Using Machine Learning to Detect COVID-19 from Chest X-Rays (Dec 2021)

Joshua A. Rothe, *Student, Johns Hopkins University, Whiting School of Engineering*

Table of Contents

I. Introduction

COVID-19 has become a household name in the last two years, and not in a good way.

> *The COVID-19 pandemic has led to a dramatic loss of human life worldwide and presents an unprecedented challenge to public health, food systems and the world of work. The economic and social disruption caused by the pandemic is devastating tens of millions of people are at risk of falling into extreme poverty, while the number of undernourished people, currently estimated at nearly 690 million, could increase by up to 132 million by the end of the year.*
>
> *Millions of enterprises face an existential threat. Nearly half of the world's 3.3 billion global workforce are at risk of losing their livelihoods. Informal economy workers are particularly vulnerable because the majority lack social protection and access to quality health care and have lost access to productive assets. Without the means to earn an income during lockdowns, many are unable to feed themselves and their families. For most, no income means no food, or, at best, less food and less nutritious food.* (World Health Organization)

Hospital capacity is a finite resource, and the pandemic has stretched this resource thin – for example, as of August 6, 2021, 42% of ICU beds were occupied by COVID-19 patients, putting a strain on both healthcare workers and facilities alike. (HHS) To alleviate this workload, it would be beneficial to healthcare workers to create a system that could easily detect or confirm a COVID infection without the need of additional testing by trained staff. Such a system could then free up doctors and nurses for tending to the patients that truly need it, improving healthcare service overall. Additionally, since medical errors are the third-leading cause of death in the U.S. (Hopkins), improved diagnosis accuracy can prevent unnecessary intubation and expediate patient care, helping to lower the risk of death while in the hospital or similar.

II. Project Objective

The objective of this project was to be able to detect the presence of a COVID-19 infection based on chest x-ray images.

III. Project Tasks

Images of chest x-rays from people that are both healthy and infected with COVID-19 were be used to train a deep features classifier. Different parameters and image pre-processing tasks were used, and the resulting models were evaluated for accuracy.

A. Sensor

The images were taken from chest (PA view) x-ray imagery of both healthy lungs and COVID patients' lungs. Additionally, images of patients with non-COVID viral pneumonia were also evaluated to determine Machine Learning's capabilities with discerning different types of respiratory infection.

B. Signal Capture

The data was taken from a publicly available dataset that contains thousands of chest x-ray images categorized into folders based on diagnosis – Healthy, COVID, Viral Pneumonia are the three largest ones. (Kaggle) Some of the 16-bit depth images from the Viral Pneumonia dataset needed to be thrown out, as all images used in the training and testing of the models were 8-bit depth to maintain consistent matrix size. 2,000 healthy images and 1,000 COVID infected images were used for the project training. An additional 1,000 Viral Pneumonia images were used to compare how the models performed when trained with additional respiratory images included.

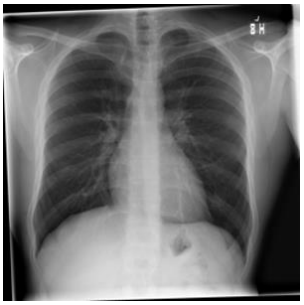

Fig. 1 – Healthy Chest X-Ray

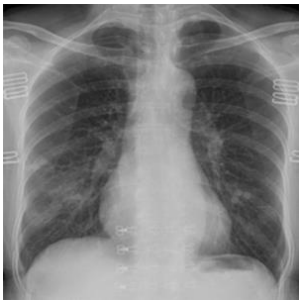

Fig. 2 – COVID Infection

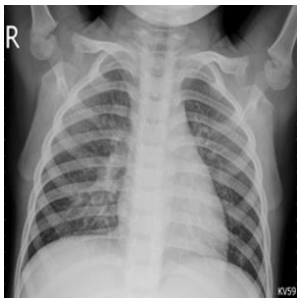

Fig. 3 – Viral Pneumonia

Visually, we can see that the healthy images have more "clear" lungs (blacker, to contrast the white of the ribcage), but Figures 2 and 3 are too visually similar to really tell apart at a glance.

