Generalizability of CNN on Predicting COVID-19 from Chest X-ray Images

**Natalia de Sousa Freire[1], Pedro Paulo de Souza Leão[1], Leonardo Albuquerque Tiago, [1]
Alberto de Almeida Campos Gonçalves[1], Rafael Albuquerque Pinto,[1]
Eulanda Miranda dos Santos [1], and Eduardo Souto[1]**

[1]Instituto de Computação (ICOMP), Universidade Federal do Amazonas (UFAM)
Av. General Rodrigo Octávio Jordão Ramos, Manaus, Amazonas, Brazil

{nsf,ppsl,lat,aacg,rafael.albuquerque}@icomp.ufam.edu.br

{emsantos,esouto}@icomp.ufam.edu.br

Abstract. *Machine learning methods have been applied to predict COVID-19 using chest X-ray images in several works. However, to be helpful, a machine learning model must be robust to give reliable predictions for any target population, rather than only for the population used to generate the training data. Despite such an important issue, testing the generalizability of machine learning models is frequently not performed in current works. To test the generalizability of three models of CNN, four different databases obtained from various data sources are investigated in this paper in an internal-and-external validation procedure. All models are trained considering lung segmentation as a pre-processing step and without lung segmentation. The results show how important an external evaluation is to avoid providing performance evaluations excessively optimistic and inaccurate.*

Resumo. *Diversos trabalhos têm utilizado métodos de aprendizagem de máquina para detectar Covid-19 a partir de imagens de raio x. Entretanto, para serem úteis, modelos de aprendizagem de máquina devem ser generalistas a fim de prover predições confiáveis para qualquer população de pacientes, não apenas para a população utilizada para gerar sua base de treinamento. Apesar da importância dessa característica, os trabalhos atuais dificilmente testam a capacidade de generalização dos modelos de aprendizagem de máquina entre diferentes populações. Neste artigo, nós estudamos a capacidade de generalização de três modelos de CNN em quatro bases de dados obtidas a partir de diversas populações de pacientes. É utilizado um processo de validação interna e externa. Todos os modelos são treinados considerando dois cenários: pré-processamento via segmentação da região do pulmão; e sem segmentação. Os resultados mostram a importância de realizar uma validação externa em uma população diferente da população que compõe a base de treinamento para evitar avaliações de desempenho excessivamente otimistas e imprecisas.*

1. Introduction

The research on automatic diagnosis of coronavirus (COVID-19) has been rapidly developed since the pandemic began in early 2020. Chest X-ray (CXR) radiographs and

abdominal computed tomography (CT) scans are considered important evidence to support clinical diagnosis of COVID-19 [Zu et al. 2020]. Besides allowing confirming the infection, it is possible to evaluate the extent of damage incurred to the lungs by screening through CXR and CT imaging [Khan et al. 2021]. In one hand, there may be visible lesions on CT that are not visible on CXR images. On the other, among CXR and CT scans, CXR images are interesting because they have a lower associated cost, are faster to acquire, and are more widely available [Pereira et al. 2020], [López-Cabrera et al. 2021]. In this work, we focus our attention on the CXR images.

Convolutional Neural Networks (CNN) have especially achieved great success in COVID-19 diagnosis using radiological imaging, such as CXR radiographs and CT [Arias-Garzón et al. 2021], [Saha et al. 2021] and [Khasawneh et al. 2021]. Despite this success, the literature indicates that several works may suffer from biases or issues in terms of generalization ability (generalizability)[Roberts et al. 2021].

Generalizability is related to the ability of a model to predict accurately on varied data sources not included in the model's training dataset. It has been suggested that the performance of deep learning models, including CNN, may show variable generalization on external data (also called out-of-distribution data) [Li et al. 2020a]. This behavior could be due to the so-called shortcut learning problem [Geirhos et al. 2020], which refers to models that learn decision rules to solve a given problem based on the simplest solution instead of on features related to the pathology to be classified. In the case of CXR images, the shortcut learning problem may result from differences in X-ray equipment manufacturers and acquisition techniques, for instance [Li et al. 2020a].

The literature indicates external validation as one of the ideal approaches to try to measure the generalizability of learning-based models. Internal validation refers to testing the model using data from the same source as that used to train on, while in external validation, the test and training datasets are from different sources. This procedure allows providing more insight about the model's generalizability. However, very few works have been devoted to cope with generalizability issues or to try to evaluate this problem in the context of diagnosis or prognosis of COVID-19 using CXR or CT images.

