DISTINGUISHING COVID-19 USING CHEST X-RAYS

Srinivas Atreya

Chief Data Scientist, RoundSqr

Malini Vyakaranam

Associate Data Scientist, RoundSqr

RoundSqr

April, 2020

Table of Content

1 Abstract

COVID-19's rate of transmission depends on the timely detection of the carrier and the immediate implementation of interventions. Even though the CT scan is more sensitive to COVID Pneumonia, our approach considered Chest X-rays, for a possible preliminary classification, due to its prevalent usage as a primary diagnostic test. As linear dimension reduction techniques (like SVD) failed to classify the COVID Pneumonia from the Normal and CAP X-Rays, we have used Uniform Manifold Approximation and Projection (UMAP) to do the trick. UMAP uses manifold learning techniques and ideas from topological data analysis for dimension reduction. Using Supervised UMAP, we have been able to separate the 3 classes in the dataset. Our current dataset is limited, and hence not representative of the larger Covid-19 positive population. We hope to remedy that and use UMAP as a feature extraction technique for a classifier in the future. The study is carried out by a team of machine learning practitioners with support from subject-matter experts.

2 Introduction

Several countries are under lockdown today because of the exponentially increasing number of COVID-19 cases. The one factor that can contain the virus, apart from an increase in hygiene and social distancing, is early diagnosis to effectively isolate carriers of the disease.

Limited availability of viral testing kits and the time-consuming nature of these tests is making radiology come to the forefront of diagnosis. The report given by them is turning out to be a key element in deciding the treatment. COVID-19's rate of transmission depends on our capacity to reliably identify infected patients, with a low percentage of false negatives. Timely detection of the disease enables the implementation of all supportive care required by the affected patients as well as isolation to prevent spread.

A study at the Department of Radiology, Wuhan stated that 'deep learning methods' can be used to distinguish COVID-19 from community-acquired pneumonia. Using the Convolution Network model, Xie et all concluded this study of 4,356 CT exams with an AUC (Area Under Curve) of 0.96 for COVID-19.

A recent study has indicated that CXR started showing signs in 4 days. Though CT scans provide more conclusive data points with regards to the diagnosis, extensive availability of CXRs, including the fact that mobile X-rays are commonly used, have convinced us that this approach of using CXR for a possible preliminary classification of COVID-19 is worth pursuing

3 Dataset

COVID-19 images are gathered from several sources, primarily the covid-chest xray-dataset. The non-COVID pneumonia images are taken from the training images in the RSNA Pneumonia Detection Challenge on Kaggle.

Demographics	Normal	Pneumonia	COVID-19
Age (Mean±Std)	Data not available	Data not available	51±7.6
Gender	Data not available	Data not available	19 Males, 11 Females and 5 Unknown
Data Source	Rajpurkar, P, et al., *Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning.* arXiv preprint arXiv:1711.05225, 2017.	Rajpurkar, P, et al., *Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning.* arXiv preprint arXiv:1711.05225, 2017.	Cohen, J.P., Morrison, P. and Dao, L., *Covid-19 image data collection.* arXiv 2020.

The opacities are vague and fuzzy clouds of white in the darkness of the lungs. As the differences between Pneumonia and COVID-19 X-Rays were extremely subtle, high contrast images were created to make it relatively easier to classify. For the same, we normalized the X-Rays for each of the patients by subtracting the mean.

We calculated the mean of all the training images as a representation of the entire training set. The average image (Figure 2 – left) roughly represents the thorax and tells us that all the images are somewhat aligned to the center and are of comparable sizes. The standard deviation of the image (Figure 2 – right) sees a higher variance & shows up whiter.

Most of the Chest Radiograph Images (CXR) are available in the Poster anterior views (PA). This is a standard chest radiograph referring to the direction of X-Ray beam travel. It is frequently used to aid the diagnosis of acute and chronic conditions in the lungs.

