Medical Image Analysis 65 (2020) 101794

Contents lists available at ScienceDirect

Medical Image Analysis

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

Deep-COVID: Predicting COVID-19 from chest X-ray images using deep transfer learning

Shervin Minaee[a,*], Rahele Kafieh[b,**], Milan Sonka[c], Shakib Yazdani[d], Ghazaleh Jamalipour Soufi[e]

[a] Snap Inc., Seattle, WA, USA
[b] Medical Image and Signal Processing Research Center, Isfahan University of Medical Sciences, Iran
[c] Iowa Institute for Biomedical Imaging, The University of Iowa, Iowa City, USA
[d] ECE Department, Isfahan University of Technology, Iran
[e] Radiology Department, Isfahan University of Medical Sciences, Isfahan, Iran

ARTICLE INFO

Article history:
Received 24 April 2020
Revised 9 July 2020
Accepted 17 July 2020
Available online 21 July 2020

Keywords:
COVID-19
X-ray imaging
Deep learning
Transfer learning

ABSTRACT

The COVID-19 pandemic is causing a major outbreak in more than 150 countries around the world, having a severe impact on the health and life of many people globally. One of the crucial step in fighting COVID-19 is the ability to detect the infected patients early enough, and put them under special care. Detecting this disease from radiography and radiology images is perhaps one of the fastest ways to diagnose the patients. Some of the early studies showed specific abnormalities in the chest radiograms of patients infected with COVID-19. Inspired by earlier works, we study the application of deep learning models to detect COVID-19 patients from their chest radiography images. We first prepare a dataset of 5000 Chest X-rays from the publicly available datasets. Images exhibiting COVID-19 disease presence were identified by board-certified radiologist. Transfer learning on a subset of 2000 radiograms was used to train four popular convolutional neural networks, including ResNet18, ResNet50, SqueezeNet, and DenseNet-121, to identify COVID-19 disease in the analyzed chest X-ray images. We evaluated these models on the remaining 3000 images, and most of these networks achieved a sensitivity rate of 98% ($\pm$ 3%), while having a specificity rate of around 90%. Besides sensitivity and specificity rates, we also present the receiver operating characteristic (ROC) curve, precision-recall curve, average prediction, and confusion matrix of each model. We also used a technique to generate heatmaps of lung regions potentially infected by COVID-19 and show that the generated heatmaps contain most of the infected areas annotated by our board certified radiologist. While the achieved performance is very encouraging, further analysis is required on a larger set of COVID-19 images, to have a more reliable estimation of accuracy rates. The dataset, model implementations (in PyTorch), and evaluations, are all made publicly available for research community at https://github.com/shervinmin/DeepCovid.git

1. Introduction

Since December 2019, a novel corona-virus (SARS-CoV-2) has spread from Wuhan to the whole China, and many other countries. By April 18, more than 2 million confirmed cases, and more than 150,000 deaths were reported in the world (https://www.worldometers.info/coronavirus/). Due to unavailability of therapeutic treatment or vaccine for novel COVID-19 disease, early diagnosis is of real importance to provide the opportunity of immediate isolation of the suspected person and to decrease the chance of infection to healthy population. Reverse transcription polymerase chain reaction (RT-PCR) or gene sequencing for respiratory or blood specimens are introduced as main screening methods for COVID-19 (Wang et al., 2020). However, total positive rate of RT-PCR for throat swab samples is reported to be 30 to 60%, which accordingly yields to un-diagnosed patients, which may contagiously infect a huge population of healthy people (Yang et al., 2020). Chest radiography imaging (e.g., X-ray or computed tomography (CT) imaging) as a routine tool for pneumonia diagnosis is easy to perform

* Corresponding author.
** Corresponding author.
E-mail addresses: sminaee@snap.com (S. Minaee), rkafieh@amt.mui.ac.ir (R. Kafieh).

https://doi.org/10.1016/j.media.2020.101794
1361-8415/

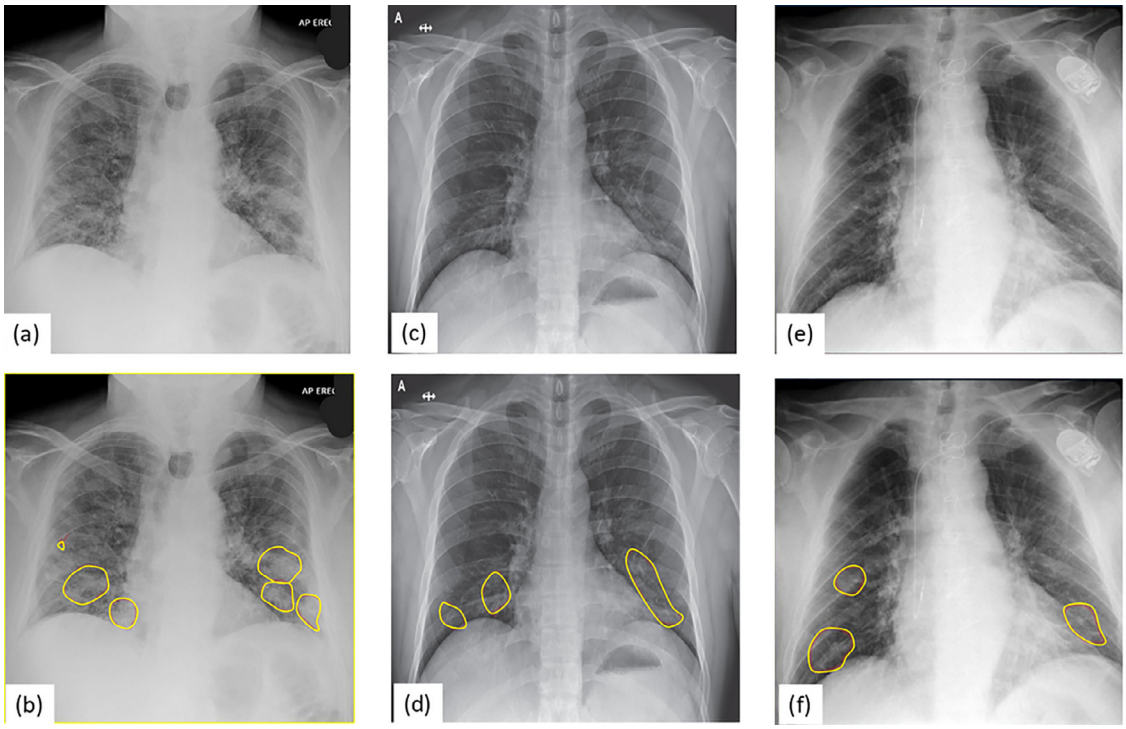

Fig. 1. Three sample COVID-19 images, and the corresponding marked areas by our radiologist.

