Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company’s public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
ARTICLE INFO

Keywords:
COVID-19
X-rays
Classification
Deep learning
Neural networks
Diagnosis

ABSTRACT

Introduction: The objective of this study was to assess seven configurations of six convolutional deep neural network architectures for classification of chest X-rays (CXRs) as COVID-19 positive or negative.

Methods: The primary dataset consisted of 294 COVID-19 positive and 294 COVID-19 negative CXRs, the latter comprising roughly equally many pneumonia, emphysema, fibrosis, and healthy images. We used six common convolutional neural network architectures, VGG16, DenseNet121, DenseNet201, MobileNet, NasNetMobile and InceptionV3. We studied six models (one for each architecture) which were pre-trained on a vast repository of generic (non-CXR) images, as well as a seventh DenseNet121 model, which was pre-trained on a repository of CXR images. For each model, we replaced the output layers with custom fully connected layers for the task of binary classification of images as COVID-19 positive or negative. Performance metrics were calculated on a hold-out test set with CXRs from patients who were not included in the training/validation set.

Results: When pre-trained on generic images, the VGG16, DenseNet121, DenseNet201, MobileNet, NasNetMobile, and InceptionV3 architectures respectively produced hold-out test set areas under the receiver operating characteristic (AUROCs) of 0.98, 0.95, 0.97, 0.95, 0.99, and 0.96 for the COVID-19 classification of CXRs. The X-ray pre-trained DenseNet121 model, in comparison, had a test set AUROC of 0.87.

Discussion: Common convolutional neural network architectures with parameters pre-trained on generic images yield high-performance and well-calibrated COVID-19 CXR classification.

1. Introduction

In December 2019, a cluster of pneumonia with unknown etiology emerged, rapidly evolving into a world-wide health crisis with significant social, health, and financial consequences. Given the rapid spread of infection, the continued concern that asymptomatic carriers are contributing to community transmission, the depletion of hospital resources due to high influxes of patients, and the current absence of specific therapeutic drugs and widely available vaccines for treatment of COVID-19 infection, it is essential to detect onset at its early stages.

Radiological examinations play an important role in the diagnosis and evaluation of this global health emergency. Common radiological findings of the infection include multiple ground glass opacity and interlobular septal thickening in the lungs, with significant correlations between the degree of pulmonary inflammation and main COVID-19 clinical symptoms. Although reverse-transcription polymerase chain reaction (RT-PCR) remains the standard to diagnose COVID-19 infection, issues with limited supply of RT-PCR assays have hindered prompt diagnosis. Complementary to RT-PCR assays, chest radiography can identify early phase lung infection and prompt larger surveillance efforts.

In particular, there has been a flurry of work concerning the use of chest X-rays (CXR) to detect COVID-19. All of these studies use models based on a number of convolutional neural network (CNN) architectures—in most instances, performance comparisons are limited to models derived from only one or a few architectures. Other studies have limitations resulting from the size or composition of the training or testing datasets, such as the absence of examples of differential diagnoses, or overwhelming class imbalance. A third drawback of existing work is a widespread...
lack of calibration. (Several relevant studies are compared in Supplementary Table 1.) In contrast to these studies, we present a more comprehensive comparison of the performance and calibration of seven models resulting from six different CNN architectures on a balanced dataset which includes multiple differential diagnoses (pneumonia, fibrosis, and emphysema). While the present study concerns COVID-19 detection, we mention that other studies use related methods to detect and predict the severity of pneumonia among patients already known to be COVID-19 positive. Additionally, a great deal of effort has been devoted to the use of computed tomography (CT) for COVID-19 detection. In the current study, we chose to use CXR images, as they are less expensive and more common.

