nature portfolio

https://doi.org/10.1038/s42256-021-00338-7

Supplementary information

AI for radiographic COVID-19 detection selects shortcuts over signal

In the format provided by the authors and unedited

Supplementary Information

Supplementary Note

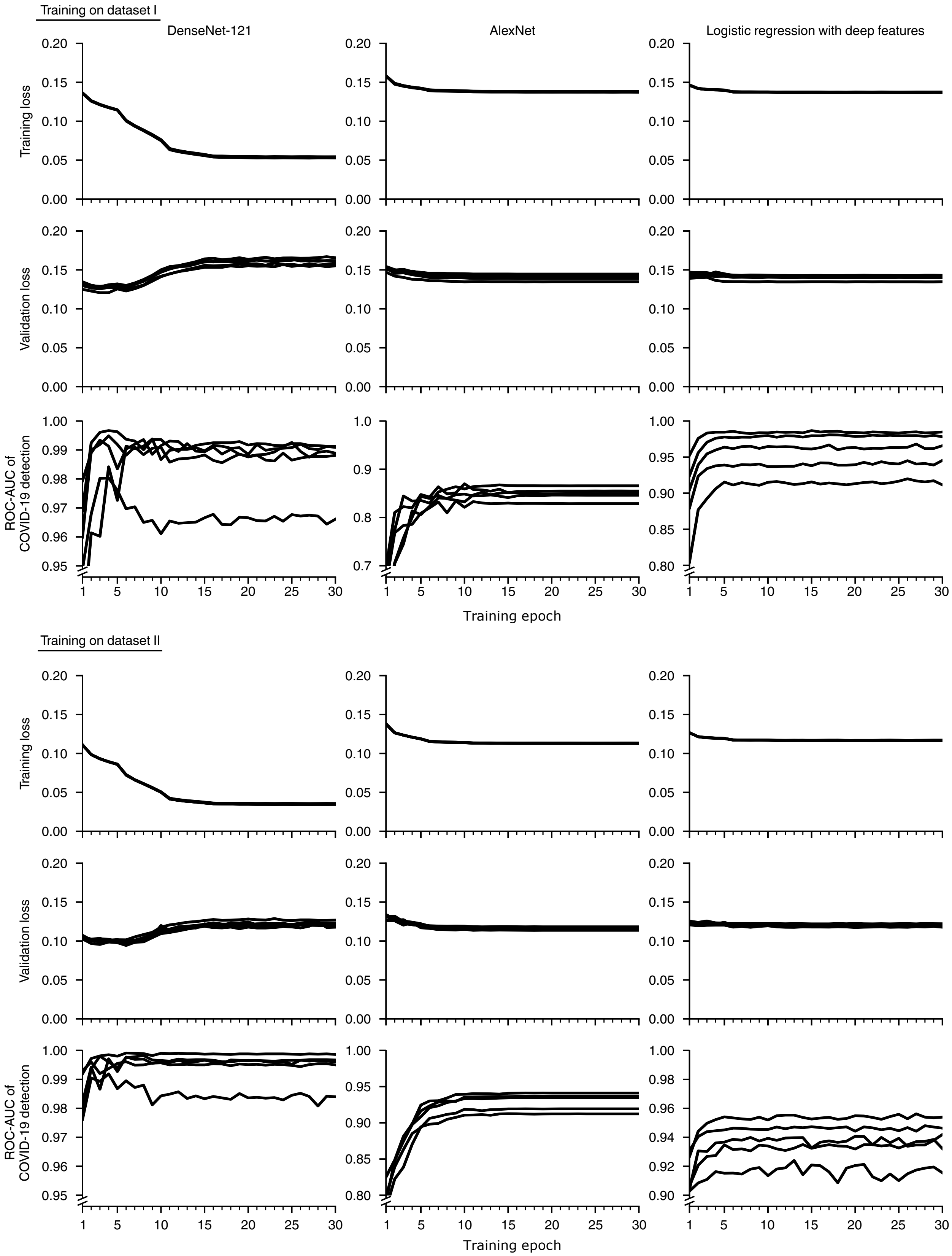
While saliency maps are widely used to interpret image-based artificial intelligence systems [1–3], the reliability of these approaches has been disputed by contemporary work, which observes that saliency maps explaining medical imaging classifiers fail to localize medically relevant pathology [4]. However, 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 [5].

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. 5).