with fast diagnosis. Chest CT has a high sensitivity for diagnosis of COVID-19 (Ai et al., 2020) and X-ray images show visual indexes correlated with COVID-19 (Kanne et al., 2020). The reports of chest imaging demonstrated multilobar involvement and peripheral airspace opacities. The opacities most frequently reported are ground-glass (57%) and mixed attenuation (29%) (Kong and Agarwal, 2020). During the early course of COVID-19, ground glass pattern is seen in areas that edges the pulmonary vessels and may be difficult to appreciate visually (Hansell et al., 2008). Asymmetric patchy or diffuse airspace opacities are also reported for COVID-19 (Rodrigues, 2020). Such subtle abnormalities can only be interpreted by expert radiologists. Considering huge rate of suspected people and limited number of trained radiologists, automatic methods for identification of such subtle abnormalities can assist the diagnosis procedure and increase the rate of early diagnosis with high accuracy. Artificial intelligence (AI)/machine learning solutions are potentially powerful tools for solving such problems.

So far, due to the lack of availability of public images of COVID-19 patients, detailed studies reporting solutions for automatic detection of COVID-19 from X-ray (or Chest CT) images are not available. Recently a small dataset of COVID-19 X-ray images was collected, which made it possible for AI researchers to train machine learning models to perform automatic COVID-19 diagnostics from X-ray images (Cohen et al., 2020). These images were extracted from academic publications reporting the results on COVID-19 X-ray and CT images. With the help of a board-certified radiologist, we re-labeled those images, and only kept ones a clear sign of COVID-19 as determined by our radiologist. Three sample images with their corresponding marked areas are shown in Fig. 1. We then used a subset of images from ChexPert (Irvin et al., 2019) dataset, as the negative samples for COVID-19 detection. The combined dataset has around 5000 Chest X-ray images (called COVID-Xray-5k), which is divided into 2000 training, and 3000 testing samples.

A machine a learning framework was employed to predict COVID-19 from Chest X-ray images. Unlike the classical approaches for medical image classification which follow a two-step procedure (hand-crafted feature extraction+recognition), we use an end-to-end deep learning framework which directly predicts the COVID-19 disease from raw images without any need of feature extraction. Deep learning based models (and more specifically convolutional neural networks (CNN)) have been shown to outperform the classical AI approaches in most of computer vision and and medical image analysis tasks in recent years, and have been used in a wide range of problems from classification, segmentation, face recognition, to super-resolution and image enhancement (LeCun, 1998; Badrinarayanan et al., 2017; Ren et al., 2015; Dong, 2014; Minaee et al., 2019).

Here, we train 4 popular convolutional networks which have achieved promising results in several tasks during recent years (including ResNet18, ResNet50, SqueezeNet, and DenseNet-161) on COVID-Xray-5k dataset, and analyze their performance for COVID-19 detection. Since so far there is a limited number of X-ray images publicly available for the COVID-19 class, we cannot simply train these models from scratch. Two strategies were adopted to address the COVID-19 image scarcity issue in this work:

- We use data augmentation to create transformed version of COVID-19 images (such as flipping, small rotation, adding small amount of distortions), to increase the number of samples by a factor of 5.
- Instead of training these models from scratch, we fine-tune the last layer of the pre-trained version of these models on ImageNet. In this way, the model can be trained with less labeled samples from each class.

The above two strategies helped train these networks with the available images, and achieve reasonable performance on the test set of 3000 images. Since the number of samples for the COVID-19 class is limited, we also calculate the confidence interval of the performance metrics. To report a summarizing performance

