Radiology

DeepCOVID-XR: An Artificial Intelligence Algorithm to Detect COVID-19 on Chest Radiographs Trained and Tested on a Large US Clinical Dataset

Article Type: Original Research

Ramsey M. Wehbe, MD[1]; Jiayue Sheng[2]; Shinjan Dutta[2]; Siyuan Chai[2]; Amil Dravid[2]; Semih Barutcu, MS[2]; Yunan Wu, MS[2]; Donald R. Cantrell, MD, PhD[3]; Nicholas Xiao, MD[4]; Bradley D. Allen, MD, MS[5]; Gregory A. MacNealy, MD[5]; Hatice Savas, MD[5]; Rishi Agrawal, MD[5]; Nishant Parekh, MD[5]; Aggelos K. Katsaggelos, PhD[2]

[1]Division of Cardiology, Department of Medicine and Bluhm Cardiovascular Institute, Northwestern Memorial Hospital, Chicago, IL

[2]Department of Electrical and Computer Engineering, McCormick School of Engineering, Northwestern University, Evanston, IL

[3]Division of Neurointerventional Radiology, [4]Division of Interventional Radiology, and [5]Division of Thoracic Imaging, [6]Department of Radiology, Northwestern Memorial Hospital, Chicago, IL

Corresponding Author: Ramsey M Wehbe, MD
676 N St. Clair St, Suite 600, Chicago, IL 60611
ramsey.wehbe@northwestern.edu
Twitter: @ramseywehbemd

Summary Statement:
DeepCOVID-XR, an artificial intelligence algorithm for detecting COVID-19 on chest radiographs, demonstrated performance similar to the consensus of experienced thoracic radiologists.

Key Results:

- DeepCOVID-XR classified 2,214 test images (1,194 COVID-19 positive) with an accuracy of 83% and AUC of 0.90 compared with the reference standard of RT-PCR.

- On 300 random test images (134 COVID-19 positive), DeepCOVID-XR's accuracy was 82% (AUC 0.88) compared to 5 individual thoracic radiologists (accuracy 76%-81%) and the consensus of all 5 radiologists (accuracy 81%, AUC 0.85).

- Using the consensus interpretation of the radiologists as the reference standard, DeepCOVID-XR's AUC was 0.95.

Abbreviations: Coronavirus Disease 2019 (COVID-19), real time polymerase chain reaction (RT-PCR), artificial intelligence (AI), area under the curve (AUC), receiver operating characteristic (ROC), convolutional neural network (CNN)

See also the editorial by van Ginneken.

Abstract:

Background:
There are characteristic findings of Coronavirus Disease 2019 (COVID-19) on chest imaging. An artificial intelligence (AI) algorithm to detect COVID-19 on chest radiographs might be useful for triage or infection control within a hospital setting, but prior reports have been limited by small datasets and/or poor data quality.

Purpose:
To present DeepCOVID-XR, a deep learning AI algorithm for detecting COVID-19 on chest radiographs, trained and tested on a large clinical dataset.

Materials and Methods:
DeepCOVID-XR is an ensemble of convolutional neural networks to detect COVID-19 on frontal chest radiographs using real-time polymerase chain reaction (RT-PCR) as a reference standard. The algorithm was trained and validated on 14,788 images (4,253 COVID-19 positive) from sites across the Northwestern Memorial Healthcare System from February 2020 to April 2020, then tested on 2,214 images (1,192 COVID-19 positive) from a single hold-out institution. Performance of the algorithm was compared with interpretations from 5 experienced thoracic radiologists on 300 random test images using the McNemar test for sensitivity/specificity and DeLong's test for the area under the receiver operating characteristic curve (AUC).

Results:
A total of 5,853 patients (58±19 years, 3,101 women) were evaluated across datasets. On the entire test set, DeepCOVID-XR's accuracy was 83% with an AUC of 0.90. On 300 random test images (134 COVID-19 positive), DeepCOVID-XR's accuracy was 82% compared to individual radiologists (76%-81%) and the consensus of all 5 radiologists (81%). DeepCOVID-XR had a significantly higher sensitivity (71%) than 1 radiologist (60%, p<0.001) and higher specificity (92%) than 2 radiologists (75%, p<0.001; 84% p=0.009). DeepCOVID-XR's AUC was 0.88 compared to the consensus AUC of 0.85 (p=0.13 for comparison). Using the consensus interpretation as the reference standard, DeepCOVID-XR's AUC was 0.95 (0.92-0.98 95%CI).

Conclusion:
DeepCOVID-XR, an AI algorithm, detected COVID-19 on chest radiographs with performance similar to a consensus of experienced thoracic radiologists.

Introduction:

Coronavirus Disease 2019 (COVID-19) is responsible for over 40 million cases and over 1.1 million deaths worldwide as of October 22, 2020(1) and has strained critical healthcare resources. Although the reference standard for diagnosis of COVID-19 is real-time polymerase chain reaction (RT-PCR) for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) viral RNA, there are characteristic findings of COVID-19 on chest imaging with CT or X-ray. This has inspired multiple efforts at developing an artificial intelligence (AI) algorithm for automated diagnosis of COVID-19 on chest imaging. But imaging findings are not sensitive, nor specific enough to be used as a *diagnostic tool* for COVID-19, and studies that suggest otherwise are limited by selection bias(2–4). Instead, a potential application of AI for chest imaging analysis is for *triage* or *infection control* programs in a hospital or emergency department setting to provide early identification of patients with suspicious findings on chest imaging for further testing and isolation.

Although there have been promising results using AI for detection of COVID-19 on CT imaging (5–7), the use of CT for this purpose is limited by concerns regarding cost/time, radiation exposure, and decontamination procedures for equipment(8). In contrast, chest radiographs can be acquired rapidly and portably, involve trivial radiation exposure, and pose a lower risk for viral spread (9,10). But previously reported AI algorithms for identification of COVID-19 using chest radiographs have been limited by small datasets and the use of publicly available images of variable quality and questionable validity(11–16). Here, we present DeepCOVID-XR, a deep learning algorithm for detecting chest radiographs suspicious for COVID-19. DeepCOVID-XR was trained and tested on a large dataset of clinical images from a major US healthcare system (to our knowledge, the largest clinical dataset of chest radiographs from the COVID-19 era used to train a published AI platform to date). In this study, we compare the performance of the DeepCOVID-XR algorithm with interpretations by experienced thoracic radiologists.

Materials and Methods:

Patients

This retrospective study was approved by the Northwestern Institutional Review Board (IRB) (STU00212323) and granted a waiver of HIPAA authorization and a waiver of written informed consent. Our study sample included consecutive patients from over 20 sites (including hospitals, standalone emergency departments, and urgent care facilities; Table E1) across the Northwestern Memorial Health Care (NMHC) System who were tested for COVID-19 from February 2020 to April 2020. Patients included adults >=18 years of age with either 1) a documented RT-PCR test result for SARS-CoV-2 (whether positive or negative), 2) a diagnosis of COVID-19 by International Classification of Diseases (ICD-10) code, or 3) a COVID-19 "definitive positive" flag in the electronic health record (EHR). COVID-19 positivity was defined as 1) any single positive RT-PCR result for SARS-CoV-2 during the associated clinical encounter (e.g. a patient with multiple tests and only one positive result would be considered positive), 2) a diagnosis of COVID-19 by ICD-10 code, or 3) COVID-19 "definitive positive" flag in the EHR (most patients with a diagnosis only by ICD-10 code or a COVID-19 "definitive positive" flag in the EHR had a prior documented positive RT-PCR test for SARS-CoV-2 at an institution outside of NMHC). Patients with only documented negative RT-PCR tests for COVID-19 during their clinical encounter were labeled as COVID-19 negative.

Image Labeling and Dataset Partitioning

Every chest X-ray obtained during the study period for patients who met inclusion criteria was included, regardless of quality. All chest radiographs acquired during a given clinical encounter were labeled as positive or negative for COVID-19 based on the above patient-level criteria, regardless of the timing of chest X-ray compared with RT-PCR results. Images were filtered to include only frontal projections (i.e. portable anteroposterior images and only posteroanterior [PA] images from PA/lateral acquisitions).

Images from NMHC's major academic teaching hospital, Northwestern Memorial Hospital (NMH), were combined with those from other sites with the exception of images from a single community hospital, Lake Forest Hospital (LFH), which were held out as a test set that the algorithm was never exposed to during training or validation. Images from NMH and the other sites were then split into training and validation sets in an 80%/20% fashion (while ensuring no cross-over of patients between groups). Figure 1 shows the breakdown of training, validation, and test datasets.

DeepCOVID-XR: An Ensemble of Deep Neural Networks for COVID-19 Prediction

Details regarding image pre-processing; the architecture of the deep convolutional neural network ensemble model; algorithm training, validation, and testing; and saliency heatmap generation are provided in Appendix E1. Briefly, DeepCOVID-XR is a weighted ensemble of deep neural networks (Figure 2). Every image in the dataset is first preprocessed to produce 4 separate images (each of 224X224 pixels and 331X331 pixels resolution, cropped and uncropped). Each image is then fed into 6 previously validated convolutional neural network (CNN) architectures - DenseNet-121(17), ResNet-50(18), InceptionV3(19), Inception-ResNetV2(20), Xception(21), and EfficientNet-B2(22) - for a total of 24 individually trained CNNs that served as members of the deep learning model ensemble. The CNNs in this ensemble were pretrained on a large publicly available dataset of over 100,000 chest X-ray images from the National Institutes of Health(23), then fine-tuned on our clinical training set of chest X-ray images from the COVID-19 era using transfer learning. The validation dataset was used to optimize hyperparameters. The final binary prediction of the neural network architecture was a weighted average of the predictions of these individual CNNs, classifying images as either "COVID-19 positive" or "COVID-19 negative" using an output threshold of >0.5 (on a scale from 0-1). Gradient class activation mapping (Grad-CAM)(24) was used to produce heatmaps to visualize feature importance at arriving at a prediction of COVID-19 positivity. Our code base, including trained weights for each of the 24 individual neural network architectures and their respective model weights for the weighted ensemble, is provided freely on GitHub at https://github.com/IVPLatNU/deepcovidxr.

