Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

COV-SNET: A deep learning model for X-ray-based COVID-19 classification[☆]

Robert Hertel [*], Rachid Benlamri

Lakehead University, 955 Oliver Rd, Thunder Bay, ON P7B 5E1, Canada

ARTICLE INFO

Keywords:
Coronavirus
COVID-19
Convolutional neural network
Deep learning
Chest X-ray
Computer vision

ABSTRACT

The AI research community has recently been intensely focused on diagnosing COVID-19 by applying deep learning technology to the X-ray scans taken of COVID-19 patients. Differentiating COVID-19 from other pneumonia-inducing illnesses is a highly challenging task as it shares many of the same imaging characteristics as other pulmonary diseases. This is especially true given the small number of COVID-19 X-rays that are publicly available. Deep learning experts commonly use transfer learning to offset the small number of images typically available in medical imaging tasks. Our COV-SNET model is a deep neural network that was pretrained on over one hundred thousand X-ray images. In this paper, we designed two COV-SNET models with the purpose of diagnosing COVID-19. The experimental results demonstrate the robustness of our deep learning models, ultimately achieving sensitivities of 95% for our three-class and two-class models. We also discuss the strengths and weaknesses of such an approach, focusing mainly on the limitations of public X-ray datasets on current COVID-19 deep learning models. Finally, we conclude with possible future directions for this research.

1. Introduction

The medical industry and researchers around the world have been urgently seeking new modalities to diagnose COVID-19. A lack of testing supplies in countries around the world has left many COVID-19 patients without a diagnosis, leading to the further spread of the illness. To help alleviate this exponentially growing need, deep learning researchers have been attempting to image lung manifestations of coronavirus disease 2019 (COVID-19) with the use of radiological techniques. COVID-19 is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and is an airborne illness that can be rapidly spread between individuals. The COVID-19 outbreak was officially recognized by the World Health Organization (WHO) as being the cause of a pandemic on March 11, 2020.

The real-time reverse transcription-polymerase chain reaction (RT-PCR) test is the current gold standard for diagnosing COVID-19 [2]. While it is the best option that is currently available for diagnosing COVID-19, there have been considerable problems reported concerning the test's sensitivity [9]. The false-negative rate of an RT-PCR test can vary significantly over the time that it is administered. In a study conducted by John Hopkins University, the best false-negative rate that RT-PCR testing achieved was 26%. This performance was reported on the eighth day since the onset of COVID-19 symptoms [20]. A large variation in RT-PCR test accuracy has motivated many researchers to find other tests that can replace or be used in addition to RT-PCR tests. A leading candidate among medical researchers has been the use of radiological imaging. In instances where a COVID-19 test is negative but the patient is strongly suspected of suffering from the disease, radiological imaging has been shown to be advantageous [26].

Chest X-rays (CXRs) and thoracic computed tomographic (CT) scans are the most common modalities radiologists use to detect COVID-19 related pneumonia in individuals. Both technologies have their merits and shortcomings. In comparing CXRs and CT scans, CXRs are generally less expensive and hence more widely used. This is especially true in developing countries where budgeting for a CT scanner can be more of a challenge. X-ray machines have another advantage over CT scanners in that they are commonly manufactured to be portable. They can be physically carted into intensive care units (ICUs) and the patient can remain where they are.

Before diving into the details of deep learning algorithms that may assist in diagnosing COVID-19, it is beneficial to first consider what imaging details radiologists have cited in determining a COVID-19

[☆] This document is the results of the research project funded by the Natural Sciences and Engineering Research Council of Canada (NSERC).

[*] Corresponding author.
E-mail address: rhertel@lakeheadu.ca (R. Hertel).

https://doi.org/10.1016/j.imu.2021.100620
Received 21 April 2021; Received in revised form 21 May 2021; Accepted 23 May 2021
Available online 27 May 2021

diagnosis. These image characteristics are of considerable importance during the process of validating COVID-19 deep learning models with saliency maps. Common features of COVID-19 in radiological imaging includes bilateral Ground-Glass Opacities (GGOs) with peripheral predominance [21]. A GGO is an infected pulmonary location in a radiological scan with increased attenuation. Song et al. [34] have additionally discovered that consolidation can commonly be observed in patients as the disease worsens. These consolidated areas in radiology represent regions where a patient's lung is filled with pus, liquid and other materials that normally would not be present. Song et al. have also reported that "patients older than 50 years had more consolidated lung lesions than did those aged 50 years or younger." [34] Older patients therefore have clinical radiological evidence that shows they are at greater risk of negative health outcomes when they are infected. Cozzi et al. have likewise published research involving X-ray scans indicating that COVID-19 patients "show patchy or diffuse reticular–nodular opacities and consolidation, with basal, peripheral and bilateral predominance." [8] The same authors have additionally established that in cases where only one lung is infected, the right lung typically is more often affected. To obtain a visual appreciation for the manifestations of COVID-19 inside an infected patient's lungs, Fig. 1 shows the chest X-rays of two COVID-19 patients with some of the visual markers that have been discussed.

Our research has been focused on the development of a new deep learning model that has been trained to classify patients suspected of suffering from COVID-19. The contributions of our work are three-fold. First, the proposed COV-SNET models we present are capable of diagnosing COVID-19 with accuracies above those reported by practicing radiologists in a related work [39]. Second, the dataset we use does not incorporate several sources of bias contained in related works. Lastly, our work presents a comprehensive study that benchmarks our new COV-SNET models with other existing COVID-19 deep learning models.

Our work commences in section 2 with a discussion of other studies that have used transfer learning for diagnosing COVID-19. In section 3, we then move on to discuss our proposed network architecture and the deep learning methods we have employed for processing the X-ray scans of COVID-19 patients. After explaining these methods, in section 4 we present the experimental results of our system. We thereafter compare the performance of our models with other existing systems and discuss the advantages of our approach. Lastly, in section 5 we conclude our discussion with possible future directions for this research.

2. Related works

There are a number of papers that have been published on using deep learning methods on X-ray images for diagnosing COVID-19. There is a variety of approaches that have been researched on the subject and a large assortment of public COVID-19 X-ray datasets in circulation. Below are some of the findings of the most important papers that have been published on the subject.

The designers of COVIDX-Net [13] compared seven 2D off-the-shelf architectures. Hemdan et al. [13] intended to compare these architectures using the same training and test methods. Apostolopoulos and Mpesiana [4] took the same approach as Hemdan et al. [13] and compared several architectures that were pretrained on ImageNet weights. Hemdan et al. [13] reported the best architecture's results came from using the VGG-19 [33] and DenseNet-201 architectures [14]. Apostolopoulos and Mpesiana [4]'s approach differed from Hemdan et al. [13] in that they reported 2-class and 3-class (COVID vs. pneumonia vs. normal) results. They found a VGG-19 obtained the highest results. There were a couple of major deficiencies in these reported studies. These studies' datasets (especially Hemdan et al. [13]) were both too small to achieve trustworthy results. They also only used ImageNet and neglected using a form of modality-specific transfer learning. Apostolopoulos and Mpesiana [4] made the mistake of using Kermany et al.'s [18] pneumonia dataset of children between the ages of one to five years old. We noticed that papers that have used this dataset tend to report unrealistic evaluation metrics.

Khalifa et al. [19] first proposed using a generative adversarial network (GAN) [10] to further augment the images input into their classifier and increase its accuracy in diagnosing patients with pneumonia. The authors increased the size of their dataset by a factor of ten. They believe this helped their classifier to avoid overfitting. They attempted to use several deep learning classifiers in their model and ultimately decided to use a ResNet-18 [12]. Waheed et al. [36] also designed their model incorporating a GAN and later released a work similar to Khalifa et al. [19]. Their model differed in that they used an auxiliary classifier generative adversarial network (AC-GAN) [25]. Their AC-GAN generated synthetic images that were input into a VGG-16 classifier [33]. Khalifa et al. [19] made the mistake of using Kermany et al.'s [18] pneumonia dataset. Waheed et al. [36] look to have made the same mistake by using the COVID-19 Radiography Database [29].

