COVID-19 PNEUMONIA DETECTION

SOFÍA VERA, LUCÍA GARCÍA, SANDRA SALCEDO

Medical Need

Coronaviruses are a family of viruses that normally only affect animals. Some of them also have the ability to be transmitted from animals to people, causing respiratory problems such as Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome, that it is known as SARS. [1]

Currently the virus that is affecting the entire world is known as **SARS-CoV-2** and the disease it causes is called **COVID-19**, that is a new type of coronavirus that can affect people and was first detected in December 2019 in Wuhan City, China. [1]

However, it is important to know the complications produced by this epidemic virus, some of them are the following [2]:

- **Pneumonia**
- Failure of various organs
- Heart problems
- Acute respiratory distress syndrome
- Blood clots
- Acute kidney injury
- Additional virus and bacterial infections

Pneumonia is one of the most important complications since it leaves many long-term sequelae. The type of pneumonia that can be caused by COVID-19 is bilateral pneumonia, that develops when pathogens inflame the lungs and lead to **severe respiratory failure**. Therefore, the sooner it is detected the better, since it will be treated from the beginning and may leave fewer sequelae.[3]

At the beginning of the pandemic, **pneumonia was one of the main important focus of the doctors**, so if anyone during the pandemic period had symptoms related to this disease, they should tell the doctor and he would first perform a chest RX, and if the doctor saw any sign of pneumonia then they did a PCR in order to check if the disease was caused by COVID-19, but these type of tests were very scarce at the beginning.

Nowadays, the **gold standard** is still the PCR, but it is a very time-consuming and complicated process. Other alternative is the screening method, where **chest X-ray** (CXR) or computed tomography (CT) imaging are analyzed by radiologists to look for indicators associated with SARS-CoV-2, as well as the pneumonia.

The main advantages of CXR is the **rapid detection** of COVID-19 and that it can be done in parallel of other screening viral methods, in addition is **available** and **accessible** in many clinical sites. Another important thing is its **portability** that means that imaging can be performed within an isolation room, thus significantly reducing the risk of COVID-19 propagation.[4]

Figure 1. Healthy patient [5]

Figure 2. Patient with pneumonia caused by COVID19 [5]

Type of images

The images we have used for this project are of patients with and without pneumonia caused by COVID-19 (more detailed in the Data Description section), therefore we are going to use these images to diagnose when a patient suffers from pneumonia or not.

For these reasons, we thought that self-detection by prediction models will be helpful for the current health emergency, so that, in our case we will use a deep learning model as prediction model.

State of the art

As it has already been explained, this project is focused on the detection of pneumonia caused by the COVID-19 with a predictive model of Deep Learning (DL). Keeping this in mind, we proceed to study which combination of hyperparameters is optimal to build a precise and fast algorithm.

To do this, we have done a previous research of scientific papers whose objectives were similar to the ones we try to achieve.

Since a few years ago, several scientists have been interested in detecting different types of pneumonia in images of chest X-rays with the help of Artificial Intelligence (AI). We can find an example in the article *CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning* [6] that was published in November 2017. They developed an algorithm called CheXNet, a Convolutional Neural Network (CNN) which consists of 121 layers, which input is chest image, and its output is the probability that this belongs to certain pathology. They obtained excellent results even exceeding the average radiologist performance.

However, although this example and our project are referred to patients with pneumonia, we have found that there are some differences between them based on their origin that will have to be considered. Furthermore, the article *Differentiating novel coronavirus pneumonia from general pneumonia based on machine learning* [7] published on August 2019 explains that for the distinction, before feature extraction, the Region Of Interest (ROI) was delineated implementing Ground-Glass Opacities (GGOs) which was the guide to differentiate both. COVID-19 presents a more bilateral extension while the general pneumonia shows one that is more unilateral and consolidated.

And even this study was carried out with CT images, and we could find inequalities when using images taken with X-Ray, it already leads us to think that our research should be focused on more recent studies in which it will be considered that pneumonia is caused by COVID-19.

