Deep learning classification of COVID-19 in chest radiographs: performance and influence of supplemental training

Rafael B. Fricks[a,b,*] **Francesco Ria**[b] **Hamid Chalian**[b] **Pegah Khoshpouri**[b] **Ehsan Abadi**[b] **Lorenzo Bianchi**[c] **William P. Segars**[b] **and Ehsan Samei**[b]

[a]National Artificial Intelligence Institute, Department of Veterans Affairs, Washington, D.C., United States

[b]Duke University, Carl E. Ravin Advanced Imaging Laboratory, Department of Radiology, Durham, North Carolina, United States

[c]ASST della Valle Olona, Medical Physics Department, Busto Arsizio, Italy

Abstract

Purpose: Accurate classification of COVID-19 in chest radiographs is invaluable to hard-hit pandemic hot spots. Transfer learning techniques for images using well-known convolutional neural networks show promise in addressing this problem. These methods can significantly benefit from supplemental training on similar conditions, considering that there currently exists no widely available chest x-ray dataset on COVID-19. We evaluate whether targeted pretraining for similar tasks in radiography labeling improves classification performance in a sample radiograph dataset containing COVID-19 cases.

Approach: We train a DenseNet121 to classify chest radiographs through six training schemes. Each training scheme is designed to incorporate cases from established datasets for general findings in chest radiography (CXR) and pneumonia, with a control scheme with no pretraining. The resulting six permutations are then trained and evaluated on a dataset of 1060 radiographs collected from 475 patients after March 2020, containing 801 images of laboratory-confirmed COVID-19 cases.

Results: Sequential training phases yielded substantial improvement in classification accuracy compared to a baseline of standard transfer learning with ImageNet parameters. The test set area under the receiver operating characteristic curve for COVID-19 classification improved from 0.757 in the control to 0.857 for the optimal training scheme in the available images.

Conclusions: We achieve COVID-19 classification accuracies comparable to previous benchmarks of pneumonia classification. Deliberate sequential training, rather than pooling datasets, is critical in training effective COVID-19 classifiers within the limitations of early datasets. These findings bring clinical-grade classification through CXR within reach for more regions impacted by COVID-19.

 [DOI: 10.1117/1.JMI.8.6.064501]

Keywords: COVID-19; radiography; DenseNet; transfer learning; classification; computer vision.

Paper 20173RRRR received Jun. 25, 2020; accepted for publication Nov. 8, 2021; published online Dec. 1, 2021.

1 Introduction

Since its initial identification in December 2019, the coronavirus SARS-CoV2 has disseminated internationally, impacting virtually every aspect of human activity. This virus causes Coronavirus Disease 2019, abbreviated as COVID-19: a highly contagious, high-mortality, acute respiratory illness. On March 11, 2020, the World Health Organization declared COVID-19

*Address all correspondence to Rafael B. Fricks, Rafael.Fricks@Duke.edu

2329-4302/2022/$28.00 © 2022 SPIE

a pandemic following a 13-fold increase in cases outside China in two weeks.[1] As of June 11, 2020, there have been 7,442,050 cases of COVID-19 confirmed globally, with 418,563 reported deaths due to the coronavirus disease.[2] The extent and severity of the COVID-19 pandemic is unprecedented in modern times, pushing the need for fast and effective diagnosis to curtail the spread of this disease.

The gold standard test for diagnosing COVID-19 is reverse transcriptase-polymerase chain reaction (RT-PCR). However, variable sensitivity in early RT-PCR tests and shortages in capacity prompted the consideration of diagnostic imaging for detection or management of COVID-19 progression.[3] As a respiratory disease, chest imaging via computed tomography (CT) or chest radiography (CXR) is a natural approach to managing potential cases. In fact, imaging played an essential role in the response at early epicenters of the pandemic.[3–5] More recently, the Fleischner Society has issued a consensus statement outlining situations in which imaging may be informative in managing patient treatment.[6] Although the consensus provides arguments for the relative merits of both CT and CXR, it leaves the choice of modality to the judgment of clinical teams based on local factors such as imaging capacity and the availability of expertise.[6]

In early studies, CT proved effective as a high-sensitivity method for detection.[3,7] Compared to CXR, CT produces higher-grade information; however, the modality presents distinct drawbacks and practical challenges in this context. Typical chest CT doses are higher than CXR, carrying a higher radiation risk to the patient.[8,9] CT scanners are also more expensive and less readily available at hospitals. Finally, turnover between patients is more difficult and poses risks to staff, who would have to clean the scanner between patients to avoid spreading the coronavirus. In comparison, portable chest x-ray offers distinct advantages pertaining to all these concerns, easing the burden on patients and facilitating the provision of care.

The use of chest x-ray raises a question of whether COVID-19 pathological findings may be distinguishable in CXR. Deep learning classification of COVID-19 in CXR may prove promising but has been less explored than CT, with a few recent works exploring transfer learning techniques to adapt an existing model to COVID-19 classification and differentiation from other pneumonia sources.[10] A universal limitation to such studies tends to be access to chest radiographs from patients with RT-PCR–confirmed cases of COVID-19. In lieu of this, a promising alternative has been to incorporate larger, publicly available chest x-ray databases and focus on proximal tasks such as general pneumonia classification. Prominent databases such as the National Institutes of Health (NIH) dataset ChestX-Ray14[11] or the pneumonia Kaggle Challenge hosted by the Radiological Society of North America (RSNA)[12] contain many samples of proximal chest radiograph findings such as pneumonia.

Considering that pneumonia and COVID-19 share common image features,[13] we hypothesized that supplemental training with pneumonia datasets may improve the performance of COVID-19 classification using convolutional neural networks. This study thus aimed to evaluate the effectiveness of a convolutional neural network informed by non-COVID pneumonia to classify COVID-19 in chest radiographs. Starting with a pretrained network, we incorporated existing databases in successive training phases to fine-tune a convolutional neural network for the coronavirus task. After pretraining, the algorithm was fine-tuned and evaluated using a database of 1060 chest radiographs from 475 patients, taken postpandemic start and containing 801 images of COVID-19 positive cases. We employed stratified sampling at the patient level to subdivide this database into a 60%/20%/20% data set split that preserves the original prevalence in the training, validation, and test sets. We report performance via various training paths on this dataset's test set to investigate the differential performance benefits of incorporating related images.

2 Methods

In this study, we trained a model by applying sequential pretraining in three successive fine-tuning phases, diagrammed in Fig. 1. In these pretraining phases, the model is presented with datasets configured to perform gradual domain adaptations. Our first phase adapted a convolutional neural network trained on ImageNet to CXR tasks, following previous approaches in end-to-end training for multilabel classification in CXR on established databases.[14,15] The