Wang et al. [37] designed "COVID-Net" for the purpose of diagnosing COVID-19. The dataset used to train this custom-designed CNN was made public and eventually used in several other research papers. This

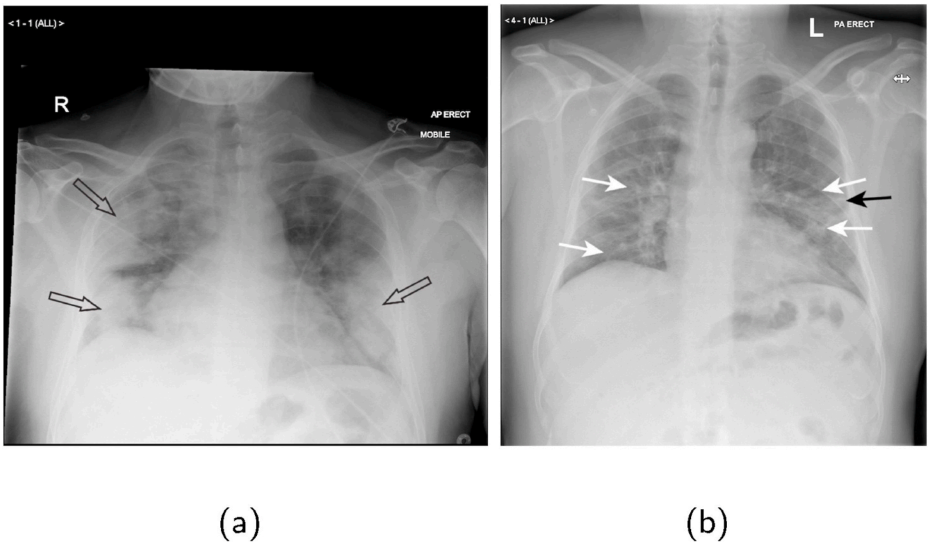


Fig. 1. Lungs of 2 men with COVID-19 pneumonia in their 50s showing (a) bilateral consolidation and (b) GGOs (white arrows) and linear opacity (black arrow) [7].

dataset is one of the largest datasets publicly available and the dataset does not contain many of the errors found in several other public datasets. Their model demonstrated promising results and achieved an accuracy of 93.3%. Their model was constructed using a "machine--driven design exploration strategy" [37] that uses generative syntheses [40]. This particular strategy was the subject of some of the authors' previous research prior to the COVID-19 pandemic. Their approach is capable of generating efficient deep neural networks automatically and designs these networks using a ResNet architecture [12]. The authors of this paper also used an explainability method called GSInquire [22] to validate their work.

Rajaraman et al. [30] created a model of iteratively pruned deep learning ensembles to diagnose COVID-19. The authors carried out their work by first training several popular CNN models (VGG-16/VGG-19 [33], Inception-V3 [35], Xception [6], DenseNet-201 [14], etc.) on a separate lung X-ray task (a modality-specific task). To use fewer model parameters and help improve the model's accuracy, the authors iteratively pruned their CNNs. They combined these iteratively pruned CNNs using several ensemble strategies. They found weighted averaging to be the most effective ensemble strategy. Like many other studies, they made the mistake of using Kermany et al.'s [18] pneumonia dataset.

Another study that deserves consideration is Wehbe et al.'s [39] publication that attempted to diagnose COVID-19 using a large private dataset from a US medical institution. This paper was similar to Rajaraman et al.'s paper [30] as the authors constructed an ensemble of many CNNs to detect COVID-19. Their dataset, however, did not suffer from the same deficiencies in size as other datasets. They also did not use Kermany et al.'s [18] dataset. The paper is noteworthy in that the authors assembled a team of five radiologists to determine the diagnosis of COVID-19 patients. They thereafter compared the predictions of the radiologists with their ensemble model. They found that the consensus of five radiologists was only able to detect COVID-19 with 81% accuracy. These results give a reasonable estimate of Bayes error for the task of determining the diagnosis of suspected COVID-19 patients. The author's ensemble model produced predictions with 83% accuracy, which is reasonable given the experts' consensus accuracy of 81%. Previous studies were unable to perform comparisons of their models against the predictions of working radiologists. The evaluation metrics mentioned in many of the previous papers were also liable to be skewed by the size of their datasets. Smaller datasets can sometimes lead to overly promising results.

Yeh et al. [41] used private datasets from several medical institutions and added them to Wang et al.'s dataset [37] when training their DenseNet-121 model [14]. They trained and tested their deep learning model initially using images from the same sources as Wang's COVIDx Dataset. They also used pneumonia, COVID-19, and normal X-ray images from two medical institutions. They obtained very promising results and achieved COVID-19 sensitivities between 95 and 100%. They held out a third much larger private dataset from a medical institution to see how their results would change with extra data. This larger dataset caused their accuracy to drop and they achieved an 81.82% COVID-19 sensitivity on their test set. This is evidence that using a small COVID-19 X-ray dataset leads to unrealistic evaluation metrics. The third private dataset only included 306 extra COVID-19 patients, but these added images caused a drastic change to the results of their deep learning model.

Mangal et al. [23] have created a computer-aided detection (CAD) system for diagnosing COVID-19 based on a ChexNet model [31]. ChexNet first gained the attention of the research community because of its ability to diagnose 14 pulmonary pathologies. The model is designed using a DenseNet-121 architecture [14] and has been trained on over 100,000 X-rays. They created 3-class and 4-class models. Mangal et al. [23] validated their model using Gradient-weighted Class Activation Mappings (Grad-CAMs) [32]. A deficiency in this model was that it used a dataset from Kermany et al. [18] when making use of Paul Mooney's Chest X-ray dataset on Kaggle [24]. The dimensions of the lungs in these X-rays that were taken from children likely caused their final classifier to produce unpredictable results. Haghanifar et al. [11] improved on Mangal et al.'s [23] original design by including a segmentation unit with their ChexNet model. They constructed a different dataset than Mangal et al. [23] for training their ChexNet model. Hagnifar et al. [11] made the same mistake as Mangal et al. [23] in including Kermany et al.'s [18] dataset. Al-Waisy et al. [3] likewise published a paper using a ChexNet model that made the same mistake. The authors obtained an even more exaggerated set of performance metrics than the previous two models mentioned. Unfortunately, the use of Kermany et al.'s [18] dataset is widespread and this has created a major flaw in all of these ChexNet models.

Islam et al. [16] developed a novel CNN-LSTM model for diagnosing COVID-19 with chest X-rays. Their model was unique in terms of its architecture in the literature. During validation, they obtained accuracies, specificities, sensitivities, and F1-scores between 98 and 100% for all classes in their results. Their model seemed to report what looked like overly optimistic performance metrics. This suspicion was confirmed when it was noticed that their model reported using Kermany et al.'s dataset [18] (also referred to as the Kaggle chest X-ray repository in their article).

Rahimzadeh et al. [28] developed a deep learning model that combined the Xception [6] and ResNet-50 [12] models together. Two '10 × 10 × 2048 feature maps' [28] forming the last feature extractor layers of both models were concatenated to improve on the final results of each classifier. This novel architecture worked quite well and the authors additionally performed five-fold cross-validation to improve the robustness of their results. Overall the authors of this article achieved reasonable success with their model as they achieved an overall accuracy of 91.4% and sensitivity of 80.5%.

Panwar [27] et al. constructed and optimized a VGG-19 model with ImageNet weights to detect COVID-19 in suspected patients. Their model was trained both on x-ray and CT scans. Their models were all binary models and these models compared COVID-19 patients vs. normal patients, COVID-19 vs. pneumonia patients, and COVID-19 patients vs. non-COVID-19 patients. The authors also focused on generating Grad-CAM heatmaps to make sure they were picking up the features of COVID-19 in X-rays and CT scans. While their CT classifier's dataset is likely adequate, their dataset for comparing COVID-19 vs pneumonia patients had a source of bias as their X-ray pneumonia images were derived from Kermany et al.'s [18] dataset.

Afshar et al. [1] published a paper that utilized a unique deep learning approach to diagnosing COVID-19. While the vast majority of models in the literature use CNNs to detect COVID-19, Afshar et al.'s [1] model utilized Capsule Networks (CapsNets). CapsNets are alternative models that can better utilize the spatial information in images by using "routing by agreement" [1]. The capsules in these networks are thereby capable of reaching "a mutual agreement on the existence of the objects" [1] in an X-ray. Like previous teams mentioned before, the authors pretrained their COVID-CAPS model on 94,323 X-rays before fine-tuning the model to a smaller COVID-19 dataset. A deficiency we found in this work is that the authors included Kermany et al.'s dataset [18] when making use the Paul Mooney's Chest X-ray dataset [24] on Kaggle.

Karthik et al. [17] presented a unique CNN in their work, which used a Channel-Shuffled Dual-Branched (CSDB) CNN that is augmented with Distinctive Filter Learning (DFL). This unique architecture learns "custom filters within a single convolutional layer for identifying specific pneumonia classes." [17] They compared their model with a variety of standard CNNs and promisingly outperformed those CNNs after training them on the same dataset. Their dataset, unfortunately, contained a deficiency whereby the authors used bacterial pneumonia and pneumonia X-rays derived from Kermany et al.'s dataset [18].