2. Materials and methods

2.1. Description of data

Studies performed on de-identified patient data constitute non-human subjects research, and this study has been determined by the Pearl Institutional Review Board to be Exempt according to FDA 21 CFR 56.104 and 45CFR46.104(b) (4): (4) Secondary Research Uses of Data or Specimens under study number 20-DASC-119. We used two datasets. The first, which we will simply refer to as the COVID-19 dataset, is a publicly accessible repository of chest radiographs from COVID-19 patients compiled by Cohen et al. (2020). The second dataset, named ChestX-ray14, is a much larger collection of 14 disease image labels including atelectasis, cardiomegaly, effusion, infiltration, mass, nodule, pneumonia, pneumothorax, consolidation, edema, emphysema, fibrosis, pleural thickening and hermia. We note that, while we reference the ImageNet dataset, we used the parameters for the six CNN architectures which were already derived from training on ImageNet.

2.2. Data preprocessing and labeling

Both datasets contained data obtained from single posteroanterior (or “front-on”) X-rays as well as from CT scans composed of multiple centered X-rays. We chose to exclusively use single CXR images, as they are less expensive and more common than CT systems. Due to the relative scarcity of COVID-19 positive images, we used all available images (294 images), despite the potential bias introduced by using multiple images from the same individual. For convenience, we selected from ChestX-ray14 as many (294 images) COVID-19 negative images as COVID-19 positive images from the COVID-19 dataset. We chose approximately equally many pneumonia, emphysema, fibrosis, and healthy images. We selected these conditions on the basis of shortness of breath and cough, which overlap with primary symptoms of COVID-19 and may therefore motivate a clinician to order a chest radiograph to determine COVID-19 status in those patients. We tabulated the two demographic pieces of information, sex and age, which were available for most of the 588 images and were used as inputs to the algorithm (Table 1).

Table 1: COVID-19 and ChestX-ray14 (NIH) dataset patient demographics

| Characteristic | Number of X-ray images | Percentage |
|----------------|------------------------|------------|
| **COVID-19 dataset (COVID-19 patients)** | | |
| Sex | | |
| Female | 98 | 33.33% |
| Male | 165 | 56.12% |
| Age | | |
| <29 | 13 | 4.42% |
| 30-39 | 29 | 9.86% |
| 40-49 | 40 | 13.61% |
| 50-59 | 39 | 13.27% |
| 60-69 | 56 | 19.05% |
| 70-79 | 67 | 22.79% |
| 80+ | 18 | 6.12% |
| **NII dataset (non-COVID-19 patients)** | | |
| Sex | | |
| Female | 136 | 46.26% |
| Male | 158 | 53.74% |
| Age | | |
| <29 | 23 | 7.82% |
| 30-39 | 31 | 10.54% |
| 40-49 | 35 | 11.90% |
| 50-59 | 65 | 22.11% |
| 60-69 | 64 | 21.77% |
| 70-79 | 62 | 21.09% |
| 80+ | 14 | 4.76% |

The size of all images was standardized to 224 × 224. We selected 220 COVID-19 positive images from 143 unique patients to the training set as well as 220 of the COVID-19 negative images from Chestx-ray14 from 153 unique patients. The final training set consisted of 220 COVID-19 positive, 55 healthy, 55 pneumonia, 55 emphysema, and 55 cystic fibrosis images for a total of 440 chest radiograph images (roughly 75% of all images) on which we trained and validated the classifiers. The remaining 148 images, 74 COVID-19 positive and 74 COVID-19 negative, were allocated to a hold-out test set. The images for the hold-out test set were selected such that no images from patients in the hold-out test set were seen by the models during training. The prevalence of COVID-19 positive images in the training and testing sets were similar to that reported in another imaging analysis study for COVID-19 screening.