Comparison with Experienced Thoracic Radiologist Interpretations

Three hundred images were selected at random from the hold-out test dataset (ensuring only one image per patient and no patient overlap with training or validation sets) for expert interpretation. Expert interpretations were independently provided by 5 radiologists – four board-certified thoracic radiologists (RA, NP, HS, and BA) with 8, 6, 6, and 1 years of post-training experience, respectively, and one board-certified diagnostic radiologist (GM) with 38 years of post-training experience. Radiologists were blinded to any identifying information or clinical characteristics and had access to the full radiologic study in our picture archiving and communications (PACS) system (i.e. radiologists were able to review lateral images for posteroanterior/lateral studies). Radiologists provided an overall interpretation of "positive for COVID-19" or "negative for COVID-19" (chest radiographs with abnormalities that were not deemed consistent with COVID-19 were graded as "negative for COVID-19") and an associated confidence level with this assessment (graded as "low", "medium", or "high"). In this way, we derived a six-point scoring system ranging from -3 (high confidence COVID-19 negative) to +3 (high confidence COVID-19 positive). In addition to the overall interpretation, this six-point scoring system was used to calculate five separate decision thresholds for each radiologist for comparison to the algorithm. Finally, a consensus interpretation for the 5 radiologists was determined by taking the majority vote (mode) of the individual interpretations, and receiver operating characteristic (ROC) curves for the consensus interpretation were produced by calculating an average of the six-point scores for all radiologists on each image.

Statistical Analysis

For comparison of DeepCOVID-XR to radiologist interpretations, 95% confidence intervals were produced for sensitivity, specificity, and area under the ROC curve (AUC, using 2,000 bootstrap samples). Sensitivity and specificity were compared using the McNemar test for paired samples(25) and AUC was compared using DeLong's test(26). A two-tailed p-value of 0.05 was considered statistically significant. Statistical analyses were performed using packages

DTComPair and pROC in R version 3.6 (R core team; R foundation for statistical computing; Vienna, Austria).

Results:

Patient Characteristics

A total of 5,853 patients (58±19 years , 3,101 females, 1,782 COVID-19 positive) were evaluated across datasets (Table 1). The rate of positivity for COVID-19 among chest radiographs in the hold-out test set (1,192/2,214; 54%) was higher than in the training (3,390/11,786; 29%) and validation (863/2,992; 29%) sets. A higher proportion of COVID-positive patients in the test set (237/324; 75%) were treated as inpatients than in the training (719/1,142; 63%) and validation (216/324; 67%) sets. Additionally, there was a higher proportion of anteroposterior images (97%, 2,141/2,214) in the test set compared with training (86%, 10,200/11,786) and validation (84%, 2,502/2,992) sets. In the test set, 263/1,192 (22%) COVID-19 positive images were acquired prior to positive RT-PCR results.

Performance of DeepCOVID-XR

A performance comparison of individual model architectures and ensemble models is provided in Appendix E1 (Tables E2 and E3). On the hold-out test set of 2,214 images, the overall accuracy for predicting COVID-19 for DeepCOVID-XR was 83% (1,846/2,214) with a sensitivity of 75% (898/1,192), specificity of 93% (948/1,022), and AUC of 0.90 (confusion matrix Figure 3A, ROC curve Figure 3B). Notably, 156/1,192 (13%) of COVID-19 positive images were acquired prior to RT-PCR results *and* accurately labeled by the algorithm as suspicious for COVID-19. As approximately 5% (44 patients contributing 151 images) of patients in the hold-out test set also received a chest X-ray at one of the institutions in our training or validation sets during the study period, we performed a sensitivity analysis where these images were dropped from the test set in which the results were unchanged (Table E4). The most representative images from each class of DeepCOVID-XR predictions (true positive, true negative, false positive, and false negative) are provided in Figure 4. Grad-CAM heatmaps of feature importance for individual chest radiographs are provided in Figure 5. Heatmaps for COVID-19 positive images highlighted features in the lung fields identifying areas of abnormalities (Figure 5A-C) in contrast to heatmaps for COVID-19 negative images (Figure 5D).

Comparison with Expert Thoracic Radiologists

A comparison of the performance of DeepCOVID-XR to expert chest radiologist interpretations on 300 patients' chest radiographs (134 COVID-19 positive) randomly selected from the hold-out test set is provided in Table 2. The overall accuracy of DeepCOVID-XR on this test set was 82% (247/300) compared with the reference standard of RT-PCR, while the accuracy of individual radiologists ranged from 76% (227/300) to 81% (242/300) and the accuracy of the consensus interpretation of all 5 radiologists was 81% (242/300). DeepCOVID-XR had a significantly higher specificity at 92% (152/166) than 2 of the radiologists (75%, 125/166; $p<0.001$ and 84%, 139/166; $p=0.009$) and significantly higher sensitivity at 71% (95/134) than 1 radiologist (60%, 81/134; $p<0.001$). A comparison of the ROC curve for DeepCOVID-XR to overall individual radiologist interpretations is provided in Figure 6A. A comparison of DeepCOVID-XR to individual radiologists on each of the 5 decision thresholds derived from the six-point scoring system is provided in Appendix E1 (Table E5 and Figure E1).

The AUC for DeepCOVID-XR was 0.88 (0.84-0.92) compared to 0.85 (0.80-0.89, p = 0.13 for comparison) for the consensus interpretation of all 5 radiologists (Figure 6B). When using the consensus interpretation as the reference standard rather than RT-PCR, the AUC for DeepCOVID-XR was 0.95 (0.92-0.98; Figure 6C). The time to analyze this subset of 300 images with DeepCOVID-XR on a single NVIDIA Titan V graphics processing unit (GPU) was approximately 18 minutes, compared with approximately 2.5-3.5 hours for each of the expert radiologists.

Discussion:

In this study, we present DeepCOVID-XR, an ensemble deep learning artificial intelligence algorithm to detect COVID-19 on chest radiographs. DeepCOVID-XR was trained and tested on, to our knowledge, the largest clinical dataset of chest radiographs from the COVID-19 era of any other published AI platform to date, including images from multiple institutions across a large US healthcare system (17,002 images from 5,853 patients total). Of note, study patients were representative of a "real world" population of patients presenting for emergency or inpatient care in the COVID-19 era – a proportion of COVID-19 negative patients likely had a spectrum non-COVID-19 related abnormalities on chest X-ray (including confounders like non-COVID viral pneumonia) that one would expect in this patient population. On a hold-out test dataset of 2,214 images (1,192 COVID-19 positive) from a single institution that the algorithm was not exposed to during model development, DeepCOVID-XR detected COVID-19 with an overall accuracy of 83% (sensitivity 75% and specificity 93%) and an AUC of 0.90. Additionally, on a random sample of 300 test images, DeepCOVID-XR's accuracy was 82% compared to 76-81% for individual experienced thoracic radiologists and 81% for the consensus interpretation of all 5 radiologists. Finally, the AUC for DeepCOVID-XR was 0.88 compared to 0.85 for the consensus interpretation (p=0.13). Using the consensus radiologist interpretation as the reference standard rather than RT-PCR, the AUC for DeepCOVID-XR was 0.95 suggesting a discriminative ability of our algorithm similar to a consensus of experts.

Errors made by the algorithm were explainable. Images categorized as positive (whether true or false positive) often had characteristic features of COVID-19 viral pneumonia previously reported in the literature including bilateral consolidations and ground glass opacities with lower lung-zone and peripheral predominance(27,28). In contrast, images categorized as negative by the algorithm (whether true or false negative) often had clear lung fields and/or concomitant pleural effusions – interestingly, pleural effusions have been previously found to be quite rare (3%) in COVID-19 related pneumonia(27). Visualization of feature importance using Grad-CAM heatmaps revealed abnormalities in the lung fields to be highly predictive of COVID-19 on chest radiographs as expected. This serves as an important sanity check to reinforce confidence in algorithm predictions.

The explainable errors in algorithm predictions likely represent limitations of chest imaging in the radiologic diagnosis of COVID-19 rather than the algorithm itself. Prior clinical studies showed COVID-19 pneumonia produces characteristic features on chest imaging, but up to 56% of symptomatic patients can demonstrate normal chest imaging, especially early in their disease course(9,27,29–31). Imaging is therefore inappropriate to “rule out” disease. Also, many of the findings seen in COVID-19 imaging are non-specific with overlap, particularly with other viral pneumonias(32). Chest imaging therefore should *not* be used as a *diagnostic* tool for COVID-19, but *could* play an important role in earlier identification of patients likely to have the disease to aid in triage and infection control. Interestingly, Wong et al.(27) found that ~9% of patients had abnormal imaging *before* positive RT-PCR, a proportion similar to the 13% (156/1,192) of COVID-19 positive images in our study obtained prior to the patient's positive RT-PCR result *and* flagged as COVID-19 positive by our algorithm.

A number of groups from industry and academic sectors have published studies and non-peer reviewed preprints with claims of extremely high sensitivity and specificity of artificial intelligence algorithms to detect COVID-19 on chest radiographs (11–14). But most of these studies have been limited by small sample sizes or have relied on images from publicly available datasets containing a mix of images from research articles and clinical reports of variable quality and questionable accuracy of image labels(15). These datasets are subject to significant bias(33) and simply not sufficient to train an algorithm ultimately intended for clinical use.

Murphy et al.(16) presented a deep learning algorithm for detection of COVID-19 on chest radiographs which included both pre-COVID era images for model pretraining and a dataset of 606 clinical images for training and 468 clinical images for testing from patients at 2 Dutch Hospitals during the COVID-19 era. The authors used a commercial patch-based convolutional neural network called CAD4COVID-Xray with an AUC of 0.81 for predicting COVID-19 on a hold-out test set from a single institution. By contrast, our model was trained and tested on greater than 15 times the number of clinical images from the COVID-19 era. DeepCOVID-XR's discriminative performance on an independent hold-out test set was superior to that reported for CAD4COVID-Xray (AUC 0.90 vs. 0.81). Although differences in patient populations may in part account for this difference in performance, our algorithm's AUC of 0.95 when compared with a consensus of radiologists was far higher than that reported by Murphy et al. (0.81-0.86) suggesting that DeepCOVID-XR more reliably produces predictions in line with the “ground truth” radiologic diagnosis as determined by a consensus of experts. Finally, although the authors made their software available for use through a cloud-based interface, no details regarding algorithm development or code were made available given the proprietary nature of their platform. We are freely providing all of our code and pretrained neural network/model ensemble weights for open source use towards a democratized approach to model development (https://github.com/IVPLatNU/deepcovidxr).

Our study has some limitations. First, the algorithm was evaluated on only those patients who were tested for COVID-19, thus there was likely some degree of selection bias. Second, the performance of our algorithm was compared to RT-PCR as a reference standard, which itself has somewhat limited sensitivity due to sampling error or viral mutation (34). Finally, it is unclear how well the algorithm performs when COVID-19 is not the dominant viral pneumonia as the study was performed at a time of considerable case load in our healthcare system.