3. Proposed network architecture

3.1. Dataset

An important aspect of developing a deep learning model in medical imaging begins with the data. The availability of X-ray images and metadata is important when considering the research directions for such a project. In our data-gathering stage, we found it difficult to find metadata accompanying X-ray images. There was an insufficient amount of metadata to assist with developing a practical COVID-19 diagnosis system. There were many publicly available datasets available, but in analyzing these datasets we found that many of them were incorrectly assembled. Many datasets on Kaggle and in various research papers used Kermany et al.'s [18] dataset. As previously mentioned, this dataset consists of chest X-rays from children between the ages of one and five years old. A child's lungs have different features than an adult's lungs and hence these datasets were taken out of consideration. We also found that the vast majority of publicly available datasets made no mention as to whether they divided their training and test sets by patient number. Most datasets incorporated COVID-19 X-rays harvested from medical research papers. In many of these datasets, multiple images from the same patient could be found. Wang et al.'s [37] 'COVIDx' dataset does not suffer from the same disadvantages. Wang et al. [37] split their training and test sets by patient number. Their COVIDx dataset is large in comparison with other datasets and is "comprised of a total of 13,975 CXR images across 13,870 patient cases" [37]. This dataset contains 358 COVID-19 images, 8066 normal images, and 5541 pneumonia images. The COVIDx dataset has been used by many other research teams and is currently a good benchmark for testing a new model's results with other papers. For these reasons, we decided to use the COVIDx dataset in our study.

We divided the COVIDx dataset into a 90% training set and 10% test set ratio. This allowed for a suitable number of COVID-19 examples in the training set given the extreme class imbalance in the COVIDx dataset. The multi-class training set, therefore, consisted of 258 COVID-19 patients, 7966 normal patients, and 5441 pneumonia patients. Ten percent of the dataset was leftover for validation, but within the test set, there was again a class imbalance. We, therefore, reduced the number of normal and pneumonia examples in the test set to match the number of COVID-19 examples. In doing so, we obtained a balanced test set for evaluating our model's performance. This three-class test set ultimately consisted of 100 COVID-19 examples, 100 normal examples, and 100 pneumonia examples. A binary classifier was also designed in this study which grouped pneumonia and normal images into a single category. Our two-class COVID-19 vs. non-COVID-19 X-ray classifier was constructed to compare our approach with other two-class studies. Our binary training set consisted therefore of 258 COVID-19 images and 13407 non-COVID-19 images. The binary classifier's test set consisted of 100 COVID-19 X-rays and 100 non-COVID-19 X-rays.

We first trained and tested our deep learning model on the aforementioned datasets but later went on to create another set of larger training sets. Given the small number of COVID-19 images available in the COVIDx dataset, we expanded the number of COVID-19 images in this dataset to examine possible overfitting. Previous studies [39,41] mention this specifically as a reason for reduced COVID-19 sensitivity in their work. We wanted to investigate if more COVID-19 images would create a significant correction to our classifier's COVID-19 sensitivity. This second training set we created started out with 517 COVID-19 images from the COVIDx5 [37] training set. This second training set also included 922 images from the MIDRC-RICORD-1C database [5] and 2474 images from the BIMCV dataset [15]. Our second training set, therefore, consisted of 3913 COVID-19 images, 7966 normal images, and 5441 pneumonia images. For binary classification, we also examined how well our model works with a training set of 3913 COVID-19 images and 13417 non-COVID-19 images. We kept the original test sets as a benchmark to test our system against our previously trained classifiers and Wang et al.'s published model [37]. Tables 1 and 2 shows the COVIDx training set dataset alongside our expanded training set as well as our shared test set.

3.2. System design

Both models in our study are designed with a DenseNet-121 [14] base feature extractor and trained on the ChestX-ray14 dataset [38]. The ChestX-ray14 dataset contains "112,120 frontal-view X-ray images of 30805 patients" [31]. This form of modality-specific transfer learning increases our model's ability to capture COVID-19 features. The DenseNet-121's earliest layers contain feature maps that have already been trained to pick up many of the tissues and patterns seen in chest X-ray images. Many architectural design options were investigated before finalizing a new architecture model based on a DenseNet-121 network. The proposed system architecture, COV-SNET network, has the following features. After loading our pretrained weights into the DenseNet-121 network we have added a dense layer with 128 units, a dropout layer with a dropout rate of 10%, and a 3-class softmax layer for multiclass classification. An illustration of our model can be observed in Fig. 2. For our binary classifier, we replaced the softmax layer with a dense layer containing a single sigmoid activation function. Table 3 shows a detailed layer by layer description of our model.

Prior to training our models, we noticed that a class imbalance existed that required correction. This mainly was due to the lack of COVID-19 X-rays publicly available. A weighted loss function was used during training to correct for this class imbalance. In addition to correcting for the class imbalance, our training required some necessary preprocessing steps. We used data augmentation methods during training to increase our model's capacity to generalize on new examples. For our multiclass models, we set image rotations to 15%, vertical/horizontal translations to 15%, image shearing to 15%, and random zooms to 15% when augmenting our training dataset. For our binary models, each of the aforementioned augmentation categories was set to 20%. In all of our models, we additionally used horizontal flips in our augmentation process. During training and testing, our batch size was set to 32. Using Kera's ImageDataGenerator class, we additionally normalized our training data so that the values in each batch had a mean of 0 and a standard deviation of 1.

The first step in training our COV-SNET models involved initially training the final layer alone. The last layer of each network was trained in TensorFlow 2.0 for 9 epochs. The Adam optimizer was used during this training. To increase the performance of our networks we unfroze all of the layers in our models for further training. For 6 epochs we left the Adam optimizer at its default learning rate. After 6 epochs we fixed the learning rate to 1×10^{-5} and trained each model until their peak sensitivities were reached. For the models trained on the COVIDx dataset alone this required 10 epochs. For the models trained on our larger training set, this took 13–14 epochs. Before unfreezing the layers in our model, we fixed the moving mean and moving variance of the batches in our model's batchnormalization layers. These batchnormalization parameters were fixed to the weights generated from training our model on the ChestX-ray14 dataset.

Table 1
Datasets - number of images in the multiclass training and test sets.

	COVID-19	Normal	Pneumonia
COVIDx Multiclass Training Set	258	7966	5451
Our Expanded Multiclass Training Set	3913	7966	5451
Multiclass COVIDx Test Set	100	100	100

Table 2
Datasets - number of images in the binary training and test sets.

	COVID-19	Non-COVID-19
COVIDx Binary Training Set	258	13417
Our Expanded Binary Training Set	3913	13417
Binary COVIDx Test Set	100	100

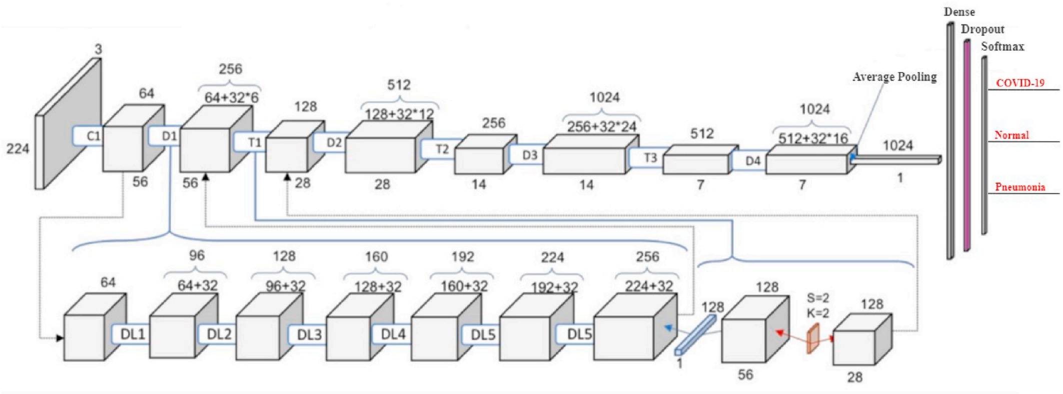


Fig. 2. Proposed network architecture for COVID-19 classification.

Table 3
Proposed network architecture for COVID-19 classification.