In the systematic review presented in [Roberts et al. 2021], the authors discuss 62 papers published from January 2020 to October 2020, which address machine learning applied to the automatic identification of COVID-19 using CXR or CT images. They point out only three works dealing with CXR images that handle external data. However, one work [Elaziz et al. 2020] does not evaluate the model on an external dataset, while the other two works [Li et al. 2020a] and [Li et al. 2020b] are not devoted to classification problems but focus on measuring the degree of severity of COVID-19.

Besides performing evaluation using out-of-distribution data, another strategy recommended in the literature is to extract features from the region containing the lung area only [Tartaglione et al. 2020], discarding possible bias sources, such as text and medical devices. In the survey presented in [López-Cabrera et al. 2021], the authors highlight five works that use an external dataset to evaluate their models [Tabik et al. 2020], [DeGrave et al. 2021], [Ahmed et al. 2021], [Tartaglione et al. 2020] and [Yeh et al. 2020]. Except for [Tabik et al. 2020], which deals with severity analysis, the remaining four works confirm a decrease in generalization in new data sources

when detecting COVID-19 in CXR images. However, lung area segmentation/crop is not performed in [DeGrave et al. 2021]. In [Ahmed et al. 2021], despite cropping the lung region, the authors conclude that the models still learn spurious features related to the data sources. However, using only segmented lung masks, instead of cropping, may be a better strategy to provide features related to the real underlying pathology of COVID-19. Finally, the authors [Tartaglione et al. 2020] and [Yeh et al. 2020] perform segmentation, but generalizability was not their focus.

Unlike the works mentioned above, in this paper we evaluate CNN generalization to diagnose COVID-19 automatically using the internal-and-external validation procedure and segmentation of the lung region. In addition, both fine-tuning of pretrained CNN and end-to-end training of a developed CNN model are used. The two approaches are investigated with and without lung segmentation.

The remainder of this paper is organized as follows. A short review of related work is provided in Section 2. The description of the datasets, as well as the models investigated in this paper, is provided in Section 3. Section 4 describes experiments and results. Finally, conclusions and future work are presented in Section 5.

2. Related Work

In [Elaziz et al. 2020], the authors propose a method to extract and reduce features of CXR images that are used as input to a KNN classifier designed to distinguish between COVID-19, normal, and pneumonia classes. The proposed method was evaluated using two datasets provided by two different sources. The first dataset consists of data collected by Joseph Paul Cohen, Paul Morrison, and Lan Dao on GitHub [Cohen et al. 2020], which groups into the COVID-19 class images extracted from 43 different publications. Images used to compose the normal and the viral pneumonia classes were obtained from the X-ray image database (pneumonia) [Kermany et al. 2018]. In turn, the second dataset is composed of data collected by a team of researchers from Qatar University, Doha - Qatar; the University of Dhaka, Bangladesh, along with collaborators from Pakistan and Malaysia [Chowdhury et al. 2020]; as well as the dataset from the Italian Society of Medical and Interventional Radiology (SIRM) COVID-19 database [1]. Despite working with datasets from different sources, the model was not evaluated on external test sets, since the authors performed training and inference using the classical holdout validation in the same data source. Therefore, possible biases in terms of generalizability were not tackled or even evaluated.

On the other hand, internal and external validation is performed by Li et al. in [Li et al. 2020a] and [Li et al. 2020b]. Their first work [Li et al. 2020a] indicated that their model was able to generalize to data obtained from a hospital different from the hospital the training data was obtained. Nevertheless, since the internal and external datasets were from urban areas of the same location, they extended their experiments in [Li et al. 2020b] to test the generalizability of the same model on four datasets acquired from different patient populations from three hospitals in two countries (Brazil and the United States). In addition, they tuned the model using outpatient data to improve model generalizability. The results showed that the model was able to generalize across distinct patient populations due to being tuned with outpatient data. It is important to mention

[1] https://www.sirm.org/category/senza-categoria/covid-19/

that the model investigated in these works is a convolutional Siamese neural network. Moreover, the task dealt with was calculating the degree of severity of COVID-19 lung disease, not the prediction of COVID-19. The work presented in [Tabik et al. 2020] also conducts external validation in the context of COVID-19 severity analysis. It is worth noting that there are additional works dealing with COVID-19 pneumonia severity scoring in CXR that address internal and external validation, such as [Frid-Adar et al. 2021], which employs CNN models.