In conclusion, DeepCOVID-XR is a deep learning artificial intelligence algorithm for detecting COVID-19 on chest radiographs trained and tested on a large US clinical dataset with performance similar to the consensus interpretation of experienced thoracic radiologists. We feel that this algorithm has the potential to benefit healthcare systems in mitigating unnecessary exposure to the virus by serving as an automated tool to rapidly flag patients with suspicious chest imaging for isolation and further testing. Planned future studies include a prospective evaluation of the algorithm (including in those patients not under investigation for COVID-19), a necessity for any AI algorithm prior to clinical implementation. Also, we plan to 1) incorporate other clinical data (e.g. demographics, vital signs, laboratory data) into the algorithm to further boost the performance and 2) adapt the algorithm for risk prediction of clinically meaningful outcomes in patients with confirmed COVID-19. By providing the DeepCOVID-XR algorithm code base as an open source resource, we hope investigators around the world will further improve, fine tune, and test the algorithm using clinical images from their own institutions.

Acknowledgements

The authors would like to acknowledge Scott M. Leonard for his assistance with transferring de-identified studies to our picture archiving and communications system for review by radiologists.

References

1. Gardner L, Ensheng D. Johns Hopkins Center for Systems Science and Engineering (CSSE) COVID-19 Dashboard. 2020. https://coronavirus.jhu.edu/map.html. Accessed August 11, 2020.
2. Weinstock MB, Echenique A, Russell JW, et al. Chest X-Ray Findings in 636 Ambulatory Patients with COVID-19 Presenting to an Urgent Care Center: A Normal Chest X-Ray Is no Guarantee. J Urgent Care Med. 2020;13–18.
3. Hope M. A role for CT in COVID-19? What data really tell us so far. Lancet. Elsevier Ltd; 2020;6736(20):30728.
4. Wynants L, Van Calster B, Collins GS, et al. Prediction models for diagnosis and prognosis of covid-19: Systematic review and critical appraisal. BMJ. 2020;369.
5. Shi F, Wang J, Shi J, et al. Review of Artificial Intelligence Techniques in Imaging Data Acquisition, Segmentation and Diagnosis for COVID-19. IEEE Rev Biomed Eng. 2020;1–13.
6. Li L, Qin L, Xu Z, et al. Using Artificial Intelligence to Detect COVID-19 and Community-acquired Pneumonia Based on Pulmonary CT: Evaluation of the Diagnostic Accuracy. Radiology. 2020;296(2):E65–E71.
7. Huang L, Han R, Ai T, et al. Serial Quantitative Chest CT Assessment of COVID-19: Deep-Learning Approach. Radiol Cardiothorac Imaging. 2020;2(2):e200075.
8. ACR Recommendations for the use of Chest Radiography and Computed Tomography (CT) for Suspected COVID-19 Infection. 2020. https://www.acr.org/Advocacy-and-Economics/ACR-Position-Statements/Recommendations-for-Chest-Radiography-and-CT-for-Suspected-COVID19-Infection. Accessed August 11, 2020.
9. Cleverley J, Piper J, Jones MM. The role of chest radiography in confirming covid-19 pneumonia. BMJ. 2020;370.
10. Jacobi A, Chung M, Bernheim A, Eber C. Portable chest X-ray in coronavirus disease-19 (COVID-19): A pictorial review. Clin Imaging. 2020;64(January):35–42.
11. Yi PH, Kim TK, Lin CT. Generalizability of Deep Learning Tuberculosis Classifier to COVID-19 Chest Radiographs: New Tricks for an Old Algorithm? J Thorac Imaging. 2020;35(4):102–104.
12. Oh Y, Park S, Ye JC. Deep Learning COVID-19 Features on CXR using Limited Training Data Sets. IEEE Trans Med Imaging. 2020;1–1.
13. Castiglioni I, Ippolito D, Interlenghi M, et al. Artificial intelligence applied on chest X-ray can aid in the diagnosis of COVID-19 infection: a first experience from Lombardy, Italy. medRxiv. 2020.
14. Wang L, Zhong QL, Wong A. COVID-Net: A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases from Chest X-Ray Images. arXiv. 2020;1–12.
15. Cohen JP, Morrison P, Dao L. COVID-19 Image Data Collection. arXiv. 2020;
16. Murphy K, Smits H, Knoops AJG, et al. COVID-19 on the Chest Radiograph: A Multi-Reader Evaluation of an AI System. Radiology. 2020;201874.
17. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. Proc - 30th IEEE Conf Comput Vis Pattern Recognition, CVPR 2017. 2017;2017-Janua:2261–2269.
18. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proc IEEE

Comput Soc Conf Comput Vis Pattern Recognit. 2016;2016-Decem:770–778.

19. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the Inception Architecture for Computer Vision. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit. 2016;2016-Decem:2818–2826.
20. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. Inception-v4, inception-ResNet and the impact of residual connections on learning. 31st AAAI Conf Artif Intell AAAI 2017. 2017;4278–4284.
21. Chollet F. Xception: Deep learning with depthwise separable convolutions. Proc - 30th IEEE Conf Comput Vis Pattern Recognition, CVPR 2017. 2017;2017-Janua:1800–1807.
22. Tan M, Le Q V. EfficientNet: Rethinking model scaling for convolutional neural networks. 36th Int Conf Mach Learn ICML 2019. 2019;2019-June:10691–10700.
23. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-ray: Hospital-Scale Chest X-ray Database and Benchmarks on Weakly Supervised Classification and Localization of Common Thorax Diseases. Adv Comput Vis Pattern Recognit. 2019;369–392.
24. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. Proc IEEE Int Conf Comput Vis. 2017;2017-Octob:618–626.
25. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika. 1947;12(2):153–157.
26. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. United States; 1988;44(3):837–845.
27. Wong HYF, Lam HYS, Fong AHT, et al. Frequency and Distribution of Chest Radiographic Findings in Patients Positive for COVID-19. Radiology. 2020;296(2):E72–E78.
28. Ng M-Y, Lee EY, Yang J, et al. Imaging Profile of the COVID-19 Infection: Radiologic Findings and Literature Review. Radiol Cardiothorac Imaging. 2020;2(1):e200034.
29. Vancheri SG, Savietto G, Ballati F, et al. Radiographic findings in 240 patients with COVID-19 pneumonia: time-dependence after the onset of symptoms. Eur Radiol. European Radiology; 2020;
30. Salehi S, Abedi A, Balakrishnan S, Gholamrezanezhad A. Coronavirus Disease 2019 (COVID-19): A Systematic Review of Imaging Findings in 919 Patients. AJR. 2020;215(1):87–93.
31. Bernheim A, Mei X, Huang M, et al. Chest CT findings in coronavirus disease 2019 (COVID-19): Relationship to duration of infection. Radiology. 2020;295(3):685–691.
32. Bai HX, Hsieh B, Xiong Z, et al. Performance of Radiologists in Differentiating COVID-19 from Non-COVID-19 Viral Pneumonia at Chest CT. Radiology. 2020;296(2):E46–E54.
33. Degrave AJ, Janizek JD, Lee S. AI for radiographic COVID-19 detection selects shortcuts over signal. 2020;1–21.
34. Watson J, Whiting PF, Brush JE. Interpreting a covid-19 test result. BMJ. 2020;369(May):1–7.

Table 1. Baseline Characteristics of Patients

Characteristics	Overall	Training			Validation			Testing		
		Total	COVID-Pos	COVID-Neg	Total	COVID-Pos	COVID-Neg	Total	COVID-Pos	COVID-Neg
Patient Level Data (N)	5,853	3,931	1,142	2,789	1,100	324	776	866	324	542
Age	58 ± 19	58 ± 18	57 ± 18	58 ± 19	59 ± 18	56 ± 16	60 ± 19	57 ± 18	56 ± 17	58 ± 19
Female Sex	3,101 (53%)	2,081 (53%)	577 (51%)	1,504 (54%)	580 (53%)	149 (46%)	431 (56%)	440 (54%)	161 (51%)	279 (55%)
Inpatient	3,629 (62%)	2,413 (61%)	719 (63%)	1,694 (61%)	672 (61%)	216 (67%)	456 (59%)	544 (66%)	237 (75%)	307 (61%)
Chest Radiographs per Patient	1 (1-3)	1 (1-3)	2 (1-3)	1 (1-3)	1 (1-2)	1 (1-3)	1 (1-2)	1 (1-2)	1 (1-2)	1 (1-2)
Chest X-ray Level Data (N)	17,002	11,796	3,390	8,406	2,992	863	2,129	2,214	1,192	1,022
AP/ PA/ Not Listed	14,843(87%)/ 1,105(6%)/ 1,054(6%)	10,200(86%)/ 781(7%)/ 815(7%)	3,024(89%)/ 111(3%)/ 255(8%)	7,176(85%)/ 670(8%)/ 560(7%)	2,502(84%)/ 264(9%)/ 226(8%)	770(89%)/ 32(4%)/ 61(7%)	1,732(81%)/ 232(11%)/ 165(8%)	2,141(97%)/ 60(3%)/ 13(0%)	1,183(99%)/ 8(1%)/ 1(0%)	958(94%)/ 52(5%)/ 12(1%)
Chest X-ray Prior to 1st Pos PCR	-	-	1,161 (34%)	-	-	241 (28%)	-	-	263 (22%)	-
Hours from Chest X-ray to 1st Pos PCR*	-	-	34 (6-273)	-	-	21 (3-76)	-	-	10 (1-40)	-

Note.—For categorical variables, values are presented as n (% of subgroup). For continuous variables, values are presented as mean ± standard deviation for normally distributed data or median (interquartile range) for non-normally distributed data. pos = positive, neg = negative, Hosp = hospital, LOS = length of stay, pts = patients, PCR = polymerase chain reaction, AP = anteroposterior, PA = posteroanterior, sd = standard deviation, N = number of patients or number of cases

*Among images that were acquired prior to first positive RT-PCR result.

