domain-adversarial training of neural networks. *The Journal of Machine Learning Research* **17,** 2096–2030 (2016).
26. 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).
27. Bressem, K. K. *et al.* Comparing different deep learning architectures for classification of chest radiographs. *arXiv:2002.08991* (2020).
28. Quionero-Candela, J., Sugiyama, M., Schwaighofer, A. & Lawrence, N. D. *Dataset shift in machine learning* (The MIT Press, 2009).
29. Rajpurkar, P. *et al.* CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. *arXiv:1711.05225* (2017).
30. Mitani, A. *et al.* Detection of anaemia from retinal fundus images via deep learning. *Nature Biomedical Engineering* **4,** 18–27 (2020).
31. 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).

32. Ng, M.-Y. *et al.* Imaging profile of the COVID-19 infection: Radiologic findings and literature review. *Radiology: Cardiothoracic Imaging* **2,** e200034 (2020).

33. 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).

34. Wong, H. Y. F. *et al.* Frequency and distribution of chest radiographic findings in COVID-19 positive patients. *Radiology,* 201160 (2020).

35. 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).

36. 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).

37. Deng, J. *et al. Imagenet: A large-scale hierarchical image database* in *2009 IEEE conference on computer vision and pattern recognition* (2009), 248–255.

38. 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`.

39. Sturmfels, P., Lundberg, S. & Lee, S.-I. Visualizing the impact of feature attribution baselines. *Distill* **5,** e22. `https://distill.pub/2020/attribution-baselines/` (2020).

40. Ruano, A. PyTorch-CycleGAN. `https://github.com/aitorzip/PyTorch-CycleGAN` (2017).

41. 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`.

42. Bien, N. *et al.* Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet. *PLoS medicine* **15,** e1002699 (2018).

43. Arun, N. *et al.* Assessing the (un) trustworthiness of saliency maps for localizing abnormalities in medical imaging. *medRxiv* (2020).

44. 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.

5 Acknowledgments

This work was funded by the National Institutes of Health [R35 GM 128638]. We thank Hugh Chen and Gabriel Erion for providing thoughtful feedback while writing the manuscript. We also thank Dr. David Janizek, MD for helpful insight into the interpretation of COVID-19 on chest radiographs.

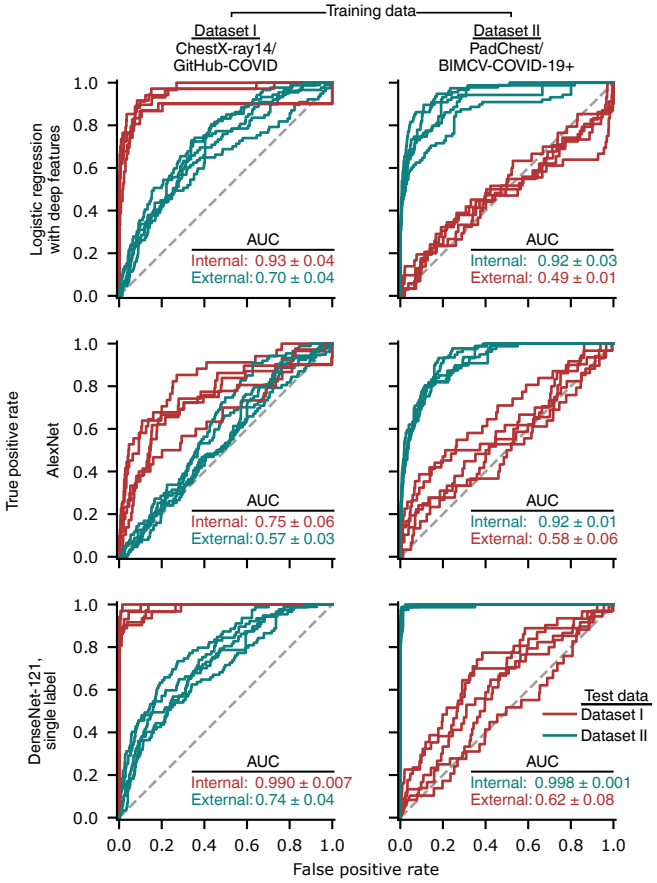
6 Author contributions

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.

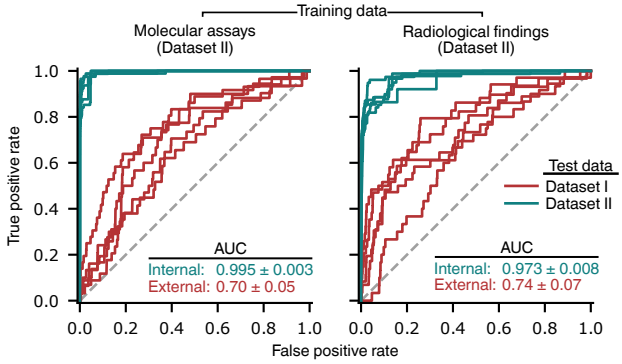
7 Competing interests

The authors declare no competing interests.