Furthermore, due to the sanitary emergency, we have been living and we are living, we find that a lot of studies have been accomplished in this last 2020. At the beginning, we research the application of AI encompassing both Machine Learning (ML) and DP methods, but owning to the problem we are facing, we have decided to focus on DP, highlighting two studies in particular.

COVIDX-Net: A framework of Deep Learning Classifiers to Diagnose COVID-19 in X-Ray Images [8] is a paper published on March, in which it explains they have created an algorithm called COVIDX-Net that includes seven different architectures of CNN models, such as modified Visual Geometry Group Network (VGG19) and the second version of Google MobileNet. Each deep neural network model is able to analyse the normalized intensities of the X-ray image to classify the patient status either negative or positive COVID-19 case. Even so, its main limitation is related to the extension of the database that has been used to train and evaluate the model, as they had only 50 images being positive 25. This can lead to the fact that the expected results are not obtained when it comes to prove it in clinical studies.

On the other hand, *COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images* [9] is other paper, this published in November, with a more extensive database that has created an initial network design prototype whose architecture is of a wide diversity with a selective long-range connectivity. In addition, since it uses a projection-expansion-projection design pattern, there is an enhanced representational capacity that at the same time allows to maintain the computational efficiency. Finally, they have reached a help for the diagnosis at the same time a way of assessment who should be prioritized in order to be tested with a PCR.

To conclude, it is worth noting to highlight the efficacy and the speed of how this setback has been handled with the help of AI and image analysis. So, our objective, on a much smaller scale due to the resources we have, will be based on verifying that we can also tackle a problem of this magnitude with the techniques learned along the course, and with this investigation.

Data Description

Our database is stored in a folder called *xray_dataset_covid19* and consists of two directories:

- *train*. This one corresponds to the images that will be used for the training of the model. Inside of it there other two directories called NORMAL and PNEUMONIA. Each one with 74 images, the NORMAL ones for the healthy patients and the PNEUMONIA ones for the positive COVID-19 patients.
- *test*. This one corresponds to the images that will be used for the evaluation of the model. As before, inside of it there other two directories called NORMAL and PNEUMONIA. Each one with 20 images, the NORMAL ones for the healthy patients and the PNEUMONIA ones for the positive COVID-19 patients.

## Methodology

- Image processing

The image processing including load of database, pre-processing methods, creation of a model, evaluation, and predictions, is created in the Jupyter Notebook.

Before creating the model, we have realized that images of our database have different sizes. This will be a problem when introducing them into the model and when predicting labels of different images.
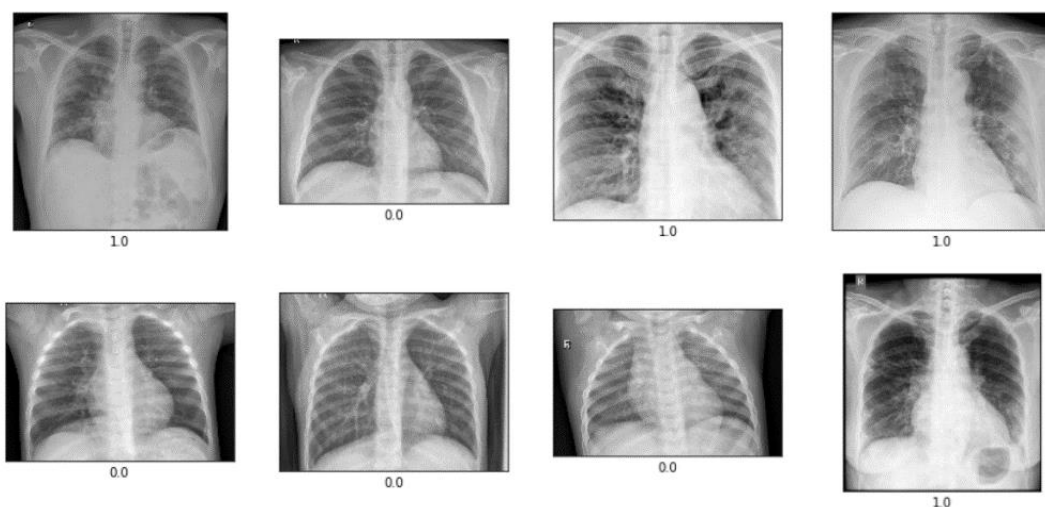
In order to normalize the images, we have selected the minimum width and height of all the images that is our case is (237, 255). The final size of the images is the following.

