ANALYSIS OF COVID-19 CHEST X-RAYS: Report 1: Exploration, data visualization, and data pre-processing report

Saniya Arfin, Yvonne Breitenbach, Alexandru Buzgan

April 2025

Contents

1 Introduction

1.1 Context

Saniya: This project is very relevant to my translational oncology and biomedical research background. It directly involves the analysis of real patient data in the form of chest radiographs, a clinically significant modality often used in diagnostic workflows. The classification of COVID-19 and related lung pathologies through machine learning aligns well with my experience in biomarker discovery, cancer diagnostics, and data-driven decision support. The hands-on application of advanced machine learning algorithms, along with techniques for data preprocessing, bias detection, and model interpretability, further strengthens my skill set in translational bioinformatics, making this project both practically valuable and conceptually aligned with my future goals in medical AI.

Yvonne: As I have no background in medicine, biology or the analysis of medical data, the context of this project is completely new to me. Furthermore, I have no experience working with X-ray images or in image processing in general. However, working on this project is very valuable to me as a meteorologist because meteorology also often works with image data, such as satellite images or radar images. In recent years, the use of machine learning and deep learning models has become more and more important in meteorology as in many industries. For example, to classify cloud types in satellite images and overall to improve the weather forecast. My goal is to combine what I will learn from the work on this project with my meteorological background to apply this in future jobs.

Alexandru:I chose this project because of my interest in image recognition. I have a background as a Software Engineer in the automotive industry, and the development of self-driving automobiles makes use of this type of technologies quite often. The skills I will acquire on this project will suit me in the projects that I hope to work on in the future. Image recognition and classification may hold the key to the future of autonomous vehicles, which will have to identify and classify possible obstacles on their path. Both supervised and unsupervised learning algorithms are key components of these technologies. My goal is to add these skills to my previous experience and be better prepared for the technologies of tomorrow.

1.2 Objectives

- Early detection of COVID-19 from X-rays
- Importance of radiology in pandemics
- Compare COVID-19 with other diseases (e. g., viral pneumonia)
- Real-world implications (e. g., hospital triage, AI for diagnostics)

1.3 Goals

Identifying the problem

Supervised learning: Here we try to predict the target variable (dependent): COVID, Normal, Viral Pneumonia, Lung Opacity using supervised learning. Since we are looking to predict a qualitative variable it is a classification problem. To achieve this, the following steps must be carried out:

- Classify chest X-ray images
- Understand dataset distribution
- Identify class imbalance, noise, or biases
- Prepare data for a ML/DL model

2 Understanding and manipulation of data

2.1 Framework

Dataset We obtained COVID-19 Radiography Dataset.zip (777 MB) from Kaggle (https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database). This dataset is open access.

Dataset Description The dataset includes X-ray images of lungs and their corresponding masks. Table 1 shows the distribution of the X-ray images and the lung masks by class. In all four classes there are as many X-ray images as lung masks.

class	**number of X-ray images**	**number of lung masks**
Normal	10192	10192
Lung Opacity	6012	6012
COVID	3616	3616
Viral Pneumonia	1345	1345

Table 1: Distribution of X-ray images and lung masks by class

The normal class contains X-ray images of healthy lungs. The other three classes contain X-ray images of diseased lungs. The diseases are divided into COVID-19, viral pneumonia and lung opacity (non-COVID lung infections).

Metadata Additionally to the X-ray images and lung masks there are four metadata files available on the Kaggle website.

- COVID.metadata.xlsx
- Lung_Opacity.metadata.xlsx
- Normal.metadata.xlsx
- Viral Pneumonia.metadata.xlsx

These metadata files contain additional information, e. g. the size of the images and their source.

2.2 Pre-processing and feature engineering

2.2.1 Exploratory Data Analysis (EDA)

Class Distribution

As already shown in table 1 the four classes contain different numbers of images. Figure 1 illustrates the distribution of images and masks in the different classes. Almost half of all images/ masks (48.2 %) belong to the ”Normal” class. Second most images/ masks are in the ”Lung Opacity” class (28.4 %), followed by the ”COVID” class (17.1 %). The ”Viral Pneumonia” class contains the least images/ masks (6.4 %).