C. Data Conditioning

K-means was applied at different K values and the accuracy of the models were evaluated (as well as with no K-means applied). Images were taken into the model, and the resulting image was flattened into a single vector of pixels. Since all images were 8-bit PNG files, reading in the image produced the same sized vector for each image. These images were matched with a corresponding one-row matrix that held the label of the corresponding input image.

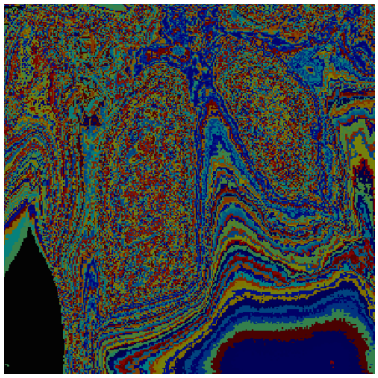

Fig. 4 – COVID Chest X-Ray with a K-Means of K=5 Applied

D. Feature Extraction

PCA was applied to each input image to increase performance by transforming the data onto a new coordinate system where the main axis emphasizes the greatest variation of the data. This makes the meaningful variation more visible to the model and allows the model to perform better.

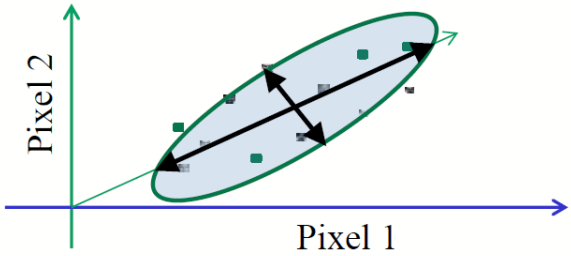

Fig. 5 – Visual Summary of PCA

PCA and K-Means complement each other – PCA helps with highlighting the "useful" features of the data by maximizing the variance and realigning the axis, and K-Means helps highlight the natural groupings of the data (given the proper K value – too high and it will start to think that noise is actual data, too low and there is not enough variance to make useful models with). They are both used for dimensionality reduction, and when used right can assist models with highlighting useful features needed to classify data properly.

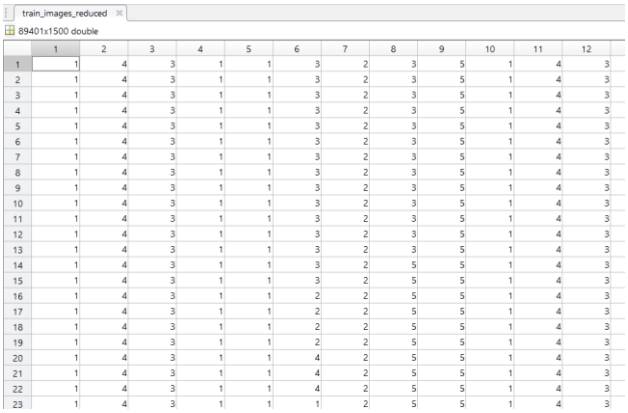

Fig. 6 – Data Matrix Prior to PCA Applied

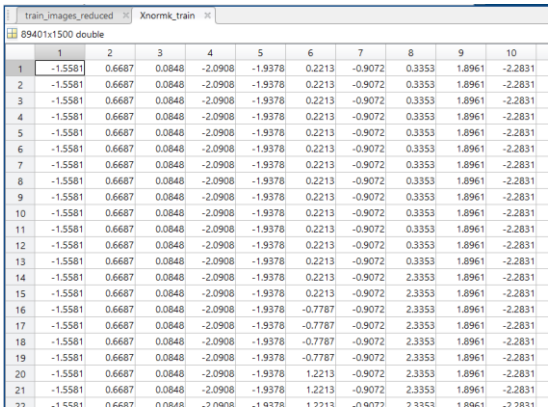

Fig. 7 – Data Matrix After PCA is Applied

E. *Feature Extraction*

The machine learning models were trained on 2,000 healthy x-ray images and 1,000 COVID-infected images for the main body of the project. Additionally, a 3-class model was evaluated and was trained on an additional 1,000 images of Viral Pneumonia patients, and a deeplearning model was also trained on the same. Testing was done against 1,000 healthy images and 1,000 COVID images, with accuracy percentages obtained for each separately to evaluate false positives and false negatives. For the 3-class model, 100 Viral Pneumonia images were also used for testing (this low number was due to the number of images available). The main model was evaluated at K values of 3, 5, and 10, and with no K-means applied. The main model was also evaluated based on normal and optimized parameters.