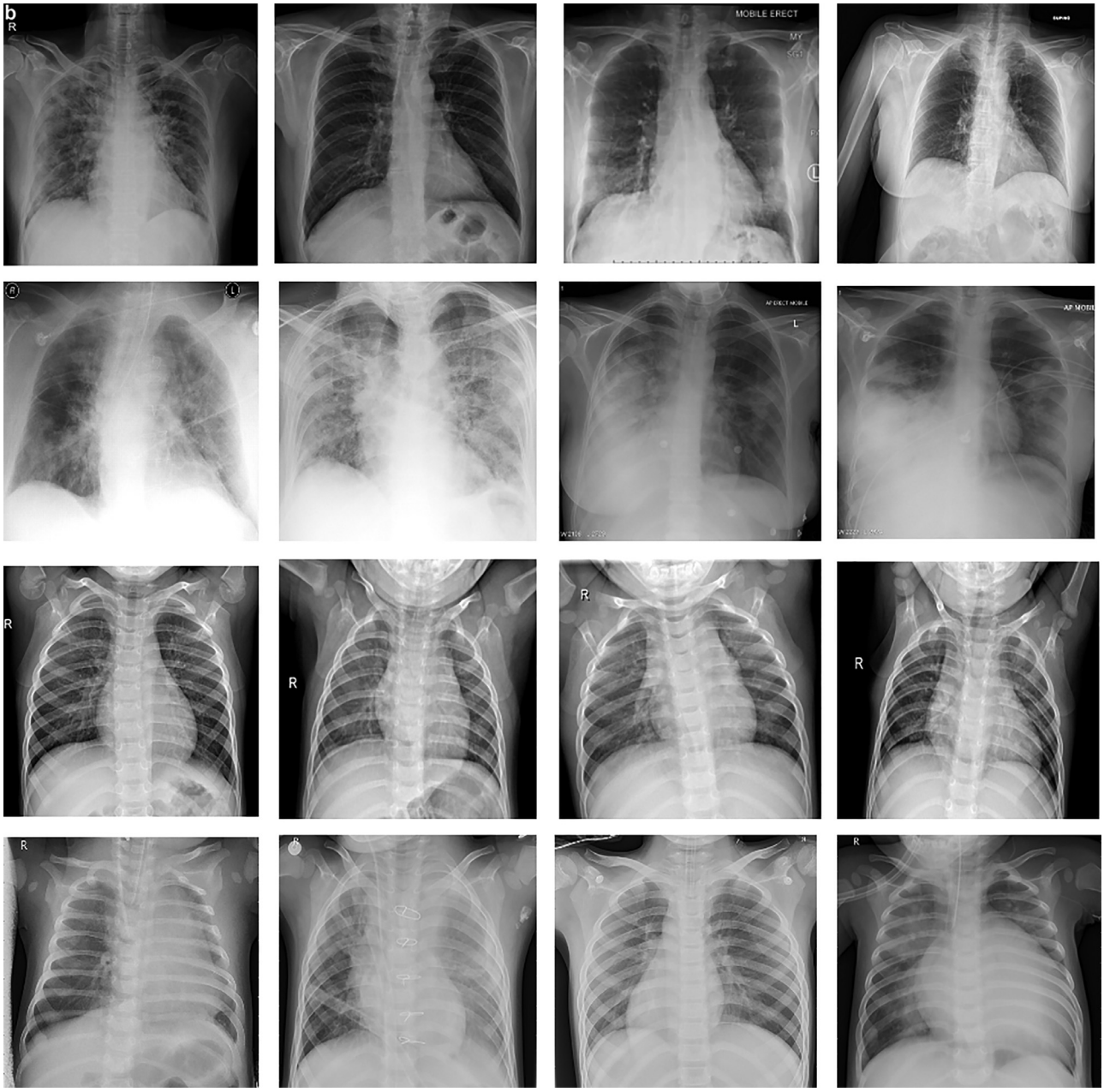

Fig. 2. Sample images from COVID-Xray-5k dataset. The images in the first row show 4 COVID-19 images. The images in the second row are 4 sample images of no-finding category in Non-COVID images from **ChexPert**. The images in the third and fourth rows give 8 sample images from other sub-categotries in **ChexPert**.

of these models, we provide the Receiver operating characteristic (ROC) curve, and area under the curve (AUC) for each of these models.

Here are the main contributions of this paper:

- We prepared a dataset of 5000 images with binary labels, for COVID-19 detection from Chest X-ray images. This dataset can serve as a benchmark for the research community. The images in COVID-19 class, are labeled by a board-certified radiologist, and only those with a clear sign are used for testing purpose.
- We trained four promising deep learning models on this dataset, and evaluated their performance on a test set of 3000 images. Our best performing model achieved a sensitivity rate of 98%, while having a specificity of 92%.
- We provided a detailed experimental analysis on the performance of these models, in terms of sensitivity, specificity, ROC curve, area under the curve, precision-recall curve, and histogram of the predicted scores.
- We provided the heatmaps of the most likely regions, which are infected due to Covid-19, using a deep visualization technique.
- We made the dataset, the trained models, and the implementation publicly available.

It is worth to mention that while very encouraging, given the amount of the labeled data the result of this work is still preliminary and more concrete conclusion requires further experiments on a larger dataset of COVID-19 labeled X-ray images. We believe this work can serve as a benchmark for future works and comparisons.

The structure of the rest of this paper is as follows. Section 2 provides a summary of the prepared **COVID-Xray-5k Dataset**. Section 3 presents the description of the overall proposed framework. Section 4 provides the experimental studies and com-

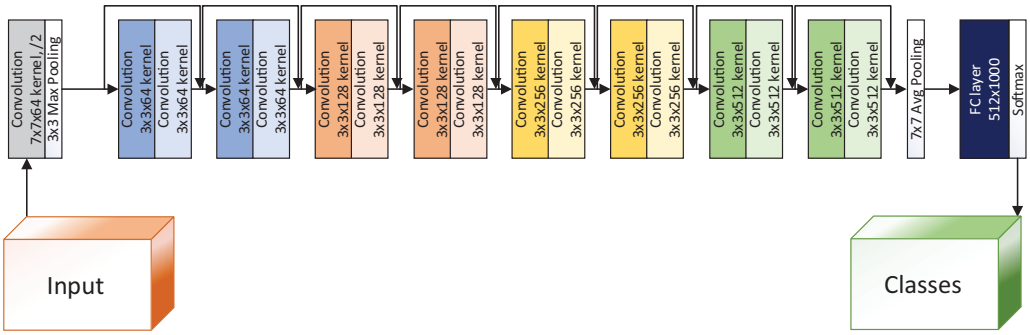

Fig. 3. The architecture of ResNet18 model (He, 2016).

parison with previous works. And finally the paper is concluded in Section 5.

2. COVID-Xray-5k Ddataset

Chest X-ray images from two datasets formed the COVID-Xray-5k dataset that contains 2084 training and 3100 test images.

One of the used datasets is the recently published **Covid-Chestxray-Dataset**, which contains a set of images from publications on COVID-19 topics, collected by https://github.com/ieee8023/covid-chestxray-dataset, Cohen et al. (2020). This dataset contains a mix of chest X-ray and CT images. As of May 3, 2020, it contained 250 X-ray images of COVID-19 patients, from which 203 images are anterior-posterior view. It is mentioned that this dataset is continuously updated. It also contains some meta-data about each patients, such as sex and age. Our COVID-19 images are all coming from this dataset. Based on our board-certified radiologist advice, only anterior-posterior images are kept for Covid-19 prediction, as the lateral images are not suitable for this purpose. The anterior-posterior images were examined by our board-certified radiologist, and the ones without even the slightest radiographic signs of Covid-19 were removed from dataset. Out of 203 interior-exterior X-ray images of COVID-19, 19 of them were excluded, and 184 images (which showed clear signs of COVID-19) were kept by our radiologist. This way, we can provide the community a more cleanly labeled dataset. Out of these images, we chose 100 COVID-19 images to include in the test set (to meet some maximum confidence interval value), and 84 COVID-19 images for the training set. Data augmentation is applied to the training set to increase the number of COVID-19 samples to 420 as described above. We made sure all images for each patient go only to one of the training or test sets. It is worth mentioning that our radiologist marked the regions with specific signs of Covid-19.

Since the number of Non-Covid images was very small in the (https://github.com/ieee8023/covid-chestxray-dataset) dataset, additional images were employed from the **ChexPert** dataset (Irvin et al., 2019), a large public dataset for chest radiograph interpretation consisting of 224,316 chest radiographs of 65,240 patients, labeled for the presence of 14 sub-categories (no-finding, Edema, Pneumonia, etc.). For the non-COVID samples in the training set, we only used images belonging to a single sub-category, composed of 700 images from the no-finding class and 100 images from each remaining 13 sub-classes, resulting in 2000 non-COVID images.

As for the Non-COVID samples in the test dataset, we selected 1700 images from the no-finding category and around 100 images from each remaining 13 sub-classes in distinct sub-folders, resulting in 3000 images in total. The exact number of images of each class for both training and testing is given in Table 1.

Fig. 2 shows 16 sample images from COVID-Xray-5k dataset, including 4 COVID-19 images (the first row), 4 normal images from ChexPert (the second row), and 8 images with one of the 13 diseases in ChexPert (third and fourth rows).

Table 1
Number of images per category in COVID-Xray-5k dataset.