Layers	Output Size	Model
Convolution	112 × 112	7 × 7 conv, stride 2
Pooling	56 × 56	3 × 3 max pool, stride 2
Dense Block (1)	56 × 56	$\begin{bmatrix} 1x1 \text{ conv} \\ 3x3 \text{ conv} \end{bmatrix} x6$
Transition Layer (1)	56 × 56 28 × 28	1 × 1 conv 2 × 2 average pool, stride 2
Dense Block (2)	28 × 28	$\begin{bmatrix} 1x1 \text{ conv} \\ 3x3 \text{ conv} \end{bmatrix} x12$
Transition Layer (2)	28 × 28 14 × 14	1 × 1 conv 2 × 2 average pool, stride 2
Dense Block (3)	14 × 14	$\begin{bmatrix} 1x1 \text{ conv} \\ 3x3 \text{ conv} \end{bmatrix} x24$
Transition Layer (3)	14 × 14 7 × 7	1 × 1 conv 2 × 2 average pool, stride 2
Dense Block (4)	7 × 7	$\begin{bmatrix} 1x1 \text{ conv} \\ 3x3 \text{ conv} \end{bmatrix} x16$
Average Pooling	1 × 1	7 × 7 global average pool
DNN	–	128 units, relu
Dropout	–	10%
Classification	–	3 category softmax

4. Experimental results

4.1. Performance evaluation

The results reported in the COVID-19 deep learning literature are typically based on a variety of evaluation metrics. Accuracy, specificity, sensitivity, precision, recall, negative predictive value (NPV), positive predictive value (PPV), F1-Score, and area under the ROC curve (AUC) are all evaluation metrics used in the literature and included in our final results.

After training the last layer of each model for 9 epochs, the overall validation accuracy for each model was between 75 and 80%. While this was close to the performance of practicing radiologists in a previous study [39], we knew this result could be further improved upon by unfreezing layers in each model. After the models were unfrozen, all of the models achieved COVID-19 sensitivities of at least 95%. The entire set of class-wise performance statistics that were calculated for each classifier can be seen in Tables 4 – 7. Their corresponding confusion matrices can also be seen in Figs. 3–6. Our three-class model trained on the original COVIDx training set ultimately hit a final validation accuracy of 84.3%. Our 3-class model trained on our expanded training set obtained a validation accuracy of 86%. The final accuracy of the two-class model trained on the original COVIDx training set was 88.5%. The two-class model trained on our expanded training set obtained a validation accuracy of 87.5%. The AUC curves of all four of our models generated comparable results as can be seen in Figs. 7–8.

The evaluation metrics of a deep learning model should never alone be relied upon while validating the model's performance. Small datasets may only contain hundreds of images of the particular pathology under investigation. They tend to be prone to generating unrealistic evaluation metrics. To ensure a deep learning model is picking up correct features, saliency maps are widely employed in medical imaging. Saliency maps are important in that they can inform a designer whether a deep learning algorithm is being deceived by image characteristics that are unrelated to the pathology being imaged. Deep learning algorithms often incorrectly lock onto necklaces, medical devices, and text appearing in X-ray images. In our study, a Grad-CAM [32] was used to determine whether our COV-SNET model is fixing onto the correct features of COVID-19 in

Table 4
Three-class model performance metrics after training on the COVIDx multiclass training set.

	TP	TN	FP	FN	Acc.	Sens.	Spec.	PPV	NPV	F1
COVID-19	95	166	34	5	0.870	0.95	0.830	0.736	0.971	0.84
Normal	86	192	8	14	0.926	0.86	0.960	0.915	0.977	0.88
Pneumonia	72	195	5	28	0.890	0.72	0.975	0.935	0.874	0.82

Table 5
Two-class model performance metrics after training on the COVIDx binary training set.

	TP	TN	FP	FN	Acc.	Sens.	Spec.	PPV	NPV	F1
COVID-19	96	81	19	4	0.885	0.96	0.81	0.835	0.959	0.89
Non-COVID-19	81	96	4	19	0.885	0.81	0.96	0.953	0.835	0.876

Table 6
Three-class model performance metrics after training on our expanded multiclass training set.

	TP	TN	FP	FN	Acc.	Sens.	Spec.	PPV	NPV	F1
COVID-19	95	170	30	5	0.833	0.95	0.850	0.760	0.971	0.86
Normal	93	189	11	7	0.940	0.93	0.945	0.894	0.964	0.91
Pneumonia	70	199	1	30	0.897	0.70	0.995	0.989	0.869	0.82

Table 7
Two-class model performance metrics after training on our expanded binary training set.

	TP	TN	FP	FN	Acc.	Sens.	Spec.	PPV	NPV	F1
COVID-19	95	80	20	5	0.875	0.95	0.80	0.826	0.941	0.883
Non-COVID-19	80	95	5	20	0.875	0.80	0.95	0.941	0.826	0.865

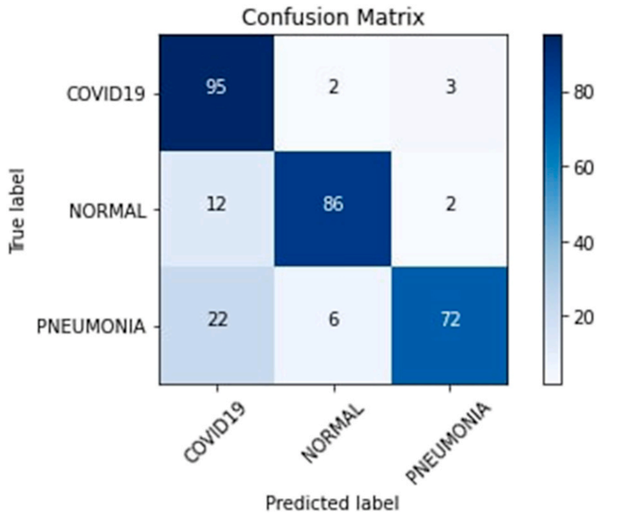


Fig. 3. Confusion matrix from three-class model after training on the COVIDx multiclass training set.

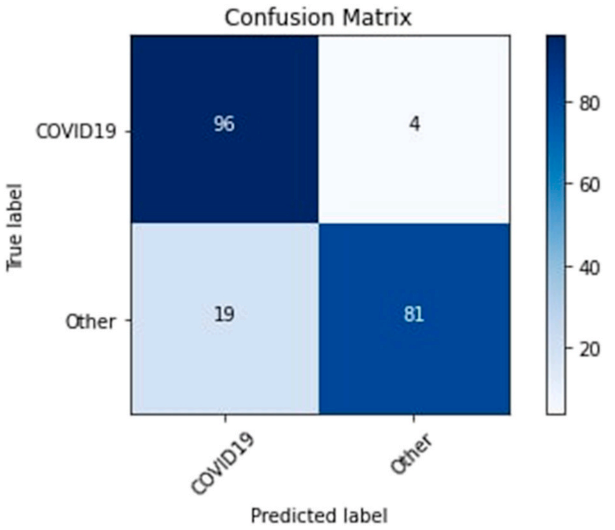


Fig. 4. Confusion matrix from two-class model after training on the COVIDx binary training set.

frontal chest X-rays. The heatmaps produced by a Grad-CAM contain color encoded information that highlights the features of an image that are the most relevant to a CNN's final classification. Fig. 9 shows the performance of our model on COVID-19 patients using Grad-CAM generated heatmaps. The red and orange regions of these Grad-CAM heatmaps are the most relevant parts of each image that contributed to a COVID-19 diagnosis in both patients. These colors transition into blue regions that are the least relevant portions of each image in contributing to our CNN's final classification. The Grad-CAM we employed uses the final feature maps in the last convolutional layers of our model to generate these regions of importance. As can be seen from our two examples, our Grad-CAM is locating the opacities in both images that would normally be picked by a radiologist when assessing these patients.

4.2. Discussion

All of our COV-SNET models achieved higher evaluation metrics than the consensus performance of the five radiologists in Wehbe et al.'s study [39] on a related dataset. While their dataset is not available publicly at this time, Wehbe et al.'s [39] study on the performance of five radiologists provides a good approximation for Bayes error. The best performing radiologist in Wehbe et al.'s [39] study only achieved an accuracy of 81% in diagnosing COVID-19 correctly. The best sensitivity among the radiologists was 76%. All of our models beat their best-performing radiologists by a substantial margin. Their work has been useful in that it provides designers with beneficial insights as to whether a deep learning model is providing reasonably grounded performance metrics. The consensus and best/worst performances of the five radiologists in Wehbe et al. [39] are provided in Table 8.