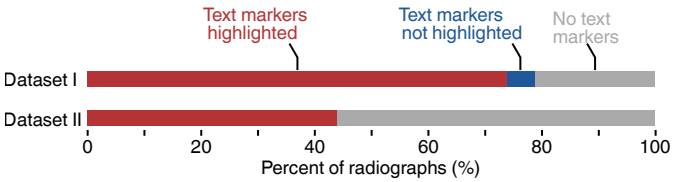
8 Extended Data



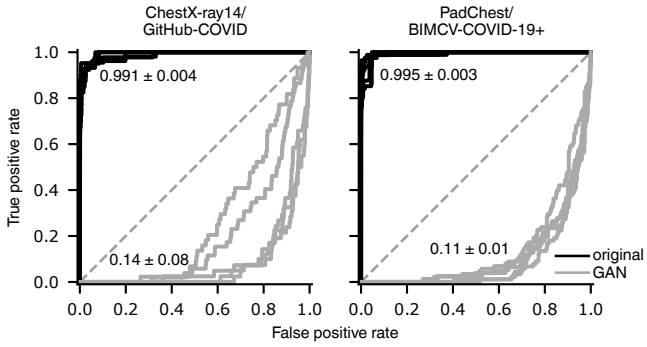
Extended Data Fig. 1 | Generalization performance of alternative models, as measured by receiver-operating characteristic (ROC) curves. The first two rows correspond to models in which the capacity to overfit, which has been implicated in learning of spurious associations [26], has been reduced. The 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 [12] 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).



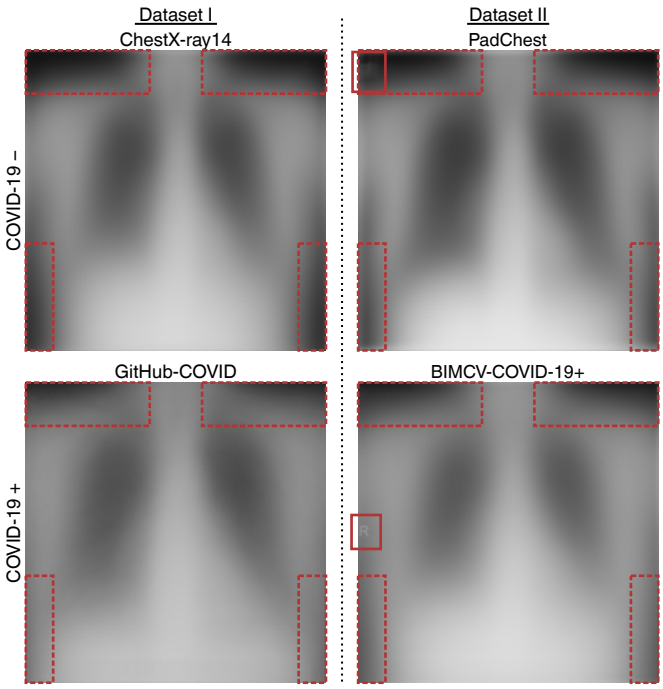
Extended Data Fig. 2 | Evaluation of the impact on generalization performance of *concept shift*, a change in the classification task between the training and testing datasets. In addition to the learning of spurious correlations that do not remain constant between datasets, generalization performance 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 [17], 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 [20] 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. 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).



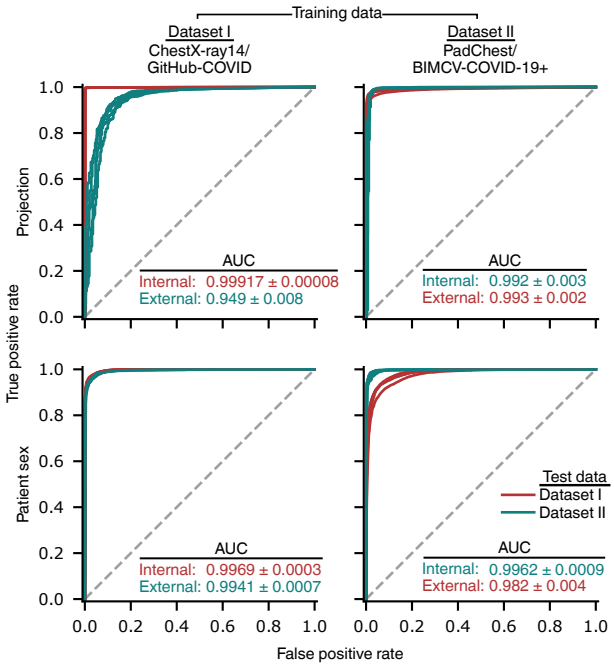
Extended Data Fig. 3 | Analysis of the frequency at which saliency maps highlight laterality markers as important features. To assess the frequency, a random sample of 100 radiographs and their corresponding saliency maps was chosen from each dataset, and each radiograph was manually categorized as (i) 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.