For the 3-class model, the models were separately trained on KNN, SVM and Naïve Bayes learners. The 2-class model was trained only on SVM.

IV. Results

For optimization of the 2-class models, MATLAB's fitcecoc function itself was used to brute force different values and settle on optimal parameters for BoxConstraint (C) and KernelScale (kernel).

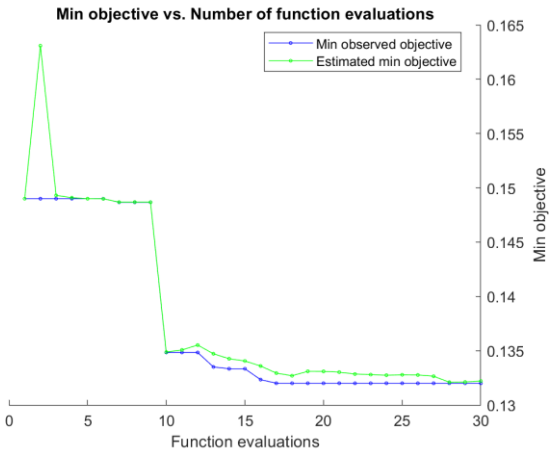

Fig. 8 – Evaluated Objective Per Iteration

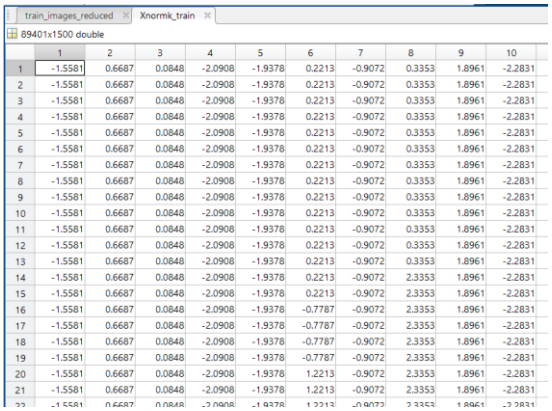

Fig. 9 – Best Results for Mdl_opt

	Basic Model (Mdl2)	Optimized Model (Mdl3)
No K-Means	93.00%	99.80%
K = 3	72.80%	88.40%
K = 5	69.80%	100.00%
K = 10	92.40%	99.70%

Fig. 10 – Testing Against Healthy Images for 2-Classt

	Basic Model (Mdl2)	Optimized Model (Mdl3)
No K-Means	70.90%	100.00%
K = 3	89.20%	93.00%
K = 5	70.40%	99.70%
K = 10	89.70%	99.50%

Fig. 11 – Testing Against COVID Images for 2-Class

The unoptimized model leaned slightly towards false negatives, and the accuracy would not make it very useful for real-world applications. The optimized model had amazing accuracy however, misclassifying only two healthy images as COVID (out of a testing set of 1,000 images). It successfully identified every COVID image provided. It is also worth noting that K-Means, while speeding up the model training slightly, decreased accuracy in all cases.

Additionally, the model was retrained with an included 1,000 images of Viral Pneumonia infections. The model performed considerably worse when these were added. No K-Means was applied. Performance was as follows:

	Basic Model (Mdl2)	Optimized Model (Mdl3)
Healthy	93.20%	96.70%
COVID	22.40%	10.70%
Viral Pn.	73.00%	80.00%

Fig. 12 – Testing 3-Class Models, No K-Mean

The performance of the COVID testing set was poor – with the optimized model, 81.5% of images were falsely labeled as

healthy. The Viral Pneumonia test set did slightly better, with 80% being correctly labeled and 20% being falsely labeled as healthy. The model seems to optimize heavily into a normal versus COVID model, which is likely due to the model being trained on an SVM learner.

Applying K-Means with a K value of 5 (our best performing K value thus far) gives us:

	Basic Model (Mdl2)	Optimized Model (Mdl3)
Healthy	62.30%	62.30%
COVID	25.70%	25.70%
Viral Pn.	49.00%	49.00%

Fig. 13 – Testing 3-Class Models, K = 5

With K-Means applied, this model performs worse than with no K-Means, so there is likely an issue with feature reduction for this model (perhaps it overfits to the remaining features it has to work with?)