2.3. Model development

Our machine learning models were built from standard building blocks to create seven models that we found to be effective for this task of classification. We selected six of the most common configurations of CNN architectures: VGG16 (named for the Visual Geometry Group at the University of Oxford), DenseNet121 and DenseNet201, InceptionV3, MobileNet and NasNetMobile. We chose these six architectures due to their popularity and the accessibility of ImageNet pre-trained parameters (available in the Python Keras library), as well as their contrasting depth and number of parameters. Among the six architectures, VGG16 has the highest number of parameters while MobileNet has the lowest. In terms of topological depth (including activation layers, normalization layers and so on) DenseNet201 is the deepest and VGG16 is the shallower. We adapted each of the six CNN architectures that were pre-trained using the ImageNet dataset by removing the classifier blocks and adding a custom classifier block to each model. The final layer in the custom classifier blocks were designed with a softmax output for the two categorical outputs: covid and non-covid. We call these six models “off the shelf” (OTS) to indicate that they were pre-trained using ImageNet data.

For the seventh model, we prepared a second configuration of the DenseNet121 network. This model was pre-trained on the full ChestX-
ray14 dataset for the multi-class classification of all 14 conditions; a characterization of the network trained in this fashion was previously reported by Rajpurkar et al. This X-ray trained (XRT) DenseNet121 was fine-tuned on our training set to compare the OTS models trained on generic (non-CXR, ImageNet) images with one trained exclusively on CXRs. The output layer of this third model was replaced with a custom classifier block consisting of fully connected layers, and an output layer similar to the OTS models.

2.4. Cross validation, training and testing

After determining the network structures of the three models, the custom dense layers of each model were trained using 5-fold stratified cross-validation. The training set of 440 X-ray images were used for the 5-fold cross-validation, every fold using a random train-validation split of 80% - 20%, with splitting performed by image. The final models were trained and validated using the full training set of 440 images before being tested on the hold-out test set with 148 X-ray images. Since the dataset used for training was small, cross validation was implemented to reduce the chances of overfitting. We also selected images for the hold-out test so that no images from patients in the training set were included in the hold-out test set. The performance metrics reported in this study were obtained on the hold-out test set. During training, all the pre-trained layers were frozen (i.e. model parameters were held fixed) and only the newly added densely connected layers were trained. An initial learning rate of 0.001 with a decay in the order of a tenth of the learning rate was set. Binary cross entropy was chosen as the loss function. Each of the models were trained with the same batch size and Adam optimization algorithm. Average training time for a single epoch for the OTS VGG16 model, which had the highest number of parameters was approximately 11 s, while for the OTS MobileNet model, with the lower number of parameters, it was approximately 5 s on a linux machine with Intel(R) Xeon(R) Platinum 8175 M CPU @ 2.50GHz hosted on Amazon Web Services (AWS). We also report representative confusion matrices and area under the receiver operating characteristic (AUROC) curve values obtained on the hold-out test set.

2.5. Computing setup

Experiments were conducted using the AWS cloud infrastructure. Model classifiers were constructed using the Python Keras (https://keras.io) and Tensorflow (https://www.tensorflow.org) deep learning libraries. Pandas (https://pandas.pydata.org), NumPy (https://numpy.org), OpenCV (https://opencv.org) and Scikit-learn (https://scikit-learn.org) libraries were used for data preparation and model evaluation.

| Model            | AUROC | F1 Score | PPV | Sensitivity | Specificity |
|------------------|-------|----------|-----|-------------|-------------|
| OTS VGG16        | 0.98  | 0.93     | 0.96| 0.91        | 0.96        |
| OTS DenseNet121  | 0.95  | 0.89     | 0.87| 0.91        | 0.87        |
| OTS DenseNet201  | 0.97  | 0.95     | 0.95| 0.95        | 0.95        |
| OTS Inception    | 0.95  | 0.88     | 0.86| 0.91        | 0.85        |
| OTS MobileNet    | 0.99  | 0.94     | 0.99| 0.91        | 0.99        |
| OTS NasNetMobile | 0.96  | 0.88     | 0.85| 0.91        | 0.84        |
| XRT DenseNet121  | 0.87  | 0.83     | 0.76| 0.93        | 0.70        |

Table 2. Performance metrics of the fine-tuned models on the hold-out test set. Abbreviations: AUROC, area under the receiver operating characteristic; OTS, off-the-shelf; PPV, positive predictive value; XRT, X-ray trained.