Extended Data Fig. 4 | 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).



Extended Data Fig. 5 | Average images from the four repositories used to construct datasets in this study. 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.



Extended Data Fig. 6 | Evaluation of the generalization performance of classifiers that predict potential confounds. Inset values indicate area under the ROC curve (AUC, mean ± standard deviation, n=5).

Supplementary Information

Supplementary Note 1

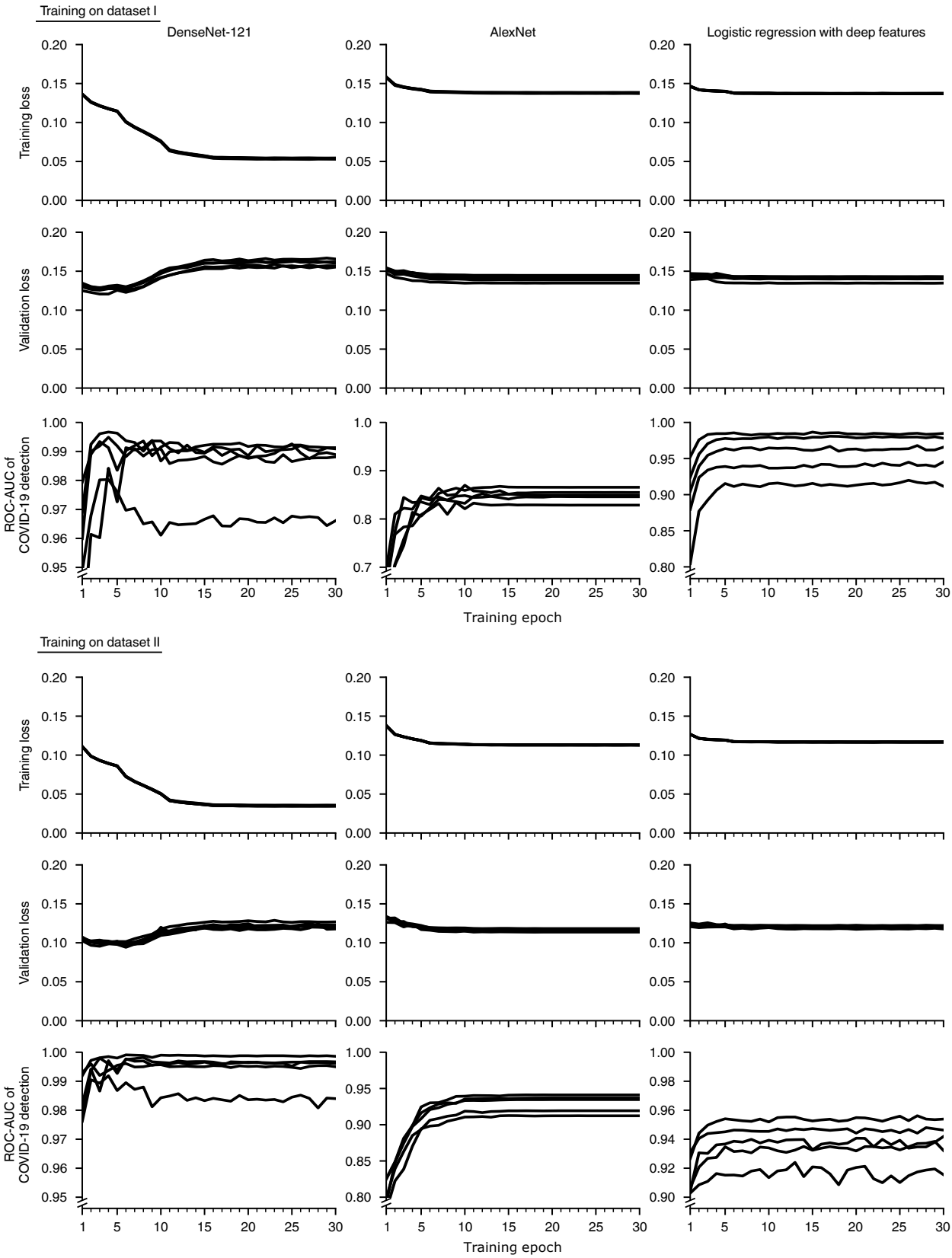
While saliency maps are widely used to interpret image-based artificial intelligence systems [29, 30, 42], the reliability of these approaches has been disputed by contemporary work, which observes that saliency maps explaining medical imaging classifiers fail to localize medically relevant pathology [43]. 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 [44].