Many deep learning models in the literature report metrics that are superior to the performance of the radiologists in Wehbe et al.'s study [39]. Some papers report evaluation metrics that are superior to our own as well. What could be the reasons for this? Many papers have incorporated Kermany et al.'s [18] dataset. This dataset contains chest X-rays from children between the ages of one and five years old. The children in these chest X-rays are all suffering from various forms of bacterial and viral pneumonia. The extra categories in Kermany et al.'s [18] dataset were used as sources for comparison when diagnosing COVID-19 in other deep learning models. Many designers thought these extra

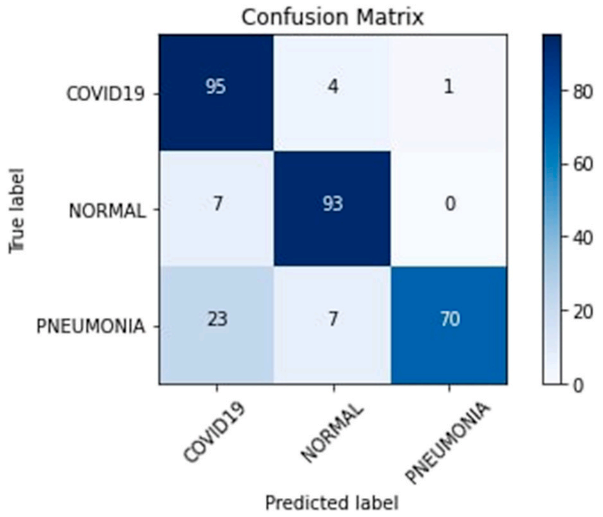


Fig. 5. Confusion matrix from three-class model after training on our expanded multiclass training set.

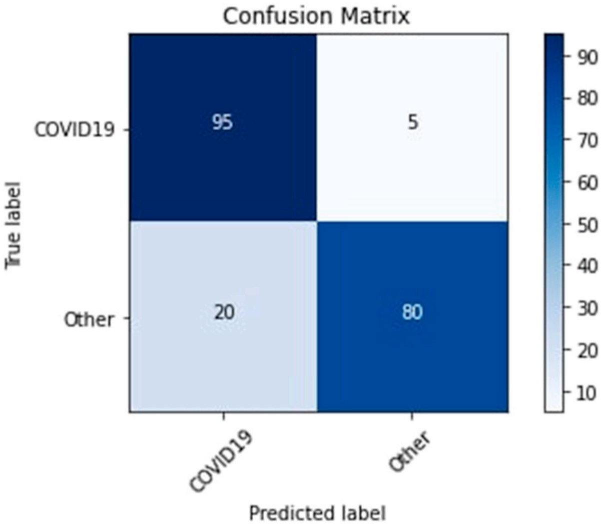


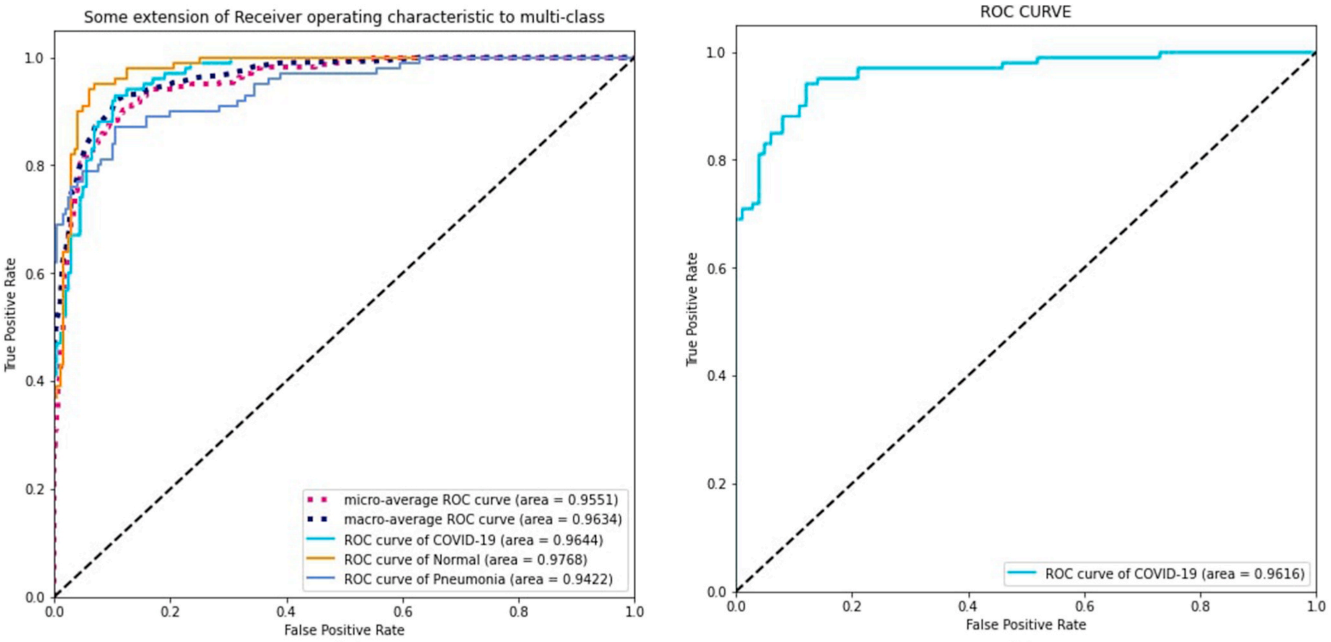
Fig. 6. Confusion matrix from two-class model after training on our expanded binary training set.

categories would be useful in clinical situations for ruling out other possible sources of infection. It is incorrect however to train a deep learning algorithm with children's lungs if that same algorithm will ultimately be deployed on adult lungs. Apostolopoulos and Mpesiana [4], Khalifa et al. [19], Waheed et al. [36], Rajaraman et al. [30], Haghanifar et al. [11], Mangal el al. [23], Al-Waisy et al. [3], and Islam et al. [16] all used Kermany et al.'s [18] dataset in their models. Many of those models reported exceedingly high-performance metrics. To the best of our knowledge there is only one other deep learning model in the existing literature that uses a COVID-19 dataset as large as our own and at the same time does not make the mistake of using Kermany et al.'s [18] dataset. That model was published by Wehbe et al. [39] and they ultimately only achieved a COVID-19 sensitivity of 75%. There is still a need therefore to explore whether a deep learning model can achieve a higher COVID-19 sensitivity while using a larger training set than has commonly been available to past authors. A correctly constructed dataset is required to perform this research. Prior to expanding Wang et al.'s [37] COVIDx dataset, we attempted to use public datasets that incorporated Kermany et al.'s dataset [18]. We trained a DenseNet-121, a DenseNet-201, and an Inception V3 architecture on these datasets. In doing so, we obtained suspiciously high-performance metrics and obtained accuracies between 98.0 and 99.6% on three-class and two-class models respectively. These performance metrics mirrored the performance metrics we have found in other studies that made the same mistake. Table 9 illustrates our point. It compares the performance of the radiologists in Wehbe et al.'s [39] study with other DenseNet-based models we have reviewed from the COVID-19 deep learning literature.

There are other possible reasons for the deep learning models in other studies to be generating unrealistic performance metrics. Many public datasets on Kaggle and various other platforms do not specifically state whether they have divided their training and test sets by patient number. If there has been cross-contamination between a deep learning model's training and test sets, there is a high probability that the trained model will have a better knowledge of the features in the test set. This data leakage leads to unrealistic performance metrics. The X-ray files in public datasets are often renamed and their original source information in many instances is lost. Many papers have combined several public datasets. They often have done so without making any mention as to how they ensured the same images from different datasets were not duplicated in their own dataset. The datasets in some papers are also difficult to reconstruct and it is challenging to trace the chain of images that ended up being included in some datasets. These are all likely factors that are contributing to the high-performance metrics of some studies which are far outside of the performance range of practicing expert radiologists. We decided to use Wang et al.'s [37] 'COVIDx' dataset because the designers of that dataset took into account these issues being discussed. The dataset, therefore, is more conservative and grounded compared to other online public datasets.

