Evaluation of multiple open-source deep learning models for detecting and grading COVID-19 on chest radiographs

Alexander Risman[a,*] **Miguel Trelles**[b] **and David W. Denning**[c]
[a]Realize, Chicago, Illinois, United States
[b]Clinica Delgado, Radiology Department, Lima, Peru
[c]The University of Manchester, Manchester Academic Health Science Centre, Manchester Fungal Infection Group, Manchester, United Kingdom

Abstract

Purpose: Chest x-rays are complex to report accurately. Viral pneumonia is often subtle in its radiological appearance. In the context of the COVID-19 pandemic, rapid triage of cases and exclusion of other pathologies with artificial intelligence (AI) can assist over-stretched radiology departments. We aim to validate three open-source AI models on an external test set.

Approach: We tested three open-source deep learning models, COVID-Net, COVIDNet-S-GEO, and CheXNet for their ability to detect COVID-19 pneumonia and to determine its severity using 129 chest x-rays from two different vendors Phillips and Agfa.

Results: All three models detected COVID-19 pneumonia (AUCs from 0.666 to 0.778). Only the COVID Net-S-GEO and CheXNet models performed well on severity scoring (Pearson's r 0.927 and 0.833, respectively); COVID-Net only performed well at either task on images taken with a Philips machine (AUC 0.735) and not an Agfa machine (AUC 0.598).

Conclusions: Chest x-ray triage using existing machine learning models for COVID-19 pneumonia can be successfully implemented using open-source AI models. Evaluation of the model using local x-ray machines and protocols is highly recommended before implementation to avoid vendor or protocol dependent bias.

 [DOI: 10.1117/1.JMI.8.6.064502]

Keywords: artificial intelligence; COVID-19; x-ray.

Paper 21016SSR received Feb. 2, 2021; accepted for publication Dec. 2, 2021; published online Dec. 21, 2021.

1 Introduction

The early radiological features of viral pneumonia can be subtle. The high transmissibility rate of SARS-CoV-2 and the wide range of disease manifestations from asymptomatic to life-threatening COVID-19 have challenged health systems throughout the world. Early CT scanning of the lungs was found to be more sensitive than molecular detection of viral RNA with polymerase chain reaction (PCR) testing,[1] but is practically unrealistic in many healthcare systems due to a lack of CT scanners, whereas chest radiography is available.[2,3] The chest radiograph is an almost immediate test, and if it were sensitive and specific enough, it could act as an immediate screen for COVID-19, as for influenza,[4] although with limited capacity to detect asymptomatic SARS-CoV-2 infection.[5]

Multiple different viruses can cause pneumonia or pneumonitis including influenza A, influenza B, swine influenza, other coronaviruses, measles, adenovirus, human metapneumovirus, and in both children and immunocompromised patients, respiratory syncytial virus, parainfluenza virus, and adenovirus are often responsible.[6–8] Such patients are usually treated unnecessarily with antibacterial agents until the diagnosis is clarified with viral PCR or culture. But the infection control aspects of hospitalization demand a rapid means of identifying such

*Address all correspondence to Alexander Risman, alex@realize.ai

patients to allow cohorting, so well illustrated by the ongoing pandemic of SARS-CoV-2 infection.

Many hospitals and health systems have a shortage of trained radiologists; in some countries, there are fewer than one radiologist per million people, severely impacting the timely delivery of radiological findings to patients and referring clinicians.[3,9]

Intelligently triaging or prioritizing a radiology work-list could help address this issue. A work-list sorted from most to least likely to have clinically significant findings, such as COVID-19, would mean that sick patients could potentially receive quick x-ray results even in resource-constrained countries, and x-rays found to be normal with high confidence by artificial intelligence (AI) could potentially be removed from worklists altogether; such prioritization could also help ensure that radiologists read the most critical studies when they are "fresh" and read progressively less critical studies as they get tired. Triage is a previously validated use case for AI in medical imaging.[10] An AI model effective at determining the likelihood of a COVID-19 diagnosis based on an x-ray could be helpful at determining which patients to admit to hospital; an AI model for COVID-19 severity could be helpful at determining which patients to admit to the ICU.

Several AI vendors have released COVID-19 detection products and are currently offering them for free, and at least one, Delft imaging, has published a peer-reviewed evaluation of their product including validation on an external dataset.[11] However, these products are closed source, which vastly limits users' ability to understand how these products work, customize them for their own needs, and ensure they fully meet security standards.[12] Use of free closed-source software also presents a risk that the vendor may demand payment or impose other unwelcome conditions for continued use in the future.