Table 2. Performance of DeepCOVID-XR on Random Sample of 100 Images from Test Set Compared with Expert Thoracic Radiologists' Interpretations and Consensus Radiologist Interpretation

	DeepCOVID-XR	Radiologist 1		Radiologist 2		Radiologist 3		Radiologist 4		Radiologist 5		Consensus	
Metrics	Performance	Performance	p-value*	Performance	p-value*	Performance	p-value*	Performance	p-value*	Performance	p-value*	Performance	p-value*
Accuracy	**82%**	79%	-	81%	-	76%	-	76%	-	79%	-	81%	-
TP (n)	95	87	-	92	-	81	-	**102**	-	99	-	94	-
TN (n)	**152**	151	-	150	-	148	-	125	-	139	-	148	-
FP (n)	**14**	15	-	16	-	18	-	41	-	27	-	18	-
FN (n)	39	47	-	42	-	53	-	**32**	-	35	-	40	-
Sensitivity (95% CI)	71% (63%-79%)	65% (57%-73%)	0.09	69% (61%-77%)	0.47	60% (52%-69%)	<0.001	**76% (69%-83%)**	0.09	74% (66%-81%)	0.37	70% (62%-78%)	0.78
Specificity (95% CI)	**92% (87%-96%)**	91% (87%-95%)	0.8	90% (86%-95%)	0.62	89% (84%-94%)	0.32	75% (69%-82%)	<0.001	84% (78%-89%)	0.009	89% (84%-94%)	0.29
AUC (95% CI)	**0.88 (0.84-0.92)**	-	-	-	-	-	-	-	-	-	-	0.85 (0.80-0.89)	0.13

Note.—TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative, AUC = Area Under the Curve, n= number, CI = confidence interval

* p-value for comparison to DeepCOVID-XR algorithm performance

Figure Legends

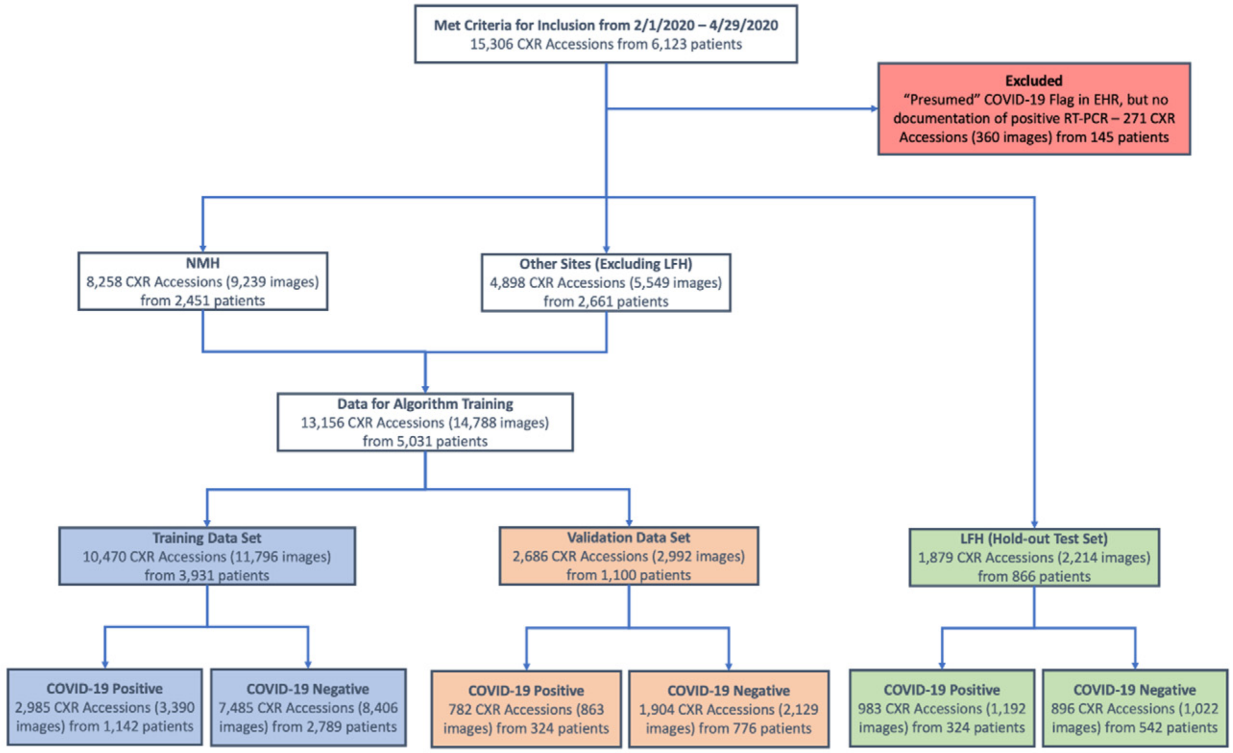

Met Criteria for Inclusion from 2/1/2020 – 4/29/2020
15,306 CXR Accessions from 6,123 patients
Excluded
"Presumed" COVID-19 Flag in EHR, but no documentation of positive RT-PCR – 271 CXR Accessions (360 images) from 145 patients
NMH
8,258 CXR Accessions (9,239 images) from 2,451 patients
Other Sites (Excluding LFH)
4,898 CXR Accessions (5,549 images) from 2,661 patients
Data for Algorithm Training
13,156 CXR Accessions (14,788 images) from 5,031 patients
Training Data Set
10,470 CXR Accessions (11,796 images) from 3,931 patients
Validation Data Set
2,686 CXR Accessions (2,992 images) from 1,100 patients
LFH (Hold-out Test Set)
1,879 CXR Accessions (2,214 images) from 866 patients
COVID-19 Positive
2,985 CXR Accessions (3,390 images) from 1,142 patients
COVID-19 Negative
7,485 CXR Accessions (8,406 images) from 2,789 patients
COVID-19 Positive
782 CXR Accessions (863 images) from 324 patients
COVID-19 Negative
1,904 CXR Accessions (2,129 images) from 776 patients
COVID-19 Positive
983 CXR Accessions (1,192 images) from 324 patients
COVID-19 Negative
896 CXR Accessions (1,022 images) from 542 patients

Figure 1. Flowchart for Patient Inclusion in the Study and Breakdown of Training, Validation, and Hold-Out Test Datasets. NMH = Northwestern Memorial Hospital, LFH= Lake Forest Hospital.

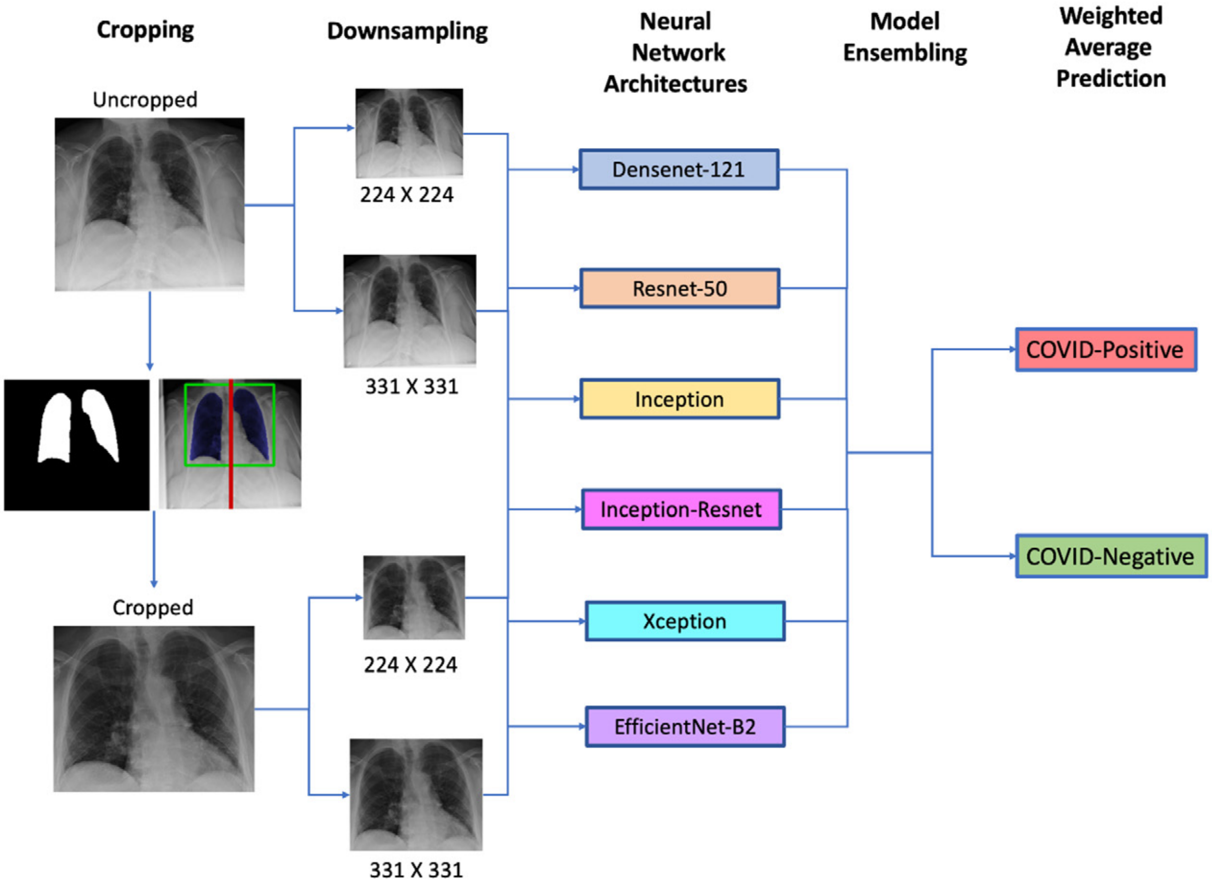


Figure 2. DeepCOVID-XR: A Weighted Average Ensemble of Deep Learning Models: Schematic showing general architecture of the DeepCOVID-XR deep learning ensemble model. Images are initially preprocessed to crop a square centered on the lung fields, then both uncropped and cropped images are downsampled to two different resolutions (224X224 and 331X331 pixels) before being fed into each of 6 different previously validated convolutional neural network architectures (4 images X 6 architectures = 24 models total). The predictions from each individual model are then ensembled using a weighted average to produce a single prediction of COVID-19 positive or COVID-19 negative per image.