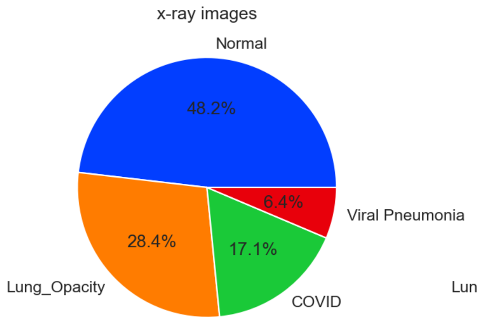

Figure 1: Precentage of X-ray images and lung masks by class

Image Visualization

In order to get an idea of the X-ray images five examples of each of the four classes are shown in figure 2.

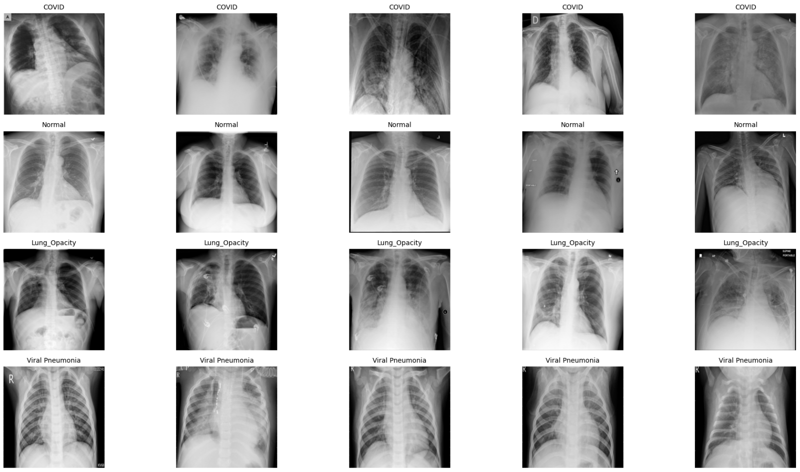

Figure 2: Examples of X-ray images of the four classes

Image Size Distribution and Type of Image

In the four files with metadata the size of each image is given. We first analyzed the image sizes from these metadata files and found out that all the images have the same size: 256 x 256 pixel.
As a next step we opened each X-ray image and the lung masks with Python and determined their sizes. The read-out image sizes are summarized in table 2. All the lung masks have a size of 256 x 256 x 3 pixel. The first two dimensions are the same as specified in the metadata files. But the masks have additionally a third dimension, which indicates that they aren't saved as grayscale images.
The analysis showed that the size of the X-ray images in the first two dimensions is 299 x 299 pixels and therefore different from the size of the lung masks. Therefore we conclude that the lung masks have to be resized before matching them with the X-ray images.
Besides that, the table 2 shows that most of the X-ray images (21025) have only two dimensions and therefore are grayscale images. But there are 140 X-ray images that have three dimensions and thus aren't grayscale images. These 140 RGB images belong to the class "Viral Pneumonia". Figure 3 shows the proportion of RGB- and grayscale images in the class "Viral Pneumonia". These RGB-images should be converted into grayscale images before using them for modeling.

Image size	**Number of X-ray images**	**Number of lung masks**
299 x 299 x 3	140	
299 x 299	21025	
256 x 256 x 3		21165

Table 2: Size of X-ray images and lung masks

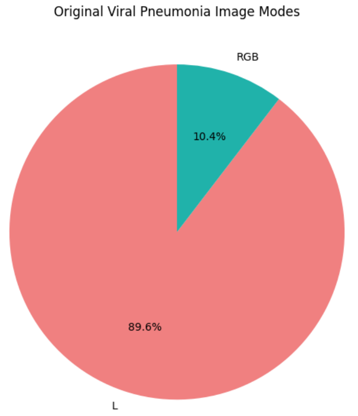

Figure 3: Proportion of RGB- and L modes in the class ”Viral Pneumonia”.

File format consistency

We confirmed format consistency to avoid read errors, and we observed all images were in PNG format as shown in Fig. 4.

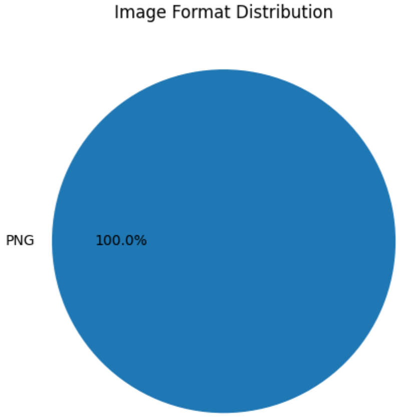

Figure 4: Image format distribution

Class Proportion by Source URL

To identify potential dataset bias by image source, we plotted a stacked bar chart of image count by class and URL source, figure 5. This figure and table 3 show that the X-ray images come from eight different sources. Most of the images come from the source https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/data.

