**Classification of Pneumonia-Related Diseases Using Chest X-Ray Images**

**Author:**

Ken Trinh, Trevor Dalton, Sudhir Suvva

**Abstract:**

Pneumonia, is a flu-like symptom lung infection that causes fluid and inflammation build up in the affected lung and is detectable by X-ray imaging. However, it is quite difficult to differentiate seasonal flu virus vs bacteria related pneumonia. Throughout the years, many chest X-rays were taken from patients, and efforts have been made to categorize them based on the type of disease causing agent. We believe that using image processing techniques, and Machine Learning technology, we can better understand the differences between healthy and infected patients. We hope our results will contribute to future research into this topic.

**Project Overview (Ken)**

Typically X-ray images are taken by applying a beam of high radiation photon toward the target. This beam, after coming into contact with human tissue, projects an image on a metal firm. Since soft tissues such as skin and organs cannot absorb high energy rays, these beams pass through, imprinting only dense material like bone onto the film. In terms of infectious diseases, pneumonia in our cases, physicians use x-ray imaging to look for white spots or infiltrates in the lung. This white spot can be seen in the red circle region in Fig. 1 right X-ray image.

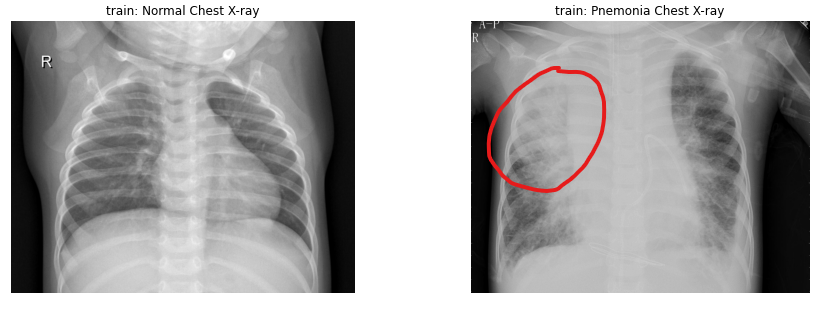
****

Fig. 1: Normal Chest X-Ray vs Pneumonia Chest X-Ray with noticeable white spot/cloud

Naturally, a trained physician would be able to distinguish these white spots easily. However, there are some cases where the white spots are not as apparent, leading to a misdiagnosis (Fig. 2). In this case, the normal chest X-ray actually shows some resemblance of white spots while the Pneumonia patient X-ray is pretty clear.

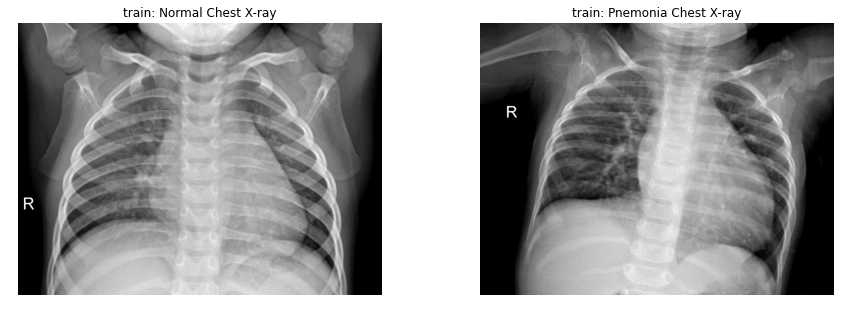


Fig. 2: Normal Chest X-Ray vs Pneumonia Chest X-Ray with no noticeable white spot/cloud

For that reason, our research wishes to answer the following question: Are there any features that we can learn through a combination of image processing and machine learning to better detect Pneumonia patients’ chest X-ray?

To address such a question, we wanted to experiment with different categorization systems to distinguish the variety of pneumonia causing agents (Covid 19, Flu Virus, or Bacteria). We believe that image processing is crucial in this project since some pneumonia patients X-rays show no discernible difference from healthy patients. First, we propose extracting HOG features for each image to characterize normal human anatomical structure. Second, we propose using local binary patterns to better represent similar regions of white clouds. Finally, since we noticed that some of the patients X-ray images have slight variation in origination (Figure 2.) and that we will be augmenting the data to create a more representative sample size, we will want to use SIFT features to capture these variations..