Considering the detection of COVID-19 in CXR using CNN, [DeGrave et al. 2021] perform an internal and external validation similar to the process presented in [Li et al. 2020b]. They investigated various CNN architectures and showed a very significant performance drop when comparing the models performance using internal and external datasets. In order to better investigate the poor performance in the external dataset, the authors employed saliency maps. Their results showed that the models were looking at non-lung regions of the images to classify the instances, consequently confirming the shortcut learning problem. Therefore, by working with the lung region only, one may potentially remove possible bias sources to reduce the shortcut learning effect.

This was the objective in [Ahmed et al. 2021]. The lung region is obtained using an approach divided into two steps: 1) segmentation of the lung field; and 2) cropping of the bounding boxes containing the lung area. The pre-trained ResNet50 was used as CNN model in a bi-class classification problem (COVID-19 and non-COVID-19 classes). The experiments results indicate that focusing only on the bounding boxes containing the lung area does not guarantee the mitigation of shortcut learning, since the models reached high performance on internal test data but low performance on external test data. Consequently, there is still need for better strategies to focus on areas within the lungs to adequately separate the classes taking into account features indeed related to the disease.

3. Methods

In this work, we study the challenges on predicting COVID-19 from CXR images considering only a multiclass problem, precisely three classes: COVID-19, Pneumonia (non-COVID19 infection, e.g., viral, bacterial, etc.) and normal (no infection), since it is well accepted in the literature that the potential utility of a model increases when it is capable of distinguishing patients with COVID-19 from patients without COVID-19 and other types of pneumonia. Two approaches are employed, (a) a lightweight CNN with end-to-end training and no transfer learning; and (2) two pre-trained deep CNN models. In both approaches, all network models are trained using original images and segmented images. The investigated methods are described in this section.

3.1. Lightweight CNN

The model used is publicly available[2]. It is designed with the following architecture: input layer accepting 224 x 224 images; 10 convolutional layers; 04 pooling layers, a flatten layer, a dropout layer and 02 fully-connected (FC) layers. Figure 1 shows the model. The pooling layers use max pooling, whilst ReLU is the activation function in

[2]https://medium.com/@estevestoni/deep-learning-e-covid-19-utilizando-de-imagens-de-raio-x-63296b5dc77a

convolutional and FC layers. Softmax is employed in the output layer. Finally, Adam (Adaptive Moment Estimation) with a learning rate of 1e-4 was used as the optimizer.

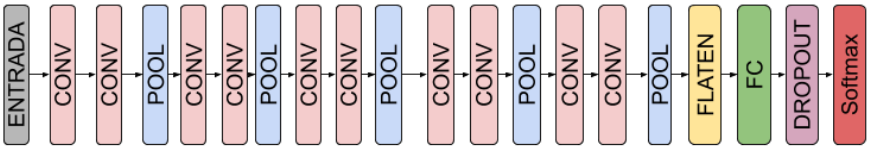

Figure 1. Architecture of the CNN with end-to-end training.

3.2. Pre-trained deep CNN

The two CNNs studied are: ResNet50v2 and VGG16. The original FC layers and the classification output layer were removed from the network architecture, while a global max pooling 2D layer, a FC layer with 1024 neurons and a classification layer were added for application to the task of predicting COVID-19. The process applied to fine-tune each model was to unfreeze all layers. In addition, data augmentation was conducted in both fine-tuning and end-to-end training. The augmentation techniques carried out are random shifting up to 20% (horizontally and vertically), rotation up to 20°, zoom in up to 10%, and flip horizontally. Finally, Nadam was the optimizer employed using learning rate 1e-5 in the first and third experimental scenarios and 1e-4 in the second. The experimental scenarios are described in the following sections.

3.3. Segmentation

As previously mentioned, both lightweight and pre-trained CNNs are evaluated considering two pipelines: 1) image segmentation + feature extraction and classification; 2) feature extraction + classification. In the first pipeline, U-Net [Ronneberger et al. 2015] is responsible for performing the segmentation step. U-Net was originally developed for biomedical image segmentation. In this paper, Montgomery County and Shenzhen Hospital[3] are the two databases used to train U-Net in lung segmentation. The pulmonary segmentation masks were dilated and the images were resized to 224x224.

