Project: Exploring the Feasibility of Using Chest X-ray Data for Diagnosis of COVID-19 Patients

Report 1: exploration, data visualization and data pre-processing report

(03.10.2024)

Members: (Data Science Bootcamp)
Hayato Yuuto Kakinuma
Alhassan Abdelsamie
Upendra Prasad Yadav

Table of Contents

Introduction:

1. Context

The COVID-19 pandemic has posed a significant global health crisis, overwhelming healthcare systems and prompting the development of rapid diagnostic tools. Traditional methods, such as PCR tests, although effective, can be resource-intensive and time-consuming. Medical imaging, specifically chest X-rays, has emerged as a potential supplementary diagnostic tool. X-rays are cost-effective, widely available, and may reveal characteristic lung changes associated with COVID-19, providing a faster diagnostic alternative.

This project focuses on exploring the feasibility of using chest X-ray data to distinguish COVID-19 patients from individuals with other respiratory conditions or healthy subjects using deep learning and machine learning techniques.

2. Objectives

The primary objective of this project is to develop a diagnostic model that can:

a) Differentiate between COVID-19 positive, non-COVID lung infections (such as pneumonia), and healthy patients using chest X-ray images.

b) Improve the interpretability of the model's predictions through visualizations and explainability techniques, ensuring that decisions are made based on relevant patterns in the X-ray data.

c) Provide a comparative analysis of traditional machine learning techniques and deep learning-based convolutional neural networks (CNNs) to find the most effective diagnostic tool.

Additionally, the project aims to provide a robust, scalable solution by leveraging data augmentation, pre-processing techniques, and performance evaluation to ensure accuracy and generalizability.

3. Framework

Data Collection:

The dataset used for this project is the **COVID-QU-Ex dataset** from Kaggle, which consists of **33,920 chest X-ray images**. The images are divided into three main categories:

- **11,263 Non-COVID infections** (including viral or bacterial pneumonia)
- **10,701 Normal/Healthy patients**
- **11,956 COVID-19 positive patients**