Supplementary Figures

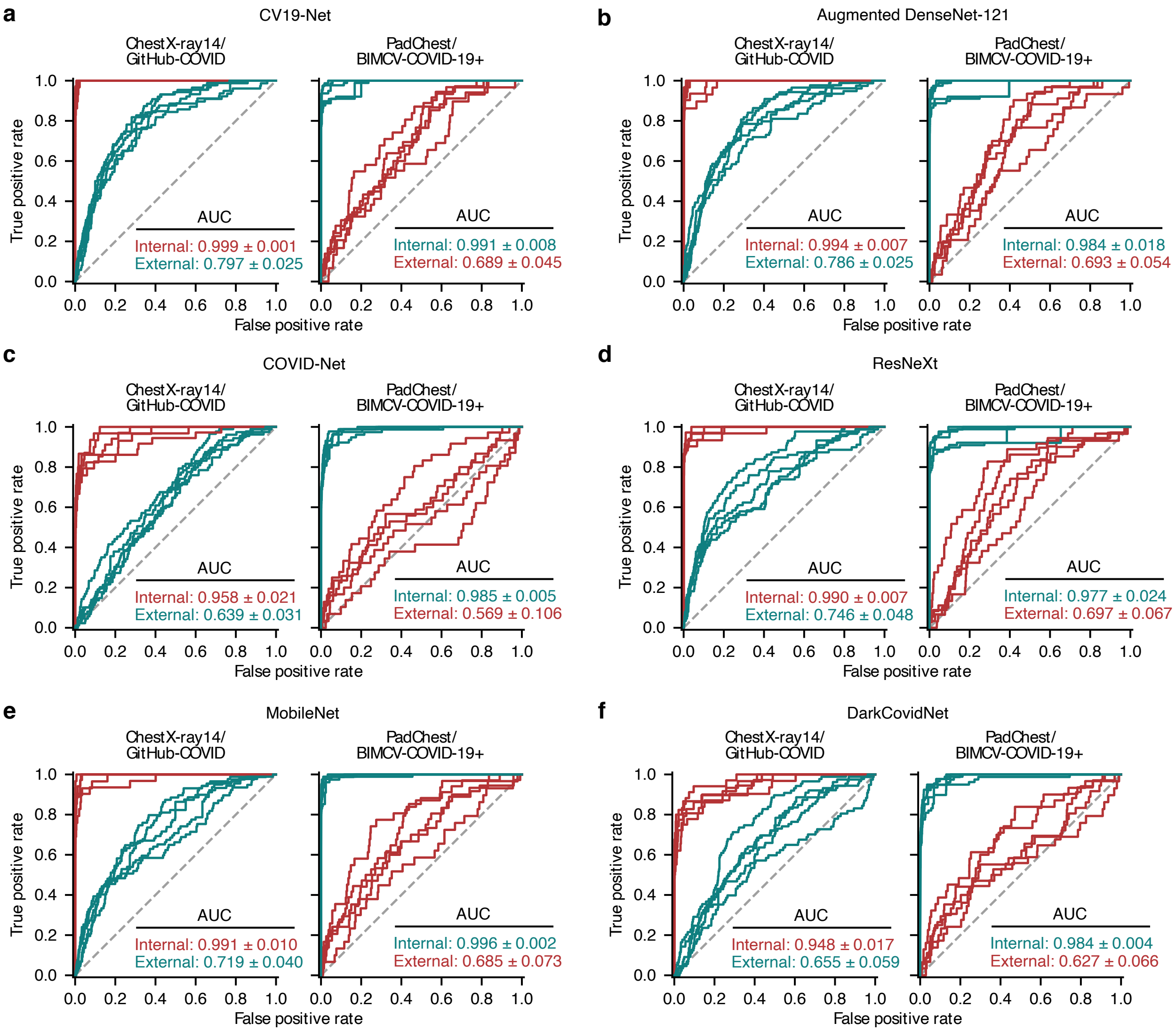
	Dataset I			**Dataset II**		
	Combined	CXR14	Cohen et al.	Combined	PadChest	BIMCV-COVID
CXR #s	112,528	112,120	408	97,866	96,270	1,596
Patients, #s	31,067	30,805	262	64,954	63,939	1,015
Age, mean (std)	46.9 (16.8)	46.9 (16.8)	57.0 (16.4)	65.4 (20.1)	65.5 (20.1)	61.2 (16.0)
Sex, N women (%)	48,926 (43.5)	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)

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.

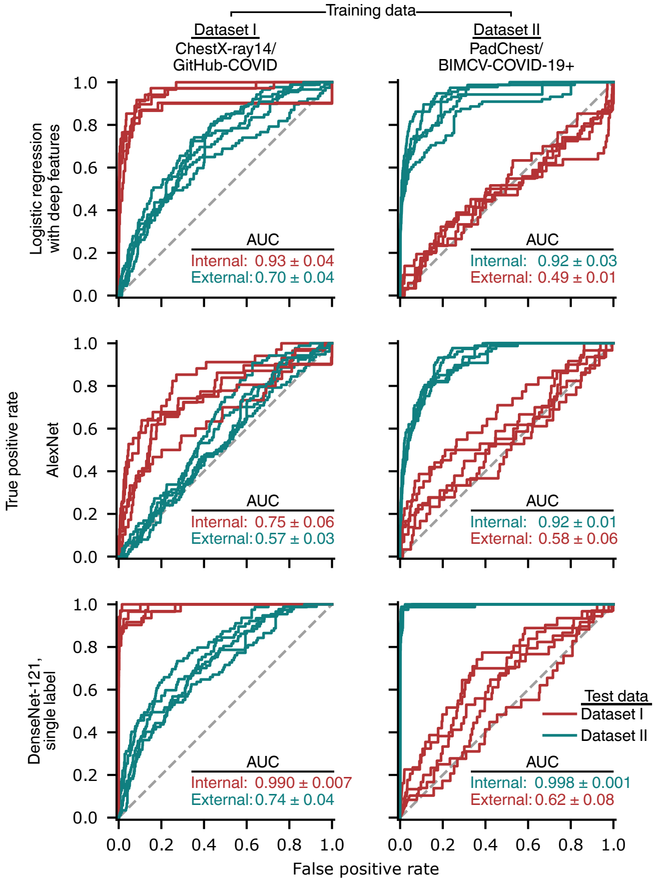


Supplementary Fig. 1 | (Caption next page.)