	Width	Height	Index
count	188.000000	188.000000	188.000000
mean	1430.377660	1725.228723	93.500000
std	633.787943	698.754955	54.415071
min	237.000000	255.000000	0.000000
25%	1039.750000	1443.750000	46.750000
50%	1445.000000	1796.000000	93.500000
75%	1767.750000	2059.250000	140.250000
max	4095.000000	4248.000000	187.000000
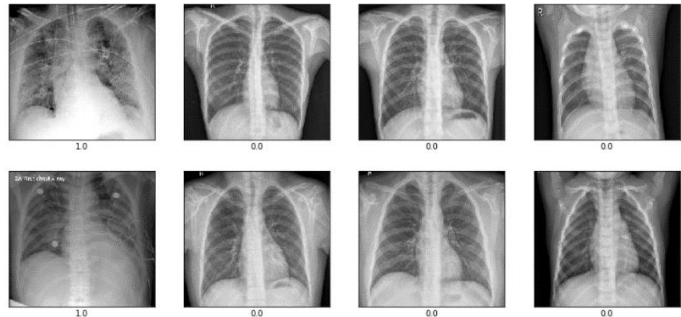
After normalizing, creation of the model can start. In order to create the model, we have decided to use Convolutional Neural Networks (CNN). The definition and reasons for using CNN instead of a simpler Deep Learning method are explained in the Jupyter Notebook.

- Model creation

The model created is composed by several layers such as two convolutional layer, two max pooling layer and two dropout layer. Then, a fully connected network is created thanks to a hidden dense layer and a dense output layer with a sigmoid function. Besides, a flatten layer is applied to transform inputs to a 1D array.

The composition of the layers and their deep explanation is explained in the Jupyter notebook.

This is the summary of the model is the following one:

```
Model: "sequential_1"
_____
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_2 (Conv2D)            (None, 235, 253, 60)      600
_____
max_pooling2d_2 (MaxPooling2 (None, 117, 126, 60)      0
_____
dropout_2 (Dropout)          (None, 117, 126, 60)      0
_____
conv2d_3 (Conv2D)            (None, 115, 124, 110)     59510
_____
max_pooling2d_3 (MaxPooling2 (None, 57, 62, 110)       0
_____
dropout_3 (Dropout)          (None, 57, 62, 110)       0
_____
flatten_1 (Flatten)          (None, 388740)            0
_____
dense_2 (Dense)              (None, 200)               77748200
_____
dense_3 (Dense)              (None, 1)                 201
=================================================================
Total params: 77,808,511
Trainable params: 77,808,511
Non-trainable params: 0
```

Finally, after applying an Adam optimizer, a loss value metric such as binary cross entropy, an accuracy metric such as accuracy and two call-backs variables (to avoid overfitting), model is trained with the train set and evaluated with the test set.