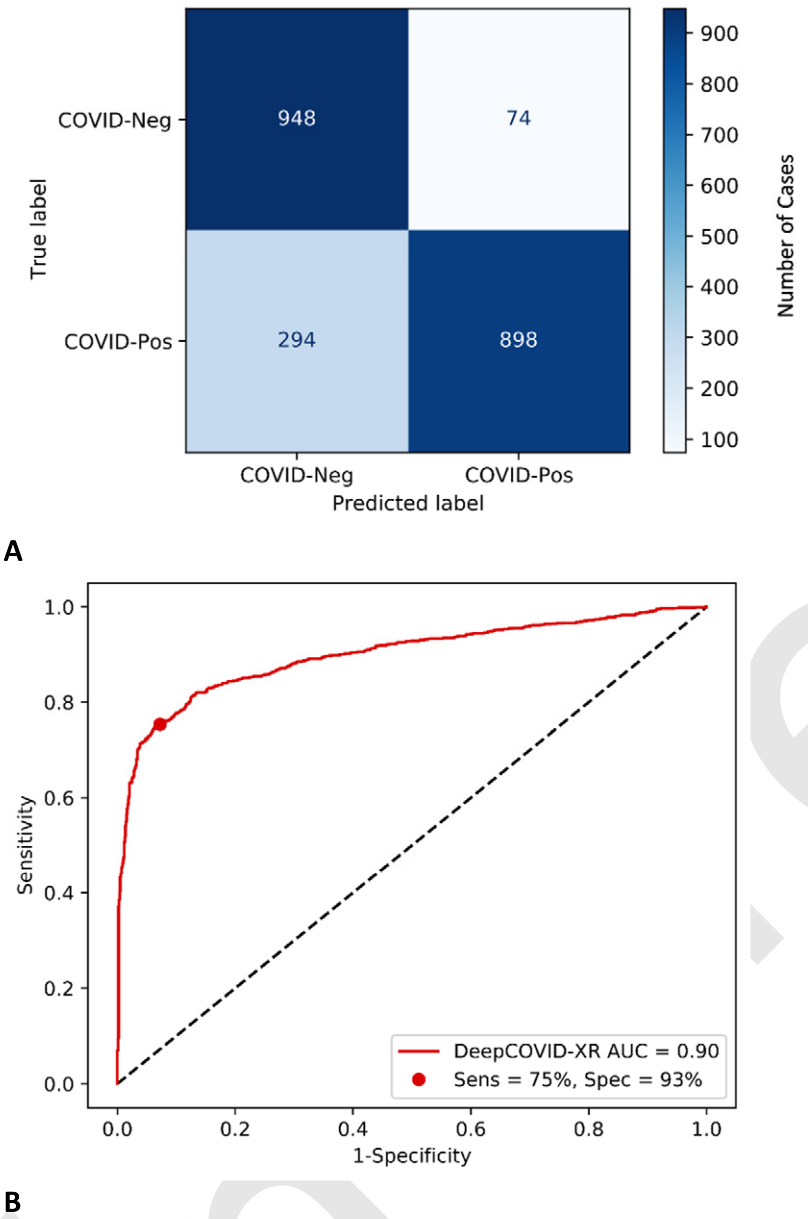


Figure 3. Performance of DeepCOVID-XR on Hold-Out Test Set of 2,214 Images. *A,* Confusion matrix of algorithm predictions and, *B,* ROC curve (red line) showing discriminative performance of the algorithm with an AUC of 0.90 and prediction threshold for COVID-19 positivity (red point) with a sensitivity of 75% (898/1,192) and specificity of 93% (948/1,022). Sens = Sensitivity, Spec = Specificity, ROC = receiver operating characteristic, AUC = area under the ROC curve.

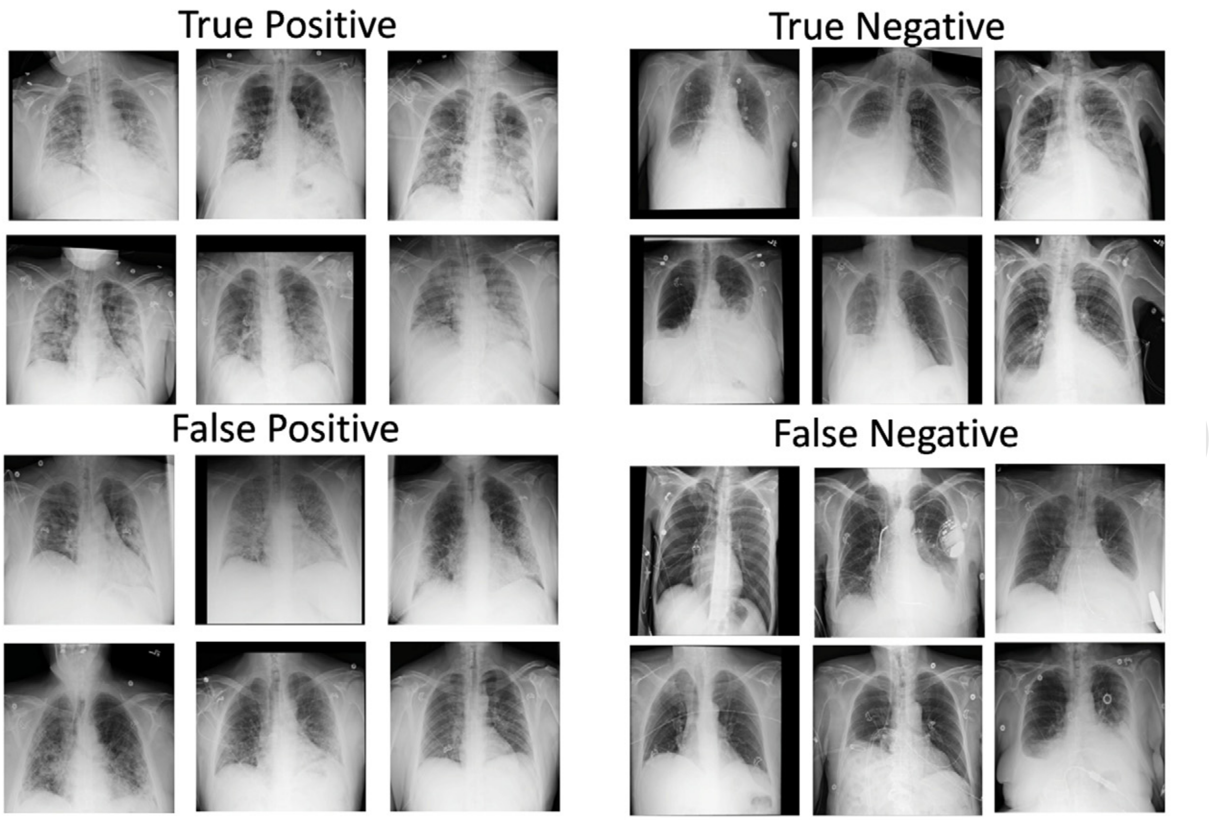


Figure 4. Sample of Most Representative Images from Different Classes of DeepCOVID-XR Predictions Relative to the Reference Standard. Images classified as positive by the algorithm (whether true or false positive) tended to have typical features of COVID-19 pneumonia described in the literature including patchy bilateral consolidations and ground glass opacities with peripheral and lower lung zone predominance. Images classified as negative by the algorithm tended to have clear lung fields and/or concomitant pleural effusions which are reported to be rare in COVID-19 pneumonia.

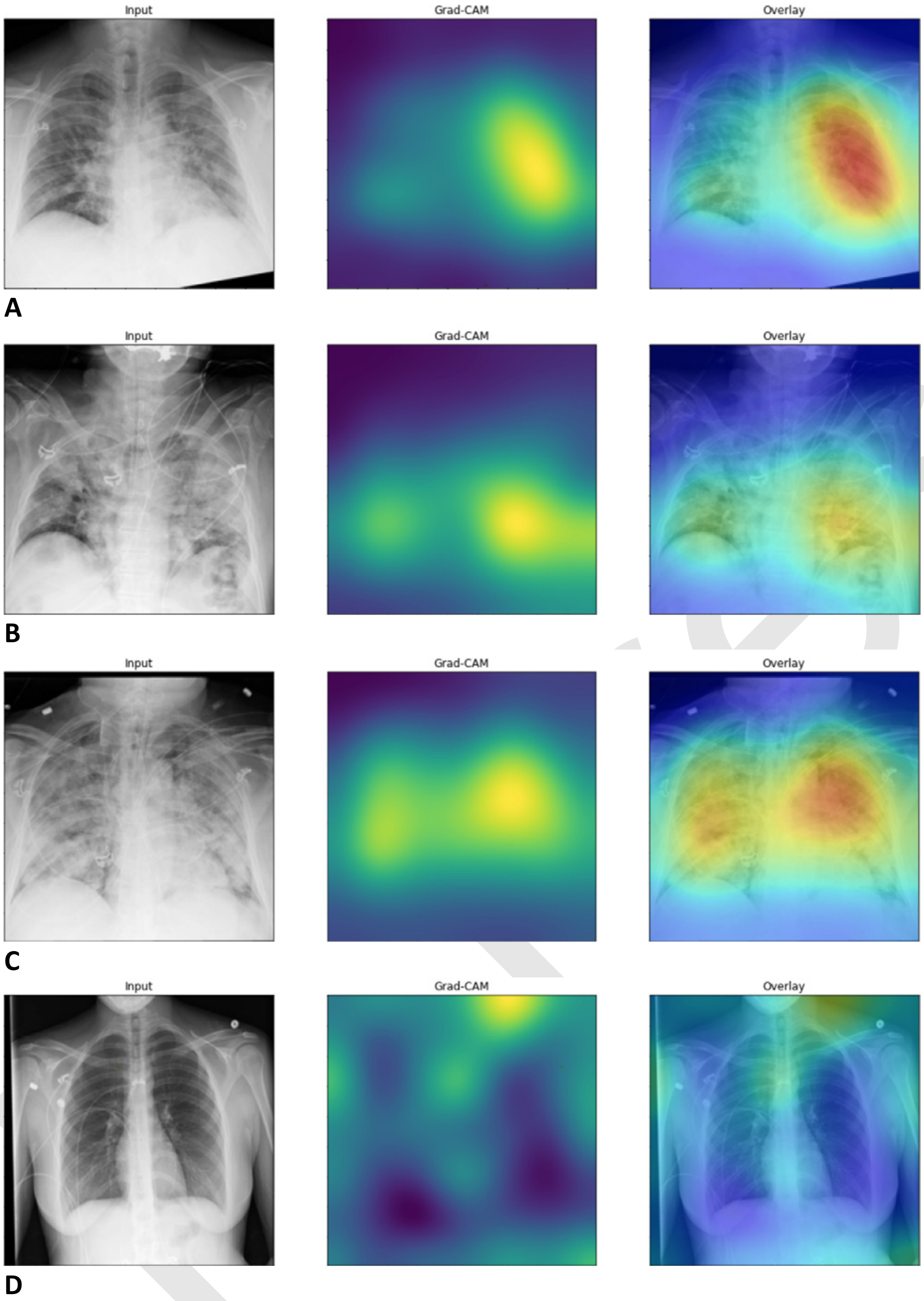
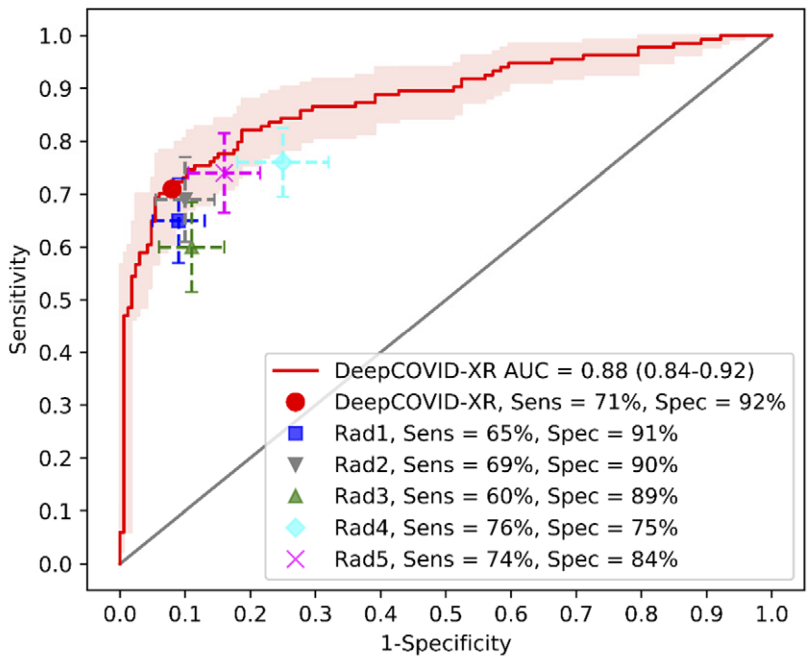
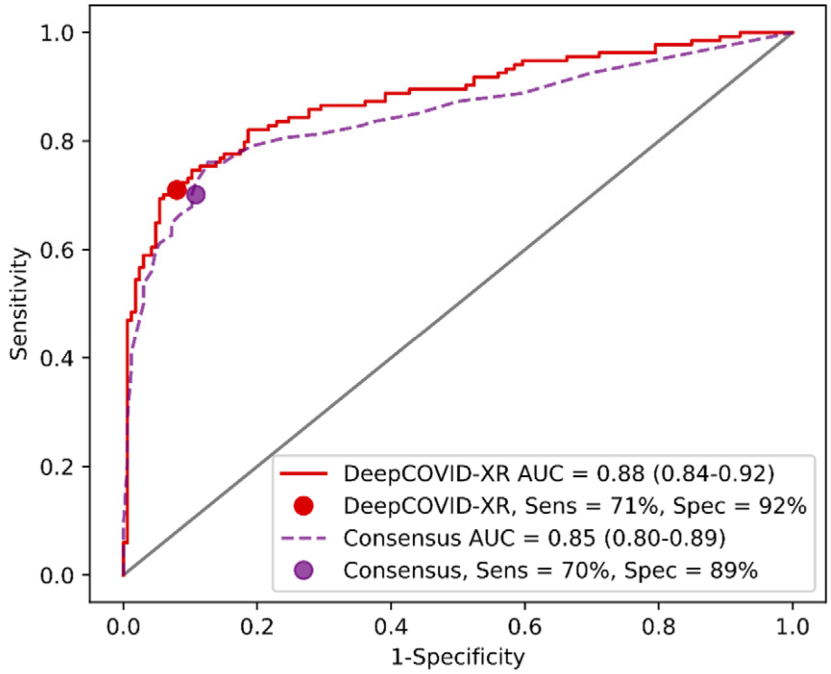