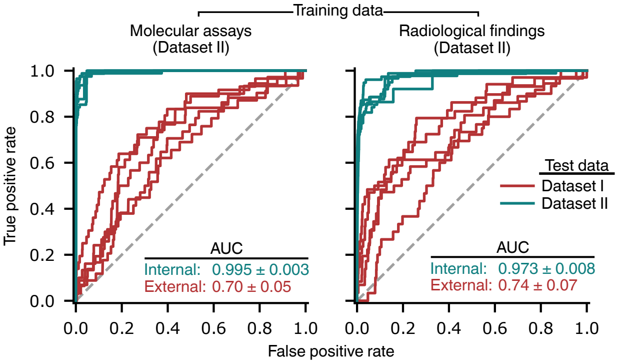
Supplementary Fig. 1 | (Previous page.) Evolution of metrics that monitor the artificial neural network training process. Training curves are shown for each of 5 random train/validation/test splits of the datasets. During the training procedure, the model is progressively optimized to decrease the training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



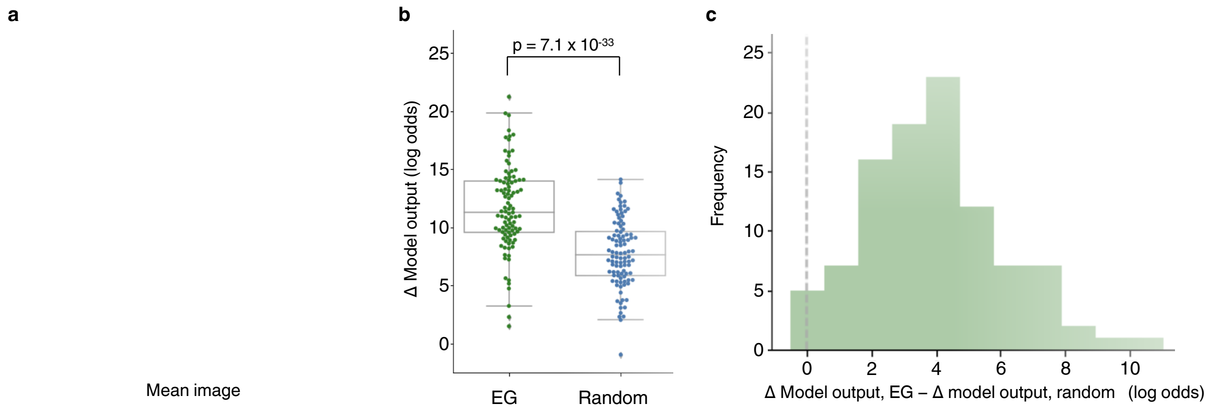
Supplementary Fig. 2 | Generalization performance of models that were specifically designed in previous studies for detection of COVID-19 in chest radiographs as well as additional "off-the-shelf" architectures. Generalization performance is examined by comparing the performance of each model on held 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[6], COVID-Net[7], and DarkCovidNet[8]. The additional "off-the-shelf" models include ResNeXT[9] and MobileNet[10]. 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).



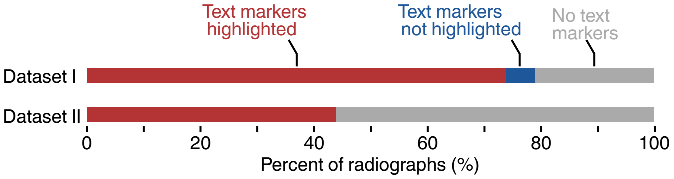
Supplementary Fig. 3 | Generalization performance of models with lower capacity or reduced label information, as measured by receiver-operating characteristic (ROC) curves. The first two rows correspond to models in which the capacity to overfit, which has been implicated in learning of spurious associations [11], 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).



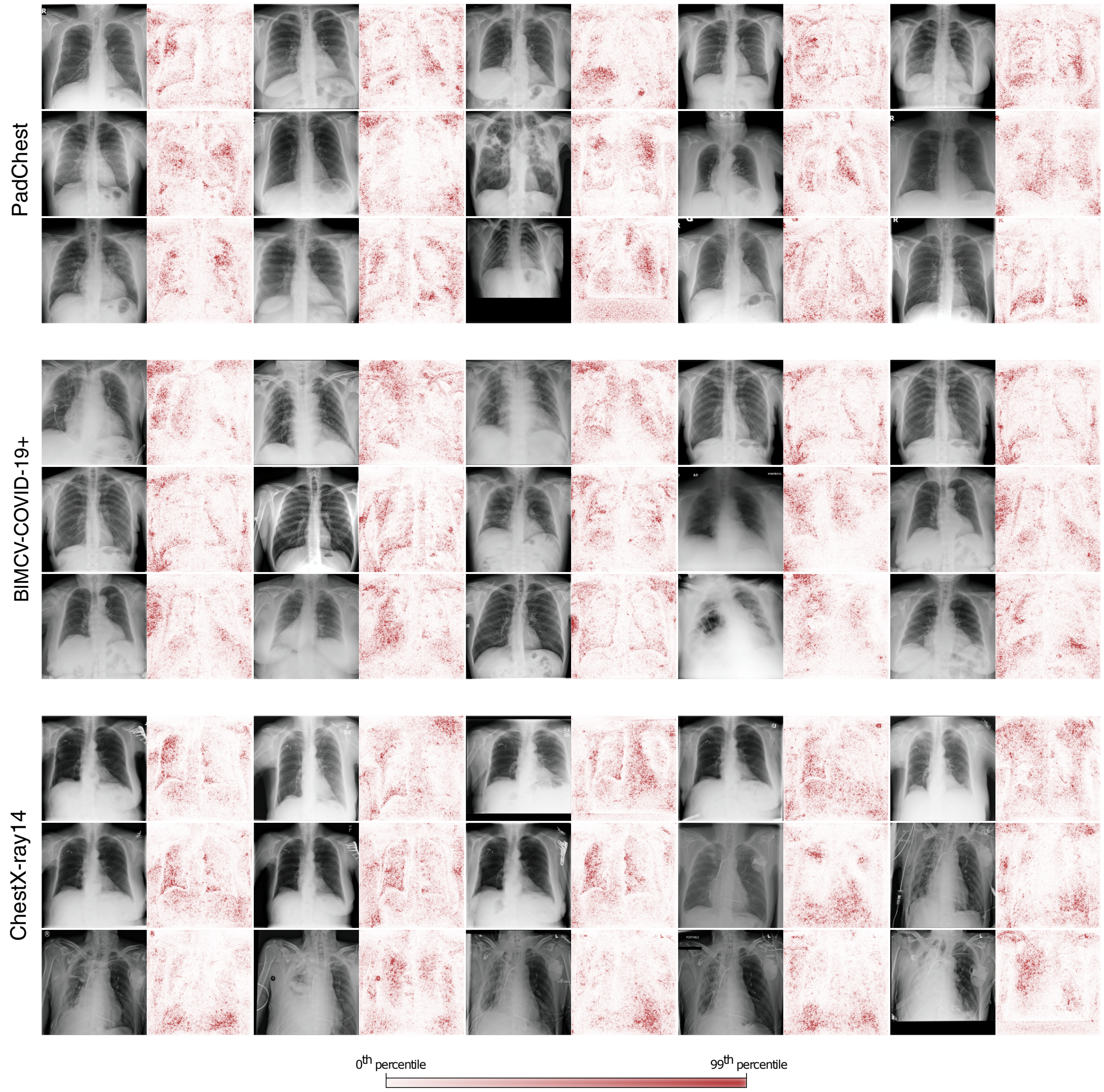
Supplementary Fig. 4 | Evaluation of the impact on generalization performance of *concept shift*, a change in the classification task between the training and testing datasets. In addition to the learning of spurious correlations that do not remain constant between datasets, generalization performance 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 [13], 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 [14] 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).