Conclusions

The result of the loss value and accuracy will change each time we run the model. In this case loss value is 0.04444122463464737 and accuracy is 1. However, loss value is always very small, and accuracy is always in a range between 0.97-1, this means a very high accuracy. These results conclude that our model is a very suitable model for our purpose.
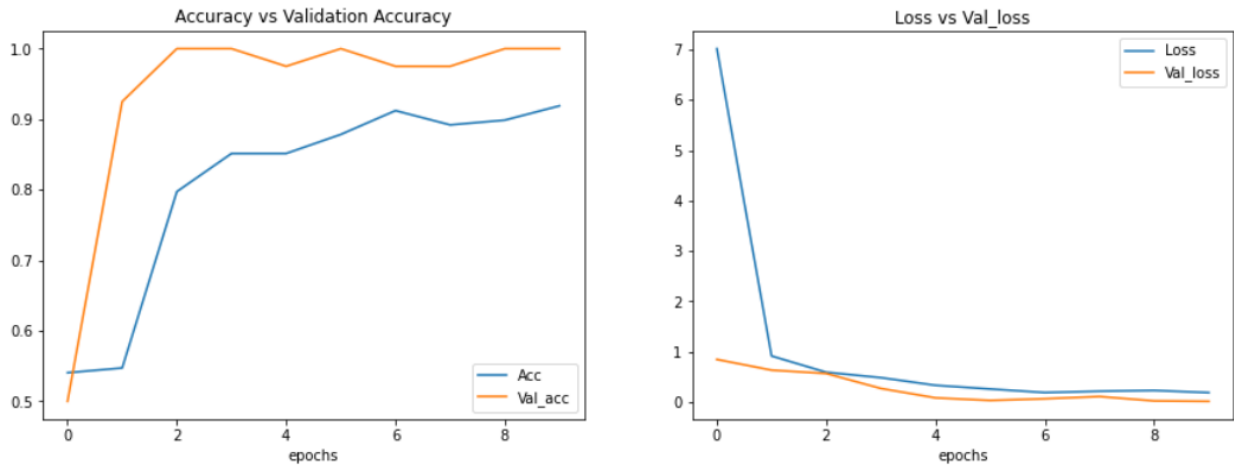
In these graphs we can visualize the performance of the model regarding the accuracy and loss value with respect to the different number of epoch.

It can be seen how accuracy (for train and validation) gets really high values with a few epochs. However, this value starts to stabilise with 8 or more epochs. For this reason, we are going to use 10 epochs.

Besides, in terms of loss value, it is seen how it reaches a very low value with only 2 epochs. However, close to 10 epochs the loss value is even smaller and the difference between loss value in train and validation is minimal.

- Model predictions

A. Predictions of our database

Prediction of the 40 images of the test set is carried out with our model. We have created two different lists; one is going to contain the correct predictions and another the incorrect ones. The results are the following.

```
print(correct_indices)
```
```
[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 3
3, 34, 35, 36, 37, 38, 39]
```

```
print(incorrect_indices)
```
```
[]
```

It can be seen as the 40 images are well-predicted. This means that our model works perfectly in this case.

Let us visualize 8 images of the 40 images of our database.
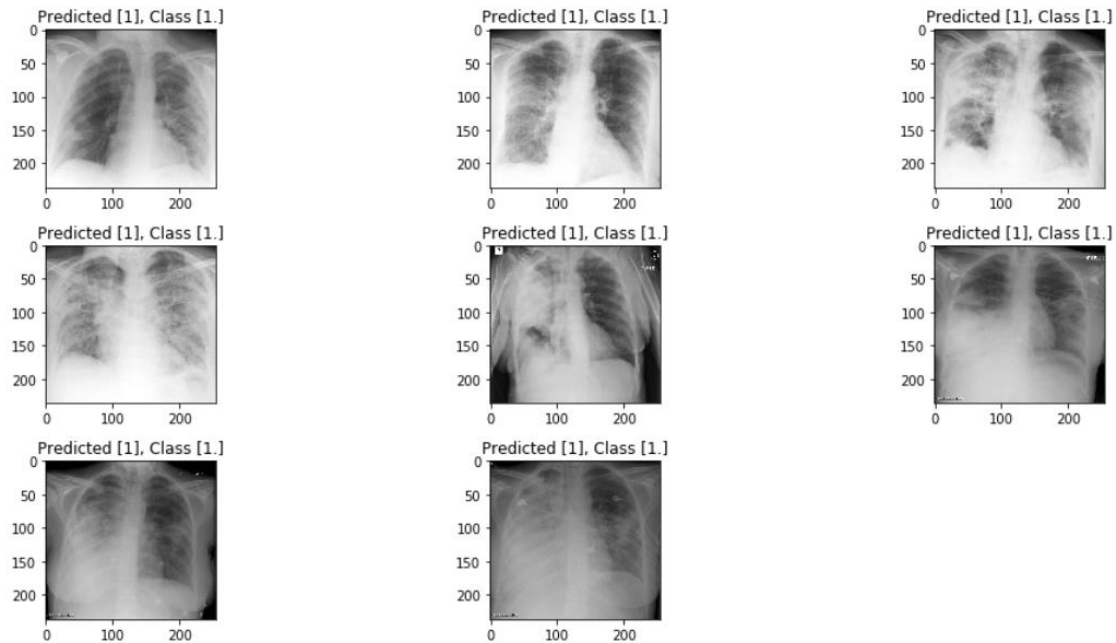
As it can be seen, predicted class corresponds to the actual class; it can be also be determined that the predictions are correct looking at the images. Images that seem to have pneumonia are predicted as 1, that means there is a disease.

