Detection of COVID-19 from Chest X-Rays

[Milestone 3] EDA, Planning, and Setting Goals

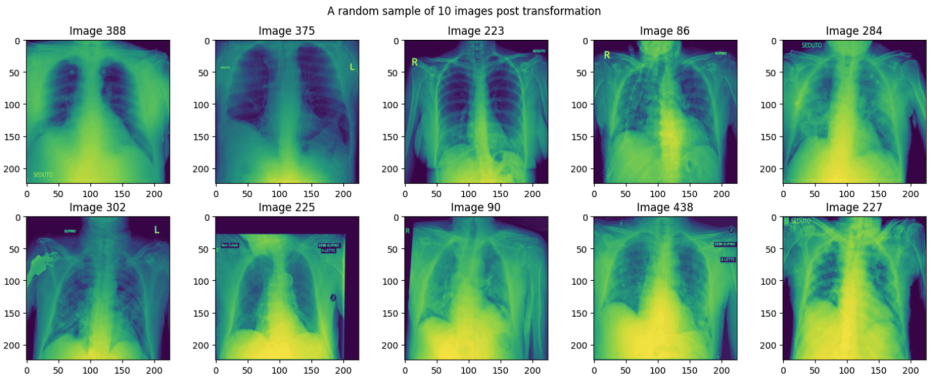
Adele Collin, Chuck Lin, Kay Wu, Xinyu Chen

The code for our project, including the EDA, can be found in our GitHub repository.

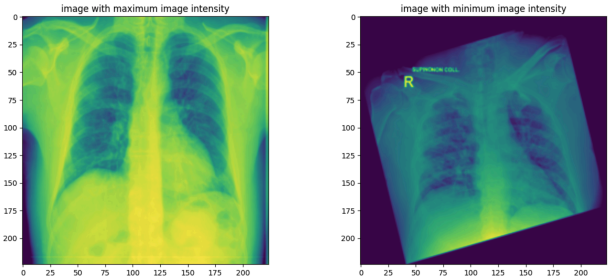
I. Summary of the Data, Data Analysis, & Clean and Labeled Visualizations

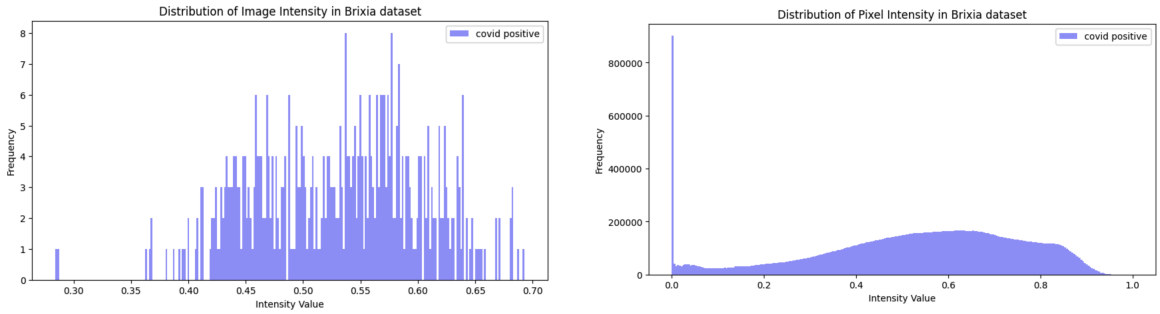
Positive cases training set: Brixia Score Covid19 dataset

- Dataset and images



> number of samples (Brixia only has covid positive): 469
> image dimension (height, width, channel): (224, 224, 1)
> image data type: float32
> average image-wise intensity: 0.53036904
> maximum image-wise intensity: 0.69281775
> minimum image-wise intensity: 0.28426707
> Standard deviation of image-wise intensity: 0.07222104