The Montgomery County dataset is composed of CXR images acquired in the context of a tuberculosis control program from the Montgomery County Department of Health and Human Services in the United States. The data set contains 138 instances divided into two classes: 80 from normal and 58 are from abnormal classes. The lungs were manually segmented in all images. In terms of the second dataset, it contains CXR images collected in the Shenzhen Hospital, Shenzhen - Guangdong province - China. This dataset is composed of 662 images, also divided into two classes: 326 images from healthy patients and 336 with abnormalities. The two datasets used to train U-Net are different from the datasets investigated in our internal-and-external validation procedure, as detailed in the next section.

3.4. Datasets

The following four datasets are investigated in this paper: (CIDC) COVID-19 Image Data collection [Cohen et al. 2020], RSNA Pneumonia Detection Challenge dataset [4], (CXRP)

[3]https://lhncbc.nlm.nih.gov/LHC-downloads/downloads.html#tuberculosis-image-data-sets
[4]https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/data

Chest X-Ray Images (Pneumonia)[5], and COVIDx dataset[6] [Wang et al. 2020]. In this section, we first describe the datasets carefully. Then, we explain how they are divided in an internal-and-external validation procedure.

3.4.1. Datasets Description

- **CIDC**: it contains 950 CXR images of COVID-19 positive or suspected patients, and patients with pneumonia due to other viral or bacterial sources from several studies and several countries.
- **RSNA**: it is composed of 29,700 CXR images divided into three classes: No Lung Opacity/Not Normal, Lung Opacity and Normal. This dataset is a sample of a larger database that contains more than 100,000 anonymized patients, published by the National Institute of Health Clinical Center hospital, United States, to be used in the RSNA Pneumonia Detection Challenge competition in 2018.
- **CXRP**: it consists of 5,863 CXR images belonging to one of two classes: Pneumonia and normal. The images are from the Guangzhou Women and Children's Medical Center, China. These are data from patients aged one to five years old and were obtained during routine examinations between July, 2013 - March, 2017.
- **COVIDx**: it currently contains 13.975 CXR images distributed among 3 classes: COVID-19, Normal and Pneumonia. This database was built and organized by Wang [Wang et al. 2020] to provide a dataset larger than the public databases available. The authors have combined and modified five public repositories to generate this dataset: (1) CIDC; (2) COVID-19 Chest X-ray Dataset Initiative; (3) ActualMed COVID-19 Chest X-ray Dataset Initiative; (4) COVID-19 radiography database[7]; and (5) RSNA.

3.4.2. Data Partition Configurations

As previously mentioned, CIDC and RSNA are among the different repositories used to compose COVIDx. Thus, to avoid reusing data, instances from these two repositories were removed from the original COVIDx. Therefore, we try to assure that all four databases used in our experiments are not overlapping datasets.

To perform and evaluate the internal-and-external validation procedure, we conduct three configurations of data combination and partition. It is important to mention that only instances from Pneumonia and normal classes from the RSNA and CXRP databases are used in our experiments. In terms of COVIDx and CIDC, only instances from their COVID-19 class were considered for data partitioning. In the first configuration, train and validation sets are composed of instances from COVIDx and RSNA, while the test set is composed of images from CIDC and CXRP. In the second configuration, COVIDx and CXRP are combined and further divided into training and validation sets, while instances from CIDC and RSNA are used to compose the test set. Finally, in the third scenario, all four datasets are put together to form a more extensive database and then divided into

[5]https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia
[6]https://github.com/lindawangg/COVID-Net
[7]https://github.com/agchung/Figure1-COVID-chestxray-dataset

training, validation, and test sets using the classical holdout validation to evaluate performance. Using datasets assembled from other datasets, as is done in the third scenario, is common in the literature [Roberts et al. 2021].

Besides varying the data source, we partitioned the instances in a balanced way. For the first and second configurations, the training and validation sets were obtained by dividing in a holdout strategy 8/10 and 2/10 of the instances, respectively. In the third configuration, holdout was performed to provide training, test, and validation partitions with 8/10, 1/10, and 1/10 of the instances, respectively. Class balancing was obtained by randomly choosing a similar number of instances per class for each partition. Tables 1, 2, and 3 highlight the number of instances in each data partition of each configuration.

Table 1. Number of instances in each data partition - first configuration.