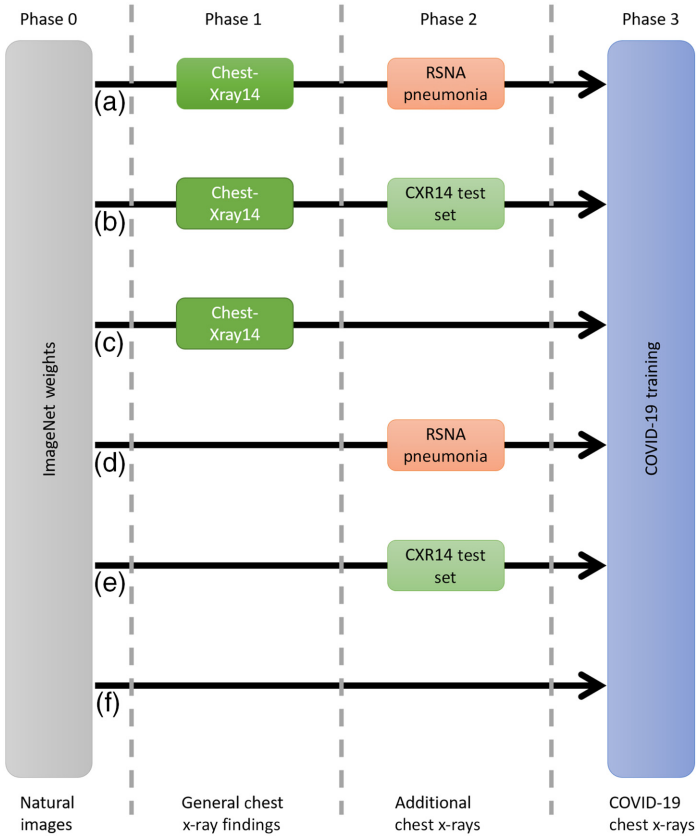

Fig. 1 Six permutations of the selected convolutional neural network architecture were refined using pretraining phases. In each variation, the network is trained sequentially on datasets containing either a mix of findings with low-pneumonia prevalence or an enhanced dataset with higher pneumonia prevalence. Path (f) serves as the control and is directly trained on COVID-19 cases with no prior training on chest radiographs. All networks are trained and evaluated on an identical COVID-19 task in phase 3.

second phase added additional training in pneumonia detection, using either a dataset with similar pneumonia prevalence to phase 1, or an enriched dataset with higher pneumonia prevalence. The third phase fine-tuned the resulting networks for classifying confirmed cases of COVID-19, using chest radiographs collected for this study.

Figure 1 diagrams the training phases in each variation; we produced six variations of a densely connected network[16] which was initially pretrained on ImageNet. Path (a) was first trained on a multilabel chest x-ray task in phase 1, before being refined with a dataset with higher pneumonia prevalence in phase 2, and finally retrained for the COVID-19 detection task. Path (b) was identical to path (a), except for employing a phase 2 dataset with lower pneumonia prevalence. Path (c) did not have a phase 2 refinement. Similarly, paths (d) and (e) omitted the first phase multilabel training and applied the two phase 2 treatments, respectively. Finally, path (f) served as a control by directly adapting from ImageNet weights to the COVID-19 classification task. We detail training conditions in these phases in the following sections.

2.1 *Phase 1: Transfer Learning from ImageNet to General Chest Radiography Findings*

The goal of phase 1 was to replicate the recent state-of-the-art in multilabel classification tasks in CXR as a foundation for a COVID-19 classifier in phase 3. We began with DenseNet121 with initial weights optimized for ImageNet[17] natural image classification and adapted it for the multilabel task using standard end-to-end transfer learning approaches.[18,19] The top classification layer was replaced with a 14-node densely connected layer with sigmoid activation and He

initialization,[20] resulting in corresponding outputs for the original labels in ChestX-ray14.[11] This database contains 112,210 images with common radiological findings which were collected prior to the emergence of COVID-19. The images were divided at the patient level into approximately a 70%/10%/20% training/validation/test split, resulting in 75,828 images for training, 10,696 for validation, and 25,596 for testing, respectively. The testing set was the list specified by the original database curators,[21] proven a useful reference set for performance benchmarking.[15] In phase 1, we estimated the area under the receiver operating characteristic curve (AUROC) for the test set, for each label in the multilabel task.

In preprocessing, we down-sampled the ChestX-ray14 images by a factor of two to 512×512 resolution. The images were normalized using values for mean and standard deviation estimated from a randomly selected set of 16,384 training images. Training images were randomly augmented at runtime, first by random horizontal flipping, then random rotations of up to 8 deg, and random cropping of up to 10% of the image size.

The network was trained with mirrored copies replicated across GPUs receiving batches of 16 images per GPU. Training employed a stochastic gradient descent with momentum optimizer, with an initial learning rate of 0.01, momentum of 0.9, and weight decay of 0.0001. We used an unweighted binary cross-entropy loss function. The learning rate was further decreased by a factor of 10 each time the validation loss did not improve in three epochs. Early stopping halted training if the validation loss did not improve after 10 epochs and reverted to the weights that achieved the highest validation loss. The model and all subsequent training were implemented in Tensorflow 2.2.0.[22]

2.2 *Phase 2: Adaptation to Pneumonia Detection*

The goal of phase 2 was to ascertain the impact of disease prevalence in an additional training stage (pneumonia as the basis given its similarity to COVID-19 features) on the classification accuracy of the algorithm. In this phase, we produced a pneumonia-centric dataset using 25,000 images from the RSNA pneumonia Kaggle Challenge training dataset.[12] This dataset contains a higher proportion of pneumonia cases, where 6012 images contain pneumonia findings. For comparison, we produced a broader finding chest x-ray set by repurposing 25,000 images from the phase 1 test set, ensuring that all 555 pneumonia cases are included. These two datasets were split at the patient level into 80%/20% sets for training and validation, using stratified sampling to generate sets with ~24% and 2% pneumonia prevalence, respectively.

For training in this phase, we retained as many of the previous training conditions as possible. The initial learning rate was reduced to 0.001, but otherwise all hyperparameters were identical to phase 1. The DenseNet121 was initialized with phase 1 weights in paths (a) and (b) or ImageNet weights in paths (d) and (e). Input images were normalized using mean and standard deviation values estimated from 2048 training images of the respective refinement set. The output layer was replaced with a densely connected layer with He initialization and a single sigmoid output for pneumonia classification. We minimized the unweighted binary cross-entropy loss. We estimated the AUROC for each of the four training permutations that received phase 2 fine tuning.

2.3 *Phase 3: Fine-Tuning from Chest Radiographs for COVID-19 Classification*

The goal of phase 3 was to fine-tune the developed classification to COVID-19 cases and evaluate COVID-19 classification performance under all pretraining conditions. With IRB approval, we collected 1060 chest radiographs consecutively from 475 patients with RT-PCR result available. All images were taken postpandemic start with 801 images of COVID-19 RT-PCR positive cases. The images were collected from multiple sites from Iran, Italy, and the US. The cases were homogenized to reduce intraset variability and approximate the ChestX-ray14 and RSNA datasets in content. This included manual cropping to remove burn-in labels and reframe wider radiograph field-of-views to center on lungs (Fig. 2). The cropped images were resized to