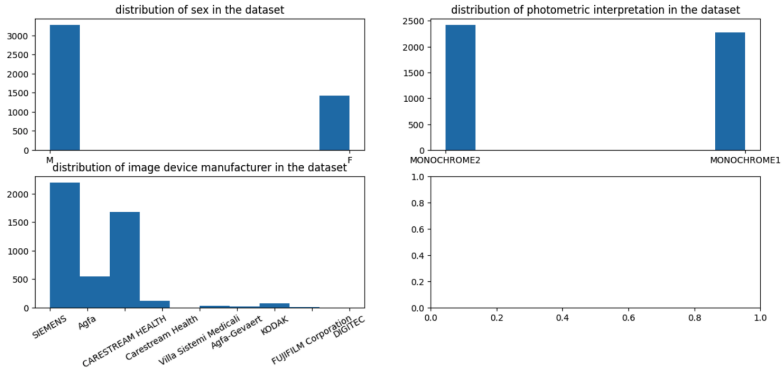
> with min intensity of 0.2842670679092407 and max intensity of 0.692817747592926, there are no outliers in this dataset in terms of extreme intensity. However, as

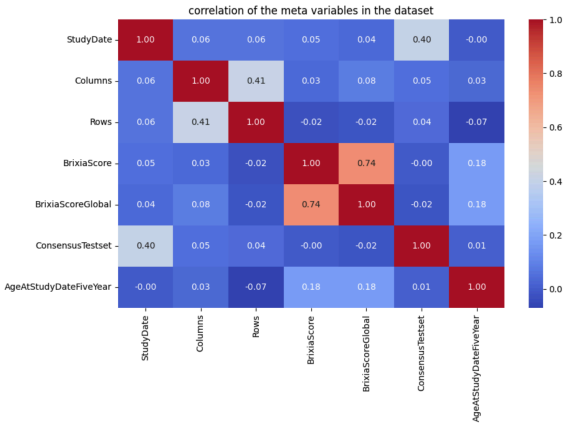
demonstrated by the sample of max and min intensity above, images with low image intensity are pre-rotated.

- Meta-data

	min	max	mean	std
StudyDate	20200304.0	20200409.0	2.020033e+07	32.166845
Columns	2019.0	3376.0	2.893871e+03	152.103555
Rows	1056.0	3050.0	2.405939e+03	142.821628
BrixiaScore	0.0	333333.0	1.021193e+05	99594.274301
BrixiaScoreGlobal	0.0	18.0	8.301171e+00	4.238625
ConsensusTestset	0.0	1.0	3.194888e-02	0.175883
AgeAtStudyDateFiveYear	1.0	19.0	1.247668e+01	2.723331

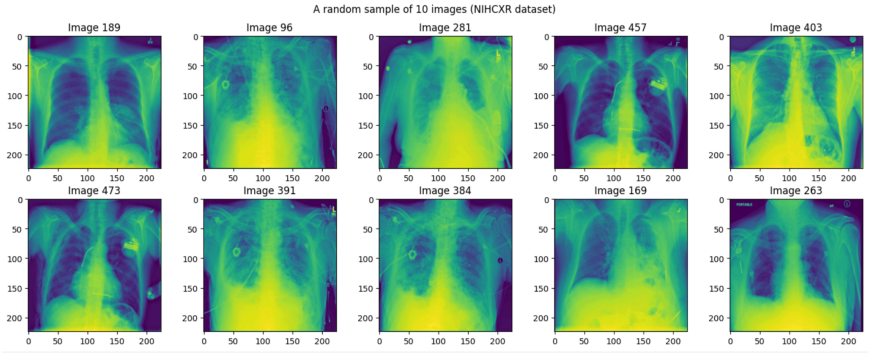
*Note: https://brixia.github.io/ states the definition of "AgeAtStudyDateFiveYear" as "Age of the patient in groups of five (i.e. 0-4yo is 0, 5-9yo is 1, ...)", no specific age is given. In other words, the minimum age group is 5-9 yrs old, maximum 95-100 yrs old, and in average in the 60-65 age group.