Several open-source software projects exist that could potentially help triage COVID-19 cases. An open-source machine learning model, COVID-Net, has been developed specifically to detect COVID-19, and the COVID-Net group has released additional models for estimating disease severity in COVID-positive cases, including COVIDNet-S-GEO for estimating geographic severity.[13,14] Additionally, in 2018, long before the pandemic began, an open-source machine learning model called CheXNet was developed by Stanford University, which they claimed could detect 14 pathologies, including pneumonia, in chest x-rays.[15] Validating on their internal test set, they found the CheXNet pneumonia model to be comparable to a number of radiologists. However, none of these models have been validated on an external test set, meaning there is currently no evidence that they generalize beyond the development context.

The present research seeks to externally validate these models. The aims of this study are to test whether these models are significantly better than chance at determining whether an x-ray image contains evidence of COVID-19 and how severe it is if present.

2 Materials and Methods

We selected COVID-Net and CheXNet for evaluation because they were the only open-source deep learning models that could be used to detect COVID-19 at the time we began this study. We believe these two projects are still worthy of our focus due to their high citation count relative to other open-source projects that have emerged since, and keeping the number of models small enables us to limit the number of hypotheses to test and reduce the possibility of false-positive results. CheXNet was trained on 112,120 images from 30,805 patients, drawn from a single institution and automatically labeled for 14 different pathologies by applying natural language processing (NLP) to their corresponding radiology reports. COVID-Net was trained on 13,975 images from 13,870 patients, drawn from five distinct sources; labeling methods varied by source and ranged from PCR to NLP. The data used to train both of these systems are publicly available but not in DICOM format, meaning that most metadata, such as imaging equipment manufacturer, is not available.

We retrospectively collected a dataset of x-ray images from a hospital in Lima, Peru between 02/23/2020 and 04/20/2020 that had reverse transcription PCR (RT-PCR) testing. The patient population included all attentions at the "fever clinic" for symptomatic patients and all hospitalized patients that had an RT-PCR test done by protocol. Chest x-rays were performed on two

different units from DX-D600 (Agfa Medical Imaging) and DigitalDiagnost (Philips). The x-rays were all subsequently deidentified and uploaded to the cloud using a Cimar Cloud Gateway. As a retrospective study using deidentified data, this study was exempt from Institutional Review Board review under institutional policy.

All of these images were in their original DICOM format, the standard clinical format for medical images; however, while both the COVID-Net and CheXNet projects have published specific instructions and/or code for running their models on images in JPEG or PNG format, neither has published instructions for running their models on DICOM or converting DICOM images to JPEG or PNG. This limitation appears to come from the fact that these models were both trained on public datasets whose creators only released JPEG or PNG formatted images, likely for privacy and security reasons. DICOM files often contain important metadata for interpreting their stored pixel values, such as value of interest tables, and the use or non-use of this metadata can significantly impact how images are finally rendered. For the purposes of this study, we use pixel data extraction functionality from the open-source Python DICOM library pydicom (version 2.0.0) to convert DICOM images to PNG;[16] the only modification to the raw pixel data we made (based on DICOM metadata when using pydicom) was to invert pixel values in images with a value of MONOCHROME1 for the DICOM field photometric interpretation, which indicates a "photonegative" image.

After conversion from DICOM to PNG, images were preprocessed according to the instructions and code published online by the COVID-Net and CheXNet projects prior to ingestion by those models.[17,18] For COVID-Net, this preprocessing involved resizing the image to 480×480 and removing the top 8% of the image; for CheXNet, this involved resizing the image to 224×224 and normalizing the image using mean and standard deviation values from the ImageNet database. We used pretrained model binary files provided by these groups online for our tests.[19–21] The code we used to preprocess the images and run the models is available at https://github.com/alexrisman/RealizeCovidStudy.