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. 7).

Supplementary Figures

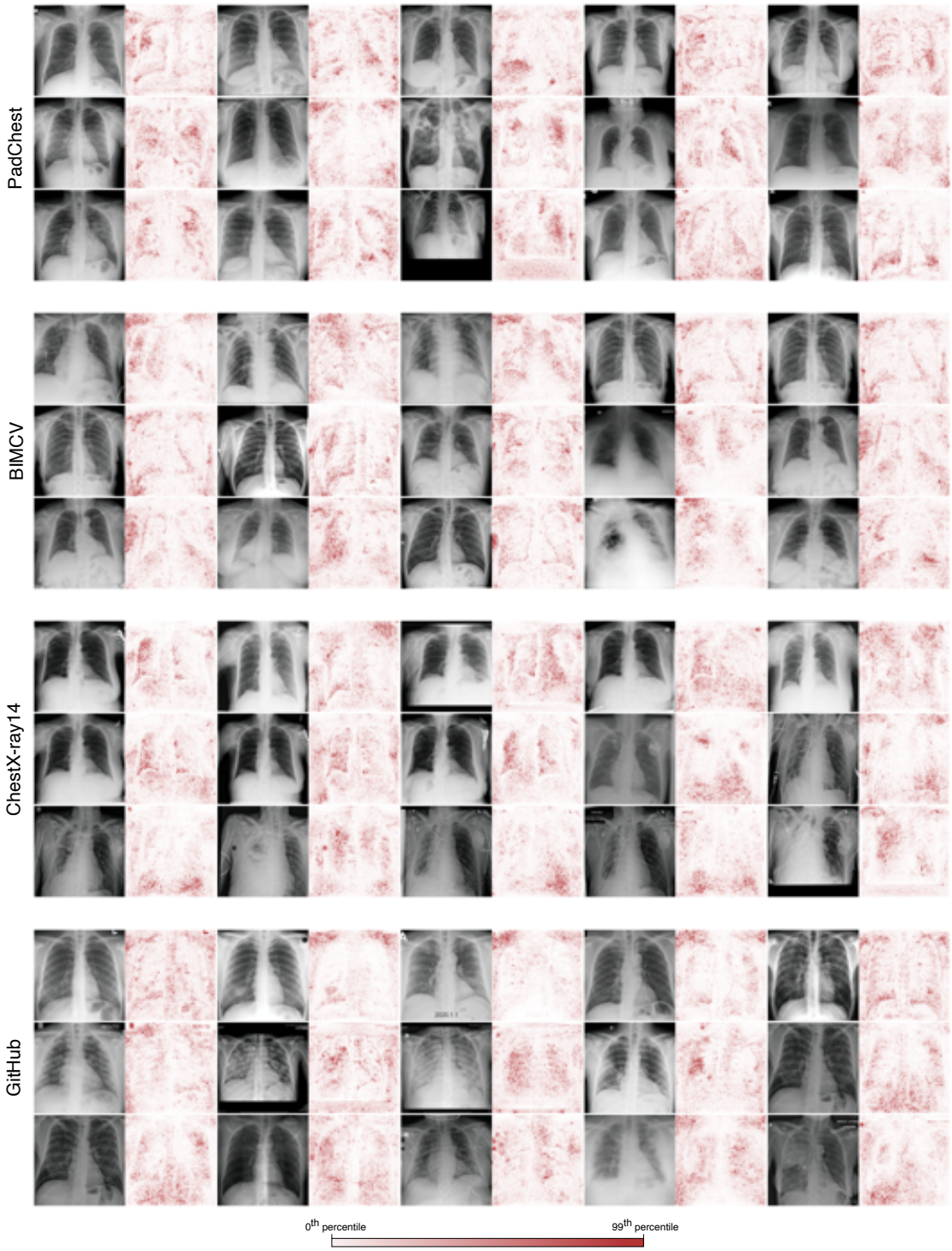
	Dataset I (multi-source)			**Dataset II (single-source)**		
	Combined	CXR14	Cohen et al.	Combined	PadChest	BIMCV-COVID
CXR #s	112,528	112,120	408	97,866	96,270	1,596
Patients, #s	31,067	30,805	262	64,954	63,939	1,015
Age, mean (std)	46.9 (16.8)	46.9 (16.8)	57.0 (16.4)	65.4 (20.1)	65.5 (20.1)	61.2 (16.0)
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)

Supplementary Table 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.