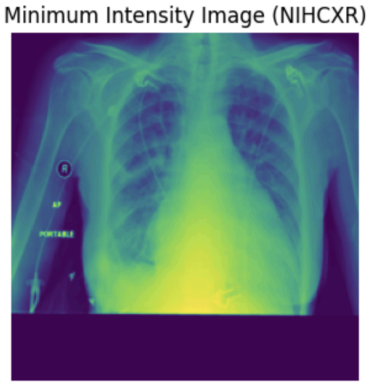
Negative cases training set: NIHCXR dataset

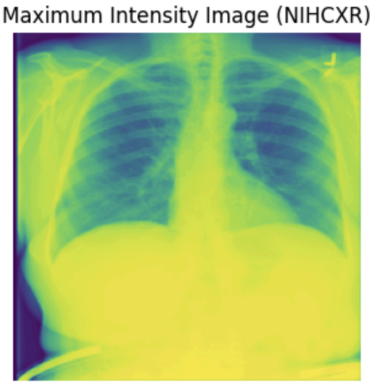
- Dataset and images

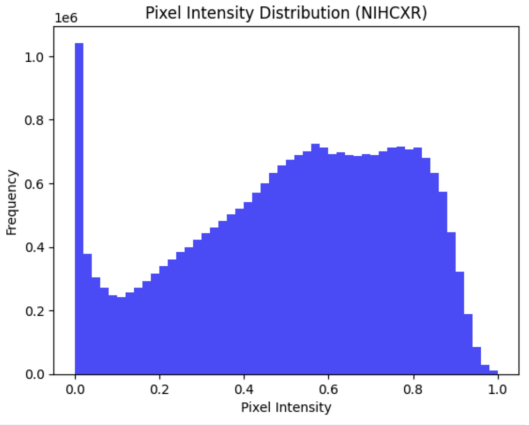


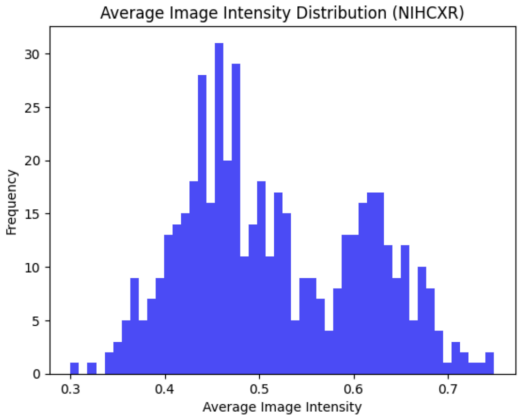
For the purposes of this milestone, 500 images were randomly sampled from this dataset.

> image dimension (height, width, channel): (224, 224, 1)
> image data type: float32
> average image intensity: 0.5158608
> maximum image intensity: 0.74893934
> minimum image intensity: 0.3004221
> standard deviation of image intensity: 0.09409551
> Standard deviation of pixel values: 0.25503743





