Deep Convolutional Neural Networks to Diagnose COVID-19 and other Pneumonia Diseases from Posteroanterior Chest X-Rays

Pierre Moutounet-Cartan

To cite this version:

Pierre Moutounet-Cartan. Deep Convolutional Neural Networks to Diagnose COVID-19 and other Pneumonia Diseases from Posteroanterior Chest X-Rays. 2020. ⟨hal-05299841⟩

HAL Id: hal-05299841

https://hal.science/hal-05299841v1

Preprint submitted on 6 Oct 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

HAL Authorization

arXiv:2005.00845v1 [eess.IV] 2 May 2020

Imperial College London

Deep Convolutional Neural Networks to Diagnose COVID-19 and other Pneumonia Diseases from Posteroanterior Chest X-Rays

Pierre G. B. Moutounet-Cartan PIERRE.MOUTOUNET-CARTAN17@IMPERIAL.AC.UK
Department of Mathematics
Imperial College London
London, SW7 2AZ, United Kingdom

Abstract

The article explores different deep convolutional neural network architectures trained and tested on posteroanterior chest X-rays of 327 patients who are healthy (152 patients), diagnosed with COVID-19[1] (125), and other types of pneumonia (48), of which Severe Acute Respiratory Syndrome (16), Streptococcus (13), Klebsiella (1), Legionellosis (2), Pneumocystis Jiroveci Pneumonia (13), Acute Respiratory Distress Syndrome not caused by COVID-19 (4), and Chlamydia Pneumoniae (1). In particular, this paper looks at the deep convolutional neural networks VGG16 and VGG19 (Simonyan and Zisserman, 2015), InceptionResNetV2 and InceptionV3 (Szegedy et al., 2016), as well as Xception (Chollet, 2017), all followed by a flat multi-layer perceptron and a final 30% drop-out.

The paper has found that the best performing network is VGG16 with a final 30% drop-out trained over 3 classes (COVID-19, No Finding, Other Pneumonia). It has a cross-validated training accuracy of $93.9(\pm 3.4)\%$, a COVID-19 sensitivity of $87.7(-1.9, +2)\%$, and a No Finding sensitivity of $96.8(\pm 0.8)\%$. The respective cross-validated values on the testing set are $84.1(\pm 13.5)\%$, $87.7(-1.9, 2)\%$, and $96.8(\pm 0.8)\%$.[2] The model optimizer was Adam (Kingma and Lei-Ba, 2015) with a 0.0001 learning rate, and categorical cross-entropy loss.

It is hoped that, once this research will be put to practice in hospitals, healthcare professionals will be able in the medium to long-term to diagnosing through machine learning tools possible pneumonia, and if detected, whether it is linked to a COVID-19 infection, allowing the detection of new possible COVID-19 foyers after the end of possible "stop-and-go" lockdowns as expected by Ferguson et al. (2020) until a vaccine is found and widespread. Furthermore, in the short-term, it is hoped practitioners can compare the diagnosis from the deep convolutional neural networks with possible RT-PCR testing results, and if clashing, a Computed Tomography could be performed as they are more accurate in showing COVID-19 pneumonia (Ai et al., 2020; Wang et al., 2020; Xu et al., 2020).

1. Referring to the coronavirus originating from Wuhan, mainland China, and discovered in 2019.
2. The values in brackets are the differences between the estimated values for the measures listed and the left or right boundary of the 95% confidence interval after Stratified 5-Fold cross-validation. The inverse cumulative distribution function value is taken from a t-distribution with 4 degrees of freedom.

Keywords: Neural Networks, Convolutional Neural Networks, Deep Learning, Coronavirus (COVID-19), Pneumonia, Posteroanterior (PA) Chest X-Rays, Machine Learning, Artificial Intelligence (AI)

Contents

Abbreviations

Listed by alphabetical order

- 2D-Conv(m, M): 2-dimensional convolution layer of M channels of $m \times m$ kernel
- AI: Artificial Intelligence
- ARDS: Acute Respiratory Distress Syndrome
- COVID-19: Novel Coronavirus 2019 SARS-n-CoV-2
- CT: Computed Tomography
- DropOut(p): drop-out between two perceptron layers with probability p
- Layer(n): flat perceptron layer made of n neurons
- MERS: Middle East Respiratory Syndrome
- MaxPool(d, s): MaxPool layer of $d \times d$ pool and $s \times s$ stride
- Output(f): output from the neural network through some activation function f
- PA (X-ray): Posteroanterior X-ray
- RT-PCR (test): Reverse Transcription Polymerase Chain Reaction test
- SARS: Severe Acute Respiratory Syndrome

1. Introduction, Aims, & State-of-the-Art

As the writing of this article, there is a significant outbreak of COVID-19, which can lead to pneumonia, visible to medical imaging methods such as X-rays or CT scans. This pneumonia has also been detected in RT-PCR COVID-19 positive patients who did not have any known underlying health conditions (Cheng et al., 2020). Epidemiology research has tried to find the effective reproduction number before and after lockdowns and school closures, as well as estimating the real number of people infected by COVID-19 through mathematical models (Ferguson et al., 2020; Flaxman et al., 2020) or seroprevalence studies (Bendavid et al., 2020). In particular, Flaxman et al. (2020) estimated that up to March 28, 2020, between 1.2% and 5.4% of the population had been COVID-19 infected in the United Kingdom (i.e., between roughly 800,000 and 3,600,000 residents of the United Kingdom). Antibody seroprevalence studies showed that between 2.24% and 3.37% of the Santa Clara, CA population contracted COVID-19 up to April 4, 2020 (Bendavid et al., 2020), and around 21% of the New York City, NY population contracted the disease according to the Governor of New York. So far, the most advanced clinical trials for a potential vaccine are located in the United States, the United Kingdom, and Germany, and such a vaccine could only be available in fall 2020 at the earliest for healthcare professionals.