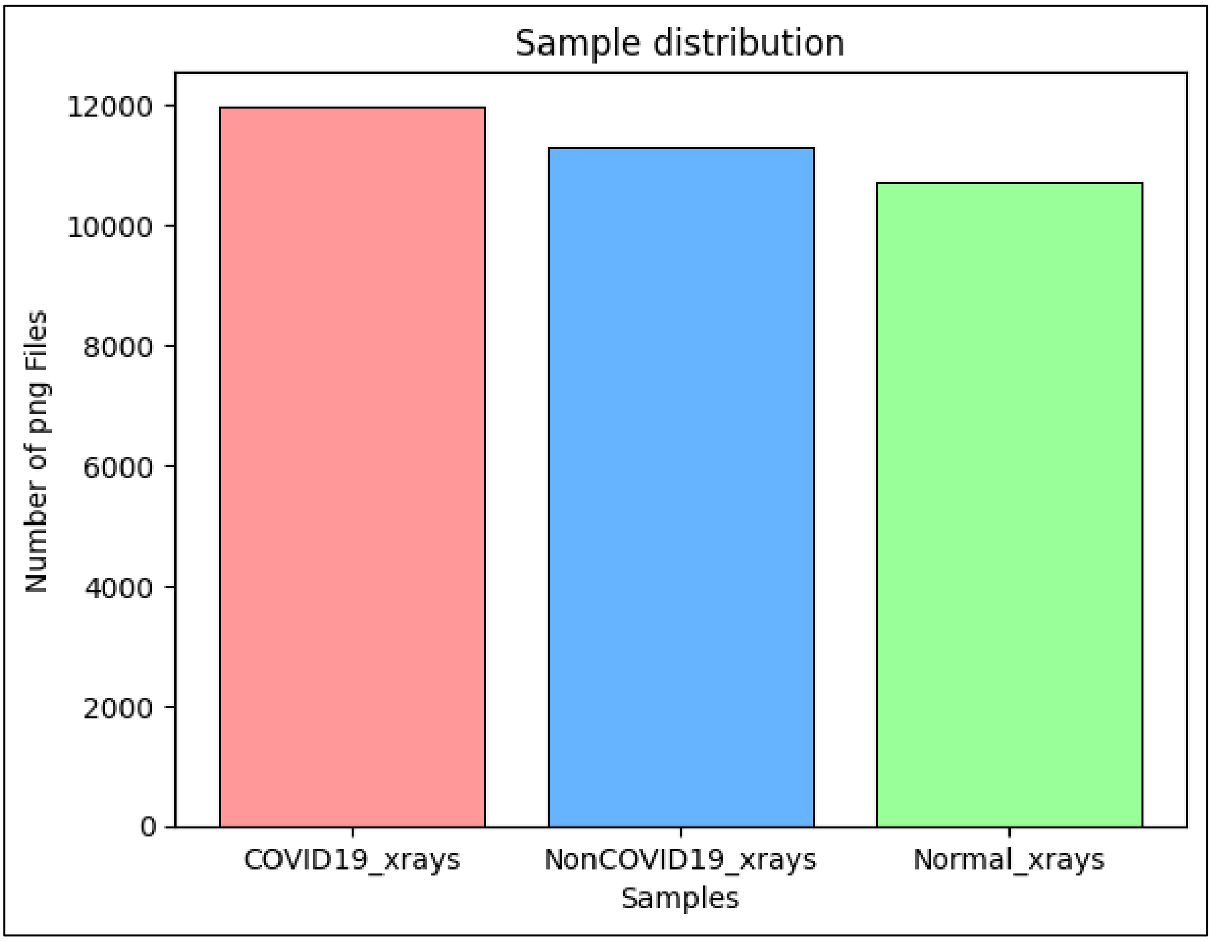

Figure 1.Distribution of X-ray Samples:
COVID-19 Patients, Non-COVID Viral Pneumonia (Non-COVID) Patients, and Healthy Individuals (Normal)

Data Preprocessing will involve:

- **Normalising, resizing, and applying colour mapping** to standardize the images for input into machine learning models. This process may involve contrast enhancement, noise reduction, and cropping to improve the clarity and focus of the images.

- **Data augmentation** techniques, such as rotation, flipping, zooming, and colour mapping variations, are used to increase the diversity of the training images. These techniques help reduce overfitting and enhance the model's robustness by introducing different visual perspectives and colour representations.

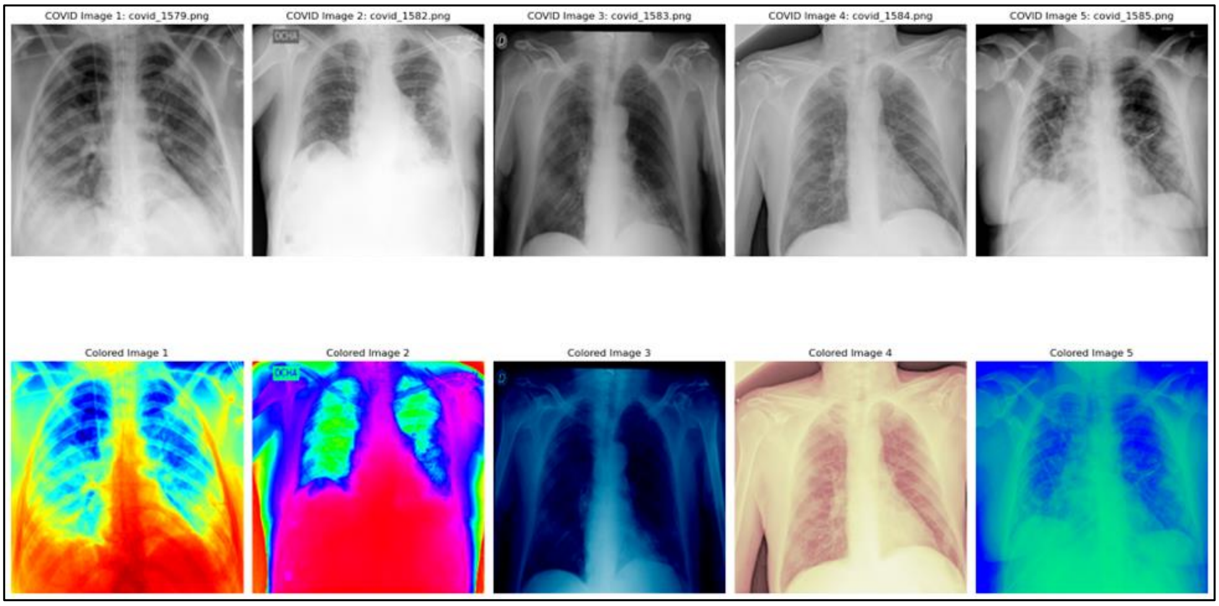
COVID Image 1: covid_1579.png
COVID Image 2: covid_1582.png
COVID Image 3: covid_1583.png
COVID Image 4: covid_1584.png
COVID Image 5: covid_1585.png
OCHA
Colored Image 1
Colored Image 2
Colored Image 3
Colored Image 4
Colored Image 5

Figure 2. Sample X-ray images from COVID-19 positive patients (top five).
Different colour maps: COLOURMAP_JET, COLOURMAP_HSV, COLOURMAP_OCEAN, COLOURMAP_PINK, COLOURMAP_WINTER (bottom five).