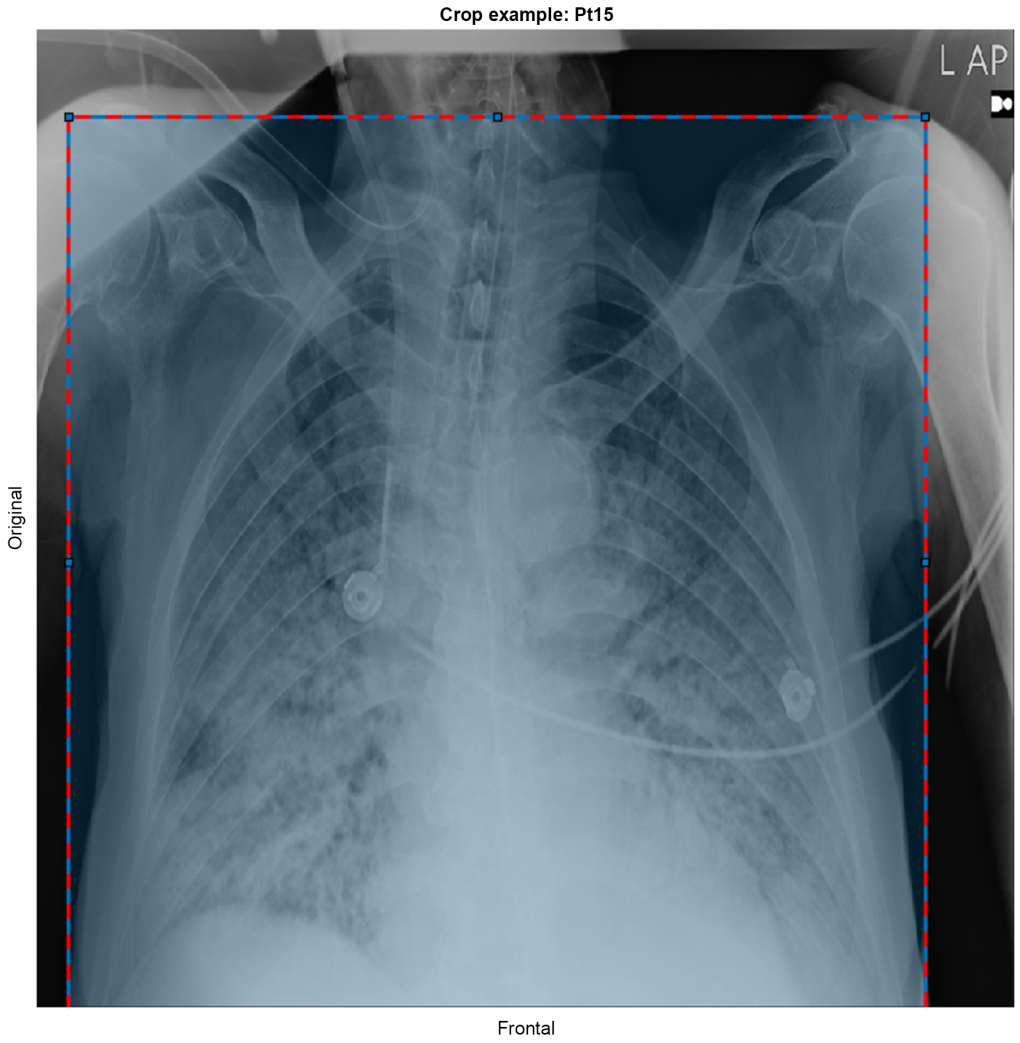

Fig. 2 An example of manual image cleaning. Wider radiograph field of views are reduced to remove burn-in annotations, and focus on chest volumes, similar to existing databases.

512×512 resolution and contrast scaled based on fields in the DICOM header to fit an 8-bit dynamic range consistent with other datasets.

The images were divided into 60%/20%/20% splits with 626 images for training, 216 images for validation, and a test set of 218 images, respectively. Images were sampled at the patient level with stratification, to ensure no patient is present across multiple sets, and that the COVID-19 prevalence is maintained in each subset at ~75% as in the total set.

As in phase 2, we maintained training parameters constant as much as possible while fine tuning the six permutations of the model. The output layer was replaced with a densely connected layer with a single output for COVID-19. Training, loss, and metrics were evaluated for each model on 20% of the data reserved as a test set and identical to previous phases except for the convolutional network weights, which were derived from previous phases or ImageNet in path (f). The initial learning rate for this phase was again reduced, starting at 0.0005, with the same learning rate reduction and early stopping rules applied. The results were again summarized in terms of the area under the ROC curve for consistency. We applied DeLong's test to evaluate the statistical significance of each possible pairwise AUROC comparison between the six models and further used DeLong's method to calculate a 95% confidence interval in the AUROC.[23,24] Finally, we adjusted p-values to account for multiple comparisons using the Bonferroni method.[25]

2.4 *Interpretability in Results*

To verify that the resulting neural network learned viable features for classifying coronavirus disease COVID-19, we generated saliency maps, then up-sampled and overlaid them on the original radiographs. We used the gradient-weighted class activation mapping (Grad-CAM) method[26] to produce saliency maps from the top-performing classifier on COVID-19 on its validation set. These resulting composite images provided insight into which image regions contributed to the network's classifications.

3 Results

3.1 *Phase 1: Transfer Learning from ImageNet to General Chest Radiography Findings*

For the pretraining task, we achieved performance on the multilabel chest x-ray task comparable to previous state-of-the-art models, measured as AUROC evaluated on the reference test set specified for Chest-ray14[11,21] (Table 1). Our model precisely matched the average AUROC of the top performing variant of the Baltruschat et al.'s study,[15] which included trained-from-scratch residual connection networks for this task. Similarly, our pneumonia performance and average AUROC were only exceeded by the Güendel et al.'s model.[28] Both of these networks took additional steps to supplement the performance of their convolutional network; we contrast these design choices in the discussion.

3.2 *Phase 2: Adaptation to Pneumonia Detection*

In phase 2, path (a) was designed to improve our phase 1 model's baseline pneumonia performance by additional training on a higher prevalence pneumonia image set. By contrast, the

Table 1 Performance comparison of our phase 1 results with existing state-of-the-art models, evaluated on a common test set.

Pathology AUROC	Wang et al.[11]	Yao et al.[27]	Güendel et al.[28]	Baltruschat et al.[15]	Our method
Cardiomegaly	0.810	0.856	0.883	0.875	0.868
Emphysema	0.833	0.842	0.895	0.895	0.924
Edema	0.805	0.806	0.835	0.846	0.834
Hernia	0.872	0.775	0.896	0.937	0.867
Pneumothorax	0.799	0.805	0.846	0.840	0.865
Effusion	0.759	0.806	0.828	0.822	0.828
Mass	0.693	0.777	0.821	0.820	0.819
Fibrosis	0.786	0.743	0.818	0.816	0.813
Atelectasis	0.700	0.733	0.767	0.763	0.761
Consolidation	0.703	0.711	0.745	0.749	0.739
Pleural thickening	0.684	0.724	0.761	0.763	0.766
Nodule	0.669	0.724	0.758	0.747	0.777
Pneumonia	0.658	0.684	0.731	0.714	0.722
Infiltration	0.661	0.673	0.709	0.694	0.696
Macroaverage	**0.745**	**0.761**	**0.807**	**0.806**	**0.806**

Table 2 Performance comparison of phase 2 refinements, arranged by pneumonia AUROC. The horizontal bar between sets denotes a change in test set.

Variant	Description	Total radiographs in training sets	Pneumonia radiographs in training sets	Pneumonia prevalence in phase 2 (%)	Pneumonia AUROC
Path (a)	Phase 1 weights + RSNA pneum.	95,828	5571	24.05	**0.885**
Path (d)	ImageNet weights + RSNA pneum.	20,000	4810	24.05	**0.877**
Path (b)	Phase 1 weights + NIH test set	95,919	1199	2.18	**0.716**
Path (e)	ImageNet weights + NIH test set	20,091	438	2.18	**0.635**

model in path (b) received the same number of additional images with no emphasis on a particular label. Paths (c) and (d) tested the necessity of phase 1 training. Table 2 shows results for phase 2 training, ordered by evaluated AUROC on the respective refinement set.

Unlike in phase 1, these results have no direct comparison; the data were split for these experiments. Although the RSNA set and the ChestX-ray14 are quite similar, given that paths (a) and (d) are evaluated on a different test set than paths (b) and (e), it is not advisable to make broad conclusions across these two groups of paths. However, we note that the pneumonia AUROC improved based on the number of pneumonia cases seen, rather than the total number of cases seen. There is a slight improvement when the pneumonia training phase was bolstered by the phase 1 training as a foundation of general x-ray cases [path (a) versus (d)]. When relying on data with few cases of pneumonia, there was a substantial improvement when more (previously unseen) cases and training were added, as expected [path (b) versus (e)]. Path (b) performance on this set did not vary significantly from the estimated generalization performance predicted in the phase 1 result. Although this may be due to nuances of the new evaluation subset, it is overall in line with the previous result and does not appear to show substantial improvement from the phase 1 baseline.