Further investigation shows that all images of the "Lung Opacity" class and most images of the "Normal" class come from this source. The remaining part of the "Normal" classes images and all images of the class "Viral Pneumonia" are from the source https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia. The images of the "COVID" class come from six other sources, but most of them from https://bimcv.cipf.es/bimcv-projects/bimcv-covid19/#1590858128006-9e640421-6711.

Further analysis has to be done to investigate if the images from different sources are somehow different.

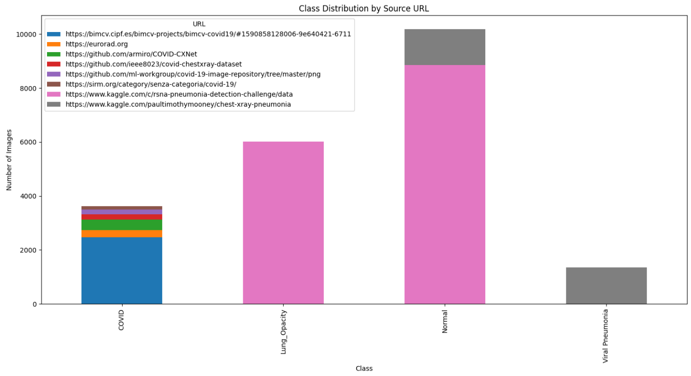

Figure 5: Class Distribution by source url

Label	URL	Count
COVID	https://bimcv.cipf.es/bimcv-projects/bimcv-covid19/#1590858128006-9e640421-6711	2474
	https://eurorad.org	258
	https://github.com/armiro/COVID-CXNet	400
	https://github.com/ieee8023/covid-chestxray-dataset	182
	https://github.com/ml-workgroup/covid-19-image-repository/tree/master/png	183
	https://sirm.org/category/senza-categoria/covid-19/	119
Lung_Opacity	https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/data	6012
Normal	https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/data	8851
	https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia	1341
Viral Pneumonia	https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia	1345

Table 3: Image source URLs and corresponding image counts for each label.

2.2.2 Exploratory Data Visualizations

We created class-wise smoothed Kernel Density Estimate (KDE) plots

- for mean pixel intensities and
- for standard deviations (std) of pixel intensities.

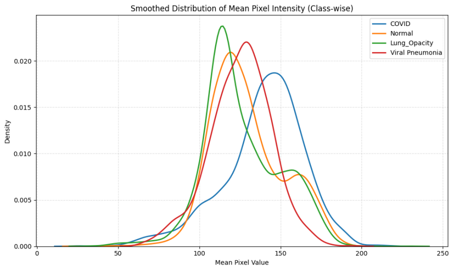

Figure 6: Distribution of mean pixel intensity by Image classes

Inference: Different classes show distinct mean intensity patterns. COVID images tend to be brighter on average, potentially due to artifacts or progression patterns, while Lung Opacity images are darker.

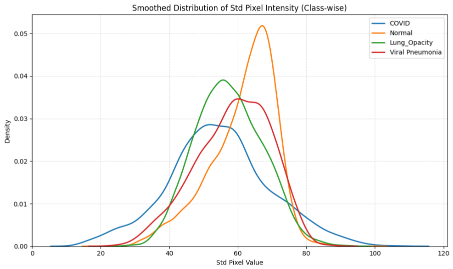

Figure 7: Distribution of standard deviation of pixel intensity by Image classes

Inference: COVID images have higher variability in pixel intensity, possibly due to complex lung patterns like patchy opacities. Normal images are consistent and smooth.

Additionally, we created sourcewise smoothed Kernel Density Estimate (KDE) plots to identify source bias

- for mean pixel intensities and
- for standard deviations (std) of pixel intensities.

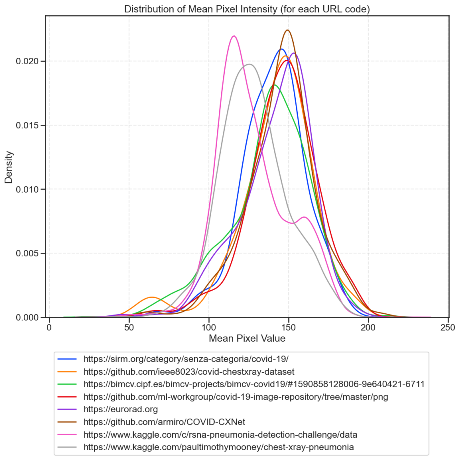

Figure 8: Distribution of mean of pixel intensities of the X-ray images depending on the different sources