The intent is to classify the X-Rays into normal lung, Pneumonia and COVID-19. From the below images (Figure 1), we can see that the lung opacities were observed in both the COVID and the pneumonia chest X-Ray images.

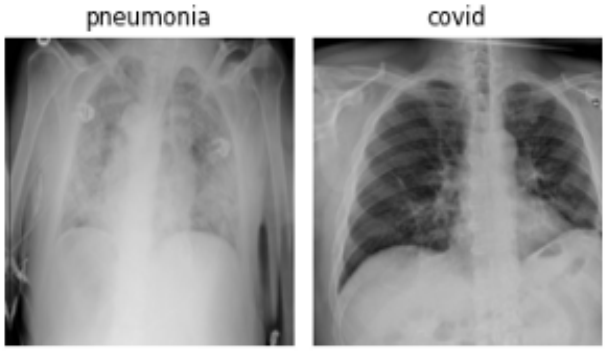

Figure 1: Chest X-Ray of a Pneumonia and COVID-19 patient

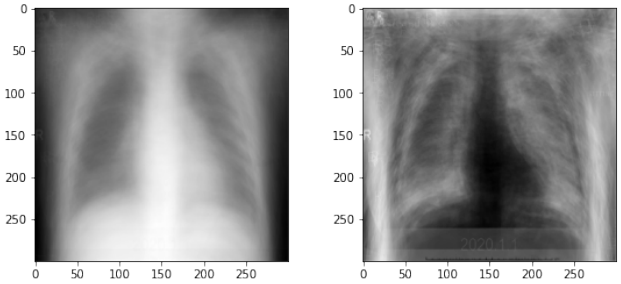

Figure 2: The average X-Ray

4 Approach

As the X-ray images were very few, the **first approach** was to use transfer learning to differentiate between the CAP and COVID pneumonia X-Rays. The idea was not to update the weights of the model's layers but to leverage the pre-trained model's weighted layers to extract features.

It was intended to use these features on a shallow classifier such as a Support Vector Machine (SVM) for disease classification.

The weights of the Chexnet model, a 121 layer Convolution Neural Network trained on the Chest X-ray 14 dataset, detects and localizes 14 kinds of diseases from Chest X-ray images. Feature maps were extracted and passed through an SVM Classifier, which achieved an AUC of only 50% on the test set.

This suggests that several samples were either insufficient and / or the 'signal to noise ratio' in these images was poor.

The **next approach** we tried was to identify important features and markers that are associated with the X-ray images.

Using Singular value decomposition, we dissociate the m × n data matrix X as follows

$$X = USV^T$$

where,

- U is an m × m orthonormal matrix whose columns are called the left singular vectors or the coefficient vectors of X.
- V is an n × n orthonormal matrix whose columns are called the right singular vectors or the expression level vectors of X.
- Matrix Sigma is an m × n diagonal matrix. The diagonal elements of matrix S are called the singular values of X which are used to create the feature vectors.

Using SVD, we've created sub images. These are more dominant features out of the X-rays.

The inner product of the first 3 modes and the entire image vector gives us 3 coordinates in the Eigen vector space.

When these points are plotted in this low dimensional feature space, we do not observe any distinct separation of clusters between COVID, Pneumonia, and Normal.

This indicates that the data is not linearly separable.