Different model types were also tested – KNN and Naïve Bayes. KNN is usually a good choice for simple image classification, but the similarity of the COVID images and the viral pneumonia images likely contributed to its poor performance as well. Naïve Bayes can also be a good choice for image classification, but this performed the worst out of all of the models tested. Results are displayed in the figure below.

	KNN	Naïve Bayes
Healthy	77.50%	0.00%
COVID	41.70%	99.80%
Viral Pn.	93.00%	0.00%

Fig. 14 – Testing KNN and Naïve Bayes

Finally, a deeplearning image classifier using Convolutional Neural Networks was tested. This also performed very poorly, with an accuracy of 28.9%. Figure 11 below shows the training output, where the learner struggled to perform with any solid degree of accuracy.

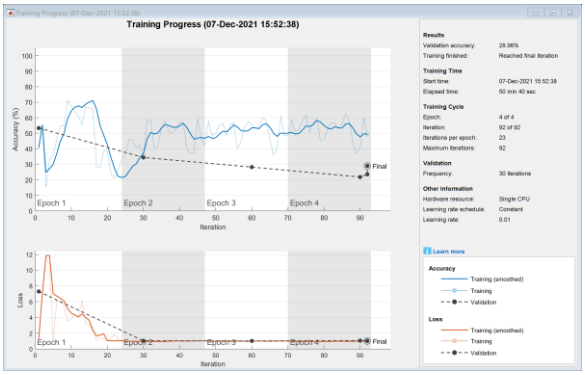

Figure 15 – Deeplearning Training Output

The visual similarities of COVID infection and viral pneumonia present unique challenges towards visual identification. This is not particularly surprising, since there are several recent whitepapers on the topic, but it is good to run through the numbers and compare how the different models handle the classification task.

V. Conclusion

The detection of normal versus COVID infection proved to be a surmountable task with a high degree of accuracy, once the models themselves were optimized with the proper parameters. It would be even more useful to classify between multiple respiratory diseases via a simple chest x-ray, but this task is clearly not as simple. Cross validating the model and running a kfoldloss leans heavily towards the model being overfitted, something that is difficult to combat with the resources available. More complex models, likely deep neural networks with large amounts of training data and ran on powerful processing servers that can handle cross validation, would likely have a better chance.

VI. Summary

When properly optimized, a trained SVM model can achieve 99.9% accuracy on labeling healthy versus COVID x-ray images despite variations in patient body type and image angle. These results can be achieved with a training set of only 3,000 images, and with only minor feature extraction applied. K-Means was seen to not improve results, but for computational reasons a K of 5 could be employed and still achieve a very high accuracy. The limitations faced when trying to discern between COVID and non-COVID respiratory infection is not related to model parameters or model type used, but appears to be primarily driven by the similarity of images and computational limitations of the standard MATLAB user. Nonetheless, a healthy versus COVID classifier could be used alongside a PCR test to help with initial diagnosis or confirmation of the disease, and is a useful starting point towards working on a more complex x-ray classification system.

VII. References

"Hospital Utilization." *HHS Protect Public Data Hub*, https://protect-public.hhs.gov/pages/hospital-utilization.

"Impact of Covid-19 on People's Livelihoods, Their Health and Our Food Systems." *World Health Organization*, World Health Organization, https://www.who.int/news/item/13-10-2020-impact-of-covid-19-on-people%27s-livelihoods-their-health-and-our-food-systems.

"Study Suggests Medical Errors Now Third Leading Cause of Death in the U.S. - 05/03/2016." *Johns Hopkins Medicine, Based in Baltimore, Maryland*, https://www.hopkinsmedicine.org/news/media/releases/study_suggests_medical_errors_now_third_leading_cause_of_death_in_the_us.

Villines, Zawn. "Coronavirus: Symptoms by Severity Level." *Medical News Today*, MediLexicon International, https://www.medicalnewstoday.com/articles/coronavirus-symptoms-severity-levels#severe-cases.

"COVID-19 Radiography Dataset." *Kaggle*, https://www.kaggle.com/preetviradiya/covid19-radiography-dataset.

VIII. Appendix

All relevant code and data for this project can be found on GitHub: https://github.com/rothej/jh_ml_covid_detect