In addition to PCR results, we had two general radiologists, with 31 and 25 years of experience, respectively, review the 129 images retrospectively. For each x-ray, each reader first judged whether the image looked normal, positive for non-COVID pneumonia, positive for COVID-19 pneumonia, or positive for some other abnormality. The cases positive for COVID-19 pneumonia were considered as positives for the purpose of radiologist's sensitivity and specificity analysis. Next, if they had judged the image to be positive for COVID or non-COVID pneumonia, they scored the left and right lung for the geographic severity of the observed opacity on a scale of 0 to 4, with 0 corresponding to no opacity, 1% to <25% involvement, 2% to 25%–50%, 3% to 50%–75%, and 4% to >75%; these scores were then added across both lungs for a final 0 to 8 geographic severity score. The averaged result across both radiologists was then used to test each model's severity scoring. This scoring methodology was adapted from the COVID-Net team.[14] Within the 22 PCR-confirmed COVID x-rays, we used Pearson correlation (r) to test interrater variability of this severity score between the two radiologists. Agreement was high, with a correlation of .937 ($p < 0.001$). There were 13 cases that both radiologists agreed were positive for COVID-19; all 13 of these cases were confirmed positive by PCR.

Finally, a subset analysis was done analyzing AUC results for each x-ray equipment used.

We use the ROC AUC score to assess classification performance and use bootstrapping to construct 95% confidence intervals for AUC scores. We formulate the null hypothesis for each AI system as follows: the value 0.5, denoting performance no better than chance, is within the AUC score 95% confidence interval. We use Pearson's correlation (r) to test each model's severity scoring. We use DeLong's test[22] to statistically compare AUCs from different systems and Steiger's test[23] to compare correlation coefficients. PCR results and radiologist readings were used as ground truth for the classification and severity analysis, respectively.

3 Results

A total of 131 chest x-ray images from 131 patients were initially collected. Two patients had multiple conflicting PCR results within several days of each other so were excluded. Of the remaining 129, 22 (17%) were confirmed to be COVID-positive by RT-PCR–based testing.

Table 1 shows the dataset characteristics. In PCR-confirmed cases with a radiological abnormality (as reported by at least one radiologist, $N = 16$) had higher severity scores (mean = 3.91, std = 2.37) than unconfirmed cases with a radiological abnormality (mean = 2.13, std = 1.66).

The AUC for detection of RT-PCR positive cases was 0.78 (0.67 to 0.87) for CheXNet, 0.67(0.52 to 0.80) for COVID-Net, and 0.75 (0.61 to 0.86) for COVID-Net-S-GEO. Sensitivity and specificity for the dataset were 0.59 and 0.90 for radiologist 1 and 0.36 and 0.93 for radiologist 2, and the low-sensitivity reflects the fact that many patients with SARS-CoV-2 infection do not have COVID-19 pneumonia. CheXNet's performance was on par with radiologist 2. Figure 1 shows the values and ROC curves. For all three models, we can reject the null hypothesis that the AUC is 0.5 with >95% confidence, which suggests that any of the open-source COVID-detection models in question would have served this clinic to at least some effect for the purpose of intelligently prioritizing an x-ray work-list based on likelihood of COVID-19.

Table 1 Summary of dataset characteristics. With respect to age, parenthetical values represent the standard deviation; with respect to all other characteristics, the parenthetical values represent proportion of the dataset.

Variable	Level	Overall
N		129
Age		44.32 (19.37)
Sex	F	73 (56.59)
	M	56 (43.41)
X-ray manufacturer	Agfa	83 (64.34)
	Philips	46 (35.66)
PCR-positive	Negative	107 (82.95)
	Positive	22 (17.05)

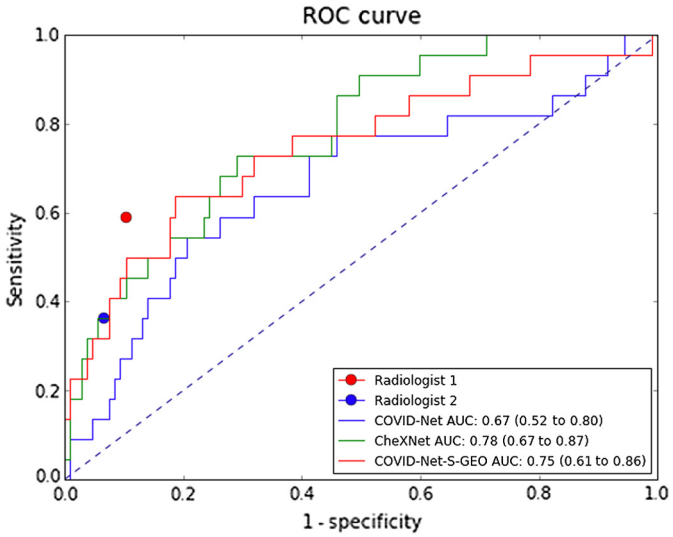

Fig. 1 AUROC curves and point estimates, with bootstrapped 95% confidence intervals, evaluating three open-source machine learning models on the classification task of distinguishing PCR-confirmed COVID-positive x-rays from COVID-negative ones. The performance of two radiologists is included for comparison.