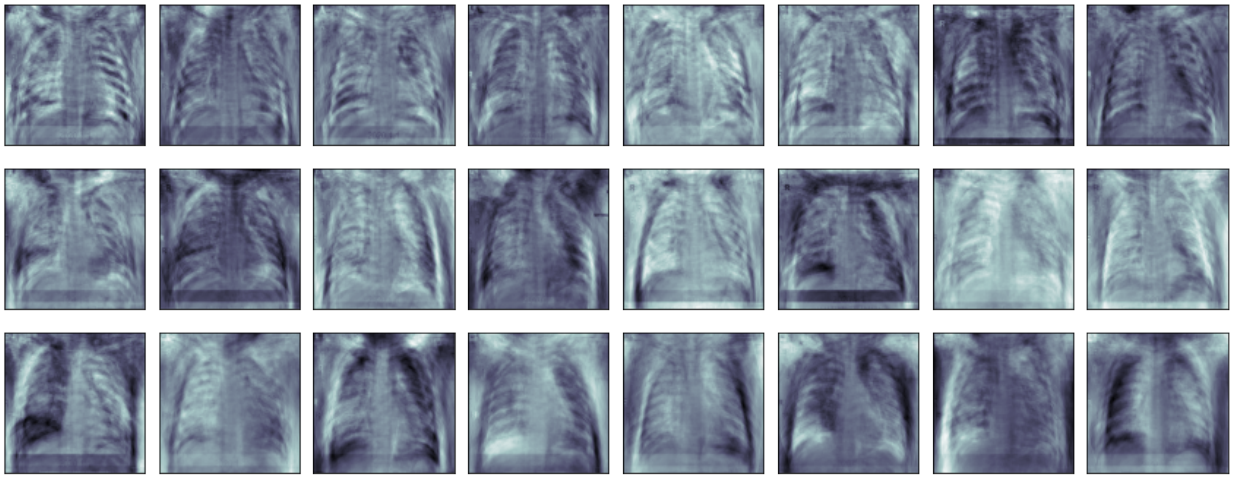

Figure 3: First 24 Eigen X-rays of the dataset

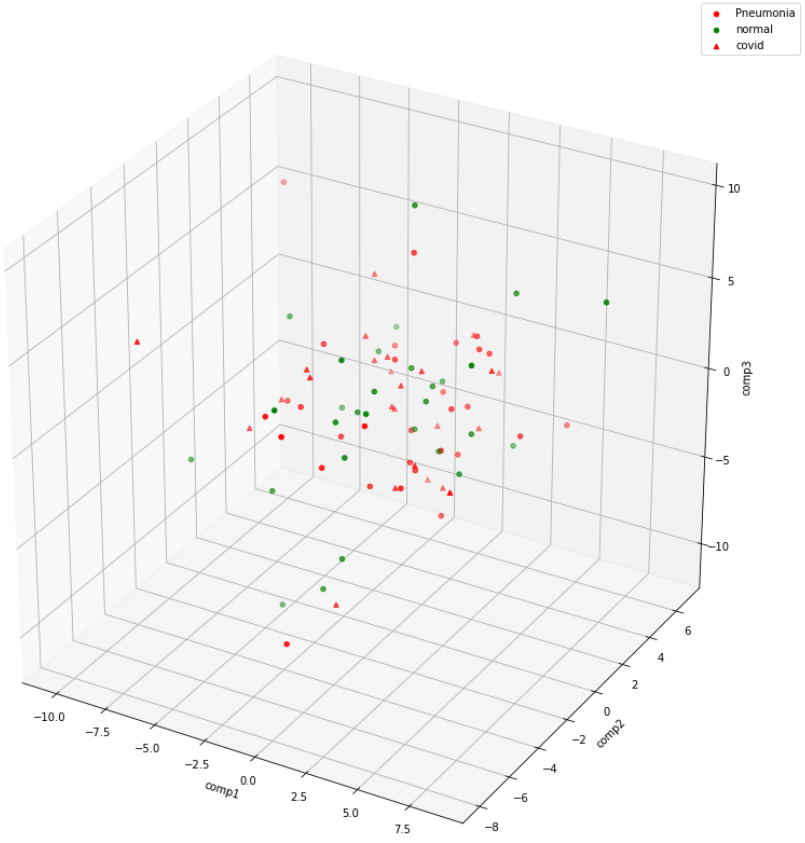

Figure 4: Scatter plot of the Normal Vs Pneumonia vs COVID-19 data points across the first three components

As the Matrix Factorization approach (SVD) was not giving good results, the **third approach** was to try and use the 'Neighborhood graph technique', like UMAP, to see if that gives better results. UMAP (Uniform Manifold Approximation and Projection) is a technique developed by McInnes et all. This graph-based technique leverages mathematical principles like Riemannian geometry and algebraic topology.

