

**Analysis of Covid-19 chest x-rays**

Data preprocessing and modelling report



July 18, 2024

Alexandra rancic, Philipp Trinh, preetha balakrishnan, Paul pourmoussavi

# INTRODUCTION

The goal of this project is to develop a deep-learning model for the detection of COVID-19 based on sample X-ray images of the chest region. As the first step towards this goal, we previously explored the data at hand and observed the following:

1. The X-ray images were 299\*299 pixels in size
2. The corresponding masks were 256\*256 pixels in size
3. The number of images and masks were equal
4. The number of images per category were not equal indicating that they must be weighted accordingly before training the model
5. The overlay between the images and corresponding masks worked well indicating that one could use it to mainly focus on the lung area of the image

In the current report, we will discuss in detail about image pre-processing that is imperative to develop a robust model, model development (deep and machine learning models) and model optimization.

# METHODLOGY

Sample X-ray image data required to train the model were obtained from Kaggle, a large machine learning and data science online platform that helps people to build their skills with various data-related challenges. It contains ‘.png’ images and corresponding masks of volunteers/patients grouped into four categories: normal (i.e. images from healthy volunteers), viral pneumonia, lung opacity, and COVID-19. Raw images and masks were studied and further processed step-by-step to make it more compatible for machine and deep learning using python.

# RESULTS

We divided the observations of this report into two parts. The first part mainly focuses on the pre-processing of the X-ray images and the second part delved deep into the various models we tested.

## Image pre-processing

### Selecting the lung area in images

Corona virus primarily resides in the lungs and leads to complications in this region. For this reason, we decided to use this area as our region of interest. To obtain this region of interest, we first resized the raw X-ray images to 256\*256 pixels in size. Using the corresponding masks of the same size, we extracted just the lung area of the X-ray of each image. This procedure was followed for every image from each category (Figure 1, shows a sample image from each category). This region of interest in grayscale was used in the subsequent pre-processing steps.

|  |
| --- |
|  |
| **Figure 1**: Selecting the lung region of the chest X-ray (a) Representative X-ray of a normal volunteer (b) X-ray image of a COVID-19 positive patient (c) X-ray image of a patient with an opaque lung (d) X-ray image of a patient with pneumonia. The scale signifies the width and height of the image in pixels. |

### Image processing

Image processing is one of the crucial steps before using the data for training any machine learning (ML) or deep learning (DL) model. Raw image data has ample noise that could unnecessarily delay the ML or DL learning time that in future could be costly for hospitals and diagnostic centres. Noise could also decrease the precision and accuracy of the model in question. Therefore, we tried various filtering techniques that could reduce the noise and improve the disease detection capacity of the model. To understand how each filter visually altered the image, we combined every filtering step with a Canny edge detection step.

#### The median blur filter

The median blur filter finds the median value in the neighbourhood of each pixel. It runs through each pixel and replaces it with the median value. We tested this filter using three different kernel sizes – (3, 3), (5, 5) and (7, 7). As shown in Figure 2a, the kernel size (5, 5) and (7, 7) blurred the image considerably and a subsequent Canny edge detection (see section 3.1.2.5 for details about Canny edge detection) of these images showed that the edges of the ribcage were not entirely detected, and it sort of introduced shapes in the images that were not represented in the original image (Figure 2b). Therefore, a kernel size of (3, 3) was the best as it retained the structures of the original image and reduced the noise in the image.

|  |
| --- |
|  |
| **Figure 2**: Median blur and Canny edge detection (a) Representative lung area of a normal volunteer and median blur filtering using a kernel size (or ksize) of (3, 3), (5, 5) or (7, 7) (b) Representative lung area of a normal volunteer and median blur filtering using the different kernels followed by a Canny edge detection (min = 0, max = 50). The scale signifies the width and height of the image in pixels. The arrow in red shows the thin fibre like bronchus edges detected with a Canny edge detector. |

The maximum threshold value used for Canny edge detection, for median blur is 50. With this value, even thin, tree like bronchus could be detected strongly (Figure 2b). This may interfere with model prediction (additional noise) in future and hence, a maximum value greater than 50 would ideally be suitable for edge detection with a kernel size of (3, 3).

#### Gaussian blur or smoothening

Another method that reduces noise and details in an image is the Gaussian blur. The visual effect of this blurring technique is a smooth blur resembling that of viewing the image through a translucent screen. It uses a Gaussian function (which also expresses the normal distribution in statistics) for calculating the transformation to apply to each pixel in the image. Values from this distribution are used to build a convolution matrix which is applied to the original image. Each pixel's new value is set to a weighted average of that pixel's neighborhood. The original pixel's value receives the heaviest weight (having the highest Gaussian value) and neighbouring pixels receive smaller weights as their distance to the original pixel increases. This results in a blur that preserves boundaries and edges better than other, more uniform blurring filters.