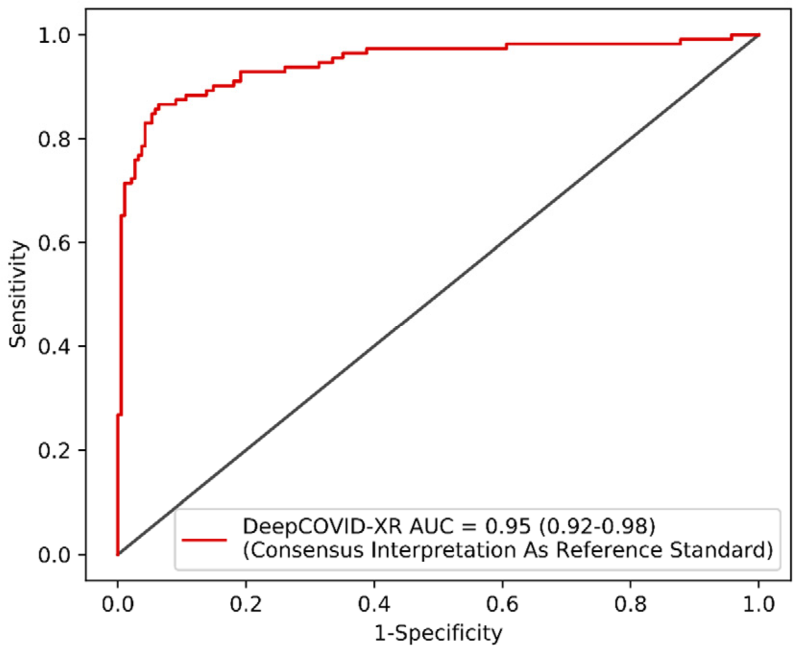
Figure 5. Grad-CAM Heatmaps of Feature Importance for Positive COVID-19 Prediction. Generated heatmaps appropriately highlighted abnormalities in the lung fields in (*A*-*C*) those images accurately labeled as COVID-19 positive in contrast to, *D,* images which were accurately labeled as negative for COVID-19. Intensity of colors on the heatmap correspond to features of the image that are important for prediction of COVID-19 positivity.


Sensitivity
1-Specificity
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
DeepCOVID-XR AUC = 0.88 (0.84-0.92)
DeepCOVID-XR, Sens = 71%, Spec = 92%
Rad1, Sens = 65%, Spec = 91%
Rad2, Sens = 69%, Spec = 90%
Rad3, Sens = 60%, Spec = 89%
Rad4, Sens = 76%, Spec = 75%
Rad5, Sens = 74%, Spec = 84%

A


Sensitivity
1-Specificity
0.0
0.2
0.4
0.6
0.8
1.0
DeepCOVID-XR AUC = 0.88 (0.84-0.92)
DeepCOVID-XR, Sens = 71%, Spec = 92%
Consensus AUC = 0.85 (0.80-0.89)
Consensus, Sens = 70%, Spec = 89%

B



C

Figure 6. Comparison with Expert Radiologist Interpretations. *A,* Comparison of the performance of DeepCOVID-XR to individual expert radiologist interpretations on random sample of 300 images from the test set. For DeepCOVID-XR, the ROC curve (red line) and decision threshold for overall interpretation of positive or negative for COVID-19 (red point) is plotted with a 95% confidence interval (red shaded area). For each radiologist, the overall interpretation sensitivity and specificity is plotted with 95% confidence intervals (dashed lines). Radiologist 1 (Rad1) = blue square, Radiologist 2 (Rad2) = grey down arrow, Radiologist 3 (Rad3) = green up arrow, Radiologist 4 (Rad4) = cyan diamond, Radiologist 5 (Rad5) = magenta "X". *B,* Comparison of DeepCOVID-XR to the consensus interpretation of all 5 radiologists. ROC curves (lines) and decision thresholds (points) for DeepCOVID-XR (red) and the consensus interpretation (purple) (AUC 0.88 vs. 0.85, p=0.13). *C,* ROC curve showing performance of DeepCOVID-XR (red line) using the consensus interpretation of all 5 radiologists as the radiologic reference standard rather than real-time polymerase chain reaction (RT-PCR).
Sens = Sensitivity, Spec = Specificity, ROC = receiver operating characteristic, AUC = area under the ROC curve.

Appendix E1

Supplemental Methods

Image Acquisition and Preprocessing

Chest Xray images were obtained from the Northwestern Memorial Health Care (NMHC) picture archiving and communications system (PACS) and vendor neutral archive (VNA) and downloaded in Digital Imaging and Communications (DICOM) format. Images were first converted to the Nifti file format to strip identifying DICOM metadata and burned in identifying annotations (e.g. date of imaging) were removed using thresholding (creating a mask at features above the 98th percentile of pixel intensity for a given bit-depth), followed by an inpainting procedure (35). This was designed to automatically remove protected health information from the image, but as a consequence all annotations (including position markers) that were above the threshold were removed. Notably, there is evidence that these annotations can lead to bias as deep learning algorithms can fit to these extraneous features based on differences in disease prevalence at imaging sites (36). Although we did not explicitly test whether removing these annotations led to a boost in algorithm performance, the grad-CAM saliency maps we produced to visualize feature importance for predictions confirmed that the algorithm ultimately focused on the lung fields to make its classifications (Figure 5).

Images were then preprocessed based on DICOM metadata tags indicating appropriate windowing and presentation state of pixel data. Images were encoded natively with bit depths ranging from 12-bits to 16-bits. In order to prepare images for input into our deep learning algorithm, images were downscaled to 8-bits and converted from grayscale to 3-channel RGB PNG files. Next, to focus on important features in the image while preserving features external to the lung fields that might be beneficial in the detection of COVID-19 (e.g. pleural effusions) and retrocardiac lung spaces (e.g. left lower lobe), we applied a cropping algorithm as follows. First, a UNet based algorithm (adapted from https://github.com/imlab-uiip/lung-segmentation-2d) that has previously been trained on images from two publicly available datasets (Montgomery (37) and JSRT (38) chest X-ray datasets) was used for semantic segmentation of the lung fields. The segmentation masks were then used to center a square cropping area around the lung fields. Finally, both cropped and uncropped images were resized from their native resolution (ranging from 2,500-3,000 X 2,500-3,000 pixels prior to cropping) to the appropriate resolution (224 X 224 pixels and 331 X 331 pixels) for input into individual neural network architectures using Lanczos (antialiasing) resampling.

DeepCOVID-XR: An Ensemble of Neural Networks for COVID-19 Prediction

The schematic of the DeepCOVID-XR weighted ensemble model is provided in Figure 2. We chose an ensemble architecture given the large body of evidence supporting ensemble methods as a strategy to resolve the bias/variance trade-off, reduce overfitting, and improve accuracy and generalizability (39–41). First, images were preprocessed as above to produce 4 images per input (each of 224X224 pixels and 331X331 pixels, cropped and uncropped), allowing individual models to focus on more general vs. more specific features and high-resolution vs low-resolution features that might be important for model predictions. This theory is based on evidence in the literature regarding the effect of differential image resolution (42) and lung segmentation/cropping to exclude irrelevant features (43–45) on performance of chest X-ray classification algorithms. Each of these images served as input into

6 separate previously validated convolutional neural network (CNN) architectures - DenseNet-121 (46), ResNet-50 (47), InceptionV3 (48), Inception-ResNetV2 (49), Xception (50), and EfficientNet-B2 (51) - for a total of 24 individually trained neural network architectures that served as members of the deep learning model ensemble. These 6 neural network architectures were chosen given extensive prior evidence of their utility for chest X-ray disease classification tasks(52–56). Each of these model architectures was modified to include a dropout layer and a single output node with a sigmoid activation function after the global average pooling layer. A Bayesian model combination approach was then used (with sampling from a Dirichlet distribution) (56,57) to derive optimal weights for individual members of the ensemble on the validation data set. The final prediction of the model ensemble was the weighted average of the predictions from individual member CNNs, yielding a final binary prediction – COVID-19 positive or COVID-19 negative – using a threshold of >0.5 for COVID-19 positivity.