4. Relevance

The relevance of this project lies in its potential to assist healthcare systems in diagnosing COVID-19 quickly and accurately using a readily available diagnostic tool—chest X-rays. This would allow medical professionals to assess patients even in resource-limited settings where PCR tests or more advanced imaging techniques are not readily available.

By leveraging machine learning models, the project aims to:

a) **Speed up the diagnostic process** by automating the detection of COVID-19 patterns in chest X-ray images.

b) **Improve diagnostic accuracy** by incorporating deep learning models trained on a large dataset of X-rays, fine-tuned with transfer learning.

c) **Contribute to explainability and trust** in AI-driven diagnoses through techniques like Grad-CAM and SHAP, which will provide visual and statistical explanations of the model's predictions.

5. Preprocessing and Feature Engineering

The preprocessing phase of this project is critical to ensuring that the data fed into the models is clean, consistent, and informative. The following steps are part of the preprocessing strategy:

a) Image Preprocessing

The X-ray images, initially provided in PNG format, were first converted to greyscale to ensure uniformity and to facilitate subsequent image analysis. Each lung X-ray image was accompanied by its corresponding ground truth mask, essential for segmentation and accurate feature extraction. To maintain consistency across the dataset, it was necessary to ensure that the dimensions of the images and their respective masks were perfectly aligned. After verifying size compatibility, the images were further processed by converting them into 8-bit unsigned integers, optimizing them for computational efficiency and reducing the overall file size while preserving critical pixel intensity information for further analysis.

Each image will be resized to a standard dimension of 224x224 pixels. This resizing ensures that all input images match the expected dimensions of popular convolutional neural networks (CNNs). Resizing will maintain the aspect ratio where possible or apply interpolation methods to prevent significant distortion.

b) Data Augmentation

Data augmentation will be used to artificially expand the size and diversity of a training dataset by applying various transformations to the existing data. This process helps improve the performance and generalization of machine learning models. It will include the methods like:

- **Rotating**, **flipping**, and **zooming** images to create variations.
- **Brightness adjustment** to simulate different X-ray conditions.

c) Feature Extraction

In addition to CNNs, traditional machine learning techniques will be tested. For these methods, feature extraction is crucial. Techniques like histogram of oriented gradients (HOG) and grey-level co-occurrence matrices (GLCM) will be used to extract key features from the images. These features will serve as input for traditional classifiers like SVM or Random Forest.

6. Visualizations and Statistics

The KDE plots below reveal that while there is overlap between the categories; normal X-rays are distinguishable by lower opacity levels, while COVID-19 and non-COVID viral pneumonia X-rays share similarities, though COVID-19 X-rays exhibit a slightly broader range of opacity levels, potentially due to more severe or varied lung changes (Figure 3).