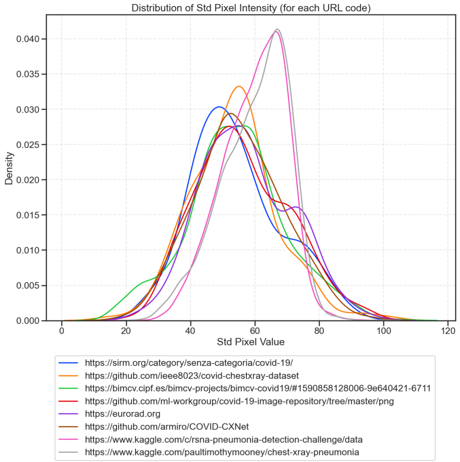

Figure 9: Distribution of standard deviation of pixel intensities of the X-ray images depending on the different sources

Inference: Images from https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia and https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia have different mean and std pixel intensities in comparison to the images from other sources.

On exploration of viral pneumonia images metadata we observed different image modes(figure 3). This may be responsible for the higher variance (std) in RGB (due to color channels) as observed in figure 9. Inconsistent image modes create hidden discrepancies in data, therefore we try to fix this in the next steps.

2.2.3 Mask Adjustment

As already described in chapter 2.2.1 we found out that the lung masks have not the same size as the X-ray images. The X-ray images have 299 x 299 pixel and the masks are smaller with only 256 x 256 pixel. Therefore the size of the masks have to be adjusted before one can overlay them with the X-ray images. We tried three different methods.

The first method we tried out, was to expand the masks by adding some black areas to the masks so that their sizes is the same as the images. We tried to add black areas to the right side and below the masks. Besides, we tried to add black areas to the left side and above the masks. The second method we tried out, was a simple resizing of the masks to 299 x 299 pixel by using the resize-function of the OpenCV Python library. As interpolation method we used "cv2.INTER_CUBIC".

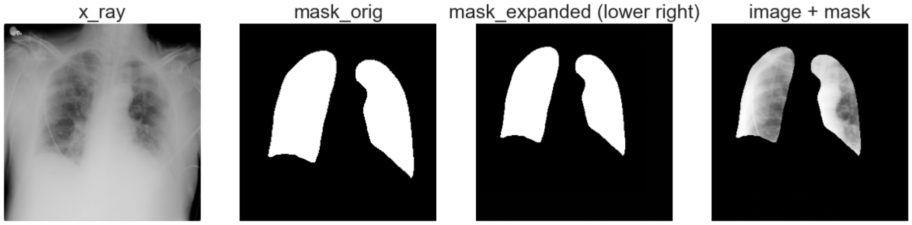

Figure 10: Modifying the lung mask by adding black areas to the right of the mask and below.

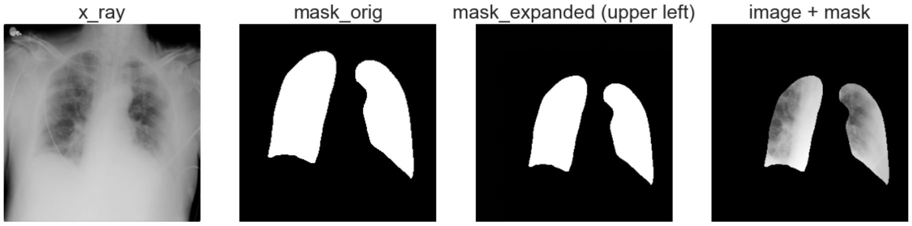

Figure 11: Modifying the lung mask by adding black areas to the left of the mask and above.

Inference:

Figures 10, 11 and 12 show (from left to right) the X-ray image, the corresponding lung mask not modified, the modified lung mask and the modified lung mask added to the X-ray image. The mask which is resized with OpenCV fits best to the area where the lung is located in the X-ray image. This can be

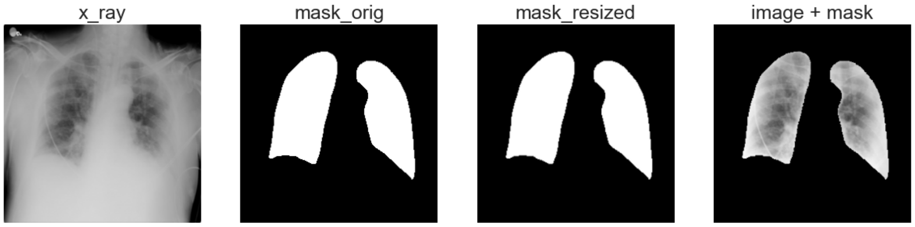

Figure 12: Modifying the lung mask by resizing them with OpenCV.

seen in detail in the overlay of some resized masks and X-ray images in figure 13.

