Analysis of Covid-19 chest x-rays

Data exploration and visualization report

JUNE 18, 2024

ALEKSANDRA RANCIC, PHILIPP TRINH, PREETHA BALAKRISHNAN,
PAUL POURMOUSSAVI

1 INTRODUCTION

COVID-19 is a disease caused by the virus SARS-CoV-2, and it was discovered by the end of 2019 in Wuhan, China. This virus has spread fast around the globe, resulting in the global pandemic, which has changed our lives. The transmission of the virus occurs mainly via droplets, through coughing, sneezing, or even speaking with an infected person [1]. Initial studies have shown that the virus can survive on surfaces, which stimulates its transmission [2]. Aged people and those with chronic diseases, e.g., diabetes and cardiovascular diseases are at a higher risk of developing serious complications. The symptoms of COVID-19 vary from mild to severe. Fever, cough, shortness of breath and tiredness are considered mild symptoms. Nevertheless, in more serious cases, the virus can lead to various lung complications such as COVID-19 Acute Respiratory Distress Syndrome (CARDS) or/and pneumonia [3,4]. CARDS, a result of the fluid build-up in the alveoli in the lungs, causes difficulty in breathing and leads to severe hypoxia. Consequently, patients with CARDS require intensive medical care and even after their recovery the lung scars may stay permanently. This reduces respiratory function and further influences the quality of life [5]. Pneumonia, which manifests with changes in breathing patterns, breathlessness, chest pain, and hypoxia, in most cases also requires hospitalization and additional oxygen therapy or mechanical ventilation in intensive care units [6]. Studies have shown that along with acute lung problems, COVID-19 can cause long-term respiratory issues known as post-COVID syndrome, leaving permanent changes in the lung tissue [7].

Current methods employed to detect COVID-19 include real-time Polymerase Chain Reaction (RT-PCR), fast antigen tests, serological tests, genome sequencing, computer tomography (CT), and radiography (X-rays) of the lungs [8]. X-rays of the lungs were indispensable during the pandemic since this method is fast, available in hospitals, and provides valuable information on patients' conditions. X-rays are of great importance for monitoring the severity of the disease and its progress. This method is used to diagnose characteristics and signs of pneumonia and gives information about lung tissue damage [9].

We live in the era of Artificial intelligence (AI) and Machine Learning (ML), and the application of these technologies has already generated revolutionary changes in medicine. One of the fields in which AI may have a strong influence is disease diagnostics- especially for the detection of COVID-19 using radiography. Diagnosis of lung diseases via X-rays largely relies on the experience of the radiologist and may be subjective depending on the workload of the medical staff and their available time for analyses. Thus, AI and ML give the possibility of automated, fast, and precise analysis of X-rays while, at the same time, increasing the accuracy

of diagnosis and decreasing the time needed for decision [10]. This project focusses on building a deep learning model that recognizes characteristic lung patterns (such as pneumonia and lung opacity) of COVID-19 patients. Once trained, we hope that it can automatically analyses new X-rays with high accuracy, helping medical staff to identify infected patients faster and with more efficacy. Consequently, this association between AI and X-rays of the lungs to detect COVID-19 improves diagnosis and reduces the workload of healthcare workers. Early detection of COVD-19 is essential as it may reduce the burden on the healthcare system, burden of the disease, and may even help to ameliorate the life-threatening respiratory complications.

2 OBJECTIVE

The main goal of this project is to develop a deep-learning model for the detection of COVID-19 based on the X-ray images of the lungs. Through this process, a precise and robust tool, which will help in the faster and more precise diagnosis of COVID-19 infection, could be developed. However, in the current report only the first step i.e. data exploration and visualization will be highlighted.

3 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 of volunteers/patients grouped into four categories: normal (i.e. images from healthy volunteers), viral pneumonia, lung opacity, and COVID-19. Raw images and metadata were studied at length. Aspects such as number of images, their format, image size, color/grey scale images, mean and standard deviation of pixel intensities were explored, corresponding visualizations were created and reported. All analysis were performed using Python version 3.12.3.

4 RESULTS

As the first step to explore the data at hand, we investigated the metadata that were a part of each category of images. Each metadata file was available as an excel sheet with data about the X-ray images. It consisted of 4 columns namely – file name, format, size and URL, that were common to all metadata files. The column 'file name' provided unique names for each image, the 'format' column specifies the format of the image file, in our case it is a '.png' file, the 'size' column specifies the size in pixels of each image and the URL column provides the link from where one could download these images. Using the file name, we calculated the number of

unique names (i.e. is the number of unique images) per category. As shown in Fig 1, the number of images from healthy people are the highest. The number of images of patients with pneumonia, COVID and those with an opaque lung are comparatively lower. This uneven number of images may affect the training of the model in the following steps, and we may have to adjust these differences.