Split	COVID-19	Non-COVID
Training Set	84 (420 after augmentation)	2000
Test Set	100	3000

It is worth mentioning that, there is wide variation inn the resolution of images in this dataset. There are some low-resolution images in Covid-19 class (below 400×400), and some high resolutions ones (more than 1900×1400). This is a positive point for the models that can achieve a reasonable high accuracy on this dataset, despite this variable image resolution and imagery methodology. Collecting all images in a super-controlled environment that results in high-resolution and super-clean images, although desired, is not always doable, and as machine learning field progresses, more and more focus is directed toward models and frameworks that can work reasonably well on variable resolution, quality, and small-scale labeled datasets. Also the images of Covid-19 class are collected from multiple sources by the original provider, and some of them may show a different dynamic range from other ones (and also from ChexPert), but during the training all images are normalized to the same distribution to make model less sensitive to that.

3. The proposed framework

To overcome the limited data sizes, transfer learning was used to fine-tune four popular pre-trained deep neural networks on the training images of COVID-Xray-5k dataset.

3.1. Transfer learning approach

In transfer learning, a model trained on one task is re-purposed to another related task, usually by some adaptation toward the new task. For example, one can imagine using an image classification model trained on ImageNet (which contains millions of labeled images) to initiate task-specific learning for COVID-19 detection on a smaller dataset. Transfer learning is mainly useful for tasks where enough training samples are not available to train a model from scratch, such as medical image classification for rare or emerging diseases. This is especially the case for models based on deep neural networks, which have a large number of parameters to train. By using transfer learning, the model parameters start with already-good initial values that only need some small modifications to be better curated toward the new task.

There are two main ways in which the pre-trained model is used for a different task. In one approach, the pre-trained model

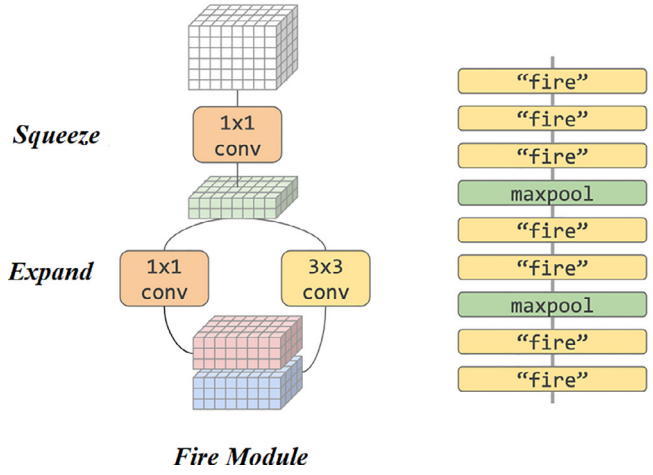

Fig. 4. The architecture of SqueezeNet based on "fire modules". Courtesy of Google (https://codelabs.developers.google.com/codelabs/keras-flowers-squeezenet/).

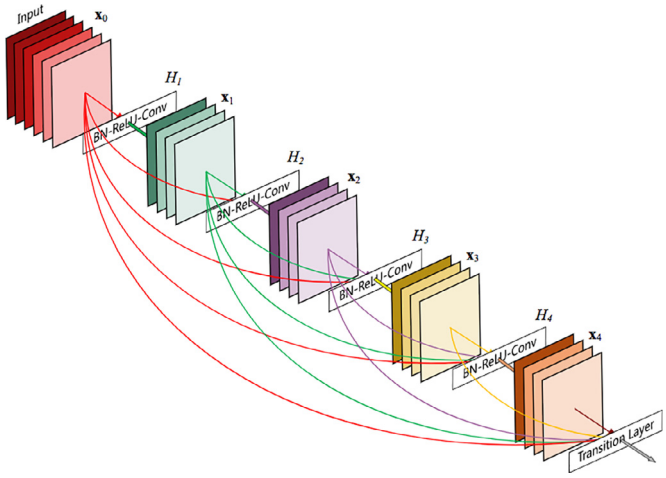

Fig. 5. The architecture of a DenseNet with 5 layers, with expansion of 4. Courtesy of model (Huang et al., 2017).

is treated as a feature extractor (i.e., the internal weights of the pre-trained model are not adapted to the new task), and a classifier is trained on top of it to perform classification. In another approach, the whole network, or a subset thereof, is fine-tuned on the new task. Therefore the pre-trained model weights are treated as the initial values for the new task, and are updated during the training stage.

In our case, since the number of images in the COVID-19 category is very limited, we only fine-tune the last layer of the convolutional neural networks, and essentially use the pre-trained models as a feature extractor. We evaluate the performance of four popular pre-trained models, ResNet18 (He, 2016), ResNet50 (He, 2016), SqueezeNet (Iandola et al., 2016), and DenseNet-121 (Huang et al., 2017). In the next section we provide a quick overview of the architecture of these models, and how they are used for COVID-19 recognition.

3.2. COVID-19 Detection using residual ConvNet – ResNet18 and ResNet50

One of the models used in this work, is the pre-trained ResNet18, trained on ImageNet dataset. ResNet is one of the most popular CNN architecture, which provides easier gradient flow for more efficient training, and was the winner of the 2015 ImageNet

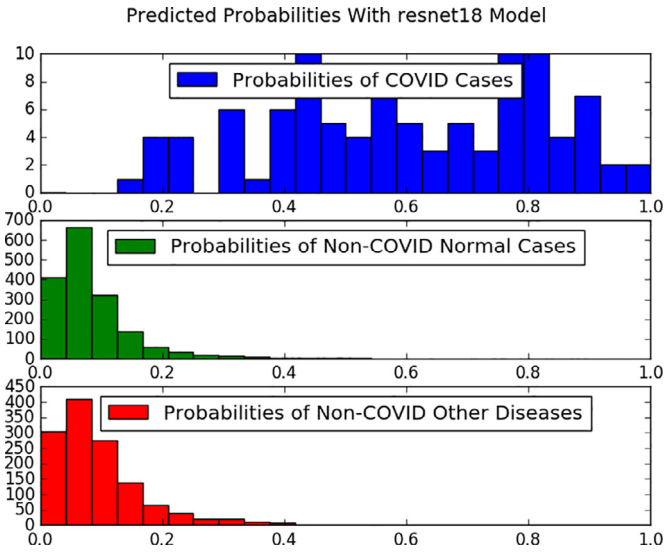

Fig. 6. The predicted probability scores by ResNet18 on the test set.