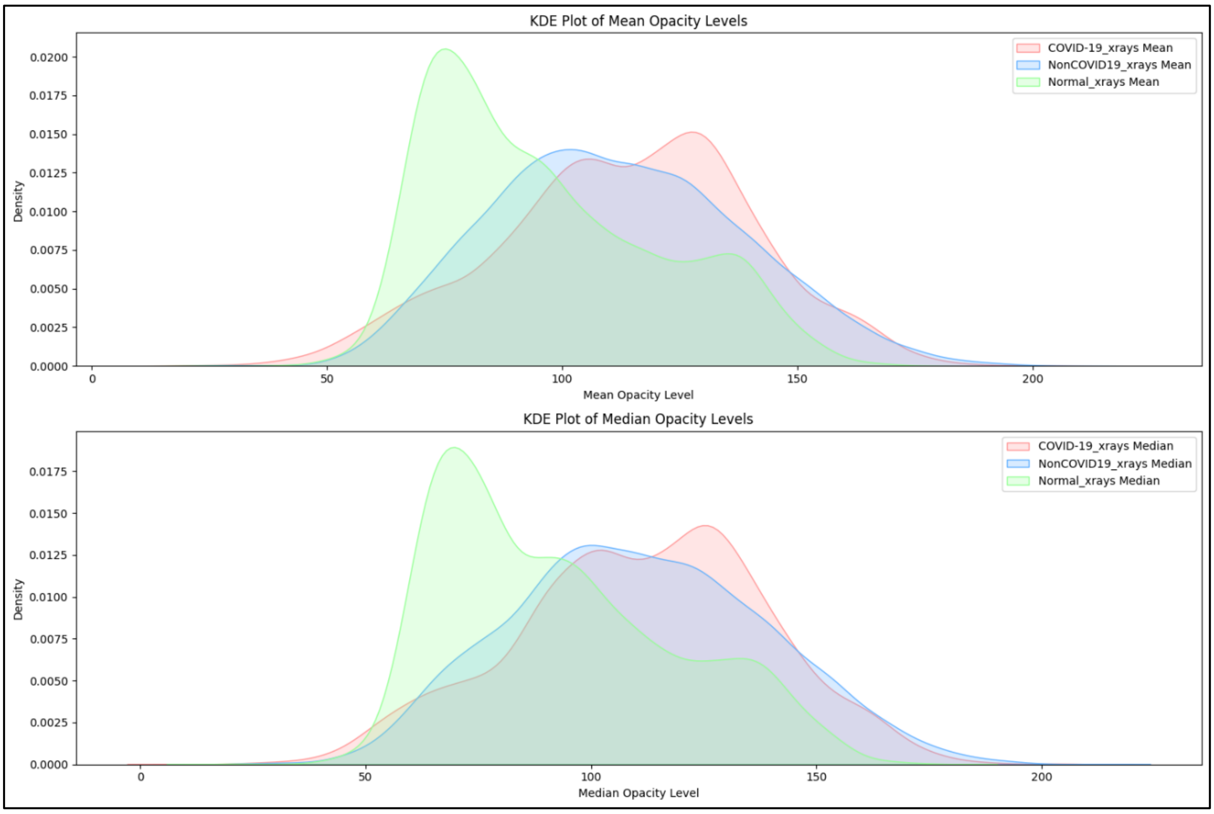

Figure 3. KDE Plot of Mean and Median Opacity Levels: COVID-19 Patients, Non-COVID Patients, and Normal X-rays.

Looking into their standard deviations (STD):

- COVID-19 and Non-COVID-19 viral pneumonia X-rays exhibit higher standard deviations in opacity levels, indicating greater variability within the lungs due to disease, which may reflect pathological features like consolidations and ground-glass opacities.
- Normal X-rays have a significantly lower standard deviation, reflecting the more homogenous nature of healthy lungs with fewer opacities and clearer structures.
- The numerous outliers in the diseased categories (COVID-19 and Non-COVID-19) suggest that some cases exhibit extreme opacity variations, which might correlate with severe disease manifestations.

The box plot highlights the more consistent and lower opacity variability in healthy lungs compared to the increased and more variable opacity seen in both COVID-19 and other viral pneumonia cases (figure 4).