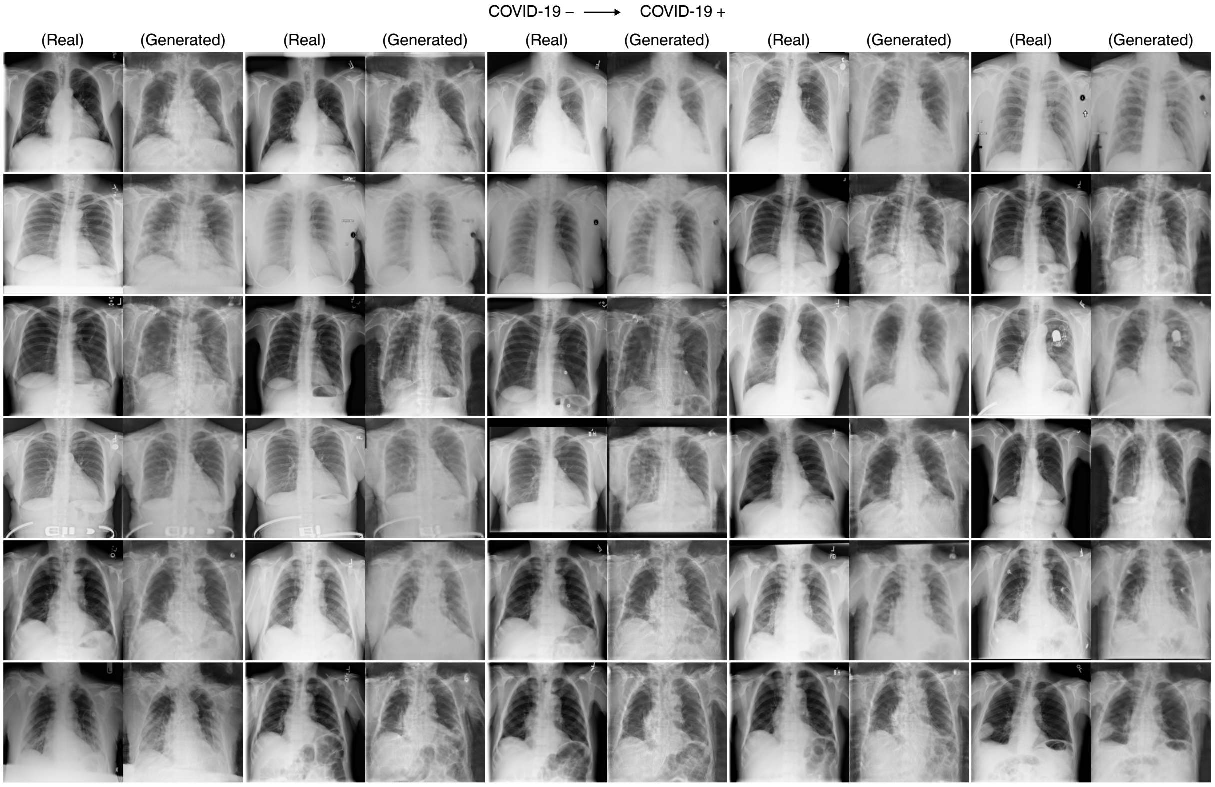
Supplementary Fig. 5 | Ablation tests to assess the importance of pixels that are highlighted by saliency maps. a, Average image of COVID-19+ radiographs from dataset I, from which pixels are drawn to "ablate", *i.e.*, hide, putatively important parts of individual radiographs in our experiment. **b**, 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.