Fig. 1. Performance of the seven models on the hold-out test set. Abbreviations: AUROC, area under the receiver operating characteristic; OTS, off-the-shelf; XRT, X-ray trained.
3. Results

Basic demographic characteristics of the patients associated with the 294 COVID-19 positive class and the 294 COVID-19 negative class images are reported in Table 1. The majority of images represented in the COVID-19 and NIH datasets are associated with male individuals who were above 50 years of age. Few images in the combined dataset are associated with individuals younger than 29 years of age. We note that, because information about sex and age was not associated to all images in the COVID-19 dataset, the corresponding percentages do not sum to 100%.

Each of the seven models had high performance in terms of AUROC values, among which OTS MobileNet had the highest AUROC of 0.99 while XRT DenseNet121 had the lowest AUROC of 0.87 on the hold-out test set (Table 2). The metrics associated with particular operating points varied more between the models.

The AUROC curves obtained by all models on the hold-out test set are presented in Fig. 1. The curves reflect the comparably sound validation performance across the models and a small decrease in test set performance relative to the validation average. This performance is not obtained at the expense of calibration, which we address using temperature scaling.

As demonstrated in Supplementary Figs. 1 and 2, after temperature scaling, the expected difference between the accuracy and confidence of OTS VGG16 model classifications is small, indicating good calibration. Grad-CAM heat maps roughly localize the regions of the X-rays which had greatest relevance to OTS VGG16 classifications (see supplementary methods, Supplementary Fig. 3).

The confusion matrices in Fig. 2 exhibit the performance of the final models on the hold-out test set. Here, label 1 corresponds to COVID-19 positive images. The support for the two classes (COVID and non-COVID) in the test set is 74 images per class.

4. Discussion

There is a bank of infectious disease literature focused on deep learning-based detection of COVID-19 from chest radiographs. Our results join the growing body of evidence that a variety of CNN architecture based models trained on CXR images can be successfully used to distinguish COVID-19 infection from conditions with similar clinical presentation (Table 2, Figs. 1 and 2) and also demonstrate that the performance need not come at the expense of calibration (Supplementary Figs. 1 and 2). To better understand the regions of CXR images giving rise to a particular classification, rough localization can be performed using a standard tool like Grad-CAM (Supplementary Fig. 3). In many cases, these architectures were not designed for COVID-19 detection in CXR and benefitted from pre-training on generic (non-CXR) image data instead of CXR data (likely due to the relative dearth of CXR data). This observation is reflected by the disparity in performance between the OTS DenseNet121 model and its XRT counterpart (Table 2). Moreover, strong performance can apparently be achieved with relatively few COVID-19 positive examples. Previous works have applied machine learning models to COVID-19 identification with CXR images, but in the absence of differential diagnoses (e.g. COVID-19 positive vs. healthy or no-finding patients) or only from pneumonia patients. Additionally, many studies applying machine learning to
evaluate X-ray images for COVID-19 diagnosis only examined small or private datasets, or datasets with large class imbalances\textsuperscript{24,28,29,30} (Supplementary Table 1).

In a 2020 study by Rubin et al.\textsuperscript{63} the authors noted that imaging had the potential to help rapidly triage patients in resource-scarce settings in which PCR testing was not widely available, or to provide additional information in cases where an apparently symptomatic patient receives a negative PCR test result. X-rays, in particular, are well suited to use in resource-scarce settings as they can be deployed rapidly for suspected COVID-19 patients. Further portable radiography units can be easily cleaned between each imaging patient, and do not require patients to enter a designated radiography room.\textsuperscript{64}