With respect to pneumonia severity, two of the three models outperform chance with respect to predicting the mean of the radiologists' assigned geographic severity scores for PCR-confirmed positive x-rays (Table 2). For the best performing model for severity, COVID-Net-S-Geo had a Pearson's r of 0.93 and is graphically shown in Fig. 2.

In order to compare the AUCs of COVID-Net and CheXNet on the PCR-positive case detection task, we performed DeLong's test, which resulted in a statistically insignificant p-value of 0.17, likely due to the small sample size. However, when comparing the correlation coefficients calculated for those two models on the severity scoring task using Steiger's test, we observed a p-value of 0.0005, indicating a statistically significant difference.

Although two of the three models outperformed chance on both tasks, one, COVID-Net, did not. In addition to failing the severity rating task, COVID-Net was also the worst performing model on the likelihood detection task. Breaking down results by manufacturer, however, we observe that COVID-Net's poor performance is largely confined to images coming from Philips machines, performing well on x-rays from Agfa machines (Table 3). The other two models do not show a significant disparity in performance related to x-ray machine manufacturer.

Table 2 Pearson's r and 95% confidence intervals, evaluating three open-source machine learning models on the regression task of rating the geographic severity of COVID-19 in PCR-confirmed positive chest x-rays, with radiologist-assigned severity scores as ground truth.

CheXNet	COVID-Net-S-GEO	COVID-Net
0.83 (0.64 to 0.93)	0.93 (0.83 to 0.97)	−0.17 (−0.55 to 0.27)

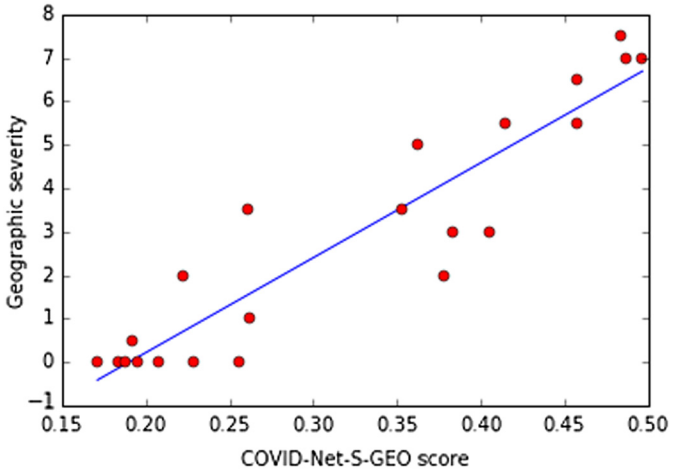

Fig. 2 Scatter plot and best fit line for the machine learning model that best predicted geographic severity scores for COVID-positive x-rays ($r = 0.93$, $p < 0.001$).

Table 3 AUCs and bootstrapped 95% CIs for each model in the likelihood detection task by x-ray machine manufacturer. COVID-Net performed poorly, but it appears that this is largely due to poor performance on x-rays from Philips' machines; the other models do not show a critical disparity in performance across x-ray machine manufacturer, while COVID-Net fails to outperform chance on Philips x-rays despite outperforming chance on Agfa ones.

COVID-Net		CheXNet		COVID-Net-S-GEO	
Agfa AUC	Philips AUC	Agfa AUC	Philips AUC	Agfa AUC	Philips AUC
0.74 (0.54 to 0.90)	0.60 (0.40 to 0.79)	0.74 (0.60 to 0.87)	0.81 (0.65 to 0.94)	0.73 (0.51 to 0.92)	0.72 (0.52 to 0.90)

COVID-Net also greatly underperformed their own published results, confirming the importance of external validation. According to the COVID-Net team, COVIDNet-CXR4-A has an accuracy of 94.3% at a sensitivity of 95%, while we found it to have an accuracy of 66.7% at a sensitivity of 59.1%. CheXNet's published performance was in line with our results: they claimed CheXNet achieved an AUC of 0.77 in the detection of pneumonia, while we found an AUC of 0.78. COVID-Net's underperformance could potentially be explained by the same reasons as their poor performance on Philips machines (e.g., if all the COVID-positive cases in their dataset were from Philips machines, and the model learned that Philips = COVID, the model would perform terrifically in their internal validations but crash in the real world.)