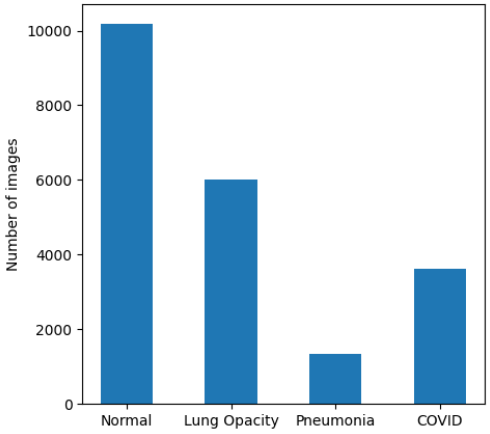

Figure 1: Number of images for each category or group of images.

We next looked into the column 'size'. All images were of the same size, 256*256 pixels (Table 1). The URL column is not relevant as we already downloaded the images. As we could gather all information the metadata files provide from the images, we did not explore these files further.

Table1. Tabulation of the unique image sizes per category

Category	**Unique sizes**
Normal	256*256
Lung Opacity	256*256
Pneumonia	256*256
COVID	256*256

Next, we explored the images and masks in each category. The number of images and masks per category were the same (Figure 2). Similar to what we observed with the metadata, the number of images in the healthy (normal) group was higher compared to all other groups (Figure 2).

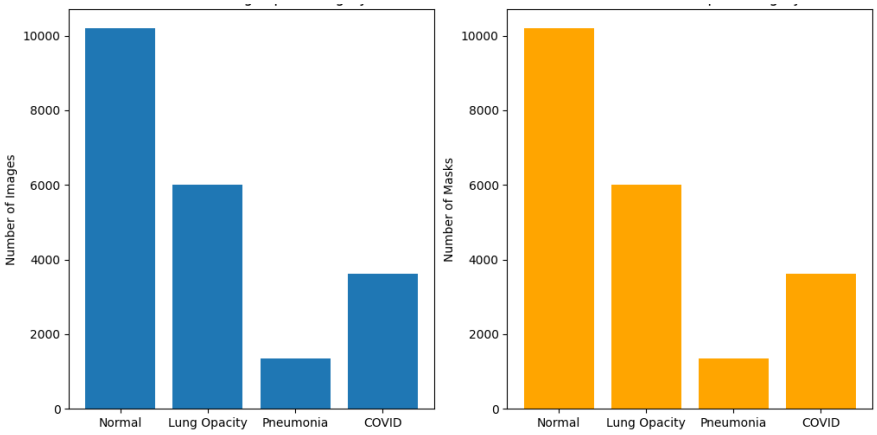

Figure 2: Number of images (left, blue bars) and number of masks (right, orange bars) for each category or group of images.

Raw images of each group were 299* 299 pixels in size (Figure 3a). The corresponding masks, however, were 256*256 pixels in size (Figure 3b). To our surprise, this information was different from what was provided in the metadata sheet.