In our project, as shown in Figure 3, we tried to smoothen sample images (normal and covid) with a kernel size of (3, 3), (5, 5) or (7, 7).

|  |
| --- |
|  |
| **Figure 3**: Gaussian blur of X-ray images (a) Representative lung area of a normal volunteer filtered with a kernel size (3, 3), (5, 5) or (7, 7) (b) Representative lung area of a COVID patient filtered with a kernel size (3, 3), (5, 5) or (7, 7). The scale signifies the width and height of the image in pixels. |

As previously observed, kernels with a higher ksize significantly blurred the image. This reduced the noise but introduced shapes due to incomplete edge detection of the ribcage and other areas (Figure 4a, b). These areas could be helpful in differentiating a normal and COVID lung. Therefore, we decided to use the kernel size (3, 3) with any filter that will be tested in our project. A Gaussian blur using a (3, 3) kernel followed by edge detection seemed best in reducing noise and detecting the haziness in the lung area of patients with COVID as structures (Figure 4b). This could be a potential filtering technique that could be used to pre-process images for machine learning.

|  |
| --- |
|  |
| **Figure 4**: Gaussian blur and Canny edge detection (a) Gaussian filtered X-ray image of a healthy individual followed by a Canny edge detection (threshold - min: 0, max: 65) (b) Gaussian filtered COVID X-ray image followed by a Canny edge detection (threshold - min: 0, max: 65) |

#### Canny edge detection

The Canny edge detection algorithm was developed by John F. Canny in 1986. It is a multi-step process consisting of:

* Applying Gaussian smoothing (or other filters in our case) to the image to help reduce noise
* Computing the image gradients using the Sobel kernel
* Applying non-maxima suppression to keep only the local maxima of gradient magnitude pixels that are pointing in the direction of the gradient
* Defining and applying the minimum and maximum thresholds for Hysteresis thresholding

In our project, we detected the edges after each filtering step described in the pre-processing to visualize the impact of each filter respectively. We used a minimum threshold value of 0 and a maximum threshold between 50-100 initially. A maximum threshold of 75 and above did not completely detect the ribcage in normal (Figure 5a). Additionally, in case of COVID lungs, the haziness was not completely captured (Figure 5b). Therefore, we decided to focus on edge detection with the maximum threshold range of 65 (see figure 4b) for pre-processing.

|  |
| --- |
|  |
| **Figure 5**: The Canny edge detection after a Gaussian blur. Representative lung area of (a) a normal volunteer (b) a COVID patient after a Gaussian blur (kernel size (3, 3)) and Canny threshold of (min: 0, max: 50, 75 or 100). The scale signifies the width and height of the image in pixels. |

#### Erosion

This technique processes images based on the shapes in the image. It computes a local minimum over the area of a given kernel. As the kernel is scanned over the image, the minimal pixel value overlapped by it is computed and the image pixel under the anchor point is replaced with that minimal value. Visually, erosion erodes away the boundaries of the foreground object. Figure 6 shows a representative X-ray image of a normal individual with/without erosion and Canny edge detection. When we eroded this image with a (3, 3) kernel followed by Canny edge detection (min = 0, max = 65), we identified that the ribcage was thinner than that observed in other filters. However, like all other filtering techniques, erosion reduced noise and detected the crucial structures of the lung.

|  |
| --- |
|  |
| **Figure 6**: Erosion and Canny edge detection. (a) Representative raw or filtered lung area of normal volunteer with/without Canny edge detection (b) Representative COVID filtered lung area with or without erosion and Canny edge detection (c) Representative raw or filtered lung area of a person with an opaque lung and with/without Canny edge detection (d) Representative raw or filtered lung area of a person with viral pneumonia with/without Canny edge detection. The scale signifies the width and height of the image in pixels. The min and max threshold values for Canny edge detection was 0 and 65. |

#### Laplacian filter

The Laplacian filter detects sudden intensity transitions in the image and highlights the edges. It convolves an image with a mask [0,1,0; 1,− 4,1; 0,1,0] and acts as a zero crossing detector that determines the edge pixels. We applied the Laplacian filter using a ddepth of 64F. As shown in Figure 7, this filter did not detect the edges well and we almost did not observe much in the lung area especially in case of COVID and hence, we did not explore or use this filter for further analysis.

|  |  |
| --- | --- |
|  |  |
| **Figure 7**: Laplacian filter. **(a)** Representative X-ray image of a normal individual with or without the Laplacian filter and **(b)** Representative X-ray image of a COVID patient with or without the Laplacian filter. The scale shows the size of the image (256\*256) in pixels. |  |

After extensively studying all filters, we decided to consider only one filter and the raw lung image for further modelling and analysis. The median blur, Gaussian blur and erosion provided desirable results with a (3, 3) kernel. However, we chose the Gaussian filter as it is widely used for X-ray pre-processing [1]. We would use filtered and raw lung images for basic classification models to see how feature extraction impacts the accuracy of the model. Only raw images would be used to train deep learning models as feature extraction is already a part of complex models such as CNN and Lenet.

# DISCUSSION

Our preliminary analysis revealed the following:



# REFERENCES

1 https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0265949