UMAP constructs a simplicial complex representation of data, and then optimizes to low dimensional space. This is a representation that is as close to the topological representation as possible using cross-entropy loss. It uses stochastic gradient descent for optimization. The explanation of the inner working of UMAP is listed in the UMAP paper.

Using the unsupervised UMAP techniques, we plot the embeddings between various combinations of normal, pneumonia and COVID19 X-Rays.

As we can clearly see in **Figure 5**, the normal and COVID-19 clusters are differentiated pretty cleanly, except for a couple of samples.

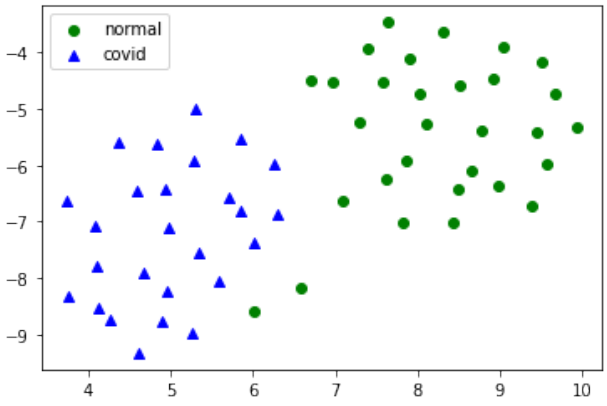

Figure 5: Unsupervised UMAP representation between Normal & COVID-19 X-Rays

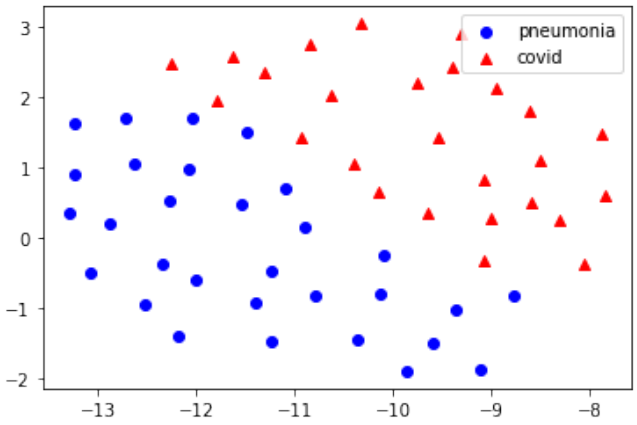

Figure 6: Unsupervised UMAP representation between Pneumonia & COVID-19 PNA

Even in **Figure 6**, the classification boundary between Pneumonia and COVID-19 is pretty distinct, with each falling in different parts of the spectrum.

In **Figure 7**, though the situation is slightly different, the clusters are pretty diffused and there does not appear to be a clear boundary. Part of this might be attributed to noise in the labelling of pneumonia cases, but we are unable to make any clear reasoning.

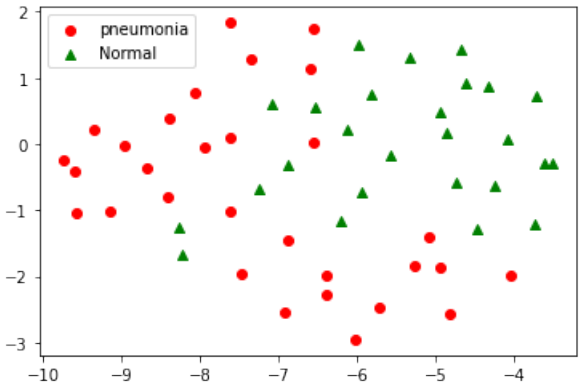

Figure 7: Unsupervised UMAP representation between Normal & Pneumonia

In the supervised UMAP, we make use of target labels to the model when fitting it to perform a supervised dimension reduction. From this visualization, of the 3 embeddings in the scatter plot, we observe that the classes are cleanly separated and isolated.