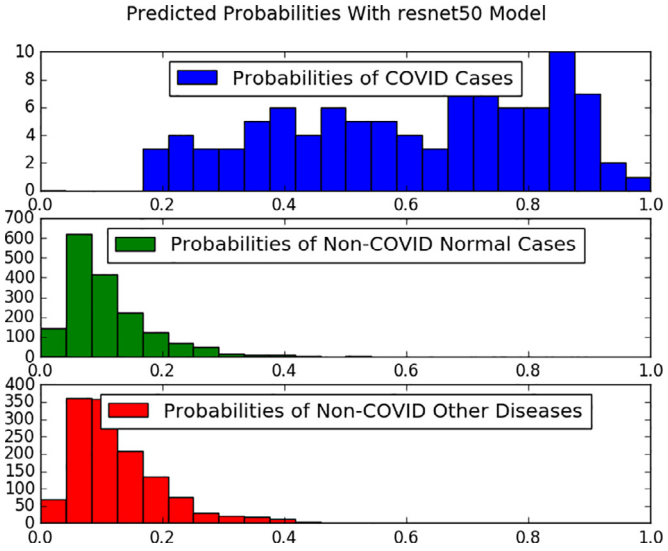

Fig. 7. The predicted probability scores by ResNet50 on the test set.

competition. The core idea of ResNet is introducing a so-called *identity shortcut connection* that skips one or more layers. This would help the network to provide a direct path to the very early layers in the network, making the gradient updates for those layers much easier.

The overall block diagram of ResNet18 model, and how it is used for COVID-19 detection is illustrated in Fig. 3. ResNet50 architecture is pretty similar to ResNet18, the main difference being having more layers.

3.3. COVID-19 Detection using SqueezeNet

SqueezeNet (Iandola et al., 2016) proposed by Iandola et al., is a small CNN architecture, which achieves AlexNet-level (Krizhevsky et al., 2012) accuracy on ImageNet with 50 $\times$ fewer parameters. Using model compression techniques, the authors were able to compress SqueezeNet to less than 0.5MB, which made it very popular for applications that require light-weight models. They alternate a 1 $\times$ 1 layer that "squeezes" the incoming data in the vertical dimension followed by two parallel 1 $\times$ 1 and 3 $\times$ 3 convolutional layers that "expand" the depth of the data again.

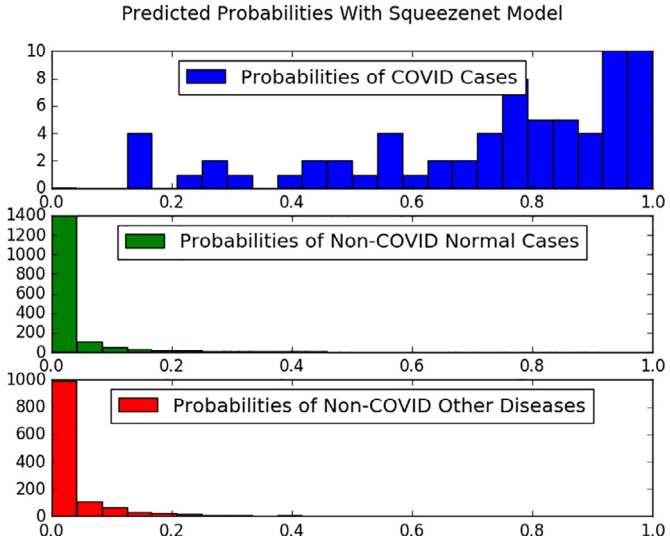

Fig. 8. The predicted probability scores by SqueezeNet on the test set.

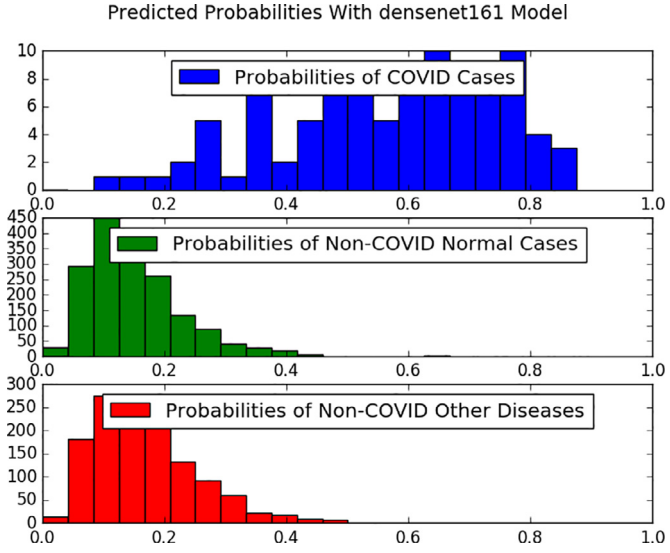

Fig. 9. The predicted probability scores by DesneNet-121 on the test set.

Three main strategies used in SqueezeNet includes: replace 3 × 3 filters with 1 × 1 filters, decrease the number of input channels to 3 × 3 filters, Down-sample late in the network so that convolution layers have large activation maps. Fig. 4 shows the architecture of a simple SqueezeNet.

3.4. COVID-19 Detection using DenseNet

Dense Convolutional Network (DenseNet) is another popular architecture (Huang et al., 2017), which was the winner of the 2017 ImageNet competition. In DenseNet, each layer obtains additional inputs from all preceding layers and passes on its own feature-maps to all subsequent layers. Each layer is receiving a âcollective knowledgeâ from all preceding layers. Since each layer receives feature maps from all preceding layers, network can be thinner and compact, i.e., number of channels can be fewer (so, it have higher computational efficiency and memory efficiency). The architecture of sample DenseNet is shown in Fig. 5.

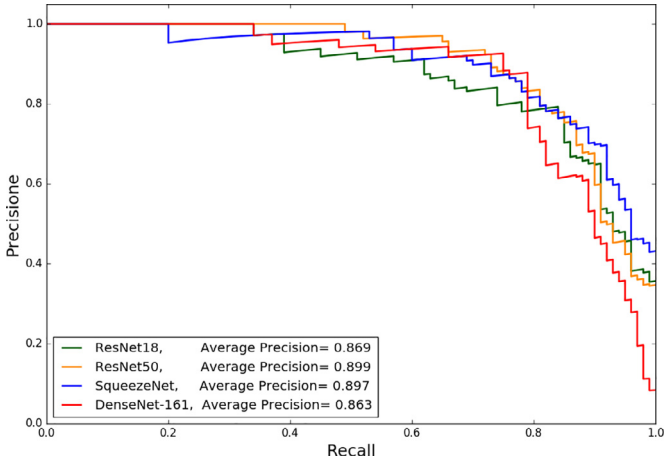

Fig. 10. The precision-recall curve of four CNN architectures on test set.