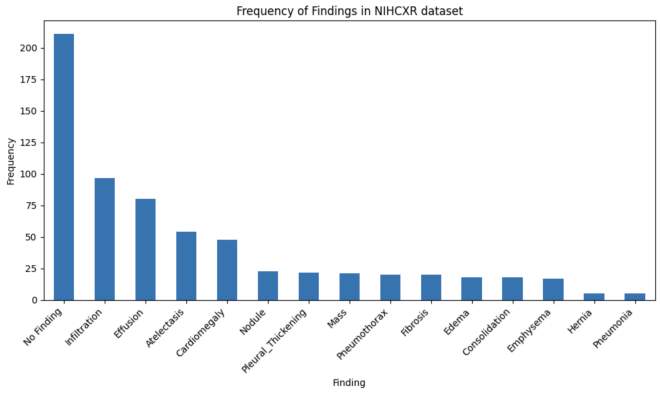


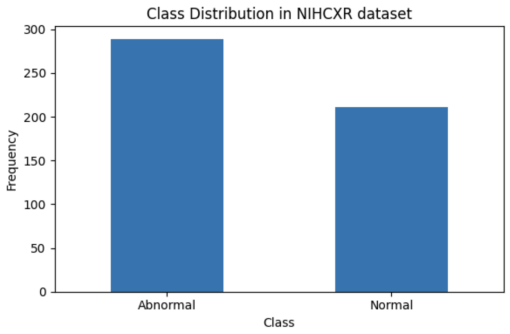


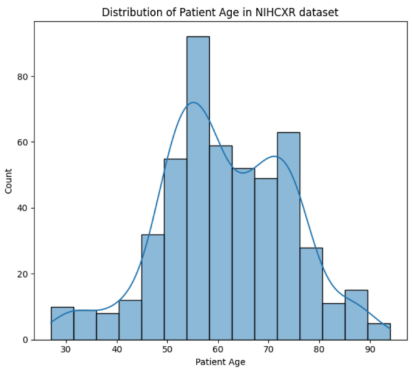
- Meta-data

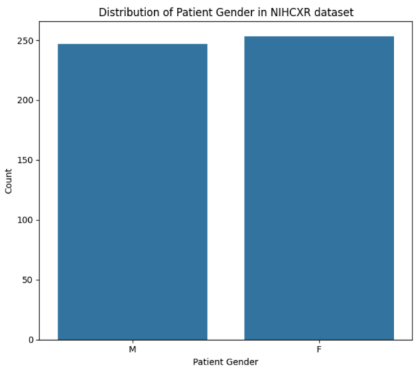
	Image Index	Finding Labels	Follow-up #	Patient ID	\
0	00000001_000.png	Cardiomegaly	0	1	
1	00000001_001.png	Cardiomegaly|Emphysema	1	1	
2	00000001_002.png	Cardiomegaly|Effusion	2	1	
3	00000002_000.png	No Finding	0	2	
4	00000003_001.png	Hernia	0	3	

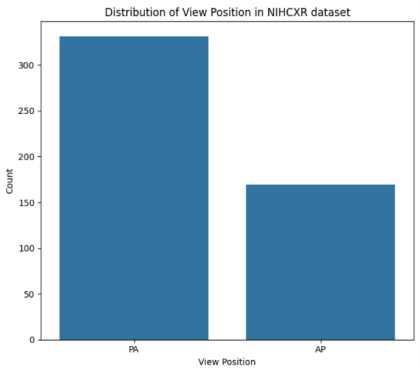
	Patient Age	Patient Gender	View Position	OriginalImage[Width	Height]	\
0	57	M	PA	2682	2749	
1	58	M	PA	2894	2729	
2	58	M	PA	2500	2048	
3	80	M	PA	2500	2048	
4	74	F	PA	2500	2048	