Algorithm Training, Validation, and Testing

Individual neural network architectures were first pretrained on images from the publicly available NIH-14 chest X-ray dataset provided by the National Institutes of Health. Details regarding this dataset have been previously published, but briefly the dataset contains 112,120 frontal chest X-ray images that are labeled with 14 separate disease classifications (e.g. cardiomegaly, pleural effusion, pulmonary nodule). We pretrained the individual component neural network architectures of DeepCOVID-XR on these images. Prior to pre-training, weights were initialized using weights trained on the ImageNet dataset, thus images were normalized to the ImageNet mean and standard deviation. We then used a transfer learning strategy to fine tune individual neural network architectures by initializing training on our clinical data with the pre-trained weights from the NIH-14 chest X-ray dataset (58). The convolutional base of each network was initially frozen, and only the final dense output layer was trained on our clinical dataset. The entire model was then unfrozen, and the model was trained end-to-end on our clinical dataset. A stochastic gradient descent optimizer was used and hyperparameter tuning was performed using a Bayesian optimization strategy (59,60) to find the optimal learning rate, momentum, and dropout rate for each individual neural network architecture. Hyperparameters used in each individual component neural network architecture are included in Table E6. A batch size of 16 images was used for training. Data augmentation, including random width and height shifts, zooming, rotation, and brightness shifts, was used to prevent overfitting. Additionally, oversampling of the minority class (COVID-positive) was performed to account for class imbalance. A binary crossentropy loss function was used, and the model was trained to optimize the receiver operator characteristic area under the curve (AUC) on the validation dataset. The AUC metric was chosen rather than accuracy given the prevalence of COVID-19 is likely to change over time and from one institution to the next. As an overall measure of the intrinsic discriminative ability of a predictive algorithm, the AUC is relatively robust to changes in disease prevalence compared with other metrics of performance (e.g. accuracy, positive predictive value, negative predictive value). The learning rate was reduced by a factor of 0.1 after a plateau in improvement of 3 epochs and early stopping was used to prevent overfitting. Finally, test-time augmentation was implemented during performance evaluation using the same random transformations used for data augmentation during algorithm training, with the final prediction of each model being the average of augmented predictions (10 iterations per image).

Grad-CAM Heatmaps for Feature Importance of Predictions

Saliency mapping with gradient class activation maps (Grad-CAM) is a popular method for visualizing feature importance in arriving at a certain prediction in deep learning computer vision classifiers (61). We applied the Grad-CAM method to individual neural network architectures to visualize heatmaps of important features for predicting positivity for COVID-19. To generate heatmaps we used uncropped 224X224 resolution images and the final heatmap used for feature importance visualization was the result of averaging individual heatmaps for each of the 6 respective convolutional neural network architectures used in this study.

Hardware and Software Stack

The GPU workstation used for model training and evaluation was a CentOs7 server with 6 Nvidia Titan V GPUs, running CUDA version 10.0. Model training was performed on Google TensorFlow's Keras (v 2.0.0). Hyperparameter tuning was performed using Keras-Tuner (https://keras-team.github.io/keras-tuner/), model ensembling was performed using DeepStack (https://github.com/jcborges/DeepStack), and Grad-CAM heatmaps were produced using Keras-Vis (https://github.com/raghakot/keras-vis). All code was written in Python (version 3.7). Our code base is provided freely on GitHub at https://github.com/IVPLatNU/deepcovidxr, including weights for each of the 24 individually trained neural network architectures and respective model weights for the weighted ensemble model.

Supplemental References

35. Telea A. An Image Inpainting Technique Based on the Fast Marching Method. J Graph Tools. 2004;9(1):23–34.
36. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLoS Med. 2018;15(11):1–17.
37. Jaeger S, Candemir S, Antani S, Wáng Y-XJ, Lu P-X, Thoma G. Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. Quant Imaging Med Surg. 2014;4(6):475–477.
38. Shiraishi J, Katsuragawa S, Ikezoe J, et al. Development of a digital image database for chest radiographs with and without a lung nodule: Receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. Am J Roentgenol. 2000;174(1):71–74.
39. Dietterich TG. Ensemble methods in machine learning. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics). 2000;1857 LNCS:1–15.
40. Sagi O, Rokach L. Ensemble learning: A survey. Wiley Interdiscip Rev Data Min Knowl Discov. 2018;8(4):1–18.
41. Opitz D, Maclin R. Popular Ensemble Methods: An Empirical Study. J Artif Intell Res. 1999;11:169–198.
42. Sabottke CF, Spieler BM. The Effect of Image Resolution on Deep Learning in Radiography. Radiol Artif Intell. 2020;2(1):e190015.
43. Candemir S, Antani S. A review on lung boundary detection in chest X-rays. Int. J. Comput. Assist. Radiol. Surg. Springer Verlag; 2019. p. 563–576.
44. Gordienko Y, Gang P, Hui J, et al. Deep learning with lung segmentation and bone shadow exclusion techniques for chest X-ray analysis of lung cancer. Adv Intell Syst Comput. 2019;754:638–647.
45. Baltruschat IM, Steinmeister L, Ittrich H, et al. When does bone suppression and lung field segmentation improve chest x-ray disease classification? Proc - Int Symp Biomed Imaging. 2019;2019-April(October):1362–1366.
46. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. Proc - 30th IEEE Conf Comput Vis Pattern Recognition, CVPR 2017. 2017;2017-Janua:2261–2269.
47. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit. 2016;2016-Decem:770–778.
48. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the Inception Architecture for Computer Vision. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit. 2016;2016-Decem:2818–2826.
49. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. Inception-v4, inception-ResNet and the impact of residual connections on learning. 31st AAAI Conf Artif Intell AAAI 2017. 2017;4278–4284.
50. Chollet F. Xception: Deep learning with depthwise separable convolutions. Proc - 30th IEEE Conf Comput Vis Pattern Recognition, CVPR 2017. 2017;2017-Janua:1800–1807.
51. Tan M, Le Q V. EfficientNet: Rethinking model scaling for convolutional neural networks.

36th Int Conf Mach Learn ICML 2019. 2019;2019-June:10691–10700.

52. Pham HH, Le TT, Tran DQ, Ngo DT, Nguyen HQ. Interpreting chest X-rays via CNNs that exploit hierarchical disease dependencies and uncertainty labels. http://arxiv.org/abs/191106475. 2019;

53. Hashmi MF, Katiyar S, Keskar AG, Bokde ND, Geem ZW. Efficient pneumonia detection in chest xray images using deep transfer learning. Diagnostics. 2020;10(6):1–23.

54. Mitra A, Chakravarty A, Ghosh N, Sarkar T, Sethuraman R, Sheet D. A Systematic Search over Deep Convolutional Neural Network Architectures for Screening Chest Radiographs. Proc Annu Int Conf IEEE Eng Med Biol Soc EMBS. 2020;2020-July(1):1225–1228.

55. Rajpurkar P, Irvin J, Zhu K, et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. http://arxiv.org/abs/171105225. 2017;

56. Monteith K, Carroll JL, Seppi K, Martinez T. Turning Bayesian model averaging into Bayesian model combination. Proc Int Jt Conf Neural Networks. 2011;2657–2663.

57. Höge M, Guthke A, Nowak W. Bayesian model weighting: The many faces of model averaging. Water (Switzerland). 2020;12(2).

58. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. Proc - 30th IEEE Conf Comput Vis Pattern Recognition, CVPR 2017. 2017. p. 3462–3471.

59. Nazareth JL. An Optimization Primer. An Optim Prim. 2004;2–5.

60. Snoek J, Larochelle H, Adams RP. Practical Bayesian optimization of machine learning algorithms. Adv Neural Inf Process Syst. 2012;4:2951–2959.

61. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. Proc IEEE Int Conf Comput Vis. 2017;2017-Octob:618–626.

Table E1. Image Count by Institution

Institution	Chest X-ray Images	Hospital Beds
Northwestern Memorial Hospital	9,239	894
Lake Forest Hospital	2,214	198
Central DuPage Hospital	1,952	392
McHenry Hospital	1,209	179
Huntley Hospital	888	128
Delnor Hospital	886	159
Kishwaukee Hospital	178	98
Woodstock Hospital	109	56
Others	327	-
Total	17,002	-

Table E2. Individual Model Performance on Test Set

	Densenet-121				Resnet-50				InceptionV3			
	224 X 224		**331 X 331**		**224 X 224**		**331 X 331**		**224 X 224**		**331 X 331**	
	Uncropped	**Cropped**	**Uncropped**	**Cropped**	**Uncropped**	**Cropped**	**Uncropped**	**Cropped**	**Uncropped**	**Cropped**	**Uncropped**	**Cropped**
AUC No TTA*	0.86	0.86	0.86	0.87	0.85	0.85	0.80	0.88	0.85	0.86	0.83	0.84
AUC TTA*	0.88	0.86	0.87	0.87	0.86	0.87	0.83	0.88	0.87	0.88	0.84	0.84
TP (n)	895	765	997	926	847	678	729	837	857	1070	720	937
TN (n)	906	964	703	862	903	992	928	942	917	617	962	734
FP (n)	116	58	319	160	119	30	94	80	105	405	60	288
FN (n)	297	427	195	266	345	514	463	355	335	122	472	255
Accuracy	81%	78%	77%	81%	79%	75%	75%	80%	80%	76%	76%	75%
Sensitivity	75%	64%	84%	78%	71%	57%	61%	70%	72%	90%	60%	0.79
Specificity	89%	94%	69%	84%	88%	97%	91%	92%	90%	60%	94%	72%
F1Score	0.81	0.76	0.80	0.81	0.78	0.71	0.72	0.79	0.80	0.80	0.73	0.78

	Inception-ResnetV2				Xception				EfficientNet-B2			
	224 X 224		**331 X 331**		**224 X 224**		**331 X 331**		**224 X 224**		**331 X 331**	
	Uncropped	**Cropped**	**Uncropped**	**Cropped**	**Uncropped**	**Cropped**	**Uncropped**	**Cropped**	**Uncropped**	**Cropped**	**Uncropped**	**Cropped**
AUC No TTA*	0.85	0.85	0.85	0.85	0.86	0.85	0.87	0.88	0.85	0.85	0.86	0.87
AUC TTA	0.86	0.87	0.86	0.87	0.88	0.87	0.88	0.89	0.87	0.86	0.86	0.87
TP (n)	867	1022	798	815	884	875	898	882	921	930	887	896
TN (n)	889	648	924	927	900	907	891	923	837	815	883	868
FP (n)	133	374	98	95	122	115	131	99	185	207	139	154
FN (n)	325	170	394	377	308	317	294	310	271	262	305	296
Accuracy	79%	75%	78%	79%	81%	80%	81%	82%	79%	79%	80%	80%
Sensitivity	73%	86%	67%	68%	74%	73%	75%	74%	77%	78%	74%	75%
Specificity	87%	63%	90%	91%	88%	89%	87%	90%	82%	80%	86%	85%
F1Score	0.79	0.79	0.76	0.78	0.80	0.80	0.81	0.81	0.80	0.80	0.80	0.80

Note.—AUC = area under the receiver operator characteristic curve, TTA = test-time augmentation, TP = true positive, TN = true negative, FP = false positive, FN = false negative, PPV = positive predictive value, NPV = negative predictive value

* All of the other performance characteristics are calculated on the predictions produced using test-time augmentation.