	Train	**Validation**	**Test**
COVID-19	840	213	450
Normal	843	210	450
Pneumonia	843	210	411
Total	2526	633	1341

Table 2. Number of instances in each data partition - second configuration.

	Train	**Validation**	**Test**
COVID-19	839	214	450
Normal	843	210	474
Pneumonia	799	210	439
Total	2481	634	1363

Table 3. Number of instances in each data partition - third configuration.

	Train	**Validation**	**Test**
COVID-19	1202	150	151
Normal	1204	149	149
Pneumonia	1198	149	149
Total	3604	448	449

4. Results

Two series of experiments are conducted in this paper. In the first, the third configuration is investigated using three different CNN models: 1) Lightweight CNN, 2) ResNet50v2, and 3) VGG16. In the second series, the same CNN models are applied using the first and the second configuration of datasets. In all scenarios, the models were investigated using non-segmented and segmented images. The results are the average of metrics across classes: Normal, COVID-19, and Pneumonia.

4.1. Holdout Evaluation Study

We first analyze the performance of each model when segmentation is not conducted. As it can be observed in Table 4, the performance metrics (accuracy, precision, and sensitivity) show a behavior quite expected when using only internal validation, e.g: high performance is obtained - accuracy was slightly below 95%; the pre-trained CNN outperformed the lightweight model; and the pre-trained models achieved very similar results.

Table 4. Results attained using the 3th configuration without segmentation.

Model	Acc.	Prec.	Sens.
Lightweight	91.31	93.87	91.39
ResNet50v2	94.20	**97.92**	93.38
VGG16	**94.43**	94.74	**95.36**

When observing the results reached using segmented images (Table 5), it is possible to verify that they did not exceed the results obtained using non-segmented images. These results were not expected, since lung segmentation usually helps to improve the classification performance. However, the key issue for performance improvement is to obtain accurate lung segmentation. Even though we have used U-Net, which is the traditional model in lung segmentation, and the datasets commonly used to train the model [Frid-Adar et al. 2021], the segmentation step was not successful in our work. Figure 2 illustrates the result of an accurate (a) and an inaccurate segmentation (b). Therefore, improvements in this process are necessary to achieve reliable and accurate segmentation, as in [Frid-Adar et al. 2021] for instance.

Table 5. – Results attained using the 3th configuration with segmentation.

Model	Acc.	Prec.	Sens.
Lightweight	85.30	86.42	80.13
ResNet50v2	91.98	**95.83**	91.39
VGG16	**92.65**	95.24	**92.72**

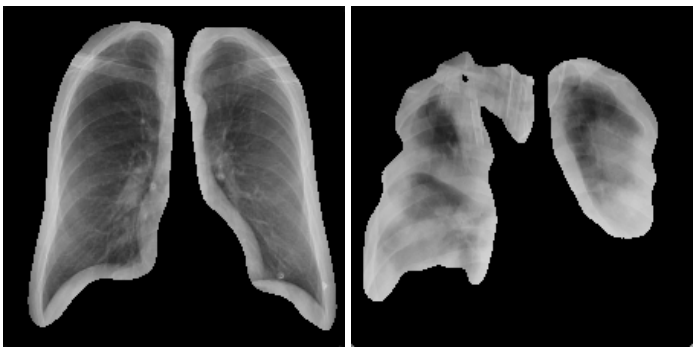

Figure 2. Two examples of segmented CXR images: (a) accurate segmentation, and (b) unsuccessful segmentation.

4.2. Generalizability Assessment

The results from our previous series of experiments were achieved without performing an external validation. In this second series of experiments, however, we investigate the performance when using external validation. This is done by employing data configurations 1 and 2.

The results shown in Tables 6 and 8 were attained by the models using no segmentation on configurations 1 and 2 respectively. These results highlight that all performance metrics drastically decreased when compared to the results obtained in the first series. We observed that all three investigated models failed to generalize to other data sources. Despite high sensitivity rates in the third configuration, accuracy and precision rates show that the models reached errors lower than the expected error of a randomized predictor in both scenarios.

Unexpectedly, lung segmentation as a pre-processing step did not help to increase the performance metrics, as it can be seen in Tables 7 and 9. Again, the reason for this behavior is probably an inaccurate lung segmentation step performed.

Table 6. Results attained using the 1st configuration without segmentation.

Model	Acc	Prec	Sens
Lightweight	31.77	30.35	70.88
ResNet50v2	48.86	42.12	46.89
VGG16	37.12	34.25	91.11

Table 7. Results attained using the 1st configuration with segmentation.