3.3 *Phase 3: Fine Tuning from Chest Radiographs to COVID-19*

Results for all six paths are presented in Table 3 with results from DeLong's test in Table 4. In general, adding any supplemental phase of chest radiograph fine tuning prior to adapting to COVID-19 classification consistently improves performance. Statistically significant improvements over standard transfer learning (at $\alpha = 0.5$) are achieved when the model receives substantial pretraining with general radiographs in paths (a)–(c) compared to paths (d)–(f). Models that receive phase 1 training using the ChestX-ray14 set uniformly perform better than alternatives with reduced-set pneumonia-focused pretraining [path (d)], a reduced-set general radiograph pretraining [path (e)], or no pretraining [path (f)]. Of the three top performing models, there are slight distinctions. Although we observe slight improvement with additional general chest x-ray training [path (b)], and a slight loss in performance when an additional high

Table 3 COVID-19 detection performance measured by AUROC.

Variant	Description	AUROC	AUROC 95% CI
Path (a)	General chest x-ray dataset + pneumonia refinement	0.849	0.786 to 0.912
Path (b)	General chest x-ray dataset + additional general radiographs	0.857	0.794 to 0.920
Path (c)	General chest x-ray dataset	0.851	0.794 to 0.907
Path (d)	Standard ImageNet weights + pneumonia refinement	0.806	0.735 – 0.876
Path (e)	Standard ImageNet weights + additional general radiographs	0.765	0.690 to 0.841
Path (f)	Standard ImageNet weights	0.757	0.683 to 0.831

Table 4 Significance (p values) of model-to-model AUROC comparisons on test set using DeLong's test. Models pairs with statistically differentiable performance at $\alpha = 0.5$ have cells holding their p-value marked in bold, likewise models that are highly similar are marked in italics.

	Path (a)	Path (b)	Path (c)	Path (d)	Path (e)	Path (f)
Path (f)	**0.047**	**0.027**	**0.034**	0.300	*0.851*	—
Path (e)	**0.033**	**0.025**	**0.025**	0.392	—	
Path (d)	0.224	0.159	0.176	—		
Path (c)	*0.960*	*0.818*	—			
Path (b)	*0.747*	—				
Path (a)	—					

pneumonia prevalence refinement set is applied [path (a)], neither effect is statistically distinguishable from the single-phase training [path (c)] in this evaluation. Pretraining a network to label with largest set of general chest x-ray findings from a single database provides the most substantial benefit as evaluated [path (b)].

There is a marked decrease in performance in the absence of prior training with a substantial set of chest radiographs, as models received in phase 1. Interestingly, a small set of proportionally higher pneumonia cases provide a more pronounced increase in performance over pretraining with a small set of general radiographs or standard transfer learning, suggesting that a high number of pneumonia cases in training contribute much of the top-class performance in COVID-19 detection. Consequently, path (d) performs well but is not clearly distinguishable from either the top performers [paths (a)–(c)] or the bottom performers [paths (e)–(f)]. Pretraining with a set of $\sim 20,000$ radiographs with low pneumonia prevalence [path (e)] provides a marginal but statistically insignificant improvement to the scenario with no radiographs in pretraining [path (f)].

In Table 5, the p-values calculated using DeLong's test are adjusted for multiple comparisons using the Bonferroni's method.[25] Bonferroni's method divides the significance level α by the number of comparisons m, or equivalently scales the p-values. We scale the p-values in Table 4 by a factor of 15, yielding Table 5. With this adjustment in place, models with a high proportion of pneumonia cases in pretraining [paths (a)–(d)] are statistically indistinguishable in performance. There is a tangible distinction between this group of models and the group that does not receive a sizable set of pneumonia cases in pretraining [paths (e)–(f)], but it is not statistically significant when adjusted by the number of point-to-point comparisons made in this experimental design. We expand on these results and approach to multiple comparisons in the discussion.

Table 5 Significance (p-values) of model-to-model AUROC comparisons on test set using DeLong's test with Bonferroni adjustments. P-values above 1 have been limited to 1 for ease of interpretation. Model pairs with statistically indistinguishable performance have cells marked in italics.

	Path (a)	Path (b)	Path (c)	Path (d)	Path (e)	Path (f)
Path (f)	0.705	0.405	0.510	*1.000**	*1.000**	—
Path (e)	0.495	0.375	0.375	*1.000**	—	
Path (d)	*1.000**	*1.000**	*1.000*	—		
Path (c)	*1.000**	*1.000**	—			
Path (b)	*1.000**	—				
Path (a)	—					

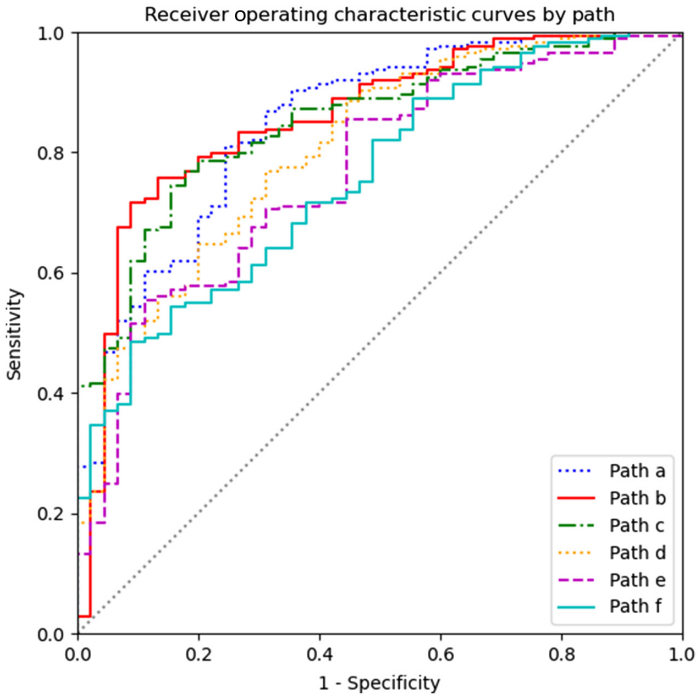

Fig. 3 Receiver operating characteristic curves for each path as evaluated on the COVID-19 test set in phase 3.

Finally, we produce ROC curves for each model evaluated on the test set in phase 3 (Fig. 3). Examining these ROCs, we see some additional nuances. Path (b) once again achieves the optimal performance in most cases, providing the highest or near-highest sensitivity for each decrease in specificity. There is a range at ~0.3 to 0.7 specificity where the path (a) model provides superior sensitivity than the path (b) model, suggesting that these complementary approaches to training may be useful for combination through ensemble methods. The three models previously distinguished for superior performance only occasionally intersect with path (d), which matches top models at low and high specificity but has decreased sensitivity at mid-level specificity. Only paths (e) and (f) show clearly inferior performance at all decision thresholds. All models are well above the center diagonal line, indicating that each model offers some effectiveness as a classifier.

3.4 *Interpretability in Results*

We used the Grad-CAM method[26] to produce saliency maps from the top-performing classifier on COVID-19 from path (b) on images in its test set. The results overall affirmed that the network appropriately uses the lung regions for classification. Figure 4 shows two selected cases and their respective overlaid saliency maps, emphasizing the image regions contributing to the classification decision, where a jet colormap spans from high-weight (red) to low-weight (blue) contribution. Note that the patient number indicated is the randomized index within the test set.