4 Discussion

Our results indicate that several open-source AI systems can be used both to detect COVID-19 in chest x-rays and to grade its severity with more success than random chance on real-world clinical data, suggesting that these systems can likely be useful for triage. Further, it appears that models trained for likelihood detection can be potentially be used for severity grading and vice versa. Since many models in the marketplace and open-source ecosystem are trained and tested for only one of those two purposes (i.e., detection of COVID-19 pneumonia), these findings suggest that a clinician with access to such a single-purpose model may be able to apply it to the other purpose (i.e., severity scoring) or vice versa, though evaluation for this purpose on their own data is recommended. This is particularly promising with respect to CheXNet, since we were able to confirm with this study that its pneumonia detection function is also useful for pneumonia severity grading, and this could possibly also be true for the 13 other pathologies CheXNet is trained to detect.

The apparent effect of x-ray machine manufacturer on model performance is troubling and suggests that doctors should make every effort to evaluate medical imaging AI models on samples of their own data that are representative of the manufacturers and acquisition protocols used within their clinical environment prior to use, as well as to re-evaluate these models when new imaging equipment or protocols are introduced to the environment. While we were not entirely sure why COVID-Net's performance was so much worse on Philips machines, differences in acquisition protocols may provide some clues. While we did not observe a major difference in exposure settings such as radiation dose between the Philips and Agfa machines, we did observe that a contrast-enhancing antiscatter grid was used for the acquisition of the bulk of the Agfa images but not for most of the Philips images. This suggests that COVID-Net may need an especially high degree of contrast to be effective. In addition, the COVIDx dataset used to train COVID-Net came from a variety of publicly available sources. If all or most of the COVID-positive x-rays in the COVIDx dataset came from Philips machines or machines without an antiscatter grid, that might lead the model to incorrectly learn that artifacts of this acquisition protocol are highly indicative of COVID, particularly if there were no normal or other non-COVID x-rays from the same sources as the COVID x-rays.

This study is subject to several limitations. Our initial data collection resulted in a relatively small sample size, particularly for the severity and manufacturer analyses. Though we are able to draw statistically significant conclusions, we acknowledge that the size of our dataset may limit how broadly applicable these findings are and suggest thinking of this work more as a successful clinic-level case study and a blueprint for other clinics to test these and other AI systems on their own data, rather than a universal test of these systems' effectiveness in every possible context. Also, while the PCR test was used as a gold standard for the purposes of this paper, it has a limited sensitivity in the range of 60% to 70%, which brings an important source of error to our analysis.[24] While a key use case for a severity scoring model would be determining whether a given patient's condition is improving or declining, we were only able to collect a single x-ray for each patient in our study, preventing us from directly testing any model's suitability for this application. This was also a retrospective study, but a prospective trial would be required to test any model's effectiveness for triage directly, for example by evaluating exam turnaround with and without use of the model. Finally, while our dataset does contain seven cases in which both of our readers reported diagnoses of non-COVID pneumonia, we judged this number too small

for meaningful statistical analysis. We are therefore unsure how well these systems differentiate between COVID and non-COVID pneumonia and therefore do not recommend using these systems to render final diagnoses in lieu of human interpretation. These limitations may be addressed in a future study. Radiological appearances cannot determine infectious aetiology, so additional work is required to determine the transferability of the findings to other viral pneumonias and indeed other lung pathologies. A significant influenza outbreak might offer this opportunity, once the current pandemic has receded.

5 Conclusions

Chest x-ray triage can be successfully implemented using existing open-source machine learning models for COVID-19 pneumonia. The local x-ray machines and protocols can affect interpretation with machine learning tools and this challenges widespread implementation to avoid vendor or protocol dependent bias. Viral pneumonia, exemplified by COVID-19 pneumonia, is amenable to AI, which should be of general benefit.

Disclosures

The authors have no conflicts of interest to report.

Acknowledgements

Research reported in this publication was supported by Realize, Inc. and IntriHEALTH Ltd. The authors would like to thank Jeffrey Swartzberg and Susan Otto for providing the radiological interpretations of the x-rays, Cimar UK Ltd. for help with image anonymization and exchange, and Eric Hart for providing helpful feedback on the initial study design. D. W. D is partly supported by the NIHR Manchester Biomedical Research Centre. Author contributions: Guarantor of integrity of entire study, A. R.; study concepts/study design, all authors; data acquisition, M. T. and A. R.; data analysis/interpretation, all authors; statistical analysis, A. R.; literature research, all authors; manuscript drafting, A. R. and D. W. D.; manuscript revision for important intellectual content, all authors; and approval of final version of submitted manuscript, all authors.