Model	Acc	Prec	Sens
Lightweight	39.34	35.10	72.00
ResNet50v2	39.04	34.26	70.89
VGG16	40.66	34.02	73.33

Table 8. Results attained using the 2nd configuration without segmentation.

Model	Acc	Prec	Sens
Lightweight	34.82	33.91	98.88
ResNet50v2	38.92	35.03	94.67
VGG16	34.82	33.89	98.44

Table 9. Results attained using the 2nd configuration with segmentation.

Model	Acc	Prec	Sens
Lightweight	36.47	34.43	97.56
ResNet50v2	34.30	33.78	96.00
VGG16	35.12	33.90	97.56

These results confirm that models trained using only one data source demonstrate high performance, while these models exhibit substantial performance degradation when

tested on external data. This indicates that the investigated models failed to generalize to other data sources even using only segmented images. Therefore, if the potential of machine learning models to predict COVID-19 focuses solely on performance on a single data source, this kind of evaluation leads to uncertainty of these models generalizability and implementation across real healthcare settings.

5. Conclusions

In this paper, we investigated the generalizability of CNNs on predicting COVID-19 from chest X-ray images when the model is tested on patient populations (data sources) different from the data sources used to train the model, i.e. when an internal-external validation procedure is conducted. Our results showed the models' inability to generalize well on external datasets, even using only segmented lungs, and reinforce the need for employing internal-external validation to reduce the risk of optimistic performance evaluations. The impact of showing that learning models provide accurate predictions across a variety of diverse data sources play a key role in the fulfillment of achieving practical application of these models in healthcare settings. However, it would be ideal to use more data sources to evaluate the models on a broader different population of patients.

6. Acknowledgements

This research, carried out within the scope of the Samsung-UFAM Project for Education and Research (SUPER), according to Article 48 of Decree nº 6.008/2006(SUFRAMA), was funded by Samsung Electronics of Amazonia Ltda., under the terms of Federal Law nº 8.387/1991, through agreement 001/2020, signed with Federal University of Amazonas and FAEPI, Brazil.

Referências

Ahmed, K. B., Goldgof, G. M., Paul, R., Goldgof, D. B., and Hall, L. O. (2021). Discovery of a generalization gap of convolutional neural networks on covid-19 x-rays classification. *IEEE Access*, 9:72970–72979.

Arias-Garzón, D., Alzate-Grisales, J. A., Orozco-Arias, S., Arteaga-Arteaga, H. B., Bravo-Ortiz, M. A., Mora-Rubio, A., Saborit-Torres, J. M., Ángel Montell Serrano, J., de la Iglesia Vayá, M., Cardona-Morales, O., and Tabares-Soto, R. (2021). Covid-19 detection in x-ray images using convolutional neural networks. *Machine Learning with Applications*, 6:100138.

Chowdhury, M. E. H., Rahman, T., Khandakar, A., Mazhar, R., Kadir, M. A., Mahbub, Z. B., Islam, K. R., Khan, M. S., Iqbal, A., Emadi, N. A., Reaz, M. B. I., and Islam, M. T. (2020). Can AI help in screening viral and COVID-19 pneumonia? *IEEE Access*, 8:132665–132676.

Cohen, J. P., Morrison, P., Dao, L., Roth, K., Duong, T. Q., and Ghassemi, M. (2020). Covid-19 image data collection: Prospective predictions are the future.

DeGrave, A. J., Janizek, J. D., and Lee, S.-I. (2021). Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, Epub May 31.

Elaziz, M. A., Hosny, K. M., Salah, A., Darwish, M. M., Lu, S., and Sahlol, A. T. (2020). New machine learning method for image-based diagnosis of COVID-19. *PLOS ONE*, 15(6):e0235187.

Frid-Adar, M., Amer, R., Gozes, O., Nassar, J., and Greenspan, H. (2021). COVID-19 in CXR: From detection and severity scoring to patient disease monitoring. *IEEE Journal of Biomedical and Health Informatics*, 25(6):1892–1903.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R. S., Brendel, W., Bethge, M., and Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Kermany, D., Zhang, K., and Goldbaum, M. (2018). Labeled optical coherence tomography (oct) and chest x-ray images for classification. *Applied Sciences*.