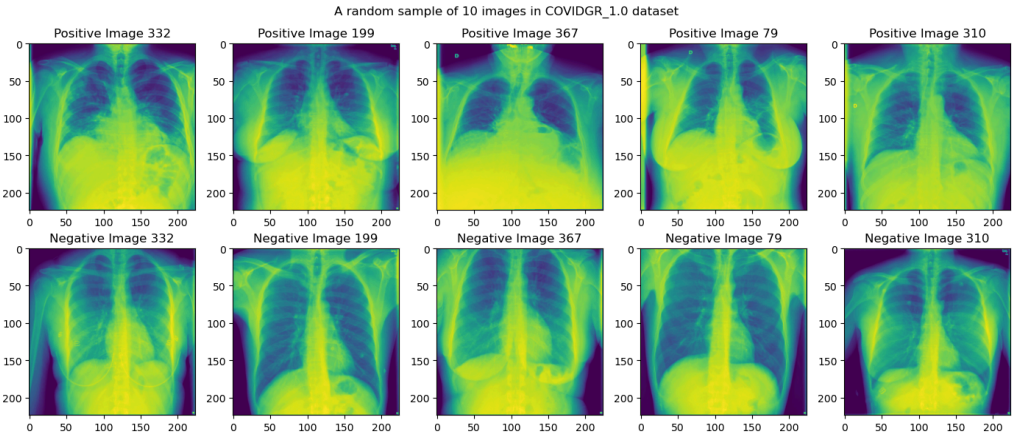






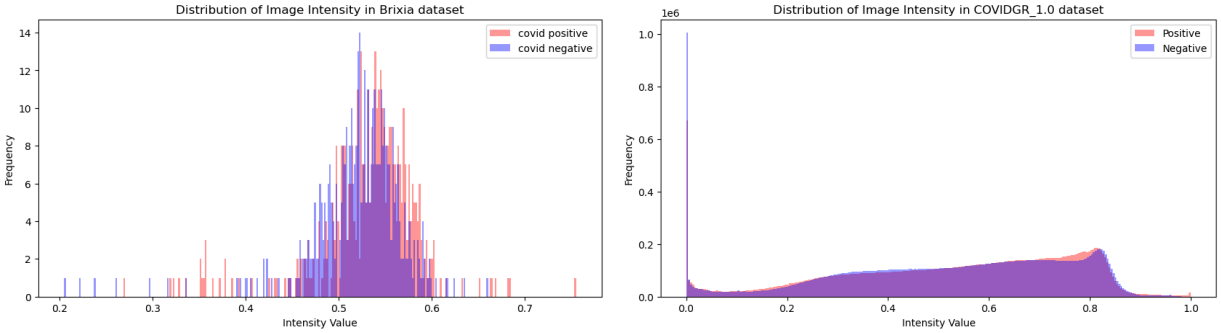


Test set: COVIDGR_1.0

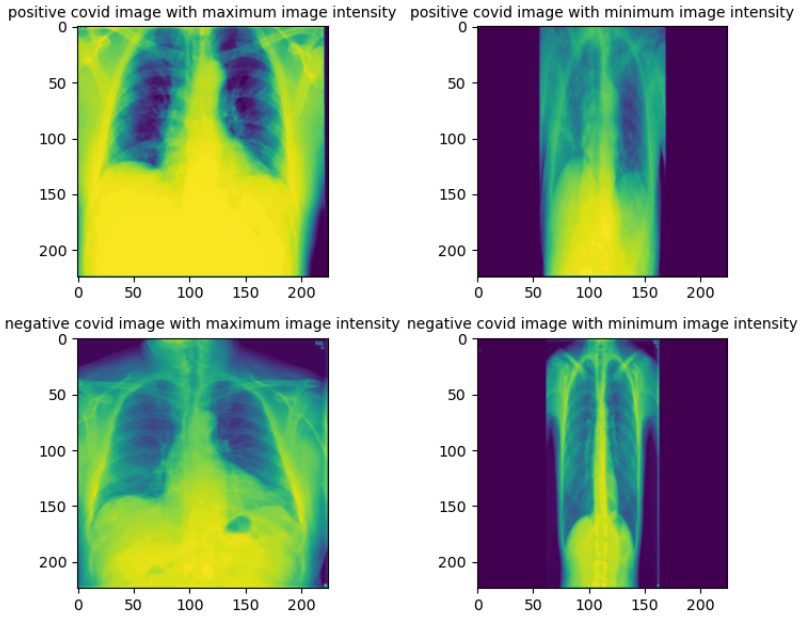


- Shape of the data:
  - > number of positive cases: 426
  - > number of negative cases: 426
  - > image dimension (height, width, channel): (224, 224, 1)
- Data types:
  - > COVIDGR_1.0 – image data type: numpy float32
- Descriptive statistics:
  - > average positve image intensity: 0.5315129
  - > average negative image intensity: 0.521097
  - > maximum positve image intensity: 0.7555852

> maximum negative image intensity: 0.6609612
> minimum positve image intensity: 0.26859397
> minimum negative image intensity: 0.20477217
> standard of positve image intensity: 0.05474118
> standard of negative image intensity: 0.05072282
> Distribution of image intensity (contrast of color)



- No class imbalance
- Outliers



> with min intensity of 0.20477217 and max intensity of 0.7555852, there seem to have a few outliers in this dataset in the above histogram. However, as demonstrated by the sample of max and min intensity above, images with low image intensity are padded with black space to form a square image. Further image augmentation can help adjust the images with extreme average intensity.