Wang et al. (2020) have applied deep convolutional neural networks on 1,065 CT images from the lungs of pathogen-confirmed COVID-19 cases (325 patients) along with those previously diagnosed with typical viral pneumonia (740 patients) not provoked by coronavirus. The internal validation in that research achieved a total accuracy of 89.5% with specificity of 88% and sensitivity of 87%, while the external testing data-set showed a total accuracy of 79.3% with specificity of 83% and sensitivity of 67%. 54 patients part of the study had the first two nucleic acid test results for COVID-19 were negative, of which 46 were predicted as COVID-19 positive by the algorithm with the probability of 85.2%, which could further prove the very low accuracy of such tests. Indeed, Ai et al. (2020) discovered that of the patients with negative RT-PCR results, 75% had positive chest CT findings of which 48% were considered as highly likely cases of COVID-19.

Similarly, Xu et al. (2020) found that the real time reverse RT-PCR detection of viral RNA from sputum or nasopharyngeal swab has a relatively low positive rate to determine COVID-19 positiveness. After training ResNet-like neural networks on CT images of patients showing COVID-19 (357 images), Other Viral Pneumonia (390), and No Finding (963), that research found respective sensitivities of 81.5%, 75.4%, and 97.8%, with an overall accuracy on the benchmark data-set of 86.7%. These are seemingly similar results to Wang et al. (2020).

An attempt of training deep neural networks on PA chest X-rays, where some of which showed a COVID-19 pathology, was made by Narin et al. (2020) with other parts of the same data-set used in this research, although they only used the classes COVID-19 (50 images) and No Finding (50), meaning their neural networks could, in principle, categorize non-COVID-19 pneumoniae as such if presented by such X-ray. This explains the high accuracies and other measures that they have found for the InceptionResNetV2, InceptionV3,

and ResNet50 architectures. Indeed, they reached an accuracy of 98%, a COVID-19 sensitivity of 96%, and a specificity value of 100% for ResNet50 on the external testing data. These are the cross-validated values through standard 5-Fold. Furthermore, the data-set was relatively small.

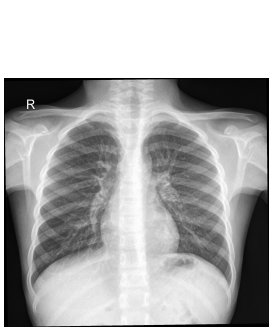
There are two main goals to this research: (1) in the medium to long-term, being able to diagnose through machine learning tools possible pneumonia, and if detected, whether it is linked to a coronavirus infection or other, allowing the detection of new possible coronavirus foyers after the end of possible "stop-and-go" lockdowns as expected by Ferguson et al. (2020) until a vaccine is found and widespread. (2) Furthermore, in the short-term, it can allow practitioners to compare the diagnosis from the deep convolutional neural networks in this paper with possible RT-PCR testing results – if clashing, a chest CT could be done as they are more accurate in showing COVID-19 pneumonia as shown by Ai et al. (2020); Wang et al. (2020); Xu et al. (2020).

This research focuses on PA chest X-rays because, although they can be less accurate to diagnose COVID-19 than CT imaging, they are more common in hospitals and relatively cheaper to perform. Hence, training neural networks on PA chest X-rays could lead to greater generalization and application in hospitals around the world, and CT imaging would only be used for cases where patients are pathogen-confirmed COVID-19 through RT-PCR testing but the AI diagnoses No Finding, or when the AI diagnoses the patients as highly-likely COVID-19 but the RT-PCR tests are negative.

2. The Data

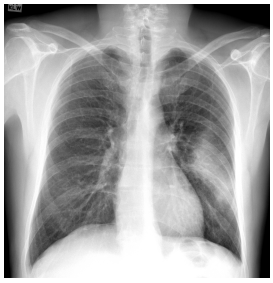
The data has been collected from Cohen et al. (2020) and Kermany et al. (2018). The first source (Cohen et al., 2020) had predominantly data from ill patients, and therefore the paper partially uses the second data-set (Kermany et al., 2018) which is made of healthy patients only (only partially otherwise there would be an over-representation).

All training X-rays have been resized to a shape 182×182. After train-test split, all training X-rays are allowed to rotate up to 50 degrees, are feature-wise standardized using statistics from the entire set, can have their widths and heights shifted by up to 20%, are allowed a shear intensity of up to .25 degrees in the counter-clockwise direction, as well as a zooming/de-zooming range of ±10%, a channel shifting range of .2, and can flip either horizontally or vertically. Missed pixels are filled through constant filling. This allows to also expand the number of training images. These allowed transformations for the training set are detailed in the Table 1.