Table E3. Comparison of Different Ensemble Models Performance on Test Set

	Weighted Average Ensemble Models								
	224 x 224 Uncropped	**331 x 331 Uncropped**	**224 x 224 Cropped**	**331 x 331 Cropped**	**224 x 224 Cropped/ Uncropped**	**331 x 331 Cropped/ Uncropped**	**Uncropped 224 x 224/ 331 x 331**	**Cropped 224 x 224/ 331 x 331**	**Final Ensemble**
AUC No TTA*	0.89	0.89	0.89	0.90	0.89	0.90	0.89	0.90	0.90
AUC TTA	0.89	0.89	0.89	0.90	0.90	0.90	0.90	0.90	0.90
TP (n)	900	913	921	914	902	905	893	879	898
TN (n)	930	928	914	930	946	940	947	956	948
FP (n)	92	94	108	92	76	82	75	66	74
FN (n)	292	279	271	278	290	287	299	313	294
Accuracy	83%	83%	83%	83%	83%	83%	83%	83%	83%
Sensitivity	76%	77%	77%	77%	76%	76%	75%	74%	75%
Specificity	91%	91%	89%	91%	93%	92%	93%	94%	93%
F1Score	0.82	0.83	0.83	0.83	0.83	0.83	0.83	0.82	0.83

Note.—AUC = area under the receiver operator characteristic curve, TTA = test-time augmentation, TP = true positive, TN = true negative, FP = false positive, FN = false negative, n = number

* All of the other performance characteristics are calculated on the predictions produced using test-time augmentation.

Table E4. Sensitivity Analysis for Test Set – Dropping 151 images from patients that also contributed images to training and validation sets

Test Set Performance	
AUC	0.90
TP (n)	859
TN (n)	855
FP (n)	68
FN (n)	281
Accuracy	83%
Sensitivity	75%
Specificity	93%
F1 Score	0.83

Note.—TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative, AUC = Area Under the Curve, n= number

Table E5. Comparison of DeepCOVID-XR to individual radiologist interpretations at each of the 5 decision thresholds determined by the six-point scoring system (from high confidence COVID-19 negative to high confidence COVID-19 positive)

Decision Thresholds	Radiologist 1				Radiologist 2				Radiologist 3				Radiologist 4				Radiologist 5			
	Specificity	Reader Sensitivity	Algorithm Sensitivity	p-value*	Specificity	Reader Sensitivity	Algorithm Sensitivity	p-value*	Specificity	Reader Sensitivity	Algorithm Sensitivity	p-value*	Specificity	Reader Sensitivity	Algorithm Sensitivity	p-value*	Specificity	Reader Sensitivity	Algorithm Sensitivity	p-value*
Threshold 1	55%	85% (79-91%)	90% (84-94%)	0.23	66%	80% (73-87%)	87% (81-93%)	0.04	68%	81% (75-87%)	87% (80-92%)	0.14	43%	90% (84-95%)	93% (87-98%)	0.32	60%	84% (78-90%)	88% (82-94%)	0.18
Threshold 2	81%	71% (63-78%)	81% (72-88%)	0.03	83%	73% (66-81%)	79% (70-87%)	0.26	85%	66% (58-74%)	77% (69-86%)	0.01	67%	79% (72-86%)	87% (81-92%)	0.04	78%	76% (69-83%)	83% (75-90%)	0.09
Threshold 3	91%	65% (57-73%)	73% (63-81%)	0.09	90%	69% (61-76%)	73% (64-81%)	0.29	89%	60% (52-69%)	74% (65-82%)	0.003	75%	76% (69-84%)	84% (77-90%)	0.047	84%	74% (66-81%)	78% (69-87%)	0.45
Threshold 4	98%	48% (40-56%)	54% (39-66%)	0.54	95%	60% (51-68%)	66% (51-77%)	0.44	95%	48% (40-56%)	66% (51-77%)	0.01	85%	68% (60-76%)	77% (69-85%)	0.08	90%	63% (54-71%)	73% (64-81%)	0.09
Threshold 5	99%	20% (14-27%)	7% (3-57%)	0.54	98%	44% (36-53%)	57% (43-71%)	0.2	98%	26% (19-34%)	57% (42-72%)	<0.001	96%	43% (34-51%)	58% (45-73%)	0.08	97%	40% (32-49%)	57% (43-71%)	0.07

*p-value for comparison of reader sensitivity to algorithm sensitivity (determined at reader specificity) for each of the 5 decision thresholds using 2,000 bootstrap samples

Table E6. Individual Model Hyperparameters for Training (Derived from Hyperparameter Tuner)

Model		Dropout Rate	Learning Rate	Momentum
Densenet-121	*224 X 224*	0.03	2.47E-03	0.85
	331 X 331	0.03	7.71E-03	0.63
Resnet-50	*224 X 224*	0.47	4.53E-03	0.54
	331 X 331	0.04	9.21E-04	0.52
InceptionV3	*224 X 224 Uncropped*	0.30	9.35E-03	0.52
	224 X 224 Cropped	0.50	1.00E-02	0.50
	331 X 331 Uncropped	0.39	7.76E-03	0.56
	331 X 331 Cropped	0.28	2.29E-03	0.93
Inception-ResnetV2	*224 X 224*	0.18	4.77E-04	0.73
	331 X 331	0.34	7.01E-04	0.94
Xception	*224 X 224*	0.20	5.79E-04	0.10
	331 X 331	0.30	1.00E-03	0.80
EfficientNet-B2	*224 X 224 Uncropped*	0.30	3.25E-03	0.70
	224 X 224 Cropped	0.40	6.4E-03	0.93
	331 X 331 Uncropped	0.30	5.60E-03	0.78
	331 X 331 Cropped	0.20	2.73E-03	0.89

Supplemental Figure

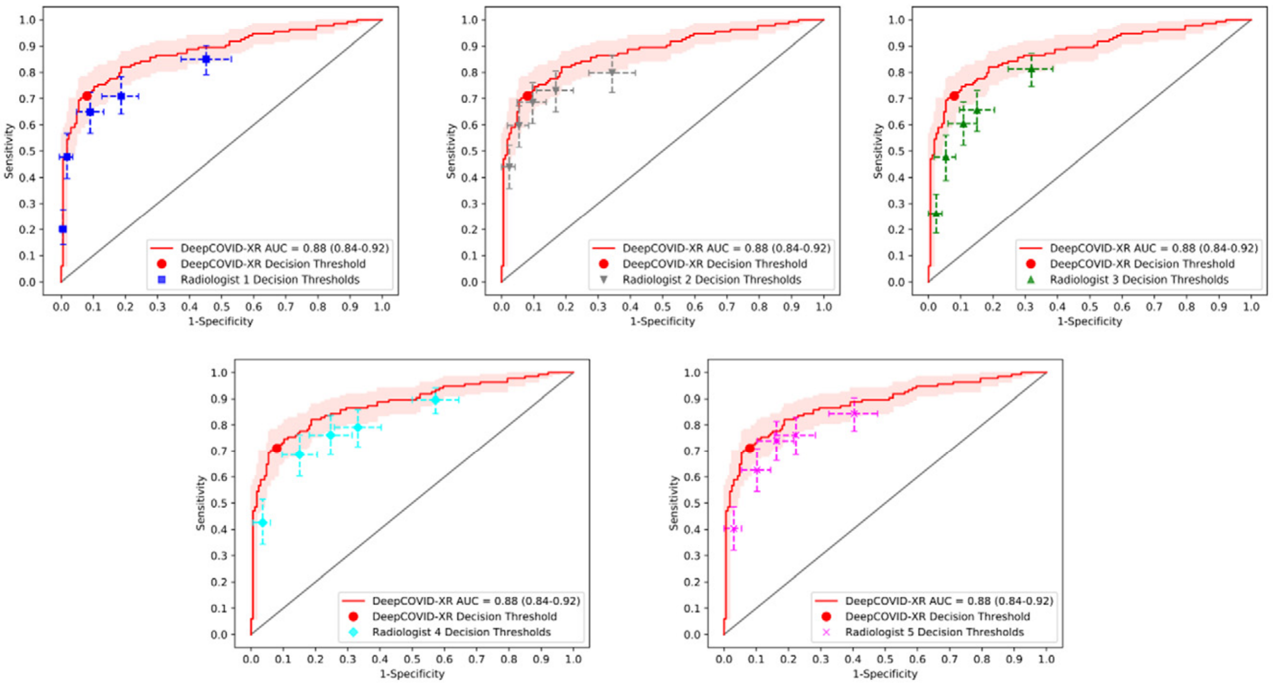


Figure E1. Comparison of DeepCOVID-XR to individual radiologist interpretations on 300 random test images at 5 separate decision thresholds. Each panel in the figure corresponds to an individual radiologist. For each radiologist, sensitivity and specificity are plotted with 95% confidence intervals (dashed lines) at each of the 5 operating points/decision thresholds within the 6-point scoring scale (based on interpretation confidence). For each panel, the ROC curve (red line), overall decision threshold (red point), and 95% confidence interval is plotted for DeepCOVID-XR. ROC = receiver operating characteristic, AUC = area under the ROC curve.