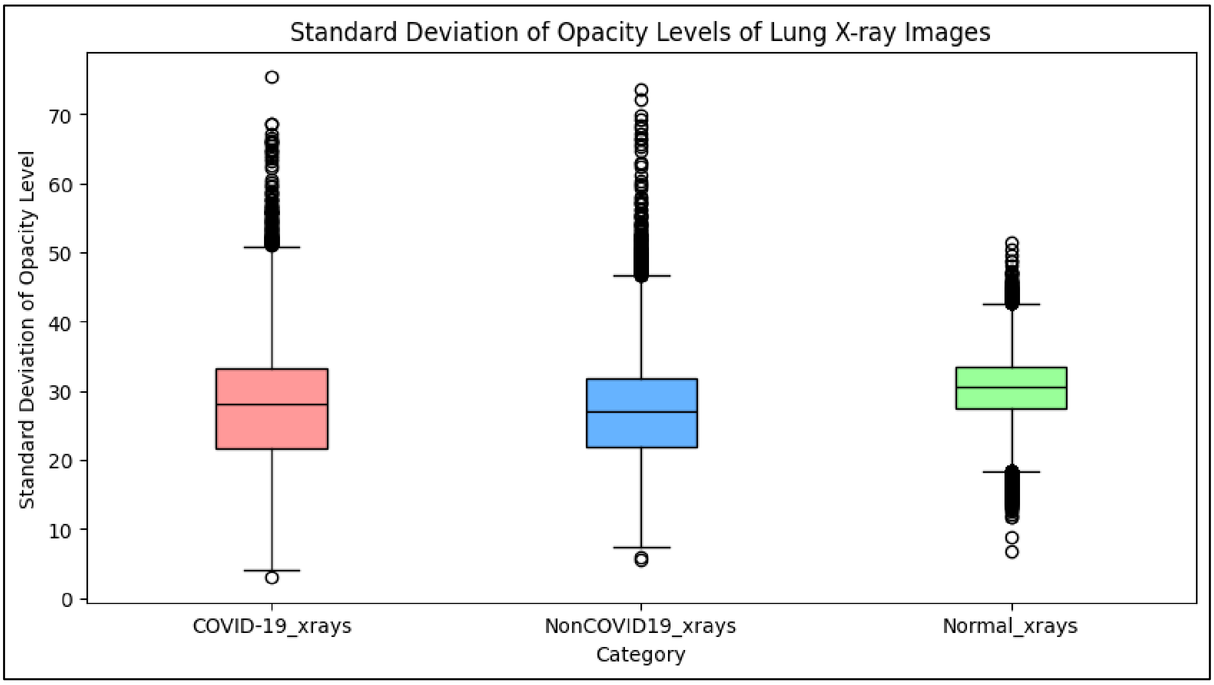

Figure 4. Box plot of STD of Opacity Levels: COVID-19 Patients, Non-COVID Patients, and Normal X-rays

In a scatter plot, we can better observe the relationship between the standard deviation of opacity level and the median opacity level of X-ray images (figure 5). We can visualise better how narrow the normal X-ray STD and median range compared to diseased X-rays.

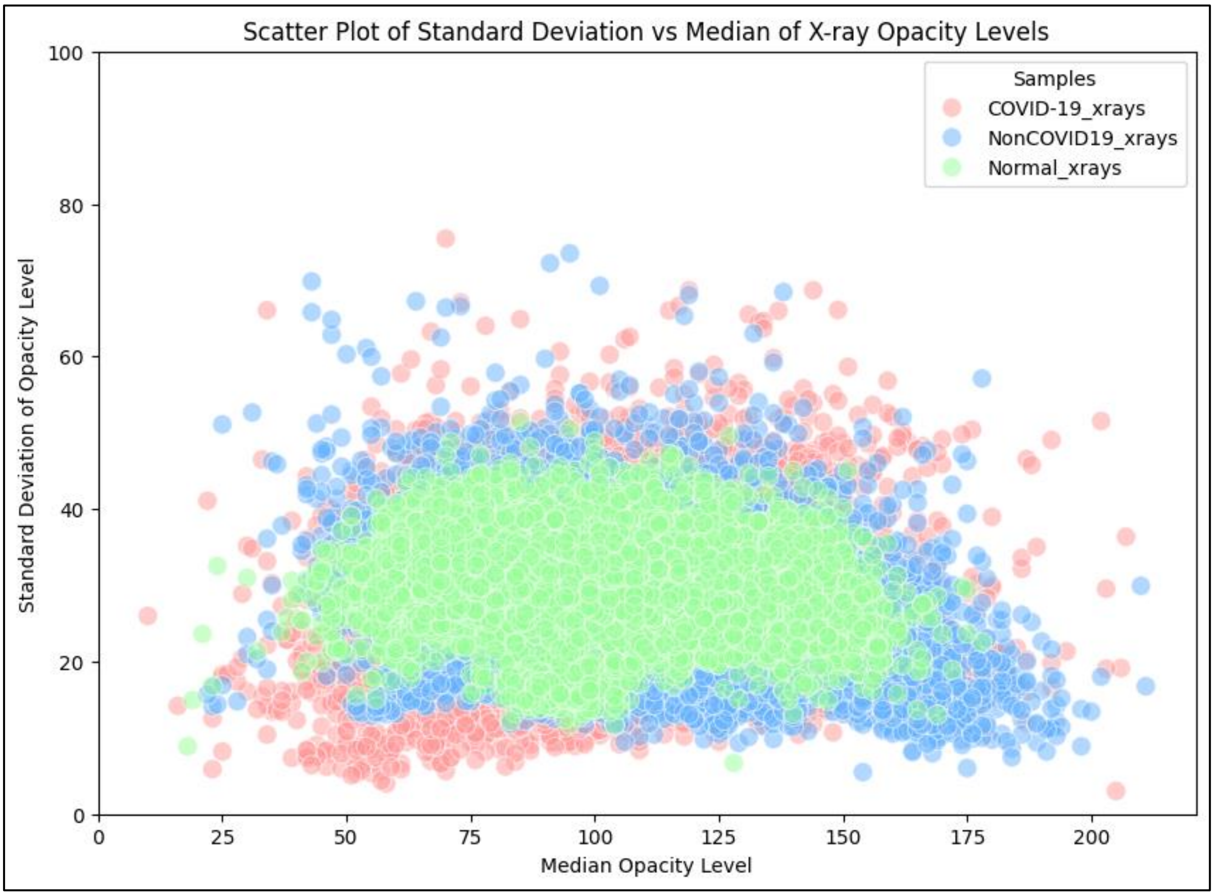

Figure 5. Scatter plot of STD against Median Opacity Levels: : COVID-19 Patients, Non-COVID Patients, and Normal X-rays