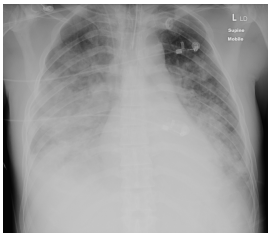


(a) Lungs of a healthy patient.

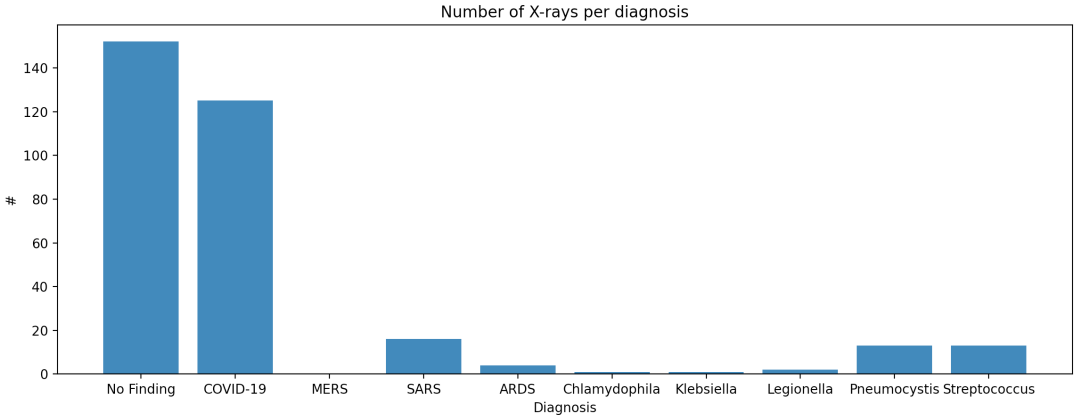
(b) Lungs of a COVID-19 infected patient.



(c) Lungs of a chlamydia infected patient.



(d) Lungs of a patient suffering ARDS.



(e) Final distribution of diagnoses by the healthcare professionals based on the PA chest X-rays in the combined data-sets.

Figure 1: Examples of PA chest X-rays in the data-set in Figures 1a to 1d, and distribution of the diagnoses based on the PA chest X-rays in Figure 1e.

Feature	Comment
Resizing	to 182×182 in RGB, resulting in $(182, 182, 3)$ arrays
Rotation	up to ± 50 degrees
Standardization	feature-wise through training set statistics
Width shift	up to $\pm 20\%$
Height shift	up to $\pm 20\%$
Shear intensity	up to .25 degrees counter-clockwise
Zooming & de-zooming	up to $\pm 10\%$
Channel shifting	in the range of 20%
Horizontal & vertical flip	True

Table 1: Possible transformations of the images in the training set before being introduced to the convolutional neural networks.

Figure 1e shows the final distribution of illness diagnoses following the reading of the PA chest X-rays by the healthcare professionals.

3. VGG16

The architecture of the first considered deep convolutional neural network follows the VGG16 architecture proposed by Simonyan and Zisserman (2015), and is detailed in Table 2. Once trained over 200 epochs on a data-set of over 14 million 224×224 images belonging to 1,000 classes based on the Large Scale Visual Recognition Challenge 2014 (ILSVRC2014), this deep convolutional neural network achieved 92.7% top-5 test accuracy. In this research, we kept the weights for pre-training following the ILSVRC2014. The optimizer is Adam (Kingma and Lei-Ba, 2015) with categorical cross-entropy loss, learning rate 0.0001, and ran over 200 epochs.

Step type	**Computations**	**Name**
First 2D-Convolutions	2D-Conv$(3, 64)$	2D-Conv_111
	2D-Conv$(3, 64)$	2D-Conv_112
First 2D-Maximum pooling	MaxPool$(2, 2)$	Pool_11
Second 2D-Convolutions	2D-Conv$(3, 128)$	2D-Conv_121
	2D-Conv$(3, 128)$	2D-Conv_122
Second 2D-Maximum pooling	MaxPool$(2, 2)$	Pool_12
Third 2D-Convolutions	2D-Conv$(3, 256)$	2D-Conv_131
	2D-Conv$(3, 256)$	2D-Conv_132
	2D-Conv$(3, 256)$	2D-Conv_133
Third 2D-Maximum pooling	MaxPool$(2, 2)$	Pool_13
Fourth 2D-Convolutions	2D-Conv$(3, 512)$	2D-Conv_141
	2D-Conv$(3, 512)$	2D-Conv_142
	2D-Conv$(3, 512)$	2D-Conv_143
Fourth 2D-Maximum pooling	MaxPool$(2, 2)$	Pool_14
Fifth 2D-Convolutions	2D-Conv$(3, 512)$	2D-Conv_151
	2D-Conv$(3, 512)$	2D-Conv_152
	2D-Conv$(3, 512)$	2D-Conv_153
Fifth 2D-Maximum pooling	MaxPool$(2, 2)$	Pool_15
Flat portion	Flatten	Flat_11
	Layer(4096)	Layer_11
	Layer(4096)	Layer_12
	Layer(1000)	Layer_13
	Batch Normalization	Norm_11
	Layer(256)	Layer_14
	DropOut(p)	Drop_11
Output	Output(softmax)	Out_1

Table 2: VGG16 configuration (Simonyan and Zisserman, 2015) with an added flat portion that has some drop-out rate p. The activation function is Rectified Linear Unit (ReLU) throughout. The total number of parameters for this architecture is 17,994,563 for 182×182 images.