B.   Prediction of an external image

In order to check that our algorithm works not only with our database, but with any kind of image, we apply to the model and external pneumonia image. The result shows that prediction is well-done, the image is predicted as 1, with disease. This is another key point to determine that our model works really well.


```
predicted_classes

array([[1]])
```

Final conclusions

As it has been tested in different cases such as with our database or with an external image, the model that we have created is very useful in order to determine if an X-ray presents pneumonia or not. This model could also be modified and applied to another medical fields or to other disease detections.

In spite of the fact that our model works well in this case, it could be work better and be applied to more important and decisive medical fields. However, due to the lack of time and the scarce resources, the project has been performed in low scale. Thanks to the application of better techniques and technologies and to the possibility of performing the problem in more time, this initial model or algorithm could be highly implemented.

The impact these algorithms can have in the health sector is really important. It could be of help for the doctors to have a tool in which they can ensure their diagnosis. Besides, it would be a help method for researching and development.

References

[1] *¿Qué es el Coronavirus SARS-CoV-2?: Hospital Clínic Barcelona.* Retrieved January 13, 2021, from
https://www.clinicbarcelona.org/asistencia/enfermedades/covid-19/definicion

[2] *Enfermedad del coronavirus 2019 (COVID-19)*. (2020, December 22). Retrieved January 13, 2021, from
https://www.mayoclinic.org/es-es/diseases-conditions/coronavirus/symptoms-causes/syc-20479963

[3]*Neumonía. (2020, June 13)*. Retrieved January 13, 2021, from https://www.mayoclinic.org/es-es/diseases-conditions/pneumonia/symptoms-causes/syc-20354204

[4] Wang, L., Lin, Z., & Wong, A. (2020, November 11). *COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images.* Retrieved January 13, 2021, from
https://www.nature.com/articles/s41598-020-76550-z#Sec3

[5] Khoong, W. (2020, March 19*). COVID-19 Xray Dataset (Train & Test Sets)*. Retrieved January 13, 2021, from https://www.kaggle.com/khoongweihao/covid19-xray-dataset-train-test-sets

[6] Rajpurkar, P. (2017, 14 noviembre). *CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning*. arXiv.org. https://arxiv.org/abs/1711.05225

[7] *Differentiating novel coronavirus pneumonia from general pneumonia based on machine learning*. (2020). PubMed Central (PMC). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7436068/

[8] Hemdan, E. E. (2020, 24 marzo). *COVIDX-Net: A Framework of Deep Learning Classifiers to Diagnose COVID-19 in X-Ray Images*. arXiv.org. https://arxiv.org/abs/2003.11055

[9] Wang, L., Lin, Z. Q., & Wong, A. (2020, 11 noviembre). *COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images*. Scientific Reports. https://www.nature.com/articles/s41598-020-76550-z?error=cookies_not_supported&code=622bf673-0a85-4307-8678-eb132d1e824f#Sec3https://www.google.es