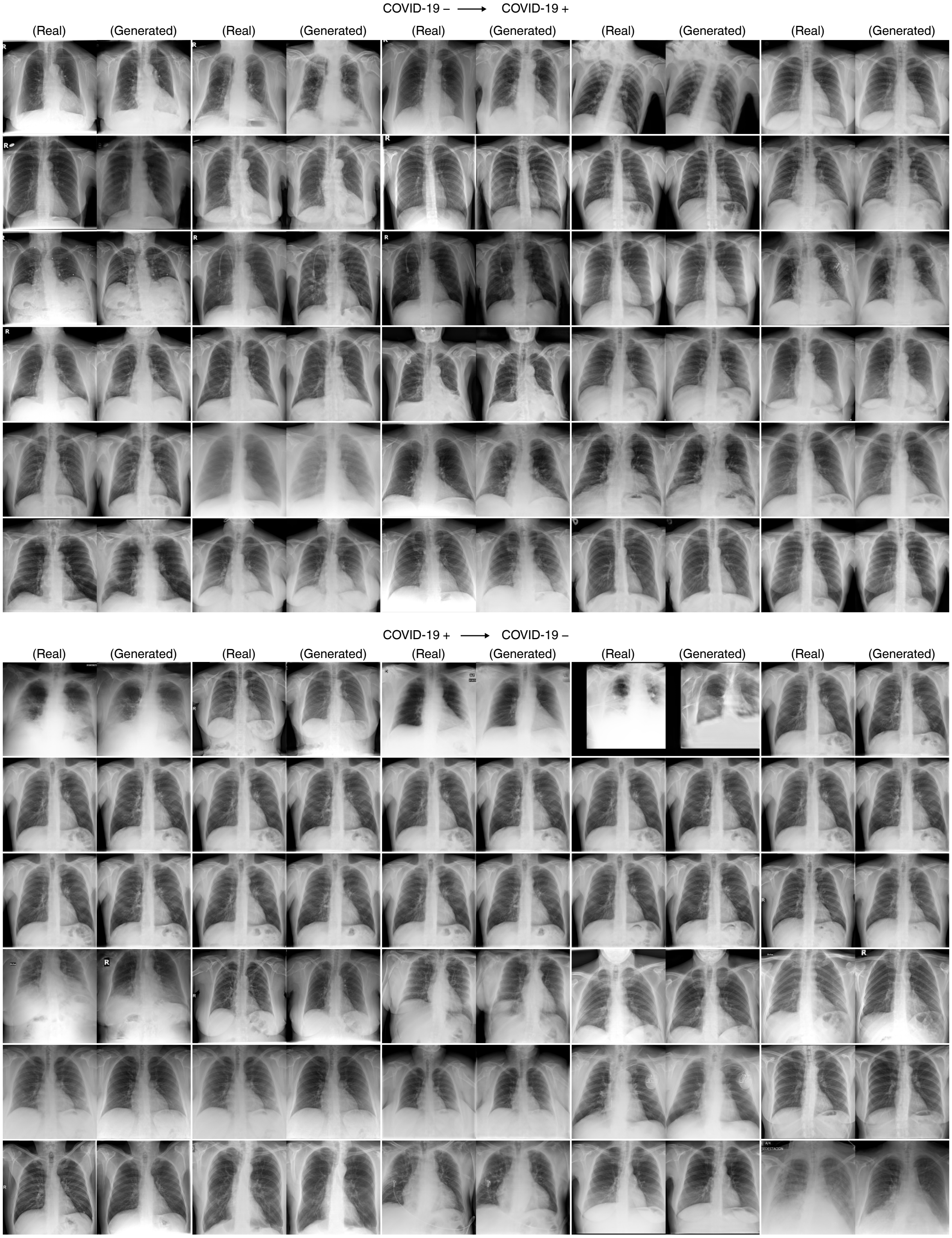
Supplementary Fig. 6 | Analysis of the frequency at which saliency maps highlight laterality markers as important features. To assess the frequency, a random sample of 100 radiographs and their corresponding saliency maps was chosen from each dataset, and each radiograph was manually categorized as (i) 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.



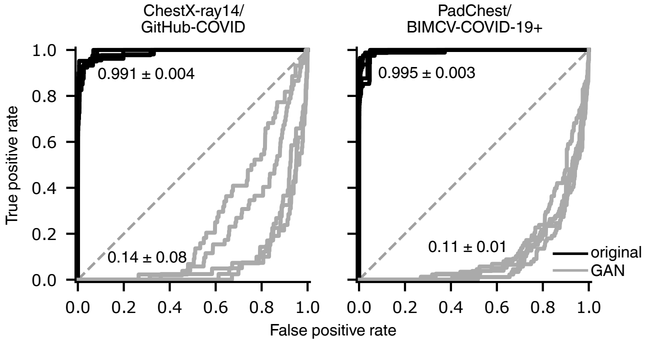
Supplementary Fig. 7 | Saliency maps for 15 radiographs from the PadChest, BIMCV-COVID-19+, and ChestX-ray14 repositories. Across the data sources, saliency maps highlight text tokens and laterality markers (e.g., the first radiograph-saliency map pair in the first row of the PadChest examples, the second-to-last 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`.