In Fig. 4, we see clear examples of what the network incorporates in a higher output score. The subdiaphragmatic region is devalued in the classification decision (blue), with larger emphasis on lung spaces with potential consolidation (red). For instance, in patient 30, an area of increased density in the middle lobe of the right lung contributes highly to the network output of 95.4% activation. Some slight emphasis is commonly present at other hyperdense regions, such as the clavicles or other bones, however, relatively low compared to activations in the lung. At most thresholds, these two images would be classified as positive findings; therefore, these are saliency maps of true positive classifications.

By contrast, Fig. 5 shows two negative cases with low-value network outputs, which typify true negative classifications. Again, the network learns to base decisions on the lungs, resulting

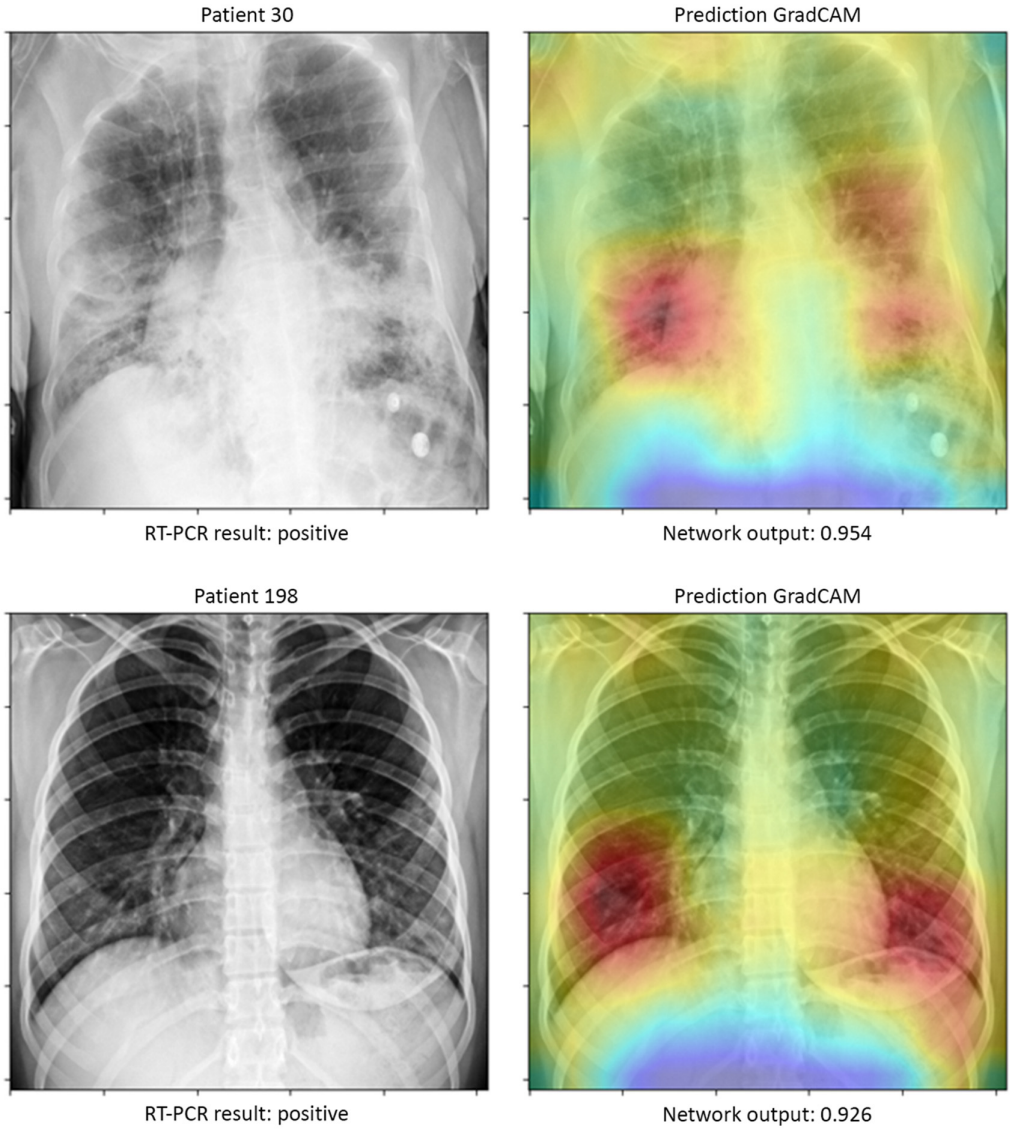

Fig. 4 Saliency maps for two exemplary cases which are true positives at most decision thresholds.

in a rough region of interest isolating the lungs. Within the lung region, no particular landmark in a non-pathological lung finding disproportionately contributes to the classification, as seen by the relatively uniform red region centered over the lungs. In patient 191, some increased density in the lower right lung especially causes some mid-level weighting, but overall the activation on this image is very low at 5.5%. Also by cropping out most burn-in labels, we can verify that the few burn-in labels present in the dataset do not contribute to the decision (blue region above patient left shoulders).

On some examples, this pattern is reversed; some additional figures showing likely false positive and true positive cases are added in the Appendix. Overall, the occurrence of these errors is proportional to the ROC metrics reported previously and depends on the selected decision threshold.

4 Discussion

The alarming pace and severity of the current COVID-19 pandemic has forced consideration of the many uses of imaging in managing treatment.[6] The utility of imaging in these situations often