References

1. Y. Fang et al., "Sensitivity of chest CT for COVID-19: comparison to RT-PCR," *Radiology* **296**(2), E115–E117 (2020).
2. B. M. Idowu and T. A. Okedere, "Diagnostic radiology in Nigeria: a country report," *J. Glob. Radiol.* **6**(1), 1072 (2020).
3. F. Ali et al., "Diagnostic radiology in Liberia: a country report," *J. Glob. Radiol.* **1**(2), 1020 (2015).
4. G. Aviram et al., "H1N1 influenza: initial chest radiographic findings in helping predict patient outcome," *Radiology* **255**(1), 252–259 (2010).
5. M. Bandirali et al., "Chest radiograph findings in asymptomatic and minimally symptomatic quarantined patients in Codogno, Italy during COVID-19 pandemic," *Radiology* **295**(3), (2020).
6. B. A. Cunha, U. Syed, and S. Strollo, "Swine influenza (H1N1) pneumonia in hospitalized adults: chest film findings," *Hear. Lung J. Acute Crit. Care* **40**(3), 253–256 (2011).
7. J. L. Mayer et al., "CT-morphological characterization of respiratory syncytial virus (RSV) pneumonia in immune-compromised adults," *RoFo* **186**(7), 686–692 (2014).
8. Y. Karimata et al., "Clinical features of human metapneumovirus pneumonia in non-immunocompromised patients: an investigation of three long-term care facility outbreaks," *J. Infect. Dis.* **218**(6), 868–875 (2018).
9. O. Bwanga, J. Mulenga, and E. Chanda, "Need for image reporting by radiographers in Zambia," *Med. J. Zambia* **46**(3), 215–220 (2019).

10. M. Annarumma et al., "Automated triaging of adult chest radiographs with deep artificial neural networks," *Radiology* **291**(1), 196–202 (2019).
11. K. Murphy et al., "COVID-19 on chest radiographs: a multireader evaluation of an artificial intelligence system," *Radiology* **296**(3), E166–E172 (2020).
12. S. Raghunathan et al., "Open source versus closed source: software quality in monopoly and competitive markets," *Syst. Hum.* **35**(6), 903–918 (2005).
13. L. Wang, Z. Q. Lin, and A. Wong, "COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest x-ray images," *Sci. Rep.* **10**(1), 19549 (2020).
14. A. Wong et al., "OVID-Net S: Towards computer-aided severity assessment via training and validation of deep neural networks for geographic extent and opacity extent scoring of chest x-rays for SARS-CoV-2 lung disease severity," https://ui.adsabs.harvard.edu/abs/2020arXiv200512855W (2020).
15. P. Rajpurkar et al., "CheXNet: radiologist-level pneumonia detection on chest x-rays with deep learning," https://ui.adsabs.harvard.edu/abs/2017arXiv171105225R (2017).
16. D. Mason, "SU-E-T-33: pydicom: an open source DICOM library," *Med. Phys.* **38**(6, Part10), 3493–3493 (2011).
17. L. Wang, "lindawangg/COVID-Net," COVID-Net open source initiative, https://github.com/lindawangg/COVID-Net (accessed 23 December 2020).
18. B. Chou, "brucechou1983/CheXNet-Keras," tool to build CheXNet-like models, https://github.com/brucechou1983/CheXNet-Keras (accessed 23 December 2020).
19. B. Chou, "brucechou1983_CheXNet_Keras_0.3.0_weights.h5," https://drive.google.com/file/d/19BllaOvs2x5PLV_vlWMy4i8LapLb2j6b/view (accessed 23 July 2020).
20. L. Wang, "COVIDNet-S-GEO," https://drive.google.com/drive/folders/1o3hOPrXAza73kWbb44iRrcSZAsAtlt7D (accessed 9 October 2020).
21. L. Wang, "COVIDNet-CXR4-A," https://drive.google.com/drive/folders/1YWaF_4ezgVZ7khB6OJD8ZqIKj7cnC02L (accessed 4 July 2020).
22. 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).
23. J. H. Steiger, "Tests for comparing elements of a correlation matrix," *Psychol. Bull.* **87**(2), 245–251 (1980).
24. H. X. Bai et al., "Performance of radiologists in differentiating COVID-19 from non-COVID-19 viral pneumonia at Chest CT," *Radiology* **296**(2), E46–E54 (2020).

Biographies of the authors are not available.