Detection of COVID-19 from Chest X-Rays

[Milestone 2] Acquire and Understand the Data

Adele Collin, Chuck Lin, Kay Wu, Xinyu Chen

The code for dataset preprocessing can be found in our GitHub repository.

1. Dataset Description

Our datasets originate from three distinct sources.

The first dataset, the Brixia Score Covid19 Dataset, curated by ASST Spedali Civili di Brescia, comprises 4,707 chest X-ray (CXR) images obtained from sub-intensive and intensive care units over a specific period from March 4th to April 4th, 2020, notably coinciding with the peak of the COVID-19 pandemic. These images, depicting real-world clinical scenarios, provide a comprehensive portrayal of variations encountered in medical practice, particularly pertaining to cases of COVID-19. This dataset is sourced from the repository of the institution's Radiology Information System-Picture Archiving and Communication System (RIS-PACS), and was acquired through both computed radiography (CR) and digital radiography (DX) modalities, encompassing both anterior-posterior (AP) and posterior-anterior (PA) projections. This dataset is going to be our positive classifications.

The second dataset, the NIH Chest X-ray Dataset, consists of 112,120 frontal-view X-ray images extracted from 30,805 individual patients, and annotated with fourteen distinct thoracic pathologies derived from radiological reports. Among these pathologies are atelectasis, pneumonia, and consolidation, thus broadening the scope of previous investigations. This dataset is going to be our negative classifications.

Lastly, the COVIDGR-1.0 dataset, developed through collaboration with specialists from Hospital Universitario San Cecilio, Spain, comprises of 852 anonymized CXR images categorized into positive and negative instances based on the results of COVID-19 tests conducted via reverse transcription-polymerase chain reaction (RT-PCR) within 24 hours of image acquisition. Consistently captured in the posterior-anterior aspect, these images are classified according to severity levels and accompanied by demographic attributes, including gender distributions, for each identifiable class. Collectively, the datasets are derived from diverse institutions to enhance the generalizability of the predictive model. This is going to be our test dataset with both positive and negative classifications.

2. Potential Data Issue & **Addressing Data Issue**

2.1 Missing data

Our data consists of numpy arrays converted from Chest X-ray (CXR) images from the COVID period, so there is no missing data.

2.2 Data imbalance and Bias

The positive CXR samples for COVID-19 are taken from the Brixia Score Covid19 Dataset. There are 4,707 CXRs positive for COVID-19, and among them there are 4,583 unique patients. The Brixia dataset contains twice the amount of data for males compared to females, and the age follows a right-skewed normal distribution with the peak around the 60-70 age group, which would create some bias in regards to the

demographics that we would employ upsampling or downsampling depending on the final dataset after speaking with the teaching team.

The negative CXR samples for COVID-19 are taken from the NIH Chest X-ray Dataset. There are 60,361 chest x-rays negative for COVID-19 among 24,907 unique patients. Among the unique patients, there are at most 59 CXRs for a single patient. Given that there are many more negative cases than positive, for the purpose of achieving a more balanced dataset we will sample at most 1 CXR from each unique patient to avoid any bias that may occur from a sample with too many cases from a single patient. Among unique patients, there are 11425 female patients and 13482 male patients. To obtain a more balanced overall dataset, we will sample 4,706 negative CXR cases from the NIH Chest X-ray Dataset, with the sample containing an approximately equal number of males and female patients.

The test set is taken from the COVIDGR-1.0 Dataset, with one image per patient and equally balanced in male and female overall. There are 426 positive cases and 426 negative cases.

We do have concerns about drawing the positive and negative cases from two different datasets as this may introduce bias unique to the acquisition and processing of the CXRs unique to the specific hospital sites and dataset creation methodology. A possible alternative is to use the test dataset for training instead.

2.3 Feature Scaling

The relative upper bound and lower bound of all values are the same, given that they are simply converted from X-ray images. We will standardize the data using standardScaler to a mean of 0 and unit variance, and that should be sufficient to avoid bias due to scaling differences.

2.4 Computational Cost

The biggest challenge presented by the proposed datasets is the size of the data. As X-ray images have high resolution (at least 1024x1024), the size of each image is fairly large (around 1.4 MB). In the Brixia Score Covid19 Dataset, the positive prediction dataset, each GB contained only around 50 images; and in the NIH CRX8 database, each GB contained around 2000 images but there are over 80,000 negative samples. We currently do not have the capacity to download and train on such large files, given that with all three datasets combined we have around 100 GB of data.

To address this challenge, first we will compress all the images to 224x224 to reduce the dimension of the matrices, and to reduce computational cost. Additionally, we propose either using a selected subset of the dataset for training and testing, or we could use the test set (~900 images, equally balanced) only and split into training and testing, or we switch to a different dataset such as this one (well balanced, appropriate data size) https://github.com/maftouni/Corona_CT_Classification. We will also consult with the teaching team, and with the TF after TF assignment, to see if there are other computational solutions to train on extensively large datasets.