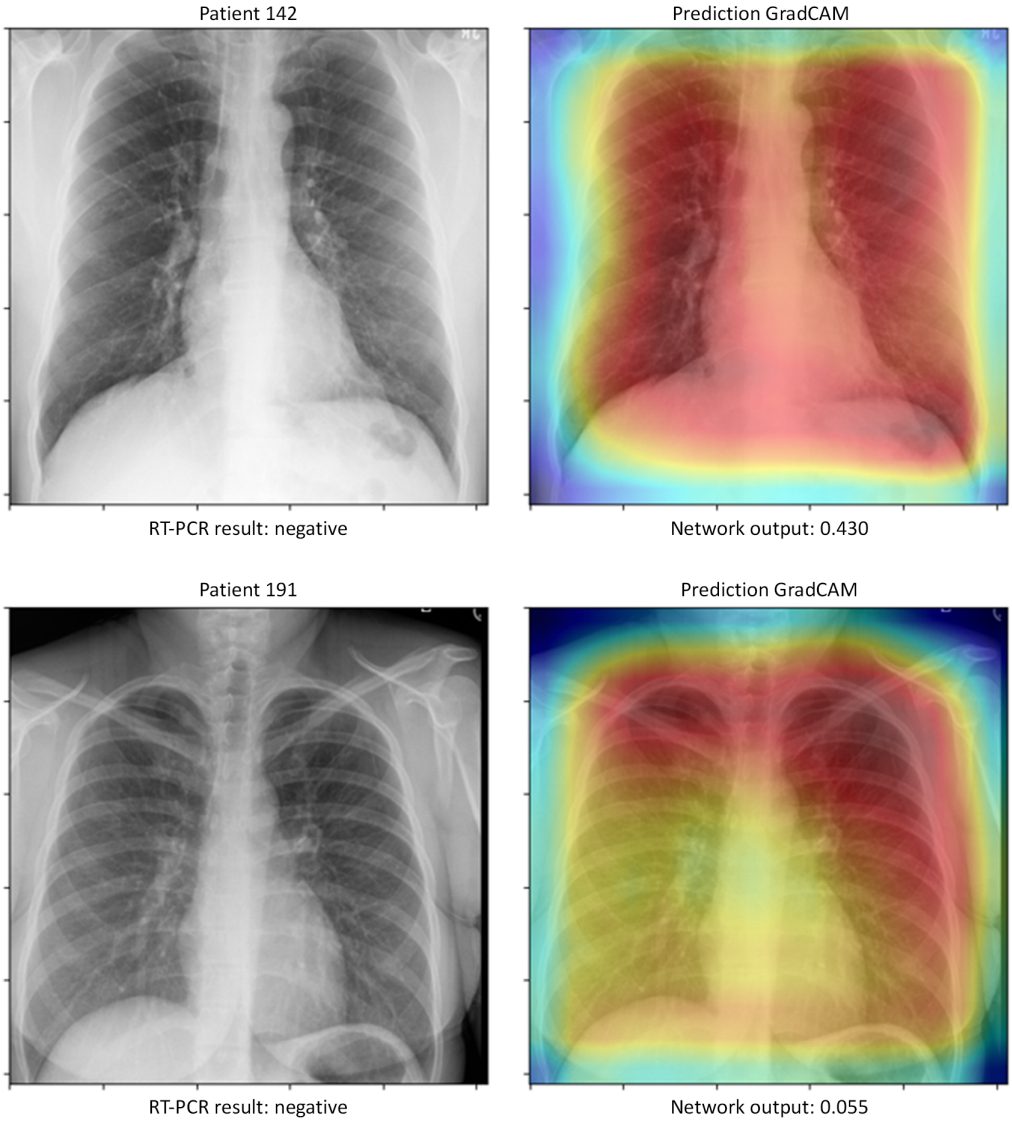

Fig. 5 Saliency maps for two exemplary cases which are true negatives at most decision thresholds.

hinges on the availability of experts to interpret the result for each patient.[6] Deep learning and other artificial intelligence techniques aim to mitigate this challenge through automated or semi-automated reading of cases. However, their quality relies on the availability of large volumes of data with known truth. It took months for COVID-19 to affect millions of people, whereas it can take years to put together effective databases for training deep learning algorithms. To address this challenge, we demonstrated that large volumes of previous data on similar conditions can tangibly improve COVID-19 classification.

Our approach builds on many previous developments in deep learning applied to CXR, which we briefly summarize here. Although the database is known to have inaccuracies in some labels,[15] the availability of a large, common set of labeled chest radiographs through the ChestX-Ray14 database has led to many high-profile successes in the multilabel task in chest x-rays.[11] Several studies iteratively improved classification performance as measured by AUROC, beginning with the originators of the database. They employed a composite model that relied on pooling together late activations of many pretrained networks.[11] Yao et al.[27] pushed performance further by relying on a simpler convolutional architecture similar to DenseNet and combining several convolutional outputs with an long-short term memory networks sequential model to capture label dependence.[27] Güendel et al.[28] simplified the model further by introducing two

convolutional layers for localization prior to a single DenseNet which classifies the input. Their performance was further boosted slightly by pooling the ChestX-Ray14 set with the prostate, lung, colorectal, and ovarian (PLCO)[29] cancer screening trial dataset for additional data in cancer-related findings primarily. From there, CheXNet famously garnered attention for initially claiming better-than-radiologist performance using a direct end-to-end transfer learning approach based on DenseNet121[14] and has since been extended in CheXpert[30] and CheXNeXt.[31] Finally, a comparison by Baltruschat's et al. introduced trained-from-scratch variants of ResNet[32] that incorporated scan parameters in its decision-making.[15]

Prior approaches all have in common convolutional architectures that employ some level of residual connection,[32] either in the use of ResNet architectures or the densely connected extensions proposed by Huang et al.[16] Our decision in using DenseNet121 as well as hyperparameter choices and in the first phase follows (with slight modifications) relatively straightforward approaches with few additions, such as CheXNet.[14] To our knowledge, the original CheXNet has not been evaluated on the official evaluation data split indicated by the ChestX-ray14 originators. Our facsimile in phase 1 achieves matching results of Güendel et al.[28] or Baltruschat et al.[15] without the addition of PLCO data or scanning parameters, respectively. Our phase 1 reproduction reaffirms the direct and simple approach, with little change from phase to phase to emphasize the effect of training data. For instance, we did not employ weighting for class imbalance, as it would need to be changed from phase to phase, and in the previous studies on chest classification and in our early testing, it was found to have minimal impact.[15] We report primarily the AUROC metric throughout in keeping with these earlier works.

One strength of the ChestX-Ray14 database and many Kaggle competitions is that the data split is common to all approaches, allowing for direct comparison of results. The selection of training set and especially evaluation set can have notable impact on the assessed classification performance of a trained model, particularly when the sample is small, even though the goal generally is to minimize this effect. It is difficult to compare metrics from different datasets, as we avoid in phase 2, due to sampling effects. For instance, our phase 3 results are roughly in range of the AUROC values reported by a recent multireader comparison study[10] undertaken concurrently with ours. However, their use of a proprietary classifier and different COVID-19 databases limit the comparisons possible. We reaffirm that using databases such as the RSNA pneumonia challenge or ChestX-Ray14 in a pretraining phase results in favorable classification performance, and anticipate classifiers based on fewer than 1000 COVID-19 cases to achieve no more than approximately the pneumonia classification performance from previous chest radiograph models. Until a standard COVID-19 radiograph database is broadly available, we cannot confidently compare COVID-19 classifiers on an equal basis, hence phase 3 emphasizes relative improvements.

Until a standard COVID-19 database is public, many approaches are limited by the number of cases available, with many preprint examples relying on the archive curated by a group as a public collaborative project.[33] Although this set provides a common, comparable data source, authors are cautioned on relying on this dataset alone for diagnostic claims.[33–35] Details such as image quality, compression method, sampling effects, and protocol may vary drastically between locations, resulting in a highly "noisy" dataset. Pooling dataset that are noisy or artefactual heterogeneously across different data groups may misleadingly yield high performance, as the classifier can choose COVID-19 cases due to imaging defects rather than underlying pathology. In our own experience, manual data cleanup including cropping was necessary as early iterations of our classifier distinguished COVID-19 easily due to detailed burn-in labels (such as Fig. 5). Even after minimizing this effect, we elected for a phased approach as COVID-19 radiographs in our set had dataset-specific, disease-related image characteristics that overwhelmed underlying pathological signals in a pooled approach. The more severe COVID-19 cases in our set were often in critical condition and imaged with portable x-rays in a supinated position, generally resulting in a lesser quality, higher density image. In some examples, the patient's head, electrocardiogram leads, or intubation equipment were visible in COVID-19 cases (such as Fig. 4), which may not be present in typical, ambulatory chest x-ray patients. This may have led to the minor emphasis on clavicular regions in the saliency maps for positive cases; however, the overall classification performance did not appear to be impacted by these cases. We minimize these confounding influences by our cleanup process. The phased approach here was effective in

deriving performance improvement from similar data that could otherwise not be used in training.

As with many other studies, our results are primarily limited by the data available. A larger dataset could improve model performance through more sample images, and likewise allow for further testing that would improve the ability to distinguish these approaches to model training. The relative improvements between certain paths, particularly (a), (b), and (c), are not distinct enough to definitively indicate one variant above all others. The Bonferroni adjustment for multiple comparisons, while known to be conservative in rejecting a null hypothesis,[25] further emphasizes how models pretrained with pneumonia-prevalent datasets are performant on COVID-19 cases and we cannot conclusively distinguish their performance with the limited dataset. The relatively small effect of training in this experiment when accounting for the number of comparisons, thereby reducing the significance in performance differences, is also consistent with the similarity between each model variant. We hypothesize that in practice an ensembling of variants such as these would likely provide the best performance. We conclude from the relative improvements in our results that the optimal paths' improved performance on COVID-19 benefitted from exposure to general chest radiograph features imparted in phase 1, as well as from fine-tuning for pneumonia-like findings from a high number of pneumonia cases in all combined pretraining.