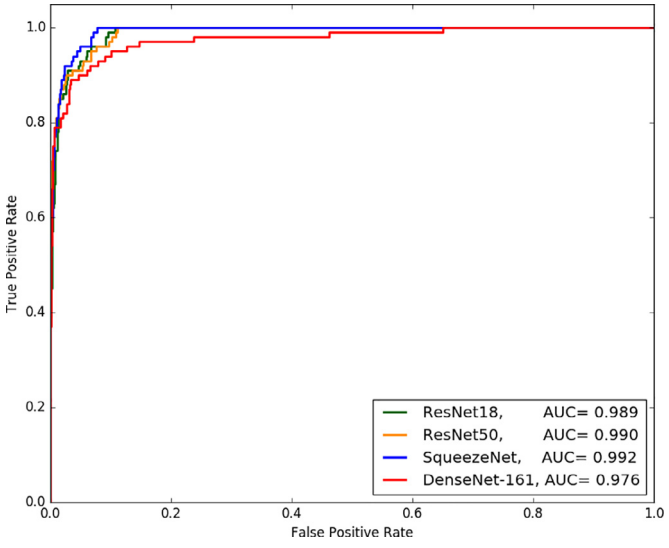

Fig. 11. The ROC curve of four CNN architectures on COVID-19 test set.

3.5. Model training

All employed models are trained with a cross-entropy loss function, which tries to minimize the distance between the predicted probability scores, and the ground truth probabilities (derived from labels), and is defined as:

$$\mathcal{L}_{CE} = -\sum_{i=1}^{N} p_i \log q_i, \tag{1}$$

where p_i and q_i denote the ground-truth, and predicted probabilities for each image, respectively. We can then minimize this loss function using stochastic gradient descent algorithm (and its variations). We attempted to add regularization to the loss function, but the resulting model was not exhibiting a better performance.

4. Experimental results

4.1. Model hyper-parameters

We fine-tuned each model for 100 epochs. The batch size is set to 20, and ADAM optimizer is used to optimize the loss function, with a learning rate of 0.0001. All images are down-sampled to 224 × 224 before being fed to the neural network

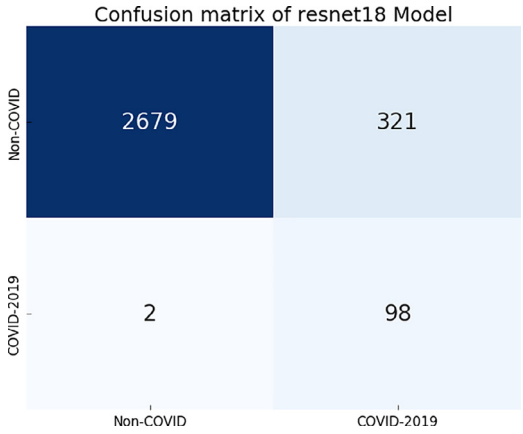

Fig. 12. The confusion matrix of the proposed ResNet18 model.

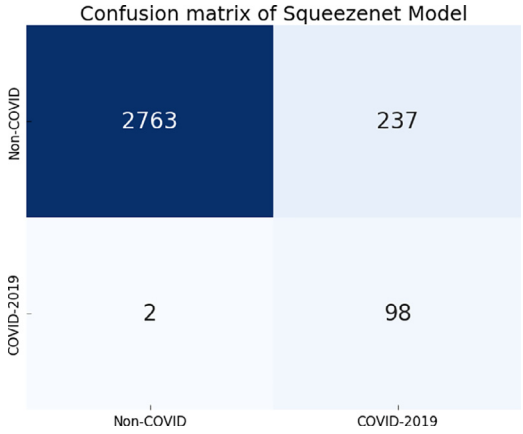

Fig. 13. The confusion matrix of the proposed SqueezeNet framework.

(as these pre-trained models are usually trained with a specific image resolution). All our implementations are done in PyTorch (https://pytorch.org/), and are publicly available at https://github.com/shervinmin/DeepCovid.git.

4.2. Evaluation metrics

There are different metrics which can be used for evaluating the performance of classification models, such as classification accuracy, sensitivity, specificity, precision, and F1-score. Since the current test dataset is highly imbalanced (100 COVID-19 images, 3000 Non-COVID image), sensitivity and specificity are two proper metrics which can be used for reporting the model performance:

$$\textbf{Sensitivity} = \frac{\#\text{Images correctly predicted as COVID-19}}{\#\text{Total COVID-19 Images}},$$
$$\textbf{Specificity} = \frac{\#\text{Images correctly predicted as Non-COVID}}{\#\text{Total Non-COVID Images}}. \quad (2)$$

4.3. Model predicted scores

As mentioned earlier, we focused on four popular convolutional networks, ResNet18, ResNet50, SqueezeNet, DenseNet121. These models predict a probability score for each image, which shows the likelihood of the image being detected as COVID-19. By comparing this probability with a cut-off threshold, we can derive a binary label showing if the image is COVID-19 or not. An ideal model should predict the probability of all COVID-19 samples close to 1, and non-COVID samples close to 0.

Table 2
Sensitivity and specificity rates of ResNet18 model, for different threshold values.

Threshold	Sensitivity	Specificity
0.1	100%	72.4%
0.17	98%	90.7%
0.2	95%	92.4%
0.25	91%	95.8%
0.35	85%	98.3%

Table 3
Sensitivity and specificity rates of ResNet50 model, for different threshold values.

Threshold	Sensitivity	Specificity
0.15	100%	78.2%
0.205	98%	89.6%
0.25	93%	94.2%
0.3	90%	97.3%
0.35	85%	98.4%

Table 4
Sensitivity and specificity rates of SqueezeNet model, for different threshold values.

Threshold	Sensitivity	Specificity
0.1	100%	89.9%
0.15	98%	92.9%
0.2	96.0%	94.6%
0.4	92%	97.6%
0.5	87%	98.3%

Figs. 6–9 show the distribution of predicted probability scores for the images in the test set, by ResNet18, ResNet50, SqueezeNet, and DenseNet-161 respectively. Since Non-COVID class in our study contains both normal cases, as well as other types of diseases, we provide the distribution of predicted scores for three classes: COVID-19, Non-COVID normal, and Non-COVID other diseases. As we can see the Non-Covid images with other disease types have slightly larger scores than the Non-COVID normal cases. This makes sense, since those images are more difficult to distinguish from COVID-19, than normal samples.