We perform the Stratified 5-Fold cross-validation of the VGG16 with 30% final drop-out, and the results are as shown in Table 3.

Internal / Training Set			
Measure	**LHS 95% CI**	**Value**	**RHS 95% CI**
Loss	-0.135	0.105	0.339
Accuracy	0.905	0.939	0.973
Flat AUC	0.971	0.974	0.976
COVID-19 Recall	0.858	0.877	0.897
No Finding Recall	0.960	0.968	0.976
Other Pneumonia Recall	0.500	0.534	0.568
External / Testing Set			
Measure	**LHS 95% CI**	**Value**	**RHS 95% CI**
Loss	-0.324	0.337	0.998
Accuracy	0.706	0.841	0.976
Flat AUC	0.970	0.974	0.977
COVID-19 Recall	0.858	0.877	0.897
No Finding Recall	0.960	0.968	0.976
Other Pneumonia Recall	0.500	0.534	0.568

Table 3: Cross-validated performance measures (rounded to the thousandth) of the VGG16 neural network outlined in Table 2 through Stratified 5-Fold and 200 epochs, learning rate of 0.0001. The 95% confidence interval were computed through the t-distribution with 4 degrees of freedom.

In this case, we find satisfactory accuracies for the training and testing set, of respectively 94% and 84%, although the 95% confidence is quite large for the latter. Interestingly, the COVID-19 sensitivity remains high with short-range confidence intervals. A 88% COVID-19 sensitivity is similar to Wang et al. (2020) for the training set, but significantly higher for the testing set. In clinical terms, around 88% of those diagnosed or tested COVID-19 positive are identified as such by the neural network outlined in Table 2. Similarly, we found that the No Finding sensitivity, i.e., the percentage of lung-healthy patients identified as such, is of 97% for both the training and testing set.

On the other hand, the sensitivity for pneumonia diseases other than COVID-19 is poor (at 53% for both the training and testing sets), meaning the neural network has difficulties differentiating COVID-19 and other pneumoniae. For such differentiation, Wang et al. (2020) found a sensitivity of 67%, which is better, although they only used two classes, namely COVID-19 and Other Pneumonia. It is possible that this improved sensitivity is due to the fact that Wang et al. (2020) trained their neural networks on CT images, and not on PA X-rays, which tend to better differentiate pathologies. Furthermore, there is an under-representation of such label in our data-set, although we consider that it could be representative of data-sets found in hospital during a COVID-19 epidemic.

As we are dealing with a multi-class problem, the Flat AUC (Area Under the ROC Curve) is not equivalent to the standard binary AUC. We here flattened the data into a single label before AUC computation. In which case, each label-prediction pair is treated as an individual data point. Therefore, this measure should not be very reliable, although does procure some information on the "overall" ROC of the neural network. It is found to be of 97.4% for both the training and testing data-sets. Due to the construction of this performance measure, it is better to compare it to other networks and may not carry significance of the network alone – which is what we will do in the sections below.

4. InceptionResNetV2 & InceptionV3

The second architecture considered is a very deep convolutional neural network named InceptionResNetV2 proposed by Szegedy et al. (2016), and it is described in the diagram shown in Figure 2. Once trained over 200 epochs on the ImageNet data-set, this very deep convolutional neural network achieved 95.3% top-5 test accuracy. In this research, we kept the weights for pre-training after the above training, we use an Adam optimizer (Kingma and Lei-Ba, 2015) with learning rate 0.0001, categorical cross-entropy loss, and 200 epochs. In the architecture, max-pooling was performed for feature extraction.