A clearer performance improvement in paths (a) and (b) compared to (c) may be possible by refining the multiphase training, where switching data sources and reinitializing the final classification layer was disruptive to initial performance in training and potentially the model parameterization. This investigation deliberately limited variant-specific optimizations to the training protocol and hyperparameters of later steps to avoid confounding effects and emphasize the effect of pretraining phases. While training conditions were reasonable in each case and variants trained until early stopping conditions were met within each phase, we cannot rule out that further optimization in earlier phases would lead to increased performance in the COVID-19 classification task. The impact of these measures on performance can be further explored.

5 Conclusions

As researchers continue to collect data on COVID-19, deliberate pretraining with similar conditions will continue to play a part in improving classification performance. We achieved COVID-19 classification accuracies comparable to previous benchmarks of pneumonia classification and a significant improvement in classification performance relative to a baseline without phased pretraining in comparing individual models. Deliberate sequential training, rather than pooling datasets, is critical in training effective COVID-19 classifiers using limited, early datasets with potentially variable image quality. These techniques improve classifier performance on currently available data, and we expect will continue to provide performance benefits that will accelerate the production of an automated COVID-19 classifier for radiography. Our results also indicate that further testing on larger datasets is required to definitively establish their relative performance gains, as when adjusted for multiple comparisons, all variants pretrained with pneumonia radiographs attain reasonable performance. These techniques bring clinical-grade radiograph classification within reach for overwhelmed healthcare operations at the frontlines of the COVID-19 pandemic.

6 Appendix

Presented here are saliency maps of cases which caused classification errors by the neural network. They are in general reversals of the previously noted pattern and RT-PCR result. Figure 6 shows a patient with some increased density in the lower right lung, resulting in a higher network output value; however, the RT-PCR result is negative. This case is a likely false positive for most decision thresholds.

Similarly, Fig. 7 shows a similar negative decision pattern from Fig. 5 on a case with a positive laboratory test result. Here a chest x-ray with minimal consolidation corresponds to a positive RT-PCR result. The network appears to base the decision primarily off the relatively non-pathological lung appearance and assigns a low-value output to this case. Subtle consolidation is

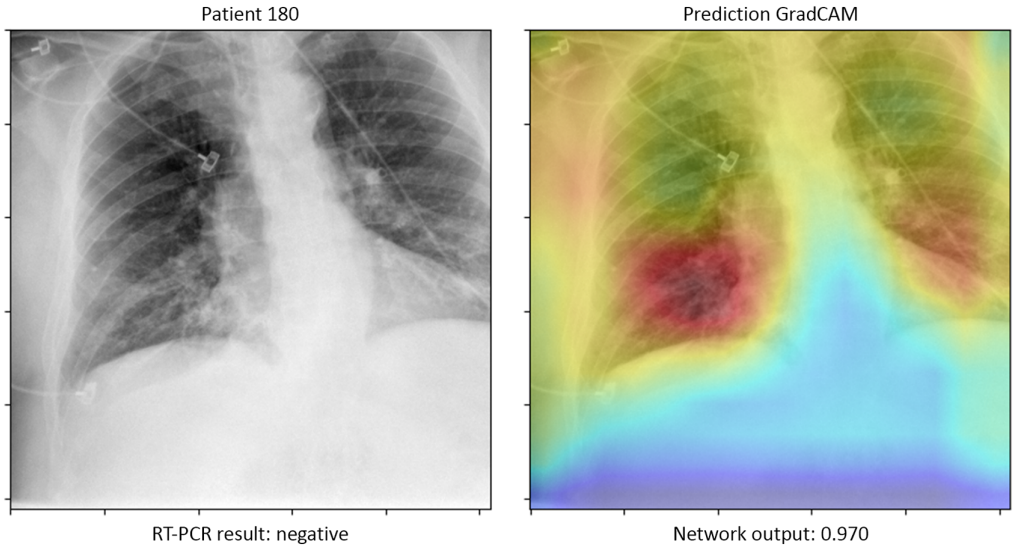

Fig. 6 Example false positive and corresponding saliency map.

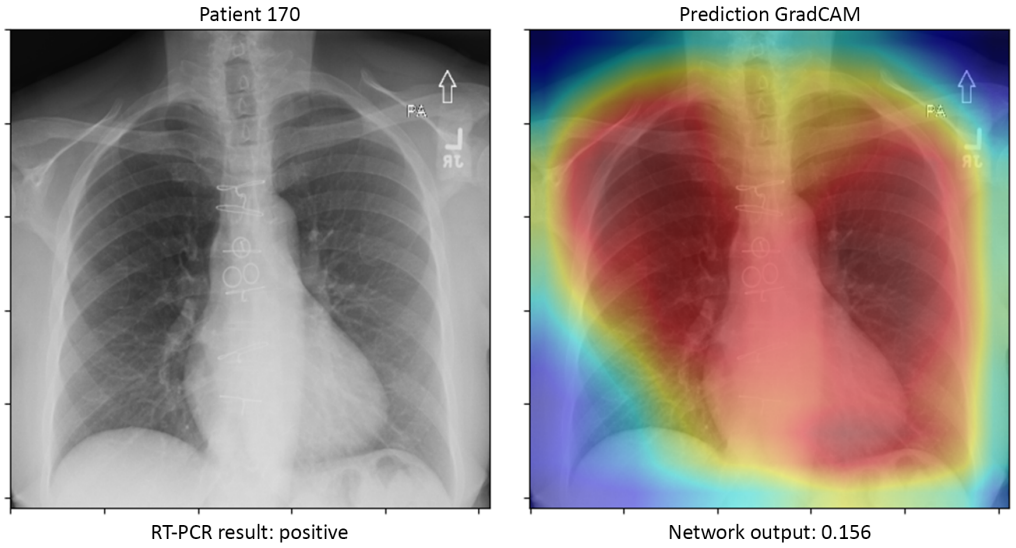

Fig. 7 Example false negative and corresponding saliency map.

difficult even for expert readers, and a subset of patients may not have visible lung involvement at time of imaging.

Disclosures

The authors have no competing interests to disclose.

Acknowledgments

We would like to thank Luca Marzoli and all our other clinical partners in collecting the COVID-19 dataset. The primary author was funded by the Department of Veterans Affairs, Big Data Science Training Enhancement Program (BD-STEP) postdoctoral program during the development of this research.