♦ Meta-data

Dataset	Class	#images	women	men	#img. per severity level
COVIDGR-1.0	Negative	426	239	187	

	COVID-19	426	190	236	Normal-PCR+: 76
					Mild: 100
					Moderate: 171
					Severe: 79

II. Meaningful Insights:

Positive training set: Brixia Score Covid19 dataset

From the first image we can first see a distribution of image variability, based on the 10 randomly sampled images. The coloring of the regions, and the specific region covered are different across the images for example some included the shoulder area but some didn't. Next the image intensities are within reasonable ranges across the images, with min around 0.28 and max around 0.69 we have no extreme outliers that went close to 0 or 1 in terms of image intensity. Upon examination of images with lower intensities against the higher intensities, we do see that the structure of the images are different, notably all the images with low intensities are pre-rotated. This is similar to data augmentation if we were to perform image rotation, for now this is contributing to the diversity within the dataset, if model performance is lacking later we can remove the image with lower intensities.

We also don't see correlation between the variables in meta-data. However, the meta-data does have some imbalance (or bias) problems, first we have twice more male patients than female patients, and the nine device manufacturers aren't balanced, with more than half of the images taken using SIEMENS and CARESTREAM Health devices, and the other seven manufacturers are underrepresented as images taken are slightly different depending on the manufacturer. Another concern is that the minimum age group is 5-9 yrs old, maximum 95-100 yrs old, and in average in the 60-65 age group, meaning the population represented in this dataset are relatively older and may prevent the model from generalizing to other age groups. As all of the concerns stem from the meta-data, and given that we will not be directly using the meta-data as part of the model training, only images will be used, the model generation should still be possible. However, the model will be constrained by the inherent data bias due to the source of the images, and limit its generalizability.

Negative training set: NIHCXR dataset

The quality of the images look appropriate with relevant areas of interest included. Image intensity distributions are in a reasonable range without any obvious outliers: inspecting the images with minimum and maximum intensities, these appear reasonable though the one with minimum intensity has more "padding" along one edge and we deem this still satisfactory for inclusion. Looking at the distribution of findings, there are more cases with abnormalities than "normal" cases or cases with no clinically significant findings. The dataset is balanced in terms of gender and the average patient age is 61.376 years.

Test set: COVIDGR_1.0

The dataset is balanced for the classification task. The image pixels are within a reasonable range of 0 and 1, and the image intensity is normally distributed with a mean of 0.53 (sd = 0.054) and 0.52 (sd = 0.051) for positive and negative images, respectively. Each x-ray image has a height and width of 224 x 224 and a color channel of 1. The images were obtained from a single center under a standardized procedure. Thus, the images in this dataset show a similar view of the chest across patients, and the 5 randomly selected images share similar sizes, orientations, and angles. We also visualized the image with maximum intensity of 0.756 and minimum intensity of 0.205 to assess the presence of outliers. Although the image with the lowest intensity seems extreme, the low intensity is the result of the padding effect to standardize images into square format. For next steps, additional image augmentation such as zooming or cropping might be helpful to correct for the padding effect.

On visual inspection of some of the negative test cases, the negative set comprises of mostly chest x-rays of healthy patients, ie. no other clinically significant findings.

Comparison between datasets

The positive training set (Brixia), negative training set (NIHCXR), and the test set (COVIDGR_1.0) all appear to have images of reasonable quality with regions of interest generally being included. The plots of the pixel intensities are similar across datasets. Although the plots of image intensity is not consistent across dataset, they still fall in a reasonable range of 0.2 - 0.8, and the few outliers are visualized which can be explained by the effect of padding.