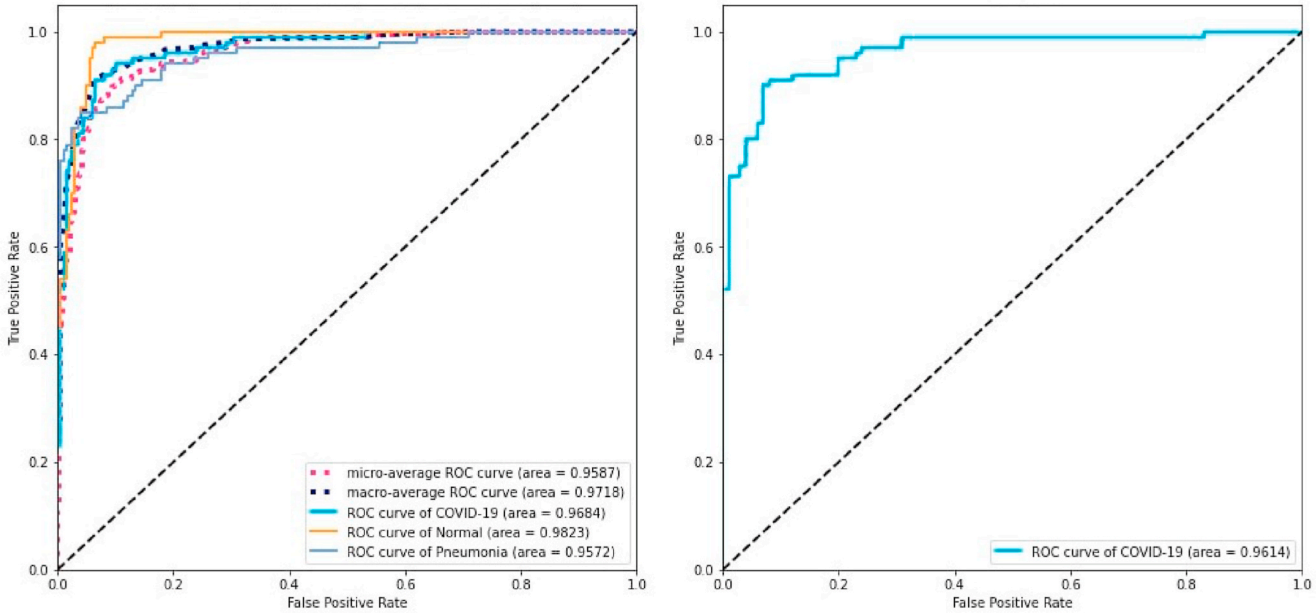
It should now be clear that the composition of the datasets used to train deep learning COVID-19 models is one of the main contributing factors to the high evaluation metrics often being reported in the literature. There is however another crucial factor that is contributing to these unrealistic evaluation metrics. Many datasets in the COVID-19 X-ray imaging literature do not have a sufficient number of COVID-19 images. This lack of COVID-19 X-ray images in medical datasets can sometimes lead to unpredictable results. When more images are added there can be a correction in a system's evaluation metrics towards the performance reported by practicing experts in the field. This is precisely what happened in Yeh et al.'s [41] study. The work in Ref. [41] commenced with using an earlier version of the COVIDx dataset. The authors of the study also initially used the private X-ray images of two medical institutions. When the authors trained a DenseNet-121 classifier on these initial datasets alone they achieved a COVID-19 sensitivity of 96.8%. This did not last however and the inclusion of a third medical institution's COVID-19 X-rays in their model's training caused a correction in its evaluation metrics. This led their model to have a final COVID-19 sensitivity of 81.82%.

Yeh et al.'s [41] final dataset contained 510 COVID-19 images. The COVIDx dataset we used had 358 COVID-19 images. Our original three-class model, therefore, contained only 70% of the number COVID-19 images that Yeh et al.'s [41] model initially trained on. Our three-class model generated a COVID-19 sensitivity of 95%. Yeh-et al.'s [41] three class-model obtained a final COVID-19 sensitivity of 81.82%. Wang et al.'s [37] three-class model used the same original dataset as ours and obtained a COVID-19 sensitivity of 91%. How do we know however that our 95% sensitivity would not correct if we trained on more COVID-19 images? After all, there are some in the research community [39] that have pointed out that overfitting is occurring in past models trained on small COVID-19 datasets. Recently a large number of COVID-19 images have become available that are independent of previous COVID-19 datasets. This led us to create an expanded dataset from the original COVIDx dataset that we used to check for overfitting. After further examination, we discovered that our evaluation metrics were not impacted by training our model on the expanded COVID-19 dataset. We were able to maintain the same COVID-19 sensitivity (95%) using this dataset on our three-class model.



(a) (b)

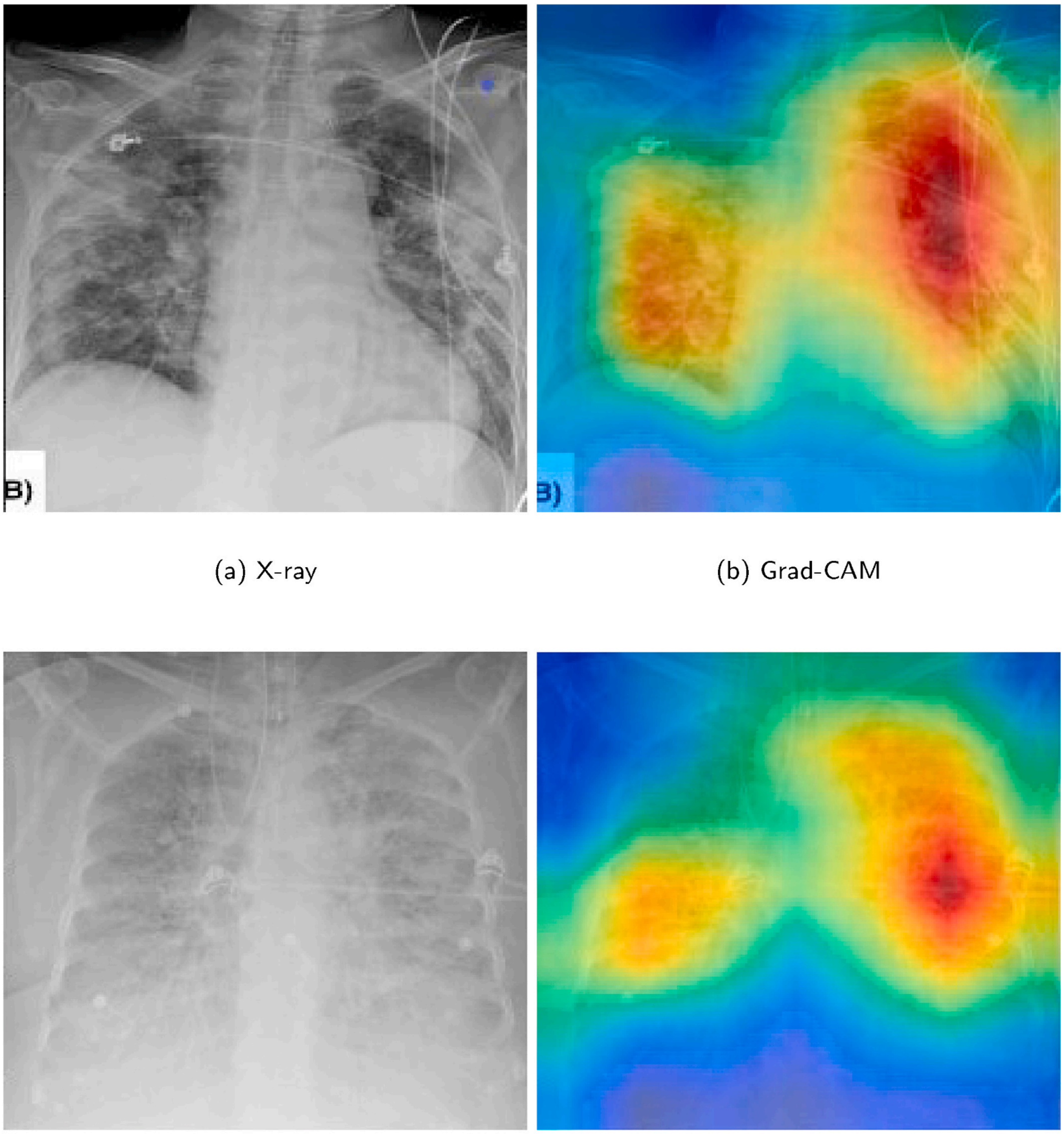
Fig. 7. ROC AUC graphs of (a) Three-class model trained on COVIDx multiclass training set and (b) Two-class model trained on COVIDx binary training set.



(a) (b)

Fig. 8. ROC AUC graphs of (a) Three-class model trained on our expanded training set and (b) Two-class model trained on our expanded training set.