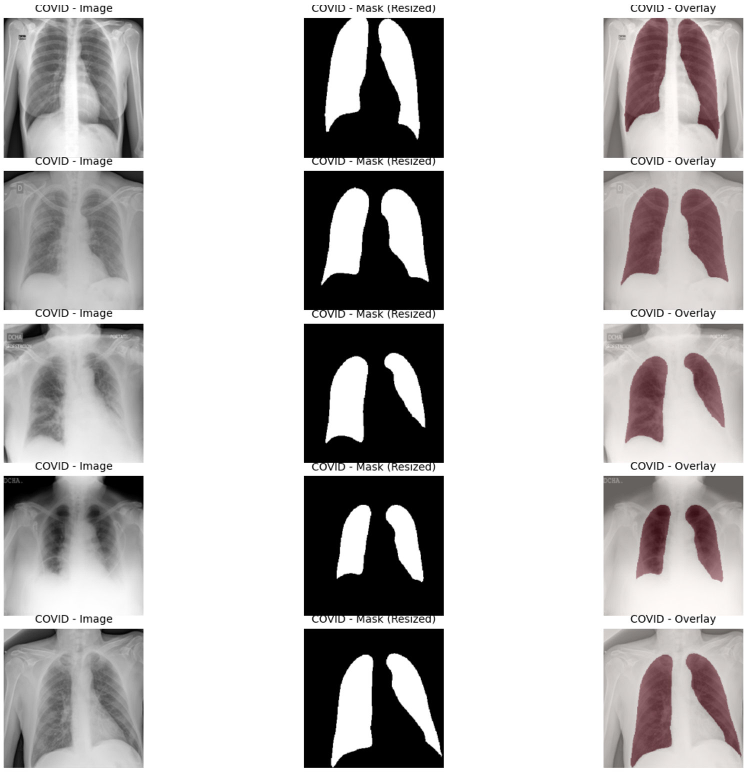

Figure 13: Overlay of resized masks and X-ray images.

2.2.4 Preprocessing Pipeline

To minimize source-related bias and standardize image quality across the dataset, we applied the following preprocessing steps:

1. Removed source bias:

- Grayscale Conversion (L-channel): All images were converted to grayscale using OpenCV (RGB conversion to L) to eliminate color-related variability and focus solely on intensity information.

- Contrast Enhancement with CLAHE: Applied CLAHE to help standardize the contrast across images, reducing the impact of varying imaging conditions. For example, images taken with different machines or in different settings may have varying brightness and contrast levels. Local enhancement ensures greater uniformity between images.

- Lung Region Isolation: Mask resizing (from 256*256 to 299*299) and applying the binary mask to the image after CLAHE (Contrast Limited Adaptive Histogram Equalization) processing, we ensured that only the lung regions are enhanced and normalized, while irrelevant areas like the background are not affected.

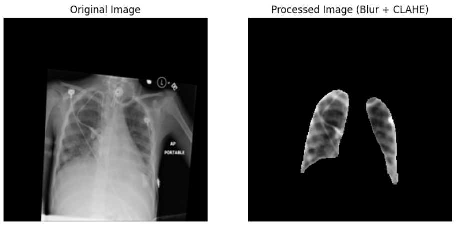

Figure 14: Original X-ray image (left) and filtered X-ray image with applied mask (right)

- Pixel Value Normalization: Since some images from one source are brighter or darker than those from another, normalizing the images ensures that they all lie within the same range, preventing models from learning biased patterns based on intensity differences. The images are converted to numpy arrays and stored as.npz files for fast loading during training and evaluation.