Unfortunately the three datasets do not consistently report the same metadata, ie. we know that the NIHCXR and COVIDGR_1.0 datasets are balanced in terms of patient gender, but the Brixia dataset does not include this information. Similarly, we know that the Brixia and NIHCXR datasets have similar average patient age but this information is not known about the COVIDGR_1.0 dataset. Some of the datasets also include interesting additional metadata, such as Brixia including machine information, but this is not provided in the other datasets. Thus, it is difficult to know if each of these three datasets are similar in terms of these other characteristics.

Looking at the "negative" cases specifically, the negative training set (NIHCXR) has what appears to be a much greater proportion of patients with other clinical findings than healthy patients compared to the negative samples in the test set (COVIDGR_1.0), which contains a few cases with clinical findings but consists of mostly what appears to be healthy patients, though there are no labels for comorbidities. Because of the difference in balance of healthy cases with cases with comorbidities in the negative training set compared to the negative samples in the test set, it may be important to only use the CXRs with no findings in the NIHCXR dataset as the negative training set for better performance on the test set.

III. Summary of findings

- The test set (COVIDGR_1.0) is balanced for positive and negative cases for COVID-19.
- Overall, the positive training set (Brixia), negative training set (NIHCXR), and the test set (COVIDGR_1.0) all appear to have images of reasonable quality and image and pixel intensities. There is variability among the datasets in terms of image padding or rotation, which add noise but may help with training to be more generalizable to the test set. Some of the variability may be due to image preprocessing when resizing the images to square dimensions, so we will plan to adjust for this via further steps such as zooming or cropping.
- Given the difference in proportion of normal vs. abnormal cases in the negative training set and the negative subset of the test set, with the negative training set appearing to have many more abnormal cases, we will filter out the healthy cases in the negative training set for training use.
- The different datasets report different metadata. Unfortunately, we cannot compare all three datasets in terms of age, gender, or other characteristics. However, it is expected that there is inherent bias in distribution of sex, device manufacturer, age group, and some other variables such as patient position while taking the x-rays which differ between the datasets. A machine learning model that we create will be constrained by the inherent bias in the training datasets, and will limit its generalizability.

IV. Project Question

- Because the images are only classified based on covid-positive and covid-negative, the underlying concurrent diseases, x-ray modalities, patient position while taking the x-ray, aren't specified in all three datasets. Therefore the model is distinguishing a variety of x-ray images with unknown concurrent conditions to identify signs of covid-postive and covid-negative.

- We propose to ask the following question: Can signs of covid-19 in X-ray scans be identified using machine learning approaches, to predict whether a patient has covid or not, with noisy image sets (e.g. patients have none or a variety of concurrent conditions other than covid, image rotation, position of patient taking the x ray, e.t.c.)?

V. Baseline Model or Implementation Plan

Further preprocessing

- We plan to filter patients with no clinical findings in the negative training dataset for training our model. We may filter datasets to one image per patient to enhance the generalization power of the model.
- We are only including a random subset of the data right now, due to computational cost.
- We would like to convert the numpy arrays to tensorflow objects, preprocess the image, and augment the images by rotating and zooming the images. To fit the input dimension of the ResNet model, we will expand the image dimension from 1 to 3.

Baseline Model

- For the baseline model, we would want to employ ResNet-50 as a feature extractor for transfer learning. We then freeze the weights of the model and replace the last fully connected layer with two densely connected layers and a dropout layer in between. The training process would update the weights of the dense layers. The final output layer should contain a single value representing the probability of COVID status, which is then converted to binary COVID status. We plan to try various filter sizes of the dense layers, different optimizers (Adam/RMSprop), different learning rates, batch sizes, and numbers of epoch. We will also employ early stopping to prevent the model from overfitting. Finally, we will report the accuracies and the AUROC of the testing data.