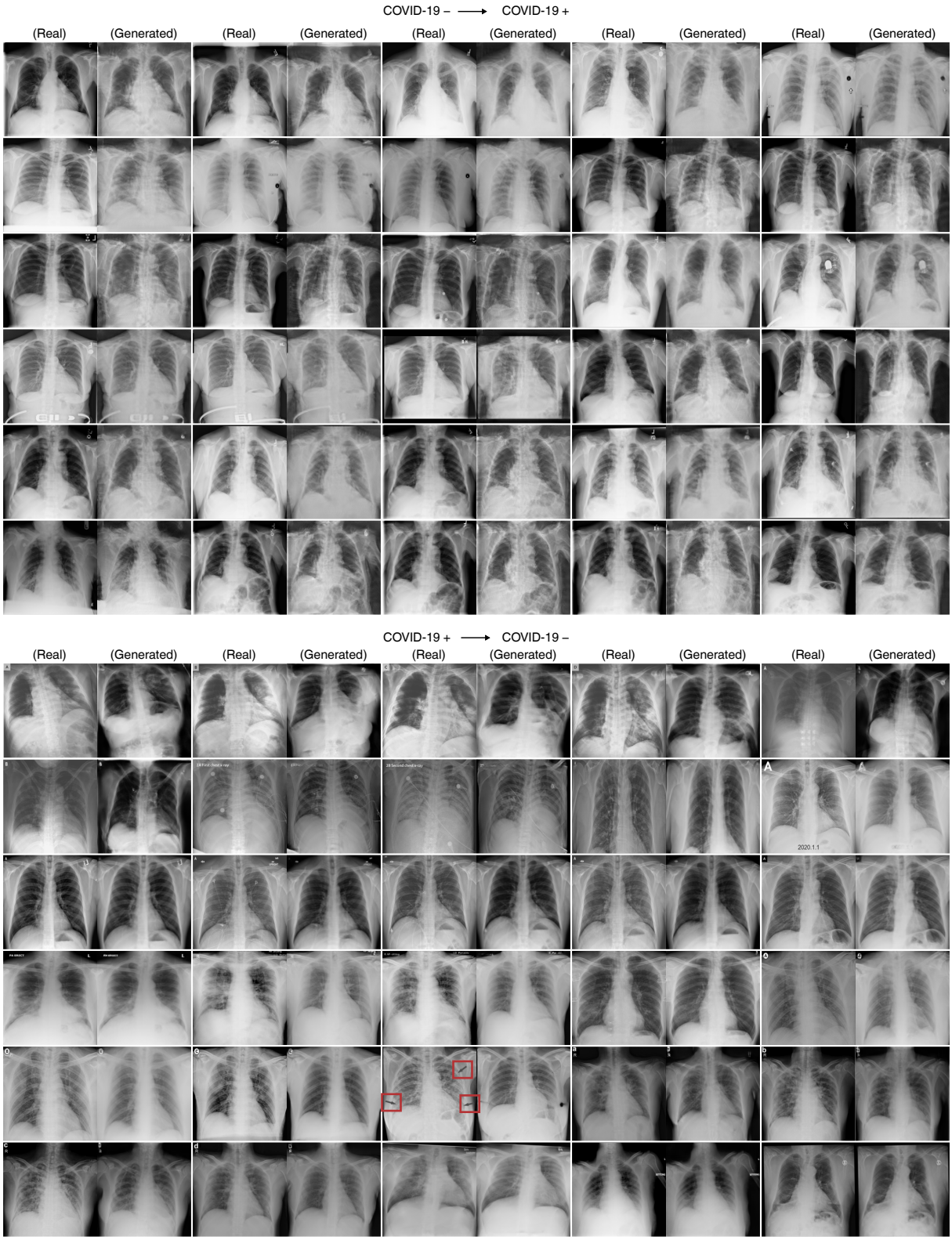


Supplementary Fig. 1 | (Caption next page.)