COVID-19 patient images are predicted to have much higher probabilities than the Non-COVID images, which is really encouraging, as it shows the model is learning to discriminate COVID-19 from non-COVID images. Among different models, it can be observed that SqueezeNet does a much better job in pushing the predicted scores for COVID-19 and Non-COVID images farther apart from each other.

4.4. Model sensitivity and specificity

Each model predicts a probability score showing the chance of the image being COVID-19. We can then compare these scores with a threshold to infer if the image is COVID-19 or not. The predicted labels are used to estimate the sensitivity and specificity of each model. Depending on the value of the cut-off threshold, we can get different sensitivity and specificity rates for each model.

Tables 2–5 show the sensitivity and specificity rates for different thresholds, using ResNet18, ResNet50, SqueezeNet, and DenseNet-121 models, respectively. As we can see, all these models achieve very promising results, and the best performing model obtains a sensitivity rate of 98% and specificity rate of 92.9%. SqueezeNet and ResNet18 achieve slightly better performance than the other models.

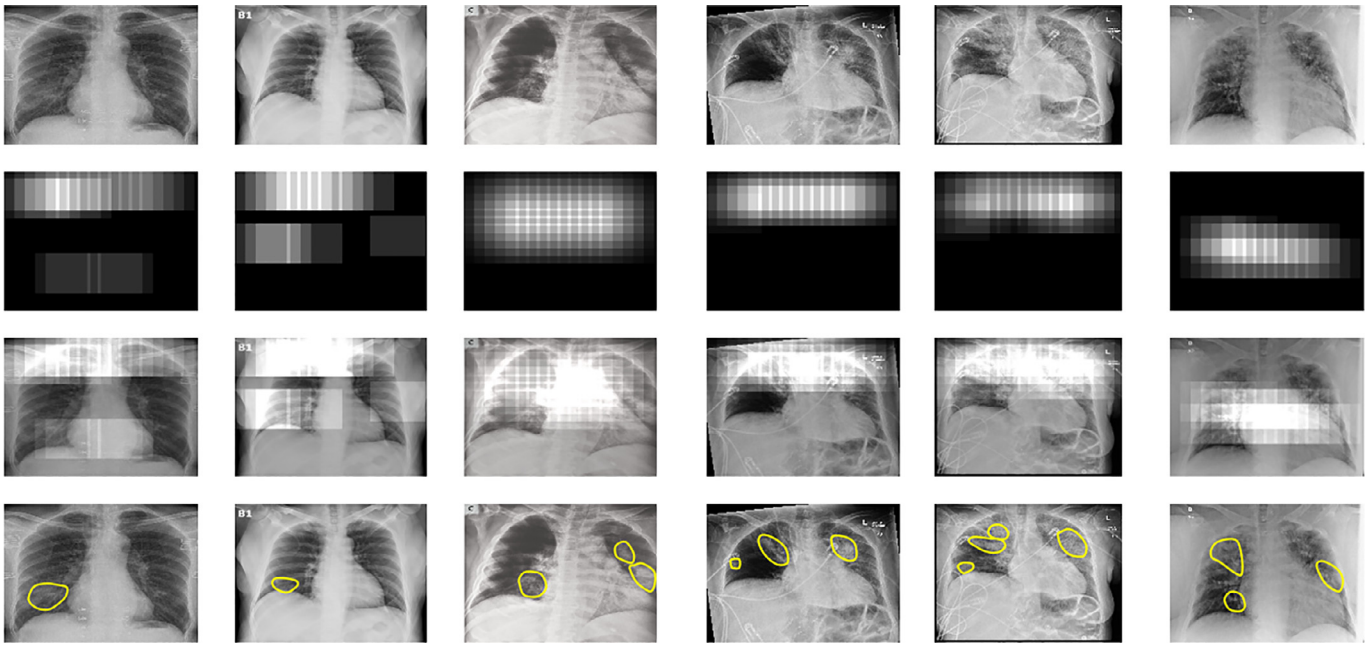

Fig. 14. COVID-19 infected regions detected by our ResNet18 model, in six chest X-ray images from the test set. Vertical sets give the Original images (top row), COVID-19 region heatmap (2nd row), heatmap overlaid on the image (3rd row), and the independent standard of radiologist-marked COVID-19 disease regions (bottom row).

Table 5
Sensitivity and specificity rates of DenseNet-121 model, for different threshold values.

Threshold	Sensitivity	Specificity
0.19	98%	75.1%
0.25	95%	88.9%
0.3	90%	94.6%
0.4	79%	98.9%

Table 6
Comparison of sensitivity and specificity of four state-of-the-art deep neural networks.

Model	Sensitivity	Specificity
ResNet18	98% ± 2.7%	90.7% ± 1.1%
ResNet50	98% ± 2.7%	89.6% ± 1.1%
SqueezeNet	98% ± 2.7%	92.9% ± 0.9%
Densenet-121	98% ± 2.7%	75.1% ± 1.5%

4.5. Small number of COVID-19 cases and model reliability

It is worth mentioning that since so far the number of reliably labeled COVID-19 X-ray images is very limited, and we only have 100 test images in COVID-19 class, the sensitivity and specificity rates reported above may not be reliable. Ideally more experiments on a larger number of test samples with COVID-19 is needed to derive a more reliable estimation of sensitivity rates. We can however estimate the 95% confidence interval of the reported sensitivity and specificity rates here, to see what is the possible range of these values for the current number of test samples in each class. The confidence interval of the accuracy rates can be calculated as:

$$r = z\sqrt{\frac{accuracy(1-accuracy)}{N}}, \tag{3}$$

where z denotes the significance level of the confidence interval (the number of standard deviation of the Gaussian distribution), accuracy is the estimated accuracy (in our cases sensitivity and specificity), and N denotes the number of samples for that class. Here we used 95% confidence interval, for which the corresponding value of z is 1.96.

As for COVID-19 diagnostic, having a sensitive model is crucial, we choose the cut-off threshold corresponding to a sensitivity rate of 98% for each model, and compare their specificity rates. Table 6 provides a comparison of the performance of these four models on the test set. As we can see the confidence interval of specificity rates are small (around 1%), since we have around 3000 samples for this class, whereas for the sensitivity rate we get slightly higher confidence interval (around 2.7%) because of the limited number of samples.

4.6. The ROC curve, precision recall curve, and confusion matrix

It is hard to compare different models only based on their sensitivity and specificity rates, since these rates change by varying the cut-off thresholds. To see the overall comparison between these models, we need to look at the comparison for all possible threshold values. One way to do this, is through the precision-recall curve, which provides the precision rate as a function of recall rate. Precision is defined as the true positive images divided by the total number of images flagged as positive by the model, and the recall is the same as sensitivity rate (defined in Eq. (2)). The precision-recall curve of these four models is shown in Fig. 10.