The models described in this study may offer improved lead time in COVID-19 identification as compared to RT-PCR assay reference standards, and these models perform favorably in the context of recent related work. Prospective validation during which the MLA is used in live settings will allow us to further demonstrate the ability of this technology to identify COVID-19 using X-rays and would benefit from collaboration with a multidisciplinary team, which may include a radiologist to label test images for comparison with the image classification made by the MLA. We are unable to claim that the models described in this study are the best possible among models resulting from deep learning architectures, as the comparison of architectures is not exhaustive and there are variants of these architectures (e.g., those with attention mechanisms) that we do not address. (For a recent review of architectures applied in this context, see Santosh et al., 2020.\textsuperscript{64} As we focus entirely on models based on deep learning, we do not directly compare with approaches for analyzing X-rays which use more traditional approaches, such as multiresolution methods.\textsuperscript{65} Additionally, we cannot determine from this retrospective study what impact the use of such an algorithm may have on clinicians and their provision of care in live clinical settings. Indeed, due to limited availability of COVID-19 chest radiographs, algorithm performance is assessed on a limited number of positive COVID-19 cases, which may not be reflective of general populations of patients with COVID-19. This limits the generalizability of our results to other patient populations and care settings. Such issues will be remedied as more COVID-19 image data become available.

5. Conclusion

In the wake of COVID-19, it has become clear that it is not merely disease itself that can contribute to death. It can also be attributed to the finite amount of medical supplies for testing and treatment of the disease, the challenge of identifying the disease course with very little known about its presentation in humans, and limited therapeutic treatment. This research joins the growing body of evidence which suggests that a variety of CNN architectures, pre-trained on generic image data, produce high performing and well calibrated models for COVID-19 detection using CXR images. These results support future prospective validation for continued optimization of ML and X-rays for COVID-19 diagnosis.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

All authors who have affiliations listed with Dascena (Houston, TX, USA) are employees or contractors of Dascena.

Acknowledgments

We gratefully acknowledge Anna Sievka and Gina Barnes for assistance with manuscript editing.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.clinimag.2021.07.004.

References

1. Sun J, He WT, Wang L, Ai J, Xi Z, et al. COVID-19: epidemiology, evolution, and cross-disciplinary perspectives. Trends Mol Med 2020. https://doi.org/10.1016/j.trends MolMed.2020.02.008.

2. Organization WH. Preparing for Large-scale Community Transmission of COVID-19: Guidance for Countries and Areas in the WHO Western Pacific Region. 2020.

3. World Health Organization (WHO). Responding to Community Spread of COVID-19: Interim Guidance, 7 March 2020. 2020.

4. Bai Y, Yao L, Wei T, Tian F, Jin D-Y, Chen L, et al. Presumed asymptomatic carrier transmission of COVID-19. JAMA 2020;323:466–7. https://doi.org/10.1001/jama.2020.2565.

5. Hu Z, Song C, Xu C. Clinical characteristics of 24 asymptomatic infections with COVID-19 screened among close contacts in Nanjing, China. Sci China Life Sci 2020. https://doi.org/10.1007/s11427-020-1661-4.

6. Kenji M, Katsushi K, Alexander Z, Gerardo C. Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the diamond princess cruise ship. Euro Surveill 2020;25(10). https://doi.org/10.2807/1560-7917.ES.2020.25.10.2000180.

7. Holdtue ML, DeBolt C, Lindquist S, Lofy KH, Wiesman J, Bruce H, et al. First case of 2019 novel coronavirus in the United States. N Engl J Med 2020;382:929–36.

8. Spector L. What Does the Coronavirus Mean for the U.S. Health Care System? Some Simple Math Offers Answering. 2020.

9. Ai T, Yang Z, Hou H, Zhan C, Chen C, Lv W, et al. Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases. Radiology 2020. https://doi.org/10.1148/radiol.2020200462.

10. Wu J, Wu X, Zeng W, Guo D, Fang Z, Chen L, et al. Chest CT findings in patients with coronavirus disease 2019 and its relationship with clinical features. Invest Radiol 2020;55:257–61. https://doi.org/10.1097/RLI.0000000000000670.

11. Zue Z, Jiang M, Xu P, Chen W, Ni Q, Lu G, et al. Coronavirus disease 2019 (COVID-19): a perspective from China. Radiology Feb 2020. https://doi.org/10.1148/ radiol.2020200490.