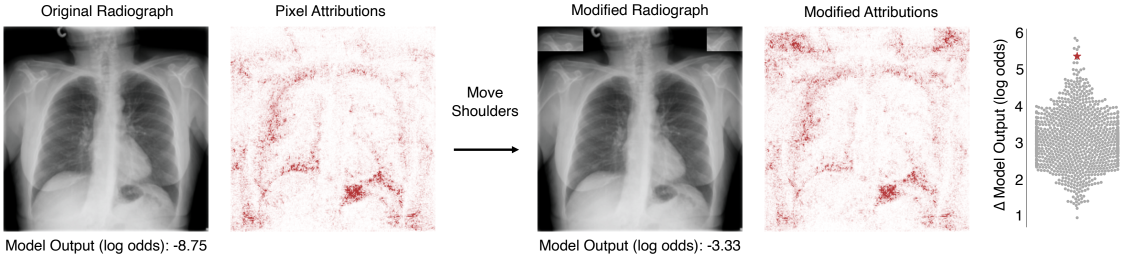
Supplementary Fig. 8 | Examples images generated by a CycleGAN that was trained to alter COVID-19 negative images from the ChestX-ray14 dataset to appear like COVID-19 positive images from the GitHub-COVID dataset and vice versa. See our GitHub repository at `https://github.com/suinleelab/cxr_covid` for a version of this figure that includes images from the GitHub-COVID repository.



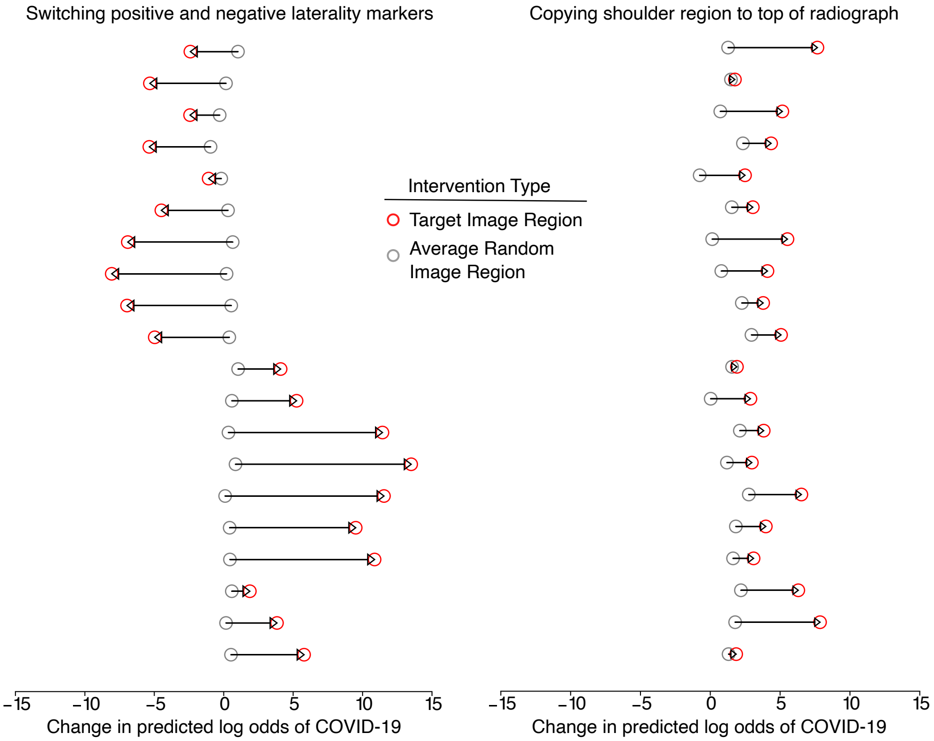
Supplementary Fig. 9 | Examples images generated by a CycleGAN that was trained to alter COVID-19 negative images from the PadCheset dataset to appear like COVID-19 positive images from the BIMCV-COVID-19+ dataset and vice versa.



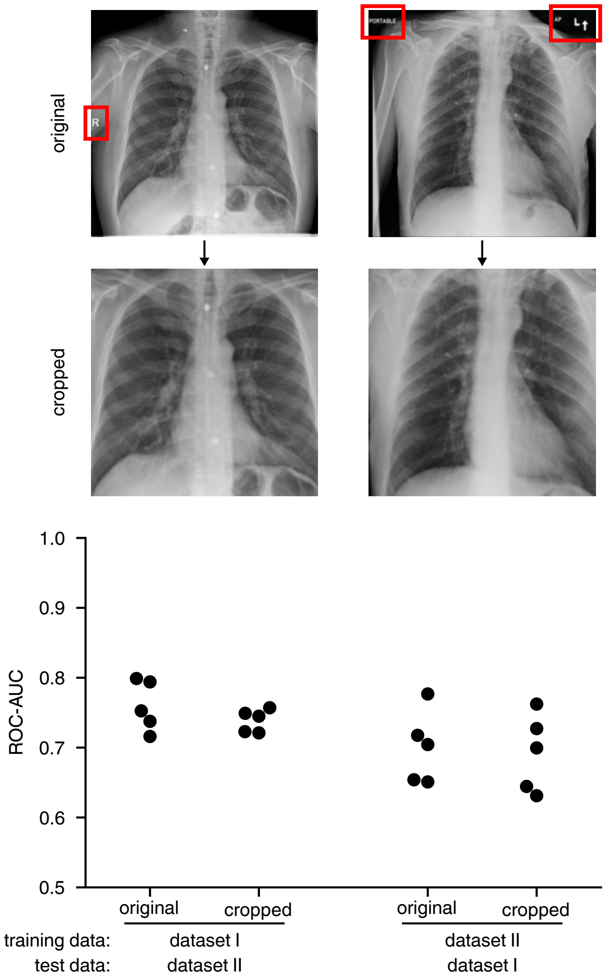
Supplementary Fig. 10 | Evaluation of the extent to which features relied upon by the COVID-19 detection models are altered by the CycleGAN, as measured by the drop in classification performance following transformation by the CycleGAN. A CycleGAN that more reliably alters images such that they appear 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).