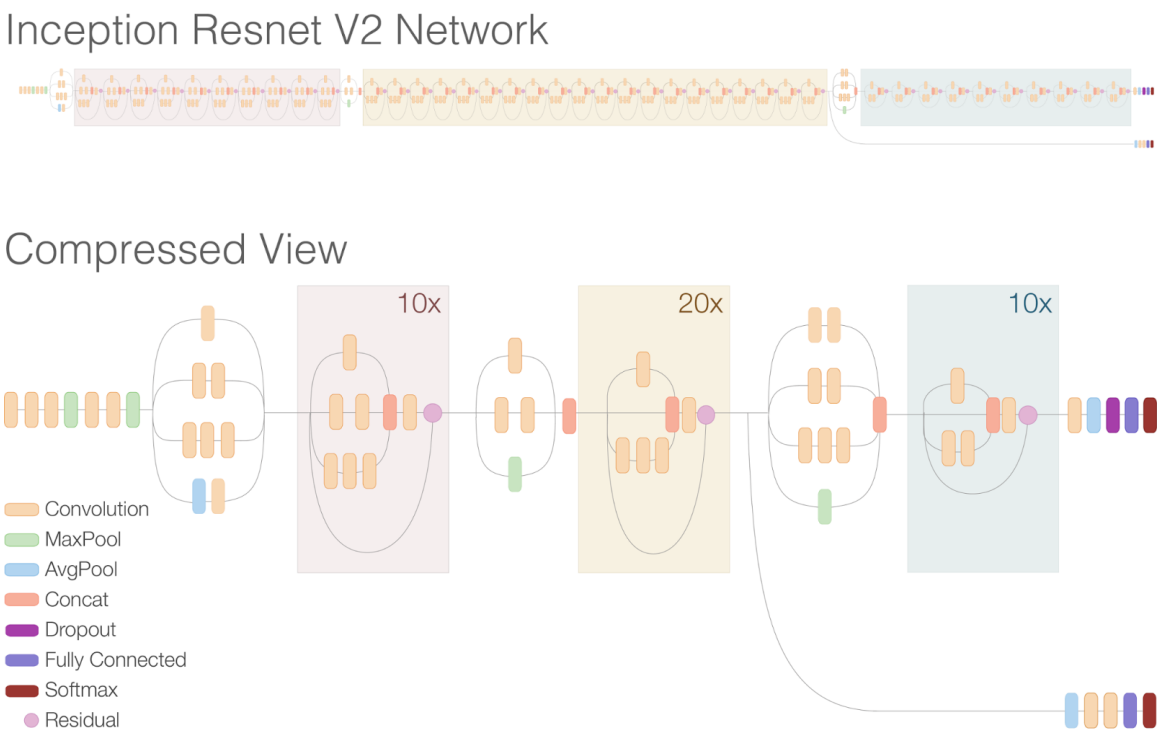


Figure 2: Description diagram of the architecture of InceptionResNetV2, as detailed in Szegedy et al. (2016). On the 182×182 data-set, this neural network has 54,737,123 parameters.

The cross-validated performance measures, obtained through Stratified 5-Fold, for the InceptionResNetV2 with final 30% drop-out are shown in Table 4.

Internal / Training Set			
Measure	**LHS 95% CI**	**Value**	**RHS 95% CI**
Loss	0.073	0.977	1.881
Accuracy	0.520	0.616	0.712
Flat AUC	0.906	0.918	0.931
COVID-19 Recall	0.673	0.708	0.744
No Finding Recall	0.911	0.918	0.925
Other Pneumonia Recall	0.351	0.433	0.514
External / Testing Set			
Measure	**LHS 95% CI**	**Value**	**RHS 95% CI**
Loss	0.488	1.411	2.335
Accuracy	0.457	0.612	0.765
Flat AUC	0.907	0.919	0.932
COVID-19 Recall	0.675	0.710	0.745
No Finding Recall	0.911	0.918	0.925
Other Pneumonia Recall	0.353	0.433	0.513

Table 4: Cross-validated performance measures (rounded to the thousandth) of the InceptionResNetV2-like neural network outlined in Figure 2 through Stratified 5-Fold and 200 epochs, learning rate of 0.0001. The 95% confidence interval were computed through the t-distribution with 4 degrees of freedom.

The performance results for InceptionResNetV2 tend here to be quite disappointing, mainly because this architecture includes many batch normalizations (Ioffe and Szegedy, 2015). During model training, the batches are normalized by their mean and variance, but in the testing phase, the batches are normalized with respect to the changing average of observed mean and variance, leading to lower accuracies for the type of data-set we are considering.

In particular, we found COVID-19 sensitivities for the InceptionResNetV2 network of $70.8(-3.5, +3.6)\%$ and $71.0(\pm 3.5)\%$ for the internal and external validations respectively. In medical terms, this is considered relatively poor, and seems indeed low even if we take into account errors from RT-PCR testing or seroprevalence studies as showcased in Xu et al. (2020). Furthermore, for the accuracies and COVID-19 sensitivities, the 95% confidence interval implied by the Stratified 5-Fold cross-validation is wide, implying greater variance. Hence, it is likely that training and testing on different but similar data-sets could lead to very different performance results.

Similarly, we can compare the Flat AUC value with the VGG16-like network trained as in the previous section. According to the measures gathered, the VGG16-like neural network improved the internal and external Flat AUCs by approximately 6% compared to the InceptionResNetV2-like neural network. Hence, we believe that the VGG16-like

neural network is more appropriate for the data-set we are studying compared to the InceptionResNetV2-like neural network.

In the case of the InceptionV3 convolutional neural network (Szegedy et al., 2015), with the same optimizer, we find similar albeit slightly enhanced performance results compared to InceptionResNetV2, as seen in Table 5.