12. Narin A, Kaya C, Pamuk Z. Automatic Detection of Coronavirus Disease (COVID-19) Using X-ray Images and Deep Convolutional Neural Networks. 2020.

13. Radiology CS. Radiological diagnosis of new coronavirus infected pneumonitis: expert recommendation from the Chinese Society of Radiology. Chin J Radiol 2020;54. https://doi.org/10.3760/cma.j.issn.1005-1201.2020.001.001–001.

14. Shi W, Peng X, Liu T, Cheng Z, Lu H, Yang S, et al. Deep-learning-based quantitative computed tomography model in predicting the severity of COVID-19: a retrospective study in 196 patients. Lancet 2020;10(2139):356089.

15. Wang X, Deng X, Fu Q, Zhou Q, Feng J, Ma H, et al. A weakly-supervised framework for COVID-19 classification and lesion localization from chest CT. IEEE Trans Med Imaging 2020;39:2615–25. https://doi.org/10.1109/TMI.2020.2995965.

16. Kanne JP. Chest CT findings in 2019 novel coronavirus (2019-nCoV) infections from Wuhan, China: key points for the radiologist. Radiology 2020. https://doi.org/10.1148/radiol.2020200234.

17. Ng M-Y, Lee EYP, Yang J, Li X, Wang H, et al. Imaging profile of the COVID-19 infection: radiologic findings and literature review. Radiol Cardiothoracic Imaging 2020;2:e200034. https://doi.org/10.1148/ryct.2020200034.

18. Wang H, Xia Y. ChestNet: A Deep Neural Network for Classification of Thoracic Diseases on Chest Radiography n.d.

19. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 2020;395:497–506.

20. Lei J, Li J, Li X, Qi X. CT imaging of the 2019 novel coronavirus (2019-nCoV) pneumonia. Radiology 2020. https://doi.org/10.1148/radiol.202020236.

21. Song F, Shi N, Shan F, Zhang Z, Shen J, Lu H, Ling Y, Jiang Y, Shi Y. Emerging 2019 novel coronavirus (2019-nCoV) in China: a review of the literature. Radiol Cardiothoracic Imaging 2020;2:e200241. https://doi.org/10.1148/ryct.2020200241.

22. Chung M, Bernheim A, Meiz X, Zhang N, Huang M, Zeng X, et al. CT imaging features of 2019 novel coronavirus (2019-nCoV). Radiology 2020;295:202–7. https://doi.org/10.1148/radiol.2020200230.

23. Wang Z, Xiao Y, Li Y, Zhang J, Lu F, Hou M, et al. Automatically discriminating and localizing COVID-19 from community-acquired pneumonia on chest X-rays. Pattern Recognit 2021;110:107613. https://doi.org/10.1016/j.patcog.2020.107613.

24. Wang L, Lin ZQ, Wong A. COVID-19: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. Sci Rep 2020;10:15959. https://doi.org/10.1038/s41598-020-76550-2.

25. Ozturk T, Talo M, Yildirim EA, Baloglu UB, Yildirim O, Rajendra Acharya U. Automated detection of COVID-19 cases using deep neural networks with X-ray images. Comput Biol Med 2020;121:103792. https://doi.org/10.1016/j.compbiomed.2020.103792.

26. Khan AI, Shah JL, Bhat MM. CoroNet: a deep neural network for detection and diagnosis of COVID-19 from chest X-ray images. Comput Methods Programs Biomed 2020;2019;160581. https://doi.org/10.1016/j.cmpb.2020.105581.

27. Yoo SH, Geng H, Chiu TL, Yu SK, Cho DC, Heo J, et al. Deep learning-based decision-tree classifier for COVID-19 diagnosis from chest X-ray imaging. Front Med 2020;7. https://doi.org/10.3389/fmed.2020.00427.

28. J. Maharjan et al. Clinical Imaging 80 (2021) 268–273