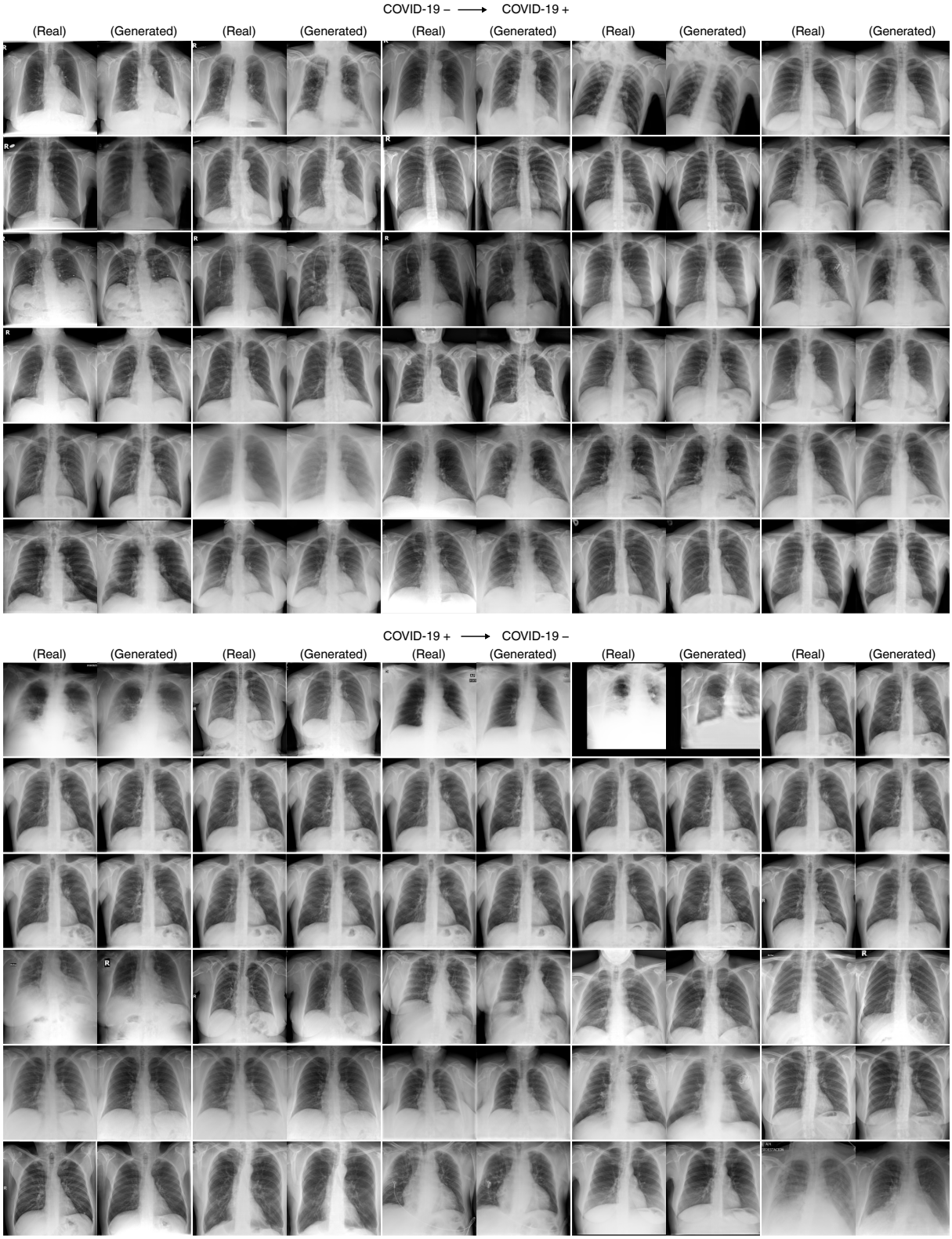
Supplementary Fig. 1 | (Previous page.) Evolution of metrics that monitor the artificial neural network training process. Training curves are shown for each of 5 random train/validation/test splits of the datasets. During the training procedure, the model is progressively optimized to decrease the training 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.



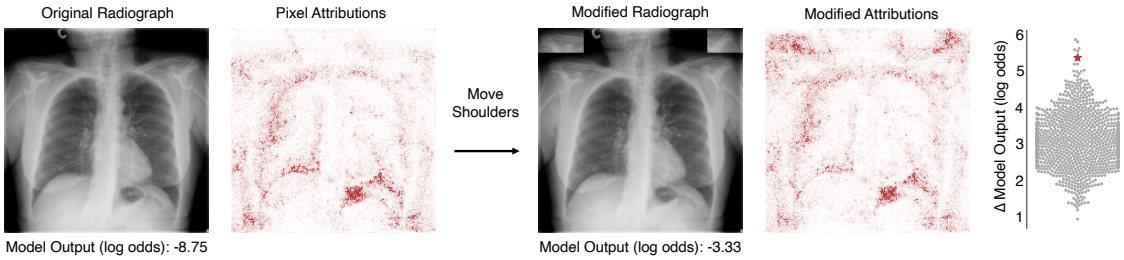
Supplementary Fig. 2 | Feature attributions saliency maps for 15 radiographs from each of the four data sources. Across all four data sources, saliency maps highlight text tokens and laterality markers (e.g., the first radiograph-saliency map pair in the first row of the PadChest examples, the second-to-last 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, all five radiograph-saliency map pairs in the third row of the ChestX-ray14 examples, and the first three radiograph-saliency map pairs in the first row of the GitHub examples).