Internal / Training Set			
Measure	**LHS 95% CI**	**Value**	**RHS 95% CI**
Loss	-2.913	2.676	8.266
Accuracy	0.529	0.705	0.880
Flat AUC	0.908	0.931	0.954
COVID-19 Recall	0.689	0.779	0.870
No Finding Recall	0.928	0.941	0.953
Other Pneumonia Recall	0.421	0.491	0.561
External / Testing Set			
Measure	**LHS 95% CI**	**Value**	**RHS 95% CI**
Loss	-0.757	3.477	7.711
Accuracy	0.477	0.691	0.905
Flat AUC	0.909	0.931	0.954
COVID-19 Recall	0.691	0.780	0.869
No Finding Recall	0.928	0.941	0.953
Other Pneumonia Recall	0.422	0.491	0.560

Table 5: Cross-validated performance measures (rounded to the thousandth) of the InceptionV3-like neural network through Stratified 5-Fold and 200 epochs, learning rate of 0.0001. The 95% confidence interval were computed through the t-distribution with 4 degrees of freedom.

5. VGG19

Due to the earlier success of VGG16, we consider the VGG19 deep convolutional neural network architecture as showcased in Simonyan and Zisserman (2015), and again with 200 epochs, the Adam optimizer (Kingma and Lei-Ba, 2015), a learning rate of 0.0001, and categorical cross-entropy loss. VGG19 is similar to VGG16, instead we add the following convolutions (Simonyan and Zisserman, 2015):

- 2D-Conv(3, 256) after 2D_Conv_133 in Table 2,
- 2D-Conv(3, 512) after 2D_Conv_143 in Table 2,
- and 2D-Conv(3, 512) after 2D_Conv_153 in Table 2.

The performance measures seen in Table 6 for the VGG19-like network are very similar than those gathered for the VGG16-like network (Table 3) due to the very similar architecture. In particular, we find that the added convolutions do not increase significantly the performance

Internal / Training Set			
Measure	**LHS 95% CI**	**Value**	**RHS 95% CI**
Loss	-0.056	0.082	0.220
Accuracy	0.902	0.927	0.953
Flat AUC	0.967	0.969	0.971
COVID-19 Recall	0.843	0.861	0.879
No Finding Recall	0.960	0.964	0.969
Other Pneumonia Recall	0.428	0.480	0.531
External / Testing Set			
Measure	**LHS 95% CI**	**Value**	**RHS 95% CI**
Loss	-0.527	0.485	1.498
Accuracy	0.702	0.820	0.937
Flat AUC	0.967	0.969	0.970
COVID-19 Recall	0.843	0.861	0.879
No Finding Recall	0.960	0.964	0.968
Other Pneumonia Recall	0.427	0.479	0.531

Table 6: Cross-validated performance measures (rounded to the thousandth) of the VGG19-like neural network through Stratified 5-Fold and 200 epochs, learning rate of 0.0001. The 95% confidence interval were computed through the t-distribution with 4 degrees of freedom.

measures, although we see that the 95% confidence intervals are shrunk. Furthermore, due to these added convolutions, VGG19 is slightly more costly than VGG16 despite no big improvement in the measures.

6. Conclusion

A summary table of some of the performance measures for the neural networks all trained over 200 epochs are shown in Table 7. Some of these measures are further shown in Figure 3.

As we can see from Table 7 and Figure 3, VGG16 performed best on all measures, although VGG19 had a similar performance with lower variability. In particular, VGG16 had Stratified 5-Fold cross-validated accuracies of $93.9(\pm 3.4)\%$ for the internal data-set and of $84.1(\pm 13.5)\%$ for the external data-set. Note that we have run all the training sessions with 200 epochs for comparability between the neural networks, and therefore this difference between the internal and external accuracies might be explained by over-fitting. Furthermore, the COVID-19 sensitivity of the VGG16 neural network is of $87.7(-1.9, 2)\%$ for both the internal and external data-sets, suggesting efficiency of that network in correctly identifying true COVID-19 positives. Idem for the No Finding sensitivity which is at $96.8(\pm 0.8)\%$.

-	Internal / Training Set			External / Testing Set		
Net	**Accuracy**	**#1 Recall**	**#2 Recall**	**Accuracy**	**#1 Recall**	**#2 Recall**
(1)	$93.9^{+3.4}_{-3.4}\%$	$87.7^{+2.0}_{-1.9}\%$	$96.8^{+0.8}_{-0.8}\%$	$84.1^{+13.5}_{-13.5}\%$	$87.7^{+2.0}_{-1.9}\%$	$96.8^{+0.8}_{-0.8}\%$
(2)	$92.7^{+2.6}_{-2.5}\%$	$86.1^{+1.8}_{-1.8}\%$	$96.4^{+0.5}_{-0.4}\%$	$82.0^{+11.7}_{-11.8}\%$	$86.1^{+1.8}_{-1.8}\%$	$96.4^{+0.4}_{-0.4}\%$
(3)	$61.6^{+9.6}_{-9.6}\%$	$70.8^{+3.6}_{-3.5}\%$	$91.8^{+0.7}_{-0.7}\%$	$61.2^{+15.3}_{-15.5}\%$	$71.0^{+3.5}_{-3.5}\%$	$91.8^{+0.7}_{-0.7}\%$
(4)	$70.5^{+17.5}_{-17.6}\%$	$77.9^{+9.1}_{-9.0}\%$	$94.1^{+1.2}_{-1.3}\%$	$69.1^{+21.4}_{-21.4}\%$	$78.0^{+8.9}_{-8.9}\%$	$94.1^{+1.2}_{-1.3}\%$
(5)	$65.3^{+21.8}_{-21.8}\%$	$75.0^{+8.9}_{-8.8}\%$	$89.5^{+2.9}_{-2.9}\%$	$61.4^{+19.6}_{-19.5}\%$	$75.1^{+8.8}_{-8.8}\%$	$89.5^{+2.9}_{-2.8}\%$