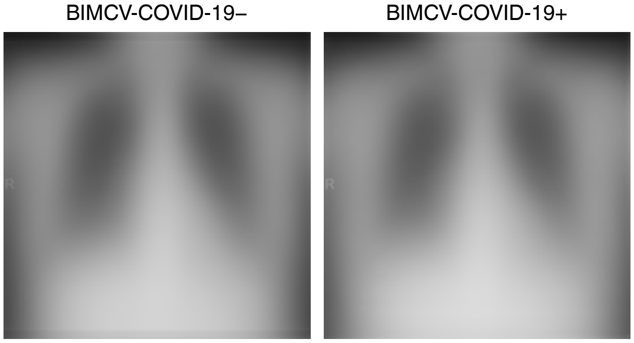
Supplementary Fig. 11 | Additional assessment of the importance of shoulder positioning to an AI model for radiographic COVID-19 detection. The procedure to generate Figure 2d was replicated with a new radiograph; *i.e.*, a patch of the radiograph containing the patient's clavicles was copied to the 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).



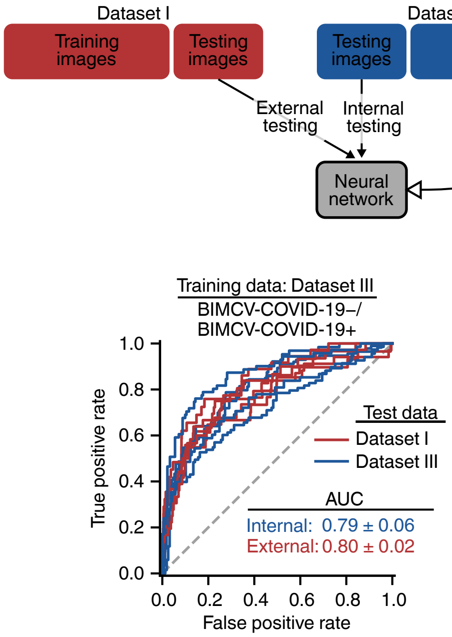
Supplementary Fig. 12 | Population-level analysis of importance of laterality markers and shoulder positioning. Each pair of dots corresponds to an radiograph sampled at random from the larger population, which enables inference of our findings to the population level, despite the infeasibility of completing 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).



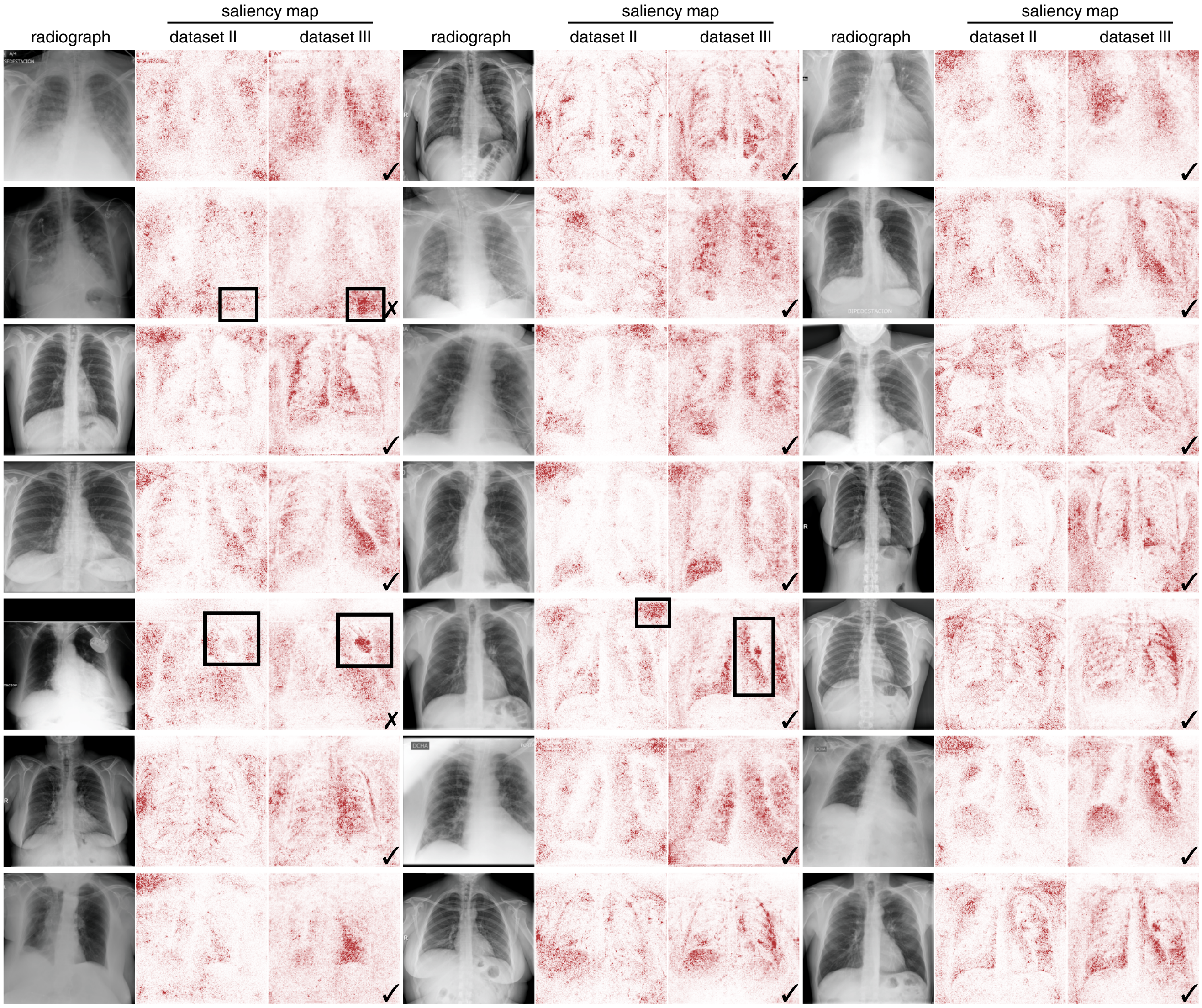
Supplementary Fig. 13 | Evaluation of the extent to which image cropping mitigates shortcut learning. For each dataset, models were trained before and after cropping to the center 75% of the radiograph, which removes from the edge of radiographs the textual markers (red boxes) that may contribute to 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).