We thereafter moved on to creating a two-class model with the same expanded dataset. Our original two-class model generated a COVID-19 sensitivity of 96%. After training this model on our expanded dataset we obtained a COVID-19 sensitivity of 95%. Wehbe et al.'s [39] two-class COVID-19 model obtained a COVID-19 sensitivity of 75%. Their ensemble model however was trained on a slightly larger dataset than ours. Their dataset contains 4253 COVID-19 images. They showed in their paper that their model's sensitivity (75%) was better than the consensus performance of the five radiologists in their study. They also argued that the high sensitivities of deep learning models presented in other studies were caused by a lack of COVID-19 images in publicly available datasets. We wrote earlier that this was indeed the case in Yeh et al.'s [41] study, but have been able to prove that it is not the case in our study. Expanding the COVIDx dataset did not significantly affect the performance of our classifier. Of all of the studies that do not improperly use Kermany et al.'s [18] dataset, our models achieve the highest sensitivities that we can find in the literature. Table 10 presents a comparison of the sensitivities among models that do not have any issues regarding dataset composition. Out of the papers in Table 10, we were able to only make a direct comparison of our work with Wang et al.'s



(a) X-ray

(b) Grad-CAM

(c) X-ray

(d) Grad-CAM

Fig. 9. Two different COVID-19 patients showing their original X-rays alongside their Grad-CAM produced heatmaps.

Table 8
Performance of five radiologists in diagnosing COVID-19 with X-rays [39].

	Acc.	Sens.	Spec.
Consensus	81%	70%	89%
Best Radiologist	81%	76%	91%
Worst Radiologist	76%	60%	75%

[37] COVID-Net model. Our models ultimately required different augmentation settings than theirs in order to achieve optimal results. Unfortunately, we were unable to replicate the other datasets in Table 10. A couple of the papers in Table 10 mention that their datasets are private. Wehbe et al. [39] currently have the largest COVID-19 dataset that we have found in the literature, but unfortunately, it's entirely private. We have however been able to assemble a dataset that is now much closer in size to Wehbe et al.'s [39] private COVID-19 dataset. In doing so, we have been able to prove that deep learning models are capable of obtaining higher COVID-19 sensitivities than has previously been reported.

Table 9
Performance of past DenseNet-Based models versus radiologists.

Paper Reviewed	F1	ACC	COVID-19 Sens.
Yeh et al. [41]			
3-class	–	–	81.82%
Haghanifar et al. [11]			
2-class	94%	98.62%	–
3-class	85%	81.04%	–
Mangal et al. [23]			
3-class	92.3%	90.5%	100%
Al-Waisy et al. [3]			
2-class	99.99%	99.99%	99.98%
Rajaraman et al. [30]			
4-class	96.77%	96.83%	96.34%
Radiologists [39]			
2-class	–	81%	70%

Note: Haghanifar et al. [11], Mangal et al. [23], Al-Waisy et al. [3], and Rajaraman et al. [30] all improperly used Kermany et al.'s dataset [18].

Table 10
Performance of papers without dataset composition issues.

Research Paper	COVID-19 Sens.
Yeh et al. [41]	
3-class	81.82%
Wang et al. [37]	
3-class	91%
Wehbe et al. [39]	
2-class	75%
Rahimzadeh et al. [28]	
3-class	80.53%
Ours	
2-class	95%
3-class	95%

Note: These papers all do not include Kermany et al.'s dataset [18].

5. Conclusion

Deep learning models trained on public datasets prior to early 2021 all have experienced dataset size limitations in terms of the number of COVID-19 X-rays available. This has made all of these models susceptible to possible overfitting. There are now however several thousand COVID-19 X-rays publicly available. This recent surge in available COVID-19 X-rays allows for past authors still working in this space to check and see if their models will ultimately correct when training with a larger dataset. At the beginning of our research, we started out hoping to benchmark our study against a popular dataset. This led us to use Wang et al.'s [37] COVIDx dataset. We ended up achieving a higher sensitivity than their model and went on to train our model with more COVID-19 images. This ultimately allowed us to ensure that our model was not overfitting on a dataset containing only a limited number of COVID-19 images. While the extra layers we added to our pretrained model increased its overall performance, the dropout layer near the end of the model additionally helped it to avoid overfitting. The data augmentation techniques we employed while training our model also improved its performance metrics and prevented overfitting. Unfreezing our pretrained model in a fashion that does not upset key batch normalization parameters also ultimately allowed our models to achieve high COVID-19 sensitivities. Our models are currently capable of obtaining a higher COVID-19 sensitivity than all other models that we have reviewed in the literature so far. We have restricted this analysis to those models that do not improperly use Kermany et al.'s [18] dataset or otherwise make any observable dataset composition mistakes.

The models constructed in this study led to promising evaluation metrics in comparison with expert radiologists in the field [39]. We achieved two-class and three-class COVID-19 sensitivities of 95%. There is room to improve on the design of our two models. In future datasets, metadata may be included alongside new COVID-19 X-ray images. Extra information regarding a patient's sex, age, blood work, temperature, and exposure history may help to increase the accuracy of COVID-19 diagnostic models. In addition to metadata, adding a segmentation unit would assist with generating better evaluation metrics and Grad-CAM heatmaps. While the results of the two models presented in this study look promising, more work is required to implement them in a clinical setting. The same can be said for Yeh et al. [41] and Wang et al.'s [37] models which are also based on different versions of the COVIDx dataset. The addition of more COVID-19 images to public databases will no doubt help to further inform the research community as to which approaches are the most promising. Medical institutions in countries all over the world are in need of new diagnostic modalities that can help increase available COVID-19 testing capacity. Deep learning X-ray technology remains a promising candidate for fulfilling this incredibly important need.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We would like to thank Linda Wang, Zhong Qiu Lin, and Alexander Wong for making their code and dataset available. We have used their dataset generator in preparing part of our dataset. We would also like to thank the Medical Imaging and Data Resource Center (MIDRC) and the Medical Imaging Databank of the Valencia Region (BIMCV) for making available COVID-19 images that we used in this study.

References

[1] Afshar P, Heidarian S, Naderkhani F, Oikonomou A, Plataniotis KN, Mohammadi A. Covid-caps: a capsule network-based framework for identification of covid-19 cases from x-ray images. Pattern Recogn Lett 2020;138:638–43. https://doi.org/10.1016/j.patrec.2020.09.010.

[2] Ai T, Yang Z, Hou H, Zhan C, Chen C, Lv W, et al. Correlation of chest ct and rt-pcr testing for coronavirus disease 2019 (covid-19) in China: a report of 1014 cases. Radiology 2020;296:E32–40. https://doi.org/10.1148/radiol.2020200642. pMID: 32101510.

[3] Al-Waisy AS, Al-Fahdawi S, Mohammed MA, Abdulkareem KH, Mostafa SA, Maashi MS, et al. Covid-chexnet: hybrid deep learning framework for identifying covid-19 virus in chest x-rays images. Soft Computing 2020. https://doi.org/10.1007/s00500-020-05424-3.

[4] Apostolopoulos I, Tzani M. Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks. Australas Phys Eng Sci Med/supported by the Australasian College of Physical Scientists in Medicine and the Australasian Association of Physical Sciences in Medicine 2020;43:635–40. https://doi.org/10.1007/s13246-020-00865-4.

[5] Bilello E. Medical imaging data resource center (midrc) - rsna international covid-19 open radiology database (ricord) release 1c - chest x-ray covid+ (midrc-ricord-1c). https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=70230281; 2021.

[6] Chollet F. Xception: deep learning with depthwise separable convolutions. arXiv URL, http://arxiv.org/abs/1610.02357; 2016.

[7] Cleverley J, Piper J, Jones MM. The role of chest radiography in confirming covid-19 pneumonia. BMJ 2020;370. https://doi.org/10.1136/bmj.m2426.

[8] Cozzi D, Albanesi M, Cavigli E, Moroni C, Bindi A, Luvarà S, et al. Chest x-ray in new coronavirus disease 2019 (covid-19) infection: findings and correlation with clinical outcome. La radiologia medica 2020;125. https://doi.org/10.1007/s11547-020-01232-9.

[9] Fang Y, Zhang H, Xie J, Lin M, Ying L, Pang P, et al. Sensitivity of chest ct for covid-19: comparison to rt-pcr. Radiology 2020;296:E115–7. https://doi.org/10.1148/radiol.2020200432. pMID: 32073353.

[10] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, et al. Generative adversarial networks. arXiv URL, https://arxiv.org/abs/1406.2661; 2014.