Table 7: Cross-validated (Stratified 5-Fold) internal and external measures for each of the models considered, rounded to the closest thousandth. Here, #1 Recall refers to the sensitivity of the neural network on the COVID-19 categorical variable, and #2 Recall the sensitivity on the No Finding categorical variable. Networks: (1) VGG16 (Simonyan and Zisserman, 2015), (2) VGG19 (Simonyan and Zisserman, 2015), (3) InceptionResNetV2 (Szegedy et al., 2016), (4) InceptionV3 (Szegedy et al., 2016), and (5) Xception (Chollet, 2017), all with a 30% final drop-out.

One topic of research that could be pursued following this paper is the creation of a specific deep neural network tailored for the data-set made of PA chest X-rays. This research would be dependent on the easy access of a greater database of PA chest X-rays COVID-19 positive patients (through CT diagnosis, RT-PCR or seroprevalence testing), as well as the computational capabilities of large-scale training of very deep convolutional networks. Furthermore, such tailored network could make it harder to share it to hospitals, and as a consequence making it less prone to front-line applications.

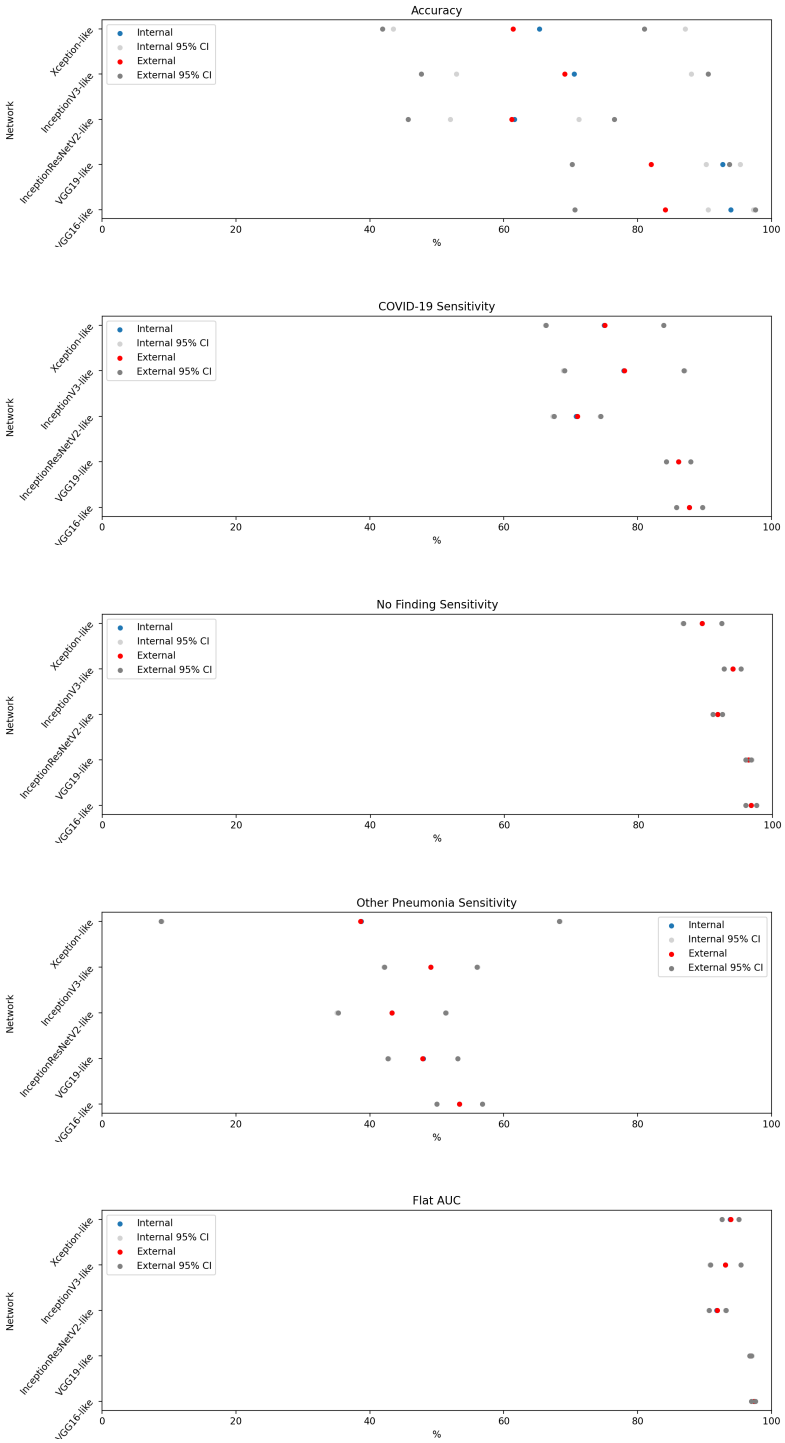


Figure 3: Performance measures of various neural networks on the data-set of PA chest X-rays.