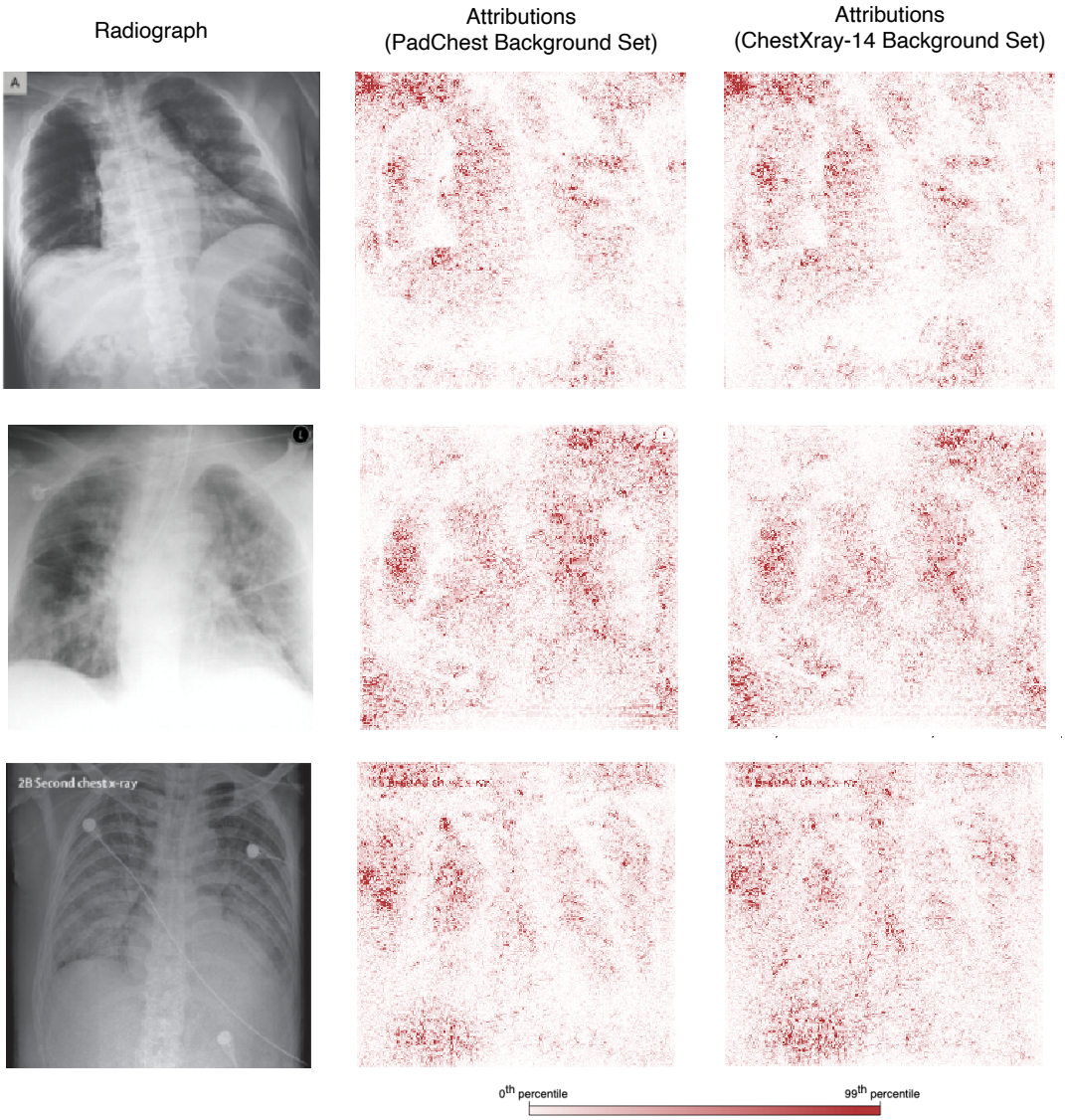
Supplementary Fig. 3 | Examples images generated by a CycleGAN that was trained to alter COVID-19 negative images from the ChestX-ray14 dataset to appear like COVID-19 positive images from the GitHub-COVID dataset and vice versa. Red boxes in lower panel mark annotations that were removed by the CycleGAN, as referenced in the main text. Images from the GitHub-COVID repository may contain annotations, as many were scraped from figures in academic publications.