2.3 Visualizations Post Processsing

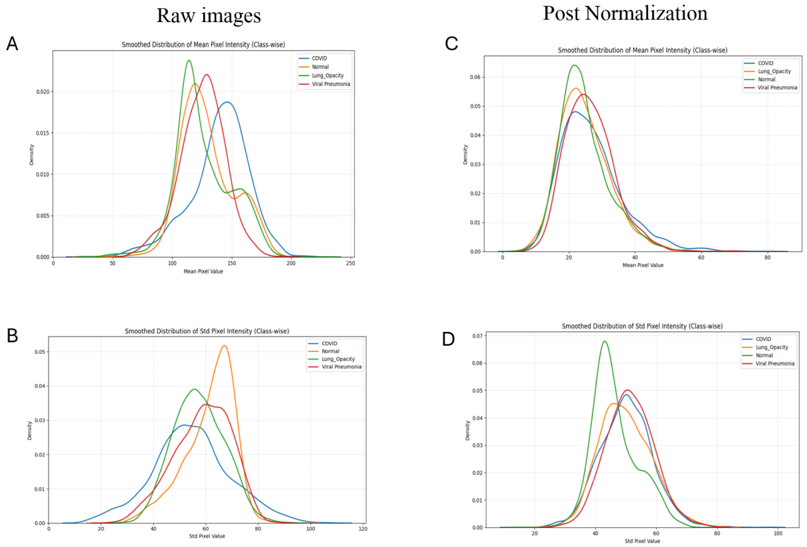

Figure 15: Distribution of mean and std of pixel intensities of the X-ray images before and after normalization by classes

Inference:

In images A and B in figure 15 we observe that raw images (left panels) have significantly different mean and standard deviation values of pixel intensities across the four classes: COVID, Normal, Lung Opacity, and Viral Pneumonia. This variation reflects inherent differences in image acquisition conditions or dataset characteristics, which could introduce bias into any machine learning model by enabling it to learn spurious class-specific brightness or contrast cues. After normalization, figure 15 (C and D) (right panels), these distributions become considerably more aligned. Both mean and standard deviation values show tighter and more overlapping density curves, indicating that normalization effectively reduces class-specific pixel-level discrepancies. This helps ensure that the model focuses on medically relevant features rather than superficial intensity differences, thus improving generalization and fairness.

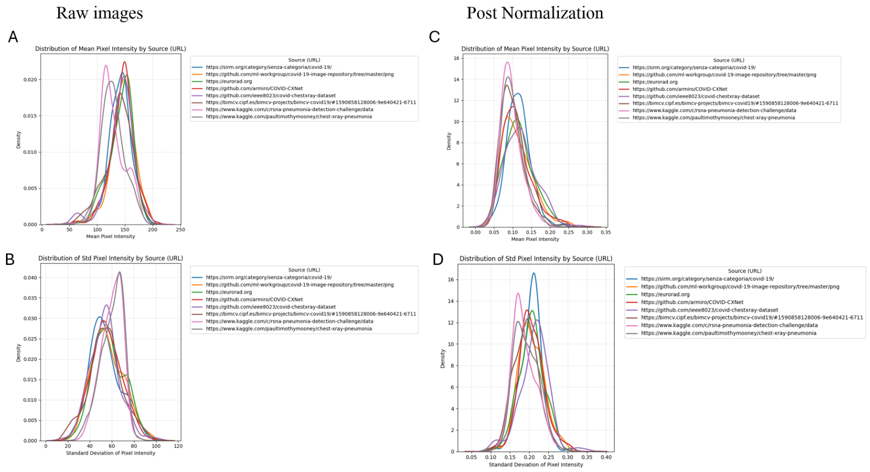

Figure 16: Distribution of mean and std of pixel intensities of the X-ray images before and after normalization by sources

Inference: In figure 16, a similar trend is analyzed but from a source-wise perspective. Each line in the plots represents data from a different image source or URL. The raw images in figure 16 (A and B) (left panel) again show substantial variation in both mean and standard deviation of pixel intensities across sources, suggesting a strong presence of domain shift — a critical issue when training models on aggregated datasets from different institutions or platforms. Such differences can hinder a model's performance on unseen data if not addressed. Post normalization (right panels), the pixel intensity distributions across all sources become much more consistent. This is a vital step in preparing the data for robust, real-world model deployment.

2.4 Next Steps

- Label Encoding Convert categorical class labels (e.g., COVID, Normal, Pneumonia, Lung Opacity) into numerical format for compatibility with machine learning algorithms.
- Dataset Splitting Partition the dataset into training, validation, and test subsets to enable robust model evaluation and prevent data leakage.
- Address Class Imbalance Since the distribution between the classes is not balanced (for example Viral Pneumonia images are far less numerous than the Normal images) we have to address this problem before the training stage. Appropriate strategies will be applied, such as:

  Class weighting during model training

  Oversampling/undersampling

  Data augmentation for minority classes

  This ensures the model does not become biased toward the majority classes and maintains fair performance across all categories.