7. Limits of this paper

The biggest limit to this research is by far the data-set. The data-set collected was small after balancing for the number of COVID-19 positive cases, and still was quite unbalanced as the class "Other Pneumonia" was under-represented. Furthermore, the data includes PA chest X-rays from a multitude of hospitals, which could have lowered the accuracies and sensitivities.

It is also clear that the predictions on a PA chest X-ray given by the trained deep neural networks cannot be relied upon alone. Its best use is in conjunction of clinical tests (such as pathogen-RT-PCR or seroprevalence testing), or professional diagnosis based on other clearer medical imaging methods such as CT scans.

8. For hospitals

Certified hospitals can request the code by contacting the author and / or Imperial College London at pierre.moutounet-cartan17@imperial.ac.uk, as well as for explanation of how to use it with their own data-set (or a collective data-set between "sister" hospitals).

There are two main goals to this research to help hospitals:

- in the medium to long-term, being able to diagnose through machine learning tools possible pneumonia, and if detected, whether it is linked to a coronavirus infection or other, allowing the detection of new possible coronavirus foyers after the end of possible "stop-and-go" lockdowns as expected by Ferguson et al. (2020) until a vaccine is found and widespread;
- in the short-term, it can allow practitioners to compare the diagnosis from the deep convolutional neural networks in this paper with possible RT-PCR testing results – if clashing, a chest CT could be done as they are more accurate in showing COVID-19 pneumonia as shown by Ai et al. (2020); Wang et al. (2020); Xu et al. (2020).

This research focuses on PA chest X-rays because, although they can be less accurate to diagnose COVID-19 than CT imaging, they are more common in hospitals and relatively cheaper to perform. Hence, training neural networks on PA chest X-rays could lead to greater generalization and application in hospitals around the world, and CT imaging would only be used for cases where patients are pathogen-confirmed COVID-19 through RT-PCR testing but the AI diagnoses No Finding, or when the AI diagnoses the patients as highly-likely COVID-19 but the RT-PCR tests are negative.

9. Conflicts of interest

The author(s) note no known conflict of interest by undertaking this research.

References

Ai, T., Yang, Z., Hou, H., Zhan, C., Chen, C., Lv, W., Tao, Q., Sun, Z., et al. (2020). Correlation of Chest CT and RT-PCR Testing in Coronavirus Disease 2019 (COVID-19) in China: A Report of 1,014 Cases. *Radiology*, Ahead of print.

Bendavid, E., Mulaney, B., Sood, N., Shah, S., Ling, E., Bromley-Dulfano, R., Lai, C., Weissberg, Z., et al. (2020). COVID-19 Antibody Seroprevalence in Santa Clara County, California. Stanford University, Stanford, CA.

Cheng, S.-C., Chang, Y.-C., Chiang, Y.-L. F., Chien, Y.-C., Cheng, M., Yang, C.-H., Huang, C.-H., and Hsuh, Y.-N. (2020). First case of Coronavirus Disease 2019 (COVID-19) pneumonia in Taiwan. *Journal of the Formosan Medical Association*, 119(3):747–751.

Chollet, F. (2017). Xception: Deep Learning with Depthwise Separable Convolutions. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

Ferguson, N. M., Laydon, D., Nedjati-Gilani, G., Imai, N., Ainslie, K., Baguelin, M., Bhatia, S., Boonyasiri, A., et al. (2020). Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand. Imperial College COVID-19 Response Team, Imperial College London.

Flaxman, S., Mishra, S., Bhatt, S., Gandy, A., Unwin, J. T., Coupland, H., Mellan, T. A., Zhu, H., et al. (2020). Estimating the number of infections and the impact of non-pharmaceutical interventions on COVID-19 in 11 European countries. Imperial College COVID-19 Response Team, Imperial College London.

Ioffe, S. and Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv:1502.03167*.

Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C. S., Liang, H., Baxter, S. L., McKeown, A., Yang, G., et al. (2018). Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell*, 172(5):1122–1131.

Kingma, D. P. and Lei-Ba, J. (2015). Adam: A Method for Stochastic Optimization. *Published as a conference paper at ICLR 2015*.

Narin, A., Kaya, C., and Pamuk, Z. (2020). Automatic Detection of Coronavirus Disease (COVID-19) Using X-ray Images and Deep Convolutional Neural Networks. *arXiv:2003.10849*.

Simonyan, K. and Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. *International Conference on Learning Representations*.

Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. (2016). Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*, Google Inc.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2015). Rethinking the Inception Architecture for Computer Vision. *arXiv:1512.00567*.

Wang, S., Kang, B., Ma, J., Zeng, X., Xiao, M., Guo, J., Cai, M., Yang, J., et al. (2020). A deep learning algorithm using CT images to screen for Corona Virus Disease (COVID-19). *Pre-publication - medRXiv*.

Xu, X., Jiang, X., Ma, C., Du, P., Li, X., Lv, S., Yu, L., Chen, Y., et al. (2020). Deep Learning System to Screen Coronavirus Disease 2019 Pneumonia. *arXiv:2002.09334*.