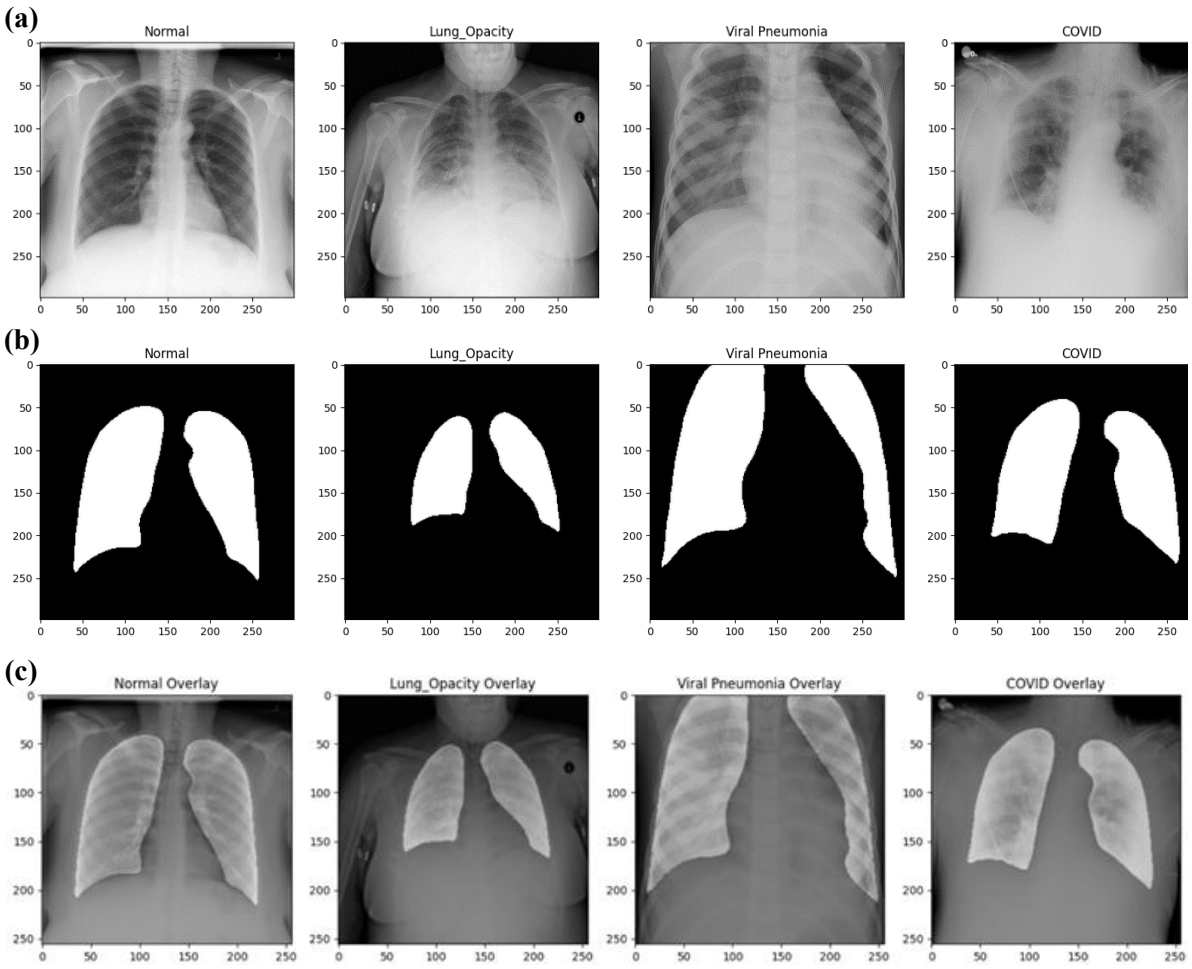

Figure 3: Representative **(a)** grayscale raw images (299x299 pixels) **(b)** corresponding masks (256x256 pixels) and (c) overlay of images and masks (256x256 pixels) one from each category. The scale shows the size of the image. The transparency level of masks in the overlay is 90

Furthermore, re-sizing images to 256x256 pixels and overlaying masks and images showed that the masks were a good approximation of the lung area, the main site corona virus targets (Figure 3c).

We further studied the pixel intensities of images and masks. To best understand the data in a short time, we limited our analysis to the first 50 images per category. As shown in Figure 4, the mean intensity range is comparable in all categories or groups. The mean intensity of X-rays from COVID patients was slightly higher compared to all other groups and the standard deviation was comparatively lower. In case of masks, although the trend is similar to that of raw X-ray images, the mean intensity of masks from patients with an opaque lung was lower compared to masks from all other categories. Here, the standard deviation was almost similar in all groups but compared to the raw images, it was higher. Although these are the preliminary impressions, we cannot use these interpretations until statistically proved.