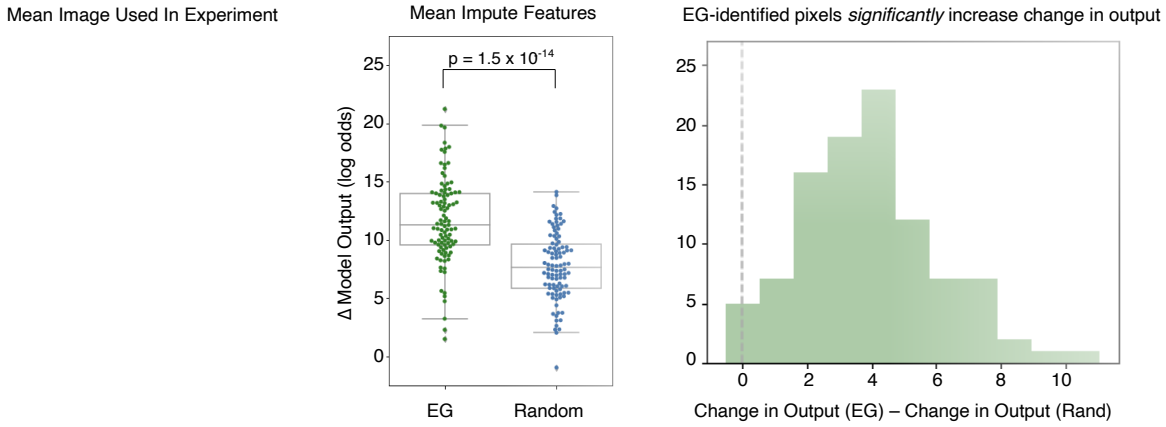
Supplementary Fig. 4 | Examples images generated by a CycleGAN that was trained to alter COVID-19 negative images from the PadCheset dataset to appear like COVID-19 positive images from the BIMCV-COVID-19+ dataset and vice versa.



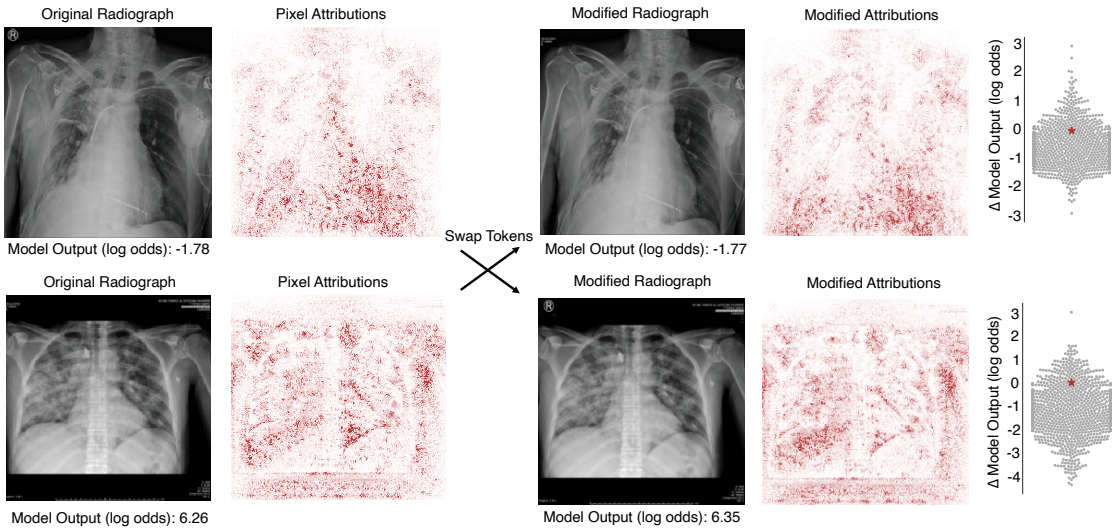
Supplementary Fig. 5 | Additional assessment of the importance of shoulder positioning to an AI model for radiographic COVID-19 detection. The procedure to generate Figure 2d was replicated with a new radiograph; *i.e.*, a patch of the radiograph containing the patient's clavicles was copied to the top 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).



Supplementary Fig. 6 | Comparison of expected gradients saliency maps generated from varied reference distributions, which provide the baseline radiographs from which the expected gradients algorithm integrates.



Supplementary Fig. 7 | 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 p-value is calculated by a two-sided Wilcoxon signed-rank test, n=100 ($W = 7.69$, $p = 1.48 \times 10^{-14}$). **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.



Supplementary Fig. 8 | Evaluation of the impact on model output of text tokens that were not highlighted as important via saliency maps. Following the procedure of Fig. 2c, text tokens of COVID-19 negative (upper row) and COVID-19 positive radiographs (lower row) were swapped, and the change in model output was compared to that produced by swapping random image patches of the same size. Neither replaced token produced a change in model output that was significantly greater than that expected from swapping random patches (top, empirical p-value = 0.25 based on Monte Carlo substitution of random image patches, n=1000; bottom, $p = 0.90$, n=1000) (see Methods Section 2.5).