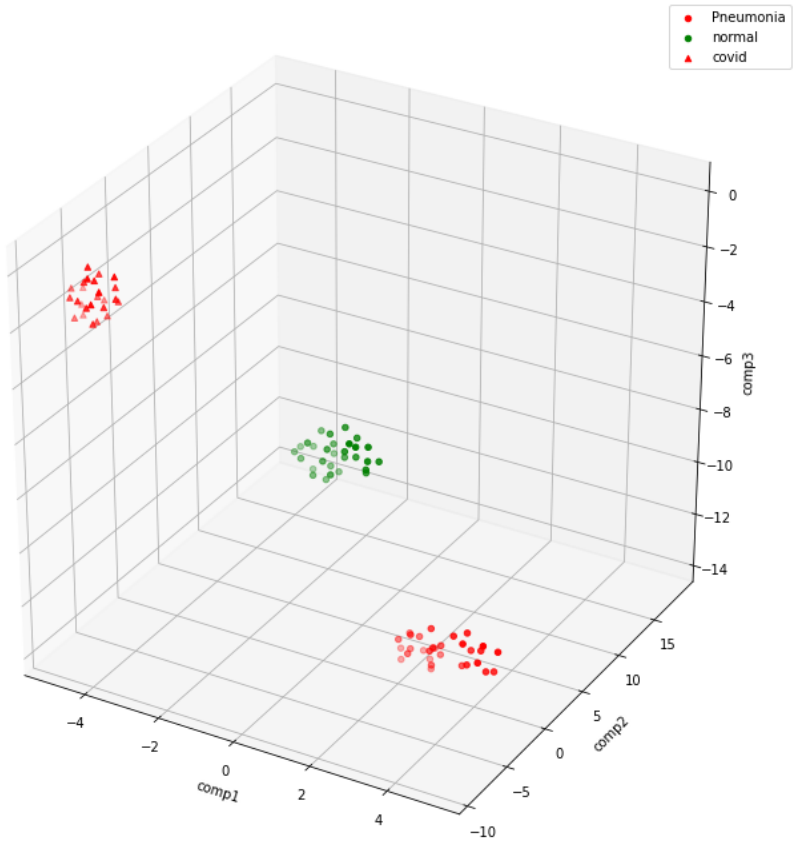

Figure 8: Scatter plot of the Normal Vs Pneumonia Vs COVID-19 data points across the first 3 embeddings in Supervised UMAP

Based on the data that is available and the approaches we have tried, we come to the conclusion that using supervised UMAP and metric learning, we are able to separate out the Normal, Pneumonia and COVID-19 chest X-Ray images (PA view).

5 Future work

UMAP could be used as a feature extraction technique on the disease classification task, using the classifier. For this to be conclusive, we will **need more data** at various stages, with the COVID-19 X-rays taken when the patients present severe symptoms.

We also plan to **develop an inference module** using UMAP that can translate an unknown X-Ray sample using the learned manifold representation.

The other line of thought, that we are pursuing using X-ray images, is around identifying the **prognosis of disease** based on the longitudinal data that is available.

We plan to publish an updated version of this paper in the next couple of weeks.

6 Reference

https://pair-code.github.io/understanding-umap/

https://umap-learn.readthedocs.io/en/latest/api.html

https://towardsdatascience.com/how-exactly-umap-works13e3040e1668

https://pubs.rsna.org/doi/pdf/10.1148/radiol.2020200905

https://www.nejm.org/doi/full/10.1056/NEJMoa2001191

https://github.com/ieee8023/covid-chestxray-dataset