[11] Haghanifar A, Majdabadi MM, Ko S. Covid-cxnet: detecting covid-19 in frontal chest x-ray images using deep learning. arXiv URL, https://arxiv.org/abs/2006.13807; 2020.

[12] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. arXiv URL, https://arxiv.org/abs/1512.03385; 2015.

[13] Hemdan EED, Shouman MA, Karar ME. Covidx-net: a framework of deep learning classifiers to diagnose covid-19 in x-ray images. arXiv URL, https://arxiv.org/abs/2003.11055; 2020.

[14] Huang G, Liu Z, Weinberger KQ. Densely connected convolutional networks. arXiv URL, https://arxiv.org/abs/1608.06993; 2016.

[15] de la Iglesia Vayá M, Saborit JM, Montell JA, Pertusa A, Bustos A, Cazorla M, et al. Bimcv-covid19, datasets related to covid19's pathology course. https://bimcv.cipf.es/bimcv-projects/bimcv-covid19/; 2020.

[16] Islam MZ, Islam MM, Asraf A. A combined deep cnn-lstm network for the detection of novel coronavirus (covid-19) using x-ray images. Informatics in Medicine Unlocked 2020;20:100412. https://doi.org/10.1016/j.imu.2020.100412.

[17] Karthik R, Menaka R, H M. Learning distinctive filters for covid-19 detection from chest x-ray using shuffled residual cnn. Appl Soft Comput 2021;99:106744. https://doi.org/10.1016/j.asoc.2020.106744.

[18] Kermany DS, Goldbaum M, Cai W, Valentim CC, Liang H, Baxter SL, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell 2018;172. https://doi.org/10.1016/j.cell.2018.02.010. 1122 – 1131.e9.

[19] Khalifa NEM, Taha MHN, Hassanien AE, Elghamrawy S. Detection of coronavirus (covid-19) associated pneumonia based on generative adversarial networks and a fine-tuned deep transfer learning model using chest x-ray dataset. arXiv URL, https://arxiv.org/abs/2004.01184; 2020.

[20] Kucirka L, Lauer S, Laeyendecker O, Boon D, Lessler J. Variation in false-negative rate of reverse transcriptase polymerase chain reaction–based sars-cov-2 tests by time since exposure. Ann Intern Med 2020;173:262–7. https://doi.org/10.7326/M20-1495. pMID: 32422057.

[21] Liang T, Liu Z, Wu C, Jin C, Zhao H, Wang Y, et al. Evolution of ct findings in patients with mild covid-19 pneumonia. Eur Radiol 2020;30:4865–73. https://doi.org/10.1007/s00330-020-06823-8.

[22] Lin ZQ, Shafiee MJ, Bochkarev S, Jules MS, Wang X, Wong A. Do explanations reflect decisions? A machine-centric strategy to quantify the performance of explainability algorithms. CoRR; 2019. http://arxiv.org/abs/1910.07387.

[23] Mangal A, Kalia S, Rajgopal H, Rangarajan K, Namboodiri V, Banerjee S, et al. Covid aid: covid-19 detection using chest x-ray. arXiv URL, https://arxiv.org/abs/2004.09803; 2020.

[24] Mooney P. Chest x-ray images (pneumonia). Kaggle; 2018. https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia.

[25] Odena A, Olah C, Shlens J. Conditional image synthesis with auxiliary classifier GANs. In: Precup D, Teh YW, editors. Proceedings of the 34th international conference on machine learning, PMLR. Sydney, Australia: International Convention Centre; 2017. p. 2642–51. http://proceedings.mlr.press/v70/odena17a.html.

[26] Pan Y, Li X, Yang G, Fan J, Tang Y, Zhao J, et al. Serological immunochromatographic approach in diagnosis with sars-cov-2 infected covid-19 patients. J Infect 2020;81. https://doi.org/10.1016/j.jinf.2020.03.051. e28 – e32.

[27] Panwar H, Gupta P, Siddiqui MK, Morales-Menendez R, Bhardwaj P, Singh V. A deep learning and grad-cam based color visualization approach for fast detection of covid-19 cases using chest x-ray and ct-scan images. Chaos, Solit Fractals 2020; 140:110190. https://doi.org/10.1016/j.chaos.2020.110190.

[28] Rahimzadeh M, Attar A. A modified deep convolutional neural network for detecting covid-19 and pneumonia from chest x-ray images based on the concatenation of xception and resnet50v2. Informatics in Medicine Unlocked 2020; 19:100360. https://doi.org/10.1016/j.imu.2020.100360.

[29] Rahman T. Covid-19 radiography database. Kaggle; 2020. https://www.kaggle.com/tawsifurrahman/covid19-radiography-database.

[30] Rajaraman S, Siegelman J, Alderson PO, Folio LS, Folio LR, Antani SK. Iteratively pruned deep learning ensembles for covid-19 detection in chest x-rays. IEEE Access 2020;8:115041–50. https://doi.org/10.1109/ACCESS.2020.3003810.

[31] Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, et al. Chexnet: radiologist-level pneumonia detection on chest x-rays with deep learning. https://arxiv.org/abs/1711.05225; 2017.

[32] Selvaraju RR, Das A, Vedantam R, Cogswell M, Parikh D, Batra D. Grad-cam: why did you say that? visual explanations from deep networks via gradient-based localization. CoRR; 2016. http://arxiv.org/abs/1610.02391.

[33] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv URL, https://arxiv.org/abs/1409.1556; 2014.

[34] Song F, Shi N, Shan F, Zhang Z, Shen J, Lu H, et al. Emerging 2019 novel coronavirus (2019-ncov) pneumonia. Radiology 2020;295:210–7. https://doi.org/10.1148/radiol.2020200274. pMID: 32027573.

[35] Szegedy C, Liu W, Jia Y, Sermanet P, Reed SE, Anguelov D, et al. Going deeper with convolutions. arXiv URL, http://arxiv.org/abs/1409.4842; 2014.

[36] Waheed A, Goyal M, Gupta D, Khanna A, Al-Turjman F, Pinheiro PR. Covidgan: data augmentation using auxiliary classifier gan for improved covid-19 detection. IEEE Access 2020;8:91916–23. https://doi.org/10.1109/ACCESS.2020.2994762.

[37] Wang L, Lin ZQ, Wong A. Covid-net: a tailored deep convolutional neural network design for detection of covid-19 cases from chest radiography images. Sci Rep 2020;10. https://doi.org/10.1038/s41598-020-76550-z.

[38] 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. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR); 2017. p. 3462–71. https://doi.org/10.1109/CVPR.2017.369.

[39] Wehbe, R.M., Sheng, J., Dutta, S., Chai, S., Dravid, A., Barutcu, S., et al., 0. Deepcovid-xr: an artificial intelligence algorithm to detect covid-19 on chest radiographs trained and tested on a large us clinical dataset. Radiology 0, 203511. doi:10.1148/radiol.2020203511. pMID: 33231531.

[40] Wong A, Shafiee MJ, Chwyl B, Li F. Ferminets: learning generative machines to generate efficient neural networks via generative synthesis. CoRR; 2018. URL, http://arxiv.org/abs/1809.05989.

[41] Yeh CF, Cheng HT, Wei A, Chen HM, Kuo PC, Liu KC, et al. A cascaded learning strategy for robust covid-19 pneumonia chest x-ray screening. arXiv URL, https://arxiv.org/abs/2004.12786; 2020.

Robert Hertel received the B.S. degree in electrical and computer engineering from Lakehead University, Canada, in 2016. He is currently pursuing an MSc degree in electrical and computer engineering at Lakehead University. Between 2014 and 2016, he worked as an Intern in ZTR Control System's research and development department. From 2017 to 2019 he worked as a Control Systems Engineer with JMP Controls Systems. His research interest includes developing deep learning techniques in computer vision applications, researching industrial controls systems for manufacturing vehicles, designing control systems for monitoring wastewater, and developing new machine learning algorithms in discovering the diagnosis and prognosis of individuals suffering from infectious diseases.

Mr. Hertel was a recipient of the Alexander Graham Bell Canada Graduate Scholarship in 2020.

Rachid Benlamri is a Professor of Software Engineering at Lakehead University - Canada. He received his Master's degree and a Ph.D. in Computer Science from the University of Manchester - UK in 1987 and 1990, respectively. He is the head of the Artificial Intelligence and Data Science Lab at Lakehead University. He served as keynote speaker and general chair for many international conferences. Professor Benlamri is a member of the editorial board for many referred international journals. His research interests are in the areas of Artificial Intelligence, Semantic Web, Data Science, Ubiquitous Computing and Knowledge Engineering.