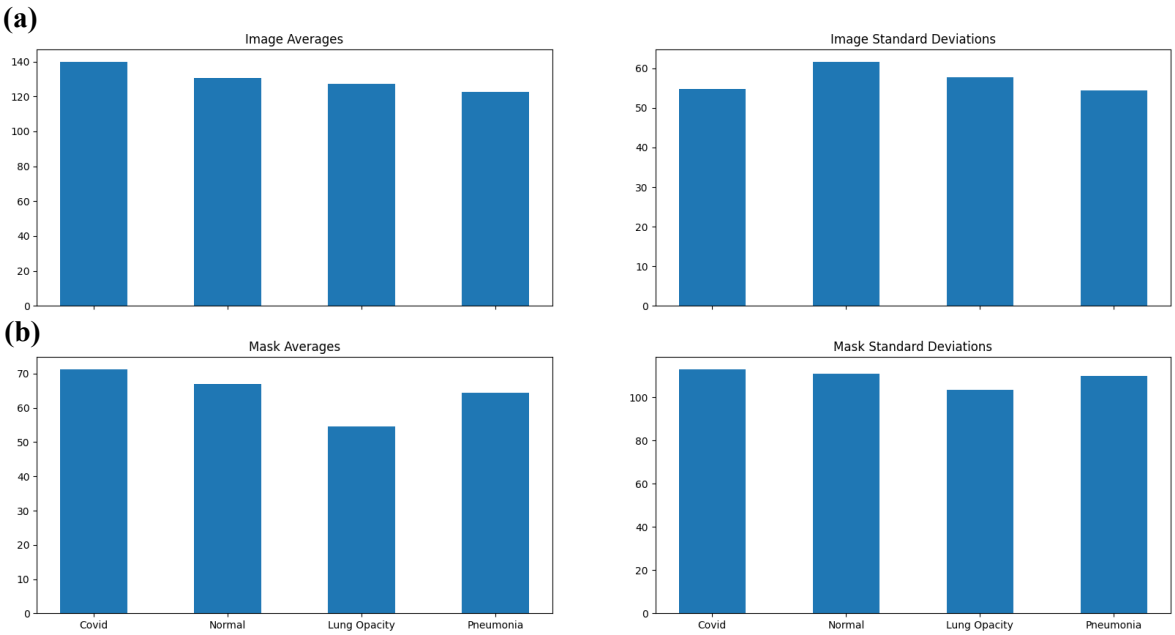

Figure 4. Mean intensity and standard deviation of **(a)** X-ray images and **(b)** corresponding masks for the first 50 images of each category.

Lastly, we explored the pixel intensities of representative images and masks per category. As expected, the intensity distribution of individual X-ray images ranged between 0 - 255 (Figure 5a). Masks are black and white derivatives of raw images and hence, would have either a 0 or 255 as its intensity. This is replicated in Figure 5b.

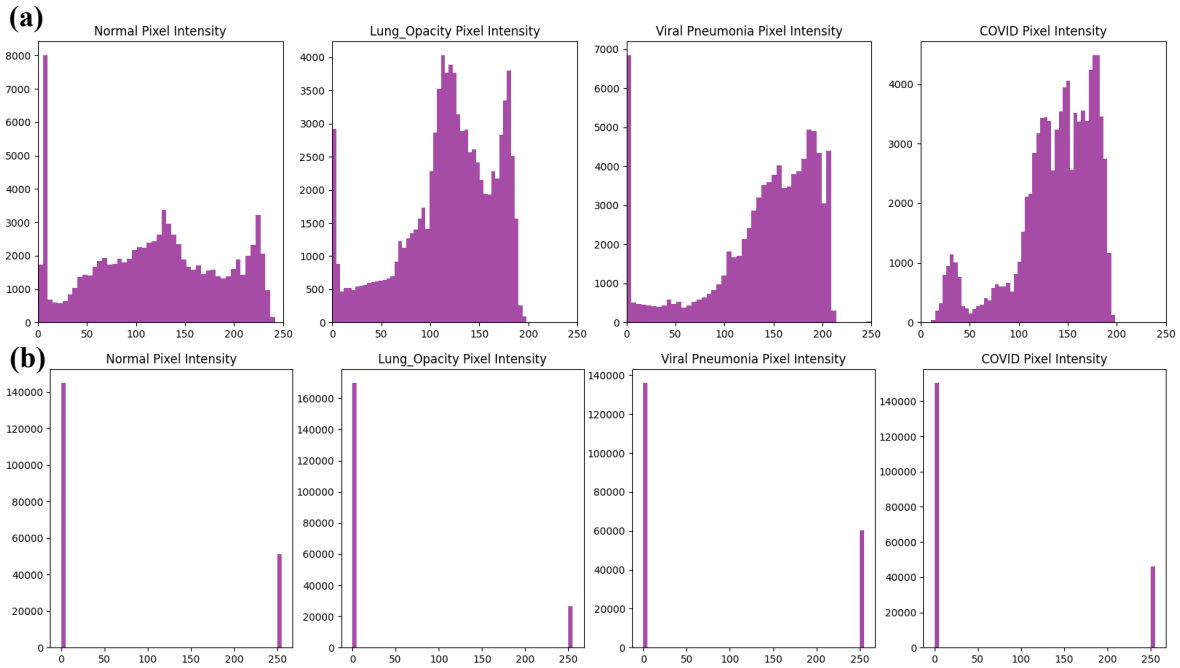

Figure 5. Pixel intensity distribution of a representative **(a)** X-ray image and **(b)** corresponding mask from each category. The x-axis defines the pixel intensity, and the y-axis defines the counts

5 DISCUSSION

Our preliminary analysis revealed the following:

1) The metadata files are not relevant for our analysis as we get similar information from the X-ray images and masks. Also, there is discrepancy in the data provided in the metadata and original images
2) The number of images is not the same for each category and may need to be adjusted before we train our model
3) The size of the image although the same across categories (299*299 pixels), is different from the size of the corresponding masks (256*256 pixels). We will re-size all images to 256*256 pixels in the pre-processing step
4) After resizing representative images to 256*256 pixels, an overlay of masks and images showed a good overlap in the lung area. This indicates that the masks best represent the lung area and it can be used for subsequent analysis if needed.
5) The pixel intensity of images is in similar range for all images across categories and their corresponding masks. The mean intensity of X-rays from COVID patients is slightly higher compared to the rest as expected. These trends suggest the accuracy of data/images and ensures that the images are labelled or named rightly. Hence, we could

use these images to train our model. Additionally, we could also think whether normalization of intensities is required for subsequent analysis.

6 REFERENCES

[1] M. Ciotti, M. Ciccozzi, A. Terrinoni, W.-C. Jiang, C.-B. Wang, S. Bernardini, The COVID-19 pandemic, Crit. Rev. Clin. Lab. Sci. 57 (2020) 365–388. https://doi.org/10.1080/10408363.2020.1783198.

[2] A. Spena, L. Palombi, M. Carestia, V.A. Spena, F. Biso, SARS-CoV-2 Survival on Surfaces. Measurements Optimisation for an Enthalpy-Based Assessment of the Risk, Int. J. Environ. Res. Public. Health 20 (2023) 6169. https://doi.org/10.3390/ijerph20126169.

[3] M.A. Matthay, A. Leligdowicz, K.D. Liu, Biological Mechanisms of COVID-19 Acute Respiratory Distress Syndrome, Am. J. Respir. Crit. Care Med. 202 (2020) 1489–1491. https://doi.org/10.1164/rccm.202009-3629ED.

[4] L. Gattinoni, D. Chiumello, S. Rossi, COVID-19 pneumonia: ARDS or not?, Crit. Care 24 (2020) 154. https://doi.org/10.1186/s13054-020-02880-z.

[5] J. Aranda, I. Oriol, M. Martín, L. Feria, N. Vázquez, N. Rhyman, E. Vall-Llosera, N. Pallarés, A. Coloma, M. Pestaña, J. Loureiro, E. Güell, B. Borjabad, E. León, E. Franz, A. Domènech, S. Pintado, A. Contra, M. del S. Cortés, I. Chivite, R. Clivillé, M. Vacas, L.M. Ceresuela, J. Carratalà, Long-term impact of COVID-19 associated acute respiratory distress syndrome, J. Infect. 83 (2021) 581–588. https://doi.org/10.1016/j.jinf.2021.08.018.

[6] C. Scelfo, M. Fontana, E. Casalini, F. Menzella, R. Piro, A. Zerbini, L. Spaggiari, L. Ghidorsi, G. Ghidoni, N.C. Facciolongo, A Dangerous Consequence of the Recent Pandemic: Early Lung Fibrosis Following COVID-19 Pneumonia – Case Reports, Ther. Clin. Risk Manag. 16 (2020) 1039–1046. https://doi.org/10.2147/TCRM.S275779.

[7] J.-M. Anaya, M. Rojas, M.L. Salinas, Y. Rodríguez, G. Roa, M. Lozano, M. Rodríguez-Jiménez, N. Montoya, E. Zapata, D.M. Monsalve, Y. Acosta-Ampudia, C. Ramírez-Santana, Post-COVID syndrome. A case series and comprehensive review, Autoimmun. Rev. 20 (2021) 102947. https://doi.org/10.1016/j.autrev.2021.102947.

[8] O. Filchakova, D. Dossym, A. Ilyas, T. Kuanysheva, A. Abdizhamil, R. Bukasov, Review of COVID-19 testing and diagnostic methods, Talanta 244 (2022) 123409. https://doi.org/10.1016/j.talanta.2022.123409.

[9] V. Nikolaou, S. Massaro, M. Fakhimi, L. Stergioulas, W. Garn, COVID-19 diagnosis from chest x-rays: developing a simple, fast, and accurate neural network, Health Inf. Sci. Syst. 9 (2021) 36. https://doi.org/10.1007/s13755-021-00166-4.

[10] A.A. Borkowski, N.A. Viswanadhan, L.B. Thomas, R.D. Guzman, L.A. Deland, S.M. Mastorides, Using Artificial Intelligence for COVID-19 Chest X-ray Diagnosis, Fed. Pract. 37 (2020) 398–404. https://doi.org/10.12788/fp.0045.