Khan, S. H., Sohail, A., Khan, A., Hassan, M., Lee, Y. S., Alam, J., Basit, A., and Zubair, S. (2021). Covid-19 detection in chest x-ray images using deep boosted hybrid learning. *Computers in Biology and Medicine*, 137:104816.

Khasawneh, N., Fraiwan, M., Fraiwan, L., Khassawneh, B., and Ibnian, A. (2021). Detection of covid-19 from chest x-ray images using deep convolutional neural networks. *Sensors*, 21(17).

Li, M. D., Arun, N. T., Aggarwal, M., Gupta, S., Singh, P., Little, B. P., Mendoza, D. P., Corradi, G. C., Takahashi, M. S., Ferraciolli, S. F., Succi, M. D., Lang, M., Bizzo, B. C., Dayan, I., Kitamura, F. C., and Kalpathy-Cramer, J. (2020a). Improvement and multi-population generalizability of a deep learning-based chest radiograph severity score for covid-19. *medRxiv*.

Li, M. D., Arun, N. T., Gidwani, M., Chang, K., Deng, F., Little, B. P., Mendoza, D. P., Lang, M., Lee, S. I., O'Shea, A., Parakh, A., Singh, P., and Kalpathy-Cramer, J. (2020b). Automated assessment and tracking of COVID-19 pulmonary disease severity on chest radiographs using convolutional siamese neural networks. *Radiology: Artificial Intelligence*, 2(4):e200079.

López-Cabrera, J. D., Orozco-Morales, R., Portal-Díaz, J. A., Lovelle-Enríquez, O., and Pérez-Díaz, M. (2021). Current limitations to identify covid-19 using artificial intelligence with chest x-ray imaging (part ii). the shortcut learning problem. *Health and Technology*, 11(6):1331–1345.

Pereira, R. M., Bertolini, D., Teixeira, L. O., Silla, C. N., and Costa, Y. M. (2020). Covid-19 identification in chest x-ray images on flat and hierarchical classification scenarios. *Computer Methods and Programs in Biomedicine*, 194:105532.

Roberts, M., , Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., Aviles-Rivero, A. I., Etmann, C., McCague, C., Beer, L., Weir-McCall, J. R., Teng, Z., Gkrania-Klotsas, E., Rudd, J. H. F., Sala, E., and Schönlieb, C.-B. (2021). Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nature Machine Intelligence*, 3(3):199–217.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241.

Saha, P., Sadi, M. S., Aranya, O. R. R., Jahan, S., and Islam, F.-A. (2021). Cov-vgx: An automated covid-19 detection system using x-ray images and transfer learning. *Informatics in Medicine Unlocked*, 26:100741.

Tabik, S., Gómez-Ríos, A., Martín-Rodríguez, J. L., Sevillano-García, I., Rey-Area, M., Charte, D., Guirado, E., Suárez, J. L., Luengo, J., Valero-González, M. A., García-Villanova, P., Olmedo-Sánchez, E., and Herrera, F. (2020). Covidgr dataset and covid-sdnet methodology for predicting covid-19 based on chest x-ray images. *IEEE Journal of Biomedical and Health Informatics*, 24(12):3595–3605.

Tartaglione, E., Barbano, C. A., Berzovini, C., Calandri, M., and Grangetto, M. (2020). Unveiling covid-19 from chest x-ray with deep learning: A hurdles race with small data. *International Journal of Environmental Research and Public Health*, 17(18).

Wang, L., Lin, Z. Q., and Wong, A. (2020). COVID-net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest x-ray images. *Scientific Reports*, 10(1).

Yeh, C.-F., Cheng, H.-T., Wei, A., Chen, H.-M., Kuo, P.-C., Liu, K.-C., Ko, M.-C., Chen, R.-J., Lee, P.-C., Chuang, J.-H., Chen, C.-M., Chen, Y.-C., Lee, W.-J., Chien, N., Chen, J.-Y., Huang, Y.-S., Chang, Y.-C., Huang, Y.-C., Chou, N.-K., Chao, K.-H., Tu, Y.-C., Chang, Y.-C., and Liu, T.-L. (2020). A cascaded learning strategy for robust covid-19 pneumonia chest x-ray screening.

Zu, Z., Jiang, M., Xu, P., Chen, W., Ni, Q., Lu, G., and Zhang, L. (2020). Coronavirus disease 2019 (covid-19): A perspective from china. *Radiology*, 296:200490.