Another way to do this, is through the Receiver operating characteristic (ROC) curve, which provides the true positive rate as a function of false positive rate. The ROC curve of these four models is shown in Fig. 11. All models have a similar performance according to the AUC with the SqueezeNet achieving a slightly higher AUC than the other models. It is worth mentioning that for highly imbalanced test sets, the AUC may not be a good indicative of model performance (as it can be very high), and looking at average-precision and precision-recall curve would be a better choice in that case. Here we provided both curves for the sake of completeness.

To see the exact number of correctly samples as COVID-19 and Non-COVID, the confusion matrices of the two top-performing models – the fine-tuned ResNet18 and SqueezeNet – when classifying the set of 3100 test images are shown in Figs. 12 and 13.

4.7. The heatmap of potentially infected regions

We used a simple technique to detect the potentially infected regions, while performing COVID-19 detection. This technique is inspired by the work of Zeiler and Fergus (2014), to visualize the result of deep convolutional networks. We start from the top-left corner of the image, and each time occluding a square region of size $N \times N$ inside the image, and make a prediction using the trained model on the occluded image. If occluding that region causes the model to mis-classify a COVID-19 image as Non-COVID, that area would be considered as a potentially infected region in chest X-ray images (mainly because removing the information of that part led to model mis-classification). On the other hand, if occluding a region does not impact the model's prediction, we infer that region is not infected. Once we repeat this procedure for different sliding windows of $N \times N$, each time shifting them with a stride of S, we can get a saliency map of the potentially infected regions in detecting COVID-19. The detected regions for six example COVID-19 images from our test set are shown in Fig. 14. The likely regions of COVID-19 disease marked by our board-certified radiologist are shown in blue on the last row. The generated heatmaps show a good agreement with the radiologist-determined regions of the COVID-19 disease.

5. Conclusion

We reported a deep learning framework for COVID-19 detection from Chest X-ray images, by fine-tuning four pre-trained convolutional models (ResNet18, ResNet50, SqueezeNet, and DenseNet-121) on our training set. We prepared a dataset of around 5k images, called COVID-Xray-5k (using images from two datasets), with the help of a board-certified radiologist to confirm the COVID-19 labels. We make this dataset publicly available for the research community to use as a benchmark for training and evaluating future machine learning models for COVID-19 binary classification task. We performed a detail experimental analysis evaluating the performance of each of these 4 models on the test set of COVID-Xray-5k Dataset, in terms of sensitivity, specificity, ROC, and AUC. For a sensitivity rate of 98%, these models achieved a specificity rate of around 90% on average. This is really encouraging, as it shows the promise of using X-ray images for COVID-19 diagnostics. This study is conducted on a set of publicly available images, which contains around 200 COVID-19 images, and 5000 non-COVID images. The presented work is reflecting one of the earliest Covid-19 chest X-ray analysis and dataset preparation attempts, which brings time-sensitive relevance in combining these two aspects. However, due to the limited number of COVID-19 images publicly available so far, further experiments are needed on a larger set of cleanly labeled COVID-19 images for a more reliable estimation of the accuracy of these models.

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.

CRediT authorship contribution statement

Shervin Minaee: Conceptualization, Methodology, Writing - original draft. **Rahele Kafieh:** Data curation, Writing - original draft. **Milan Sonka:** Supervision, Writing - original draft. **Shakib Yazdani:** Data curation, Formal analysis. **Ghazaleh Jamalipour Soufi:** Data curation.

Acknowledgment

The authors would like to thank Joseph Paul Cohen for collecting the COVID-Chestxray-dataset, and Sean Mullan for helping us with data preparation. We would also like to thank the providers of ChexPert dataset, which are used as the negative samples in our case. Milan Sonka's research effort supported, in part, by NIH grant R01-EB004640.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.media.2020.101794.

References

Ai, T., Yang, Z., Hou, H., Zhan, C., Chen, C., Lv, W., Tao, Q., Sun, Z., Xia, L., 2020. Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases. Radiology doi:10.1148/radiol.2020200642.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal.Mach. Intell 39 (12), 2481–2495.

Cohen, J. P., Morrison, P., Dao, L., 2020. COVID-19 image data collection. arXiv:2003.11597.

Dong, C., et al., 2014. Learning a deep convolutional network for image super-resolution. In: European Conference on Computer Vision. Springer, Cham.

Hansell, D.M., Bankier, A.A., MacMahon, H., McLoud, T.C., Muller, N.L., Remy, J., 2008. Fleischner society: glossary of terms for thoracic imaging. Radiology 246 (3), 697–722.

He, K., et al., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Huang, G., Liu, Z., Maaten, L.V.D., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708.

Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., Keutzer, K., 2016. SqueezeNet: Alexnet-level accuracy with 50 $\times$ fewer parameters and $<$ 0.5MB model size. arXiv:1602.07360.

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., et al., 2019. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 590–597.

Kanne, J.P., Little, B.P., Chung, J.H., Brett, M.E., Ketai, L.H., 2020. Essentials for radiologists on COVID-19: an update - radiology scientific expert panel. Radiology doi:10.1148/radiol.2020200527.

Kong, W., Agarwal, P.P., 2020. Chest imaging appearance of COVID-19 infection. Radiology: Cardiothorac. Imaging 2 (1). doi:10.1148/ryct.2020200028.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet classification with deep convolutional neural networks. Adv. Neural inform. Process.Syst..

LeCun, Y., et al., 1998. Gradient-based learning applied to document recognition. Proc. IEEE 86 (11), 2278–2324.

Minaee, S., Abdolrashidi, A., Su, H., Bennamoun, M., Zhang, D., 2019. Biometric recognition using deep learning: a survey. arXiv:1912.00271.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster R-CNN: towards real-time object detection with region proposal networks. Adv. Neural Inform. Process.Syst..

Rodrigues, J.C.L., et al., 2020. An update on COVID-19 for the radiologist - a British society of thoracic imaging statement. Clin. Radiol..

Wang, W., Xu, Y., Gao, R., Lu, R., Han, K., Wu, G., Tan, W., 2020. Detection of SARS–CoV-2 in different types of clinical specimens. JAMA.

Yang, Y., Yang, M., Shen, C., Wang, F., Yuan, J., Li, J., Zhang, M., et al., 2020. Laboratory diagnosis and monitoring the viral shedding of 2019-nCoV infections. MedRxiv.

Zeiler, M., Fergus, R., 2014. Visualizing and understanding convolutional networks. In: European Conference on Computer Vision. Springer, Cham.