When it comes to image classification, Convolution Neural Network (CNN) is considered state of the art. In the article Convolutional Neural Networks proposed by Yamashita et al. [[3](https://link.springer.com/article/10.1007/s13244-018-0639-9)], the authors described and applied CNN on various classification tasks on radio imaging. Since this proposed work is relatively similar to Yamashita et al., we will be leveraging the CNN architecture. While CNN is state of the art, we wanted to experiment with different models, particularly Logistic Regression, Support Vector Machine, and Gradient Boosting. We decided to use these lower level Machine Learning algorithms alongside the state of the art CNN since Logistic Regression is an underlying mechanism of Neural Network. Support Vector Machine algorithm, on the other hand, tends to be more effective at higher dimensional spaces [[5](https://www.nature.com/articles/nbt1206-1565#citeas)]. Finally, we can leverage the learning from weak sets of learners from Gradient Boosting algorithm to give us insights into when and where of a feature the tree splitted [[6](https://journals.sagepub.com/doi/pdf/10.1177/15330338221087828)], which could be useful for us to better understand pneumonia X-rays.

**Data (Trevor)**

The data used in this project (collected by Joseph Paul Cohen of the University of Montreal) are made available on [Kaggle](https://www.kaggle.com/datasets/praveengovi/coronahack-chest-xraydataset). This dataset contains 5910 chest X-rays of normal and Pneumonia (Corona) affected patients: 1576 from unaffected patients; 2 from stress-smoking caused Pneumonia; 1555 from viral causes; and 2781 from bacterial causes.

Several papers [[1](https://arxiv.org/abs/2004.12823), [2](https://arxiv.org/abs/2004.05405)] have found concerns with detecting Covid 19 from chest X-rays. In particular it was found that there is a risk of models learning features that are entirely independent from the lungs that are unrelated to Covid 19.

**Model (Sudhir)**

We will use the one of following three simple classification models for our purpose as baseline:

1. Logistic Regression
2. Support Vector Machines
3. Gradient Boosting

For Neural Net Models we will use the following:

1. CNN
2. CNN with pre-trained models like ResNet50 or MobileNetV2.
3. Feed Forward NN (if time permits).

**Feature Engineering / Vectors**: We will try to apply the following techniques to create an abstraction of image and extract its features and quantify them.

1. Histogram of Gradients (HOG): We will generate the custom feature vector using the magnitude and orientations of the gradient. The generated feature vector will be applied for image classification. In our case we are hoping to identify the pockets of lung infection.
2. Local Binary Pattern (LBP): LBP helps in understanding the local representation of the texture of an image.
3. Scale Invariant Feature Transformation (SIFT): It is used to identify distinctive features from the images and used to apply for matching irrespective of the image scaling, rotation and translation.

**Citation**

[1] Maguolo, Gianluca, and Loris Nanni. "A critic evaluation of methods for COVID-19 automatic detection from X-ray images." *Information Fusion* 76 (2021): 1-7. <https://doi.org/10.48550/arXiv.2004.12823>

[2] Tartaglione, Enzo, et al. "Unveiling covid-19 from chest x-ray with deep learning: a hurdles race with small data." *International Journal of Environmental Research and Public Health* 17.18 (2020): 6933. <https://doi.org/10.48550/arXiv.2004.05405>

[3]Yamashita, R., Nishio, M., Do, R.K.G. et al. Convolutional neural networks: an overview and application in radiology. Insights Imaging 9, 611–629 (2018). <https://doi.org/10.1007/s13244-018-0639-9>

[4] Dale A. Schumacher, II.1 - IMAGE SMOOTHING AND SHARPENING BY DISCRETE CONVOLUTION, Editor(s): JAMES ARVO, Graphics Gems II, Morgan Kaufmann, 1991, Pages 50-56, ISBN 9780080507545, <https://doi.org/10.1016/B978-0-08-050754-5.50021-9>.

[5] Noble, W. What is a support vector machine?. Nat Biotechnol 24, 1565–1567 (2006). https://doi.org/10.1038/nbt1206-1565

[6] Vamvakas, A., Tsivaka, D., Logothetis, A., Vassiou, K., & Tsougos, I. (2022). Breast cancer classification on multiparametric MRI – increased performance of boosting ensemble methods. *Technology in Cancer Research & Treatment*, *21*, 153303382210878. https://doi.org/10.1177/15330338221087828