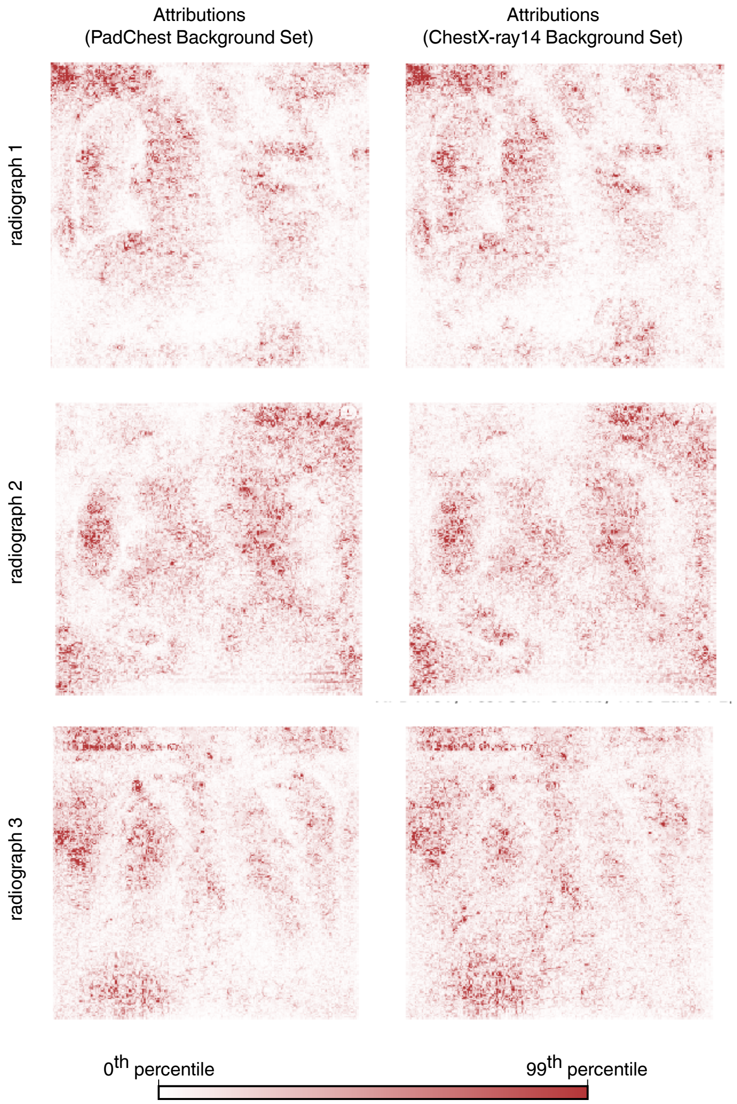
Supplementary Fig. 14 | Average images of the BIMCV-COVID-19− and BIMCV-COVID-19+ repositories. Note consistency in the laterality markers, shoulder positioning, and radiopacity of image borders.



Supplementary Fig. 15 | 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.



Supplementary Fig. 16 | 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[15], 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.



Supplementary Fig. 17 | Comparison of expected gradients saliency maps generated from varied reference distributions, which provide the baseline radiographs from which the expected gradients algorithm integrates.

References

1. Rajpurkar, P. *et al.* CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. *arXiv:1711.05225* (2017).
2. Mitani, A. *et al.* Detection of anaemia from retinal fundus images via deep learning. *Nature Biomedical Engineering* **4,** 18–27 (2020).
3. Bien, N. *et al.* Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet. *PLoS medicine* **15,** e1002699 (2018).
4. Arun, N. *et al.* Assessing the (un) trustworthiness of saliency maps for localizing abnormalities in medical imaging. *medRxiv* (2020).
5. 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.
6. 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`.
7. 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).
8. 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).
9. Xie, S., Girshick, R., Dollár, P., Tu, Z. & He, K. Aggregated Residual Transformations for Deep Neural Networks. *arXiv:1611.05431* (2016).
10. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. & Chen, L.-C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. *arXiv:1801.04381* (2019).
11. 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).
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. Cohen, J. P., Morrison, P. & Dao, L. COVID-19 image data collection. *arXiv 2003.11597.* `https://github.com/ieee8023/covid-chestxray-dataset` (2020).
14. 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).
15. Gu, J., Han, B. & Wang, J. COVID-19: gastrointestinal manifestations and potential fecal-oral transmission. *Gastroenterology* **158,** 1518–1519 (2020).