References

1. "WHO Director-General's opening remarks at the media briefing on COVID-19-11 March 2020," https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19-11-march-2020 (accessed 11 June 2020).
2. E. Dong, H. Du, and L. Gardner, "An interactive web-based dashboard to track COVID-19 in real time," *Lancet Infect Dis.* **20**(5), 533–534 (2020).
3. Y. Fang et al., "Sensitivity of chest CT for COVID-19: comparison to RT-PCR," *Radiology* **296**, 200432 (2020).
4. H. X. Bai et al., "Performance of radiologists in differentiating COVID-19 from viral pneumonia on chest CT," *Radiology* **296**, 200823 (2020).
5. T. Ai et al., "Correlation of chest CT and RT-PCR testing in Coronavirus Disease 2019 (COVID-19) in China: a report of 1014 cases," *Radiology* **296**, 200642 (2020).
6. G. D. Rubin et al., "The role of chest imaging in patient management during the COVID-19 pandemic: a multinational consensus statement from the Fleischner Society," *Radiology* **296**, 201365 (2020).
7. L. Li et al., "Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT," *Radiology* **296**, 200905 (2020).
8. Y. Zhang et al., "Comparison of patient specific dose metrics between chest radiography, tomosynthesis, and CT for adult patients of wide ranging body habitus," *Med. Phys.* **41**(2), 023901 (2014).
9. F. Ria et al., "A comparison of COVID-19 and imaging radiation risk in clinical patient populations," *J. Radiol. Prot.* **40**, 1336 (2020).
10. K. Murphy et al., "COVID-19 on the chest radiograph: a multi-reader evaluation of an AI system," *Radiology* **296**, 201874 (2020).
11. X. Wang et al., "ChestX-Ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *IEEE Conf. Comput. Vision and Pattern Recognit. (CVPR)*, pp. 3462–3471 (2017).
12. "RSNA Pneumonia Detection Challenge," https://kaggle.com/c/rsna-pneumonia-detection-challenge (accessed 19 June 2020).
13. M. Chung et al., "CT imaging features of 2019 novel coronavirus (2019-nCoV)," *Radiology* **295**(1), 202–207 (2020).
14. P. Rajpurkar et al., "CheXNet: radiologist-level pneumonia detection on chest x-rays with deep learning," https://arxiv.org/abs/1711.05225 (2017).
15. I. M. Baltruschat et al., "Comparison of deep learning approaches for multi-label chest x-ray classification," *Sci. Rep.* **9**(1), 1–10 (2019).
16. G. Huang et al., "Densely connected convolutional networks," in *IEEE Conf. Comput. Vision and Pattern Recognit. (CVPR)*, IEEE, Honolulu, Hawaii, pp. 2261–2269 (2017).
17. J. Deng et al., "ImageNet: a large-scale hierarchical image database," in *IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 248–255 (2009).
18. I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press (2016).
19. F. Chollet, *Deep Learning with Python*, 1st ed., Manning Publications Co. (2017).
20. K. He et al., "Delving deep into rectifiers: surpassing human-level performance on ImageNet classification," in *IEEE Int. Conf. Comput. Vision (ICCV)*, IEEE, Santiago, Chile, pp. 1026–1034 (2015).
21. "NIH Clinical Center provides one of the largest publicly available chest x-ray datasets to scientific community," National Institutes of Health (NIH), 2017, https://www.nih.gov/news-events/news-releases/nih-clinical-center-provides-one-largest-publicly-available-chest-x-ray-datasets-scientific-community (accessed 19 June 2020).
22. M. Abadi et al., "TensorFlow: a system for large-scale machine learning," in *Proc. 12th USENIX Conf. Operating Syst. Design and Implementation*, Association for Computing Machinery, New York, pp. 265–283 (2016).
23. E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach," *Biometrics* **44**(3), 837–845 (1988).

24. X. Sun and W. Xu, "Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves," *IEEE Signal Process. Lett.* **21**(11), 1389–1393 (2014).
25. S. Midway et al., "Comparing multiple comparisons: practical guidance for choosing the best multiple comparisons test," *PeerJ* **8**, e10387 (2020).
26. R. R. Selvaraju et al., "Grad-CAM: visual explanations from deep networks via gradient-based localization," in *IEEE Int. Conf. Comput. Vision (ICCV)*, pp. 618–626 (2017).
27. L. Yao et al., "Learning to diagnose from scratch by exploiting dependencies among labels," https://arxiv.org/abs/1710.10501 (2018).
28. S. Gündel et al., "Learning to recognize abnormalities in chest x-rays with location-aware dense networks," in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, R. Vera-Rodriguez, J. Fierrez, and A. Morales, Eds., pp. 757–765, Springer International Publishing, Cham (2019).
29. J. K. Gohagan et al., "The prostate, lung, colorectal and ovarian (PLCO) cancer screening trial of the National Cancer Institute: history, organization, and status," *Control Clin. Trials* **21**(6 Suppl.), 251S–272S (2000).
30. J. Irvin et al., "CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison," in *Proc. AAAI Conf. Artif. Intell.*, Vol. 33, pp. 590–597 (2019).
31. P. Rajpurkar et al., "Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists," *PLoS Med.* **15**(11), e1002686 (2018).
32. K. He et al., "Deep residual learning for image recognition," in *IEEE Conf. Comput. Vision and Pattern Recognit. (CVPR)*, pp. 770–778 (2016).
33. J. P. Cohen, *ieee8023/covid-chestxray-dataset*, Jupyter Notebook (2020).
34. G. Maguolo and L. Nanni, "A critic evaluation of methods for COVID-19 automatic detection from x-ray images," *Inf. Fusion* **76**, 1–7 (2020).
35. E. Tartaglione et al., "Unveiling COVID-19 from chest x-ray with deep learning: a hurdles race with small data," *Int. J. Environ. Res. Public. Health* **17**(18), 6933 (2020).

Rafael B. Fricks received his BS degree in biomedical engineering from the University of Texas at Austin in 2013 and his MS and PhD degrees in biomedical engineering from Duke University in 2017 and 2018, respectively. He is a computer engineer at MAVERIC Center and the National Artificial Intelligence Institute (NAII) in the Department of Veterans Affairs, as well as a postdoctoral fellow at Carl E. Ravin Advanced Imaging Laboratories. His current research interests include computer vision and artificial intelligence in radiology and image quality.

Francesco Ria received his medical physics doctorate degree from the University of Milan, Milan, Italy, in 2014. He is a senior research associate at Carl E. Ravin Advanced Imaging Laboratories and Clinical Imaging Physics Group at Duke University. His interests include *in vivo* quantitative analysis of radiation burden and image quality in radiology to simultaneously assess risk and clinical benefit in real patient populations.

Hamid Chalian completed his radiology training at Case Western University. He then specialized in cardiothoracic imaging at Duke University, where he stayed as an assistant professor of radiology. He subsequently moved to Salt Lake City to continue his career as an associate professor of radiology at the University of Utah. His research focus is on lung cancer, lung cancer screening, and application of innovative CT and MRI technologies to improve cardiothoracic imaging.

Pegah Khoshpouri completed her postdoctoral research fellowship from Johns Hopkins University and continued her research at Duke University. She is currently being trained as a radiology resident at the University of Utah. She is interested in the development of new biomarkers for better assessment of tumors such as liver and lung cancer.

Ehsan Abadi is an assistant professor of radiology at Duke University. He is an imaging scientist with expertise in x-ray imaging, computational human modeling, medical imaging simulation, machine learning, and quantitative image analysis. His current research focus is to identify and optimize imaging systems for accurate and precise quantifications of lung disease.

Lorenzo Bianchi received his medical physics doctorate degree from the University of Milan, Milan, Italy and, after about 10 years spent at the University Hospital of Varese, he became a chief of Medical Physics Department of ASST della Valle Olona, Italy. He is a medical physicist. His interests include nuclear medicine therapy, medical statistics, and radiation protection. Since 2010, he has been an editorial director of the Italian Association of Medical and Health Physics.

William P. Segars received his PhD in biomedical engineering from the University of North Carolina in 2001. He is an associate professor of radiology and biomedical engineering and a member of the Carl E. Ravin Advanced Imaging Laboratories (RAILabs) at Duke University. He is among the leaders in the development of simulation tools for medical imaging research where he has applied state-of-the-art computer graphics techniques to develop realistic anatomical and physiological models.

Ehsan Samei is a tenured professor and a chief imaging physicist at Duke University Health System. He is an imaging scientist with an active interest in bridging the gap between scientific scholarship and clinical practice through virtual clinical trials and clinically relevant imaging metrology and optimization. He has mentored more than 100 trainees, has published more than 300 referred journal papers, and has been the recipient of more than 30 extramural grants.