Diagnosis of Respiratory Infections from Chest X-ray Images

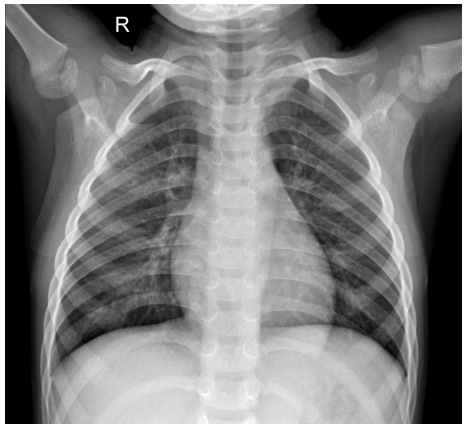
Alexandra Drossos, Julia Hossu, Anne Marshall, Hassan Saad

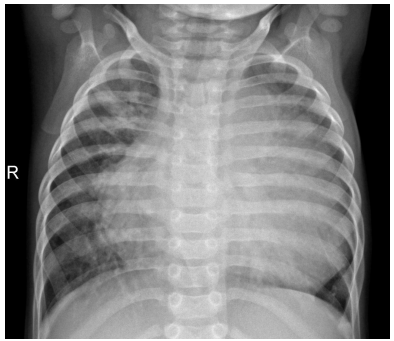
Background: Respiratory Infections

Pneumonia: bacterial, viral or fungal infection that impacts air sacs in the lungs, causing them to fill with liquid
On an x-ray there will be white spots / areas in the lungs.

Tuberculosis: bacterial infection that destroys lung tissue
On an x-ray there will be whiteness in lung lobes.

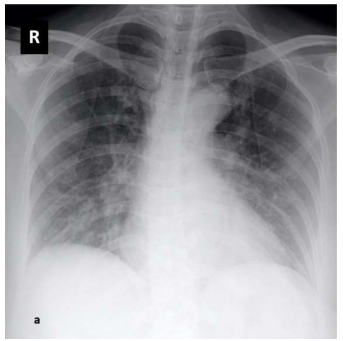
COVID-19: viral infection that causes severe inflammation to the lungs
On an x-ray there will be whiteness in the lung / obscuring of lung markings.





Pneumonia

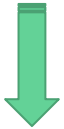
Tuberculosis



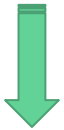
COVID-19

Question

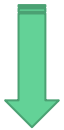
Normal vs. Pneumonia vs. Tuberculosis vs. COVID-19



Normal vs. {Pneumonia or Tuberculosis or COVID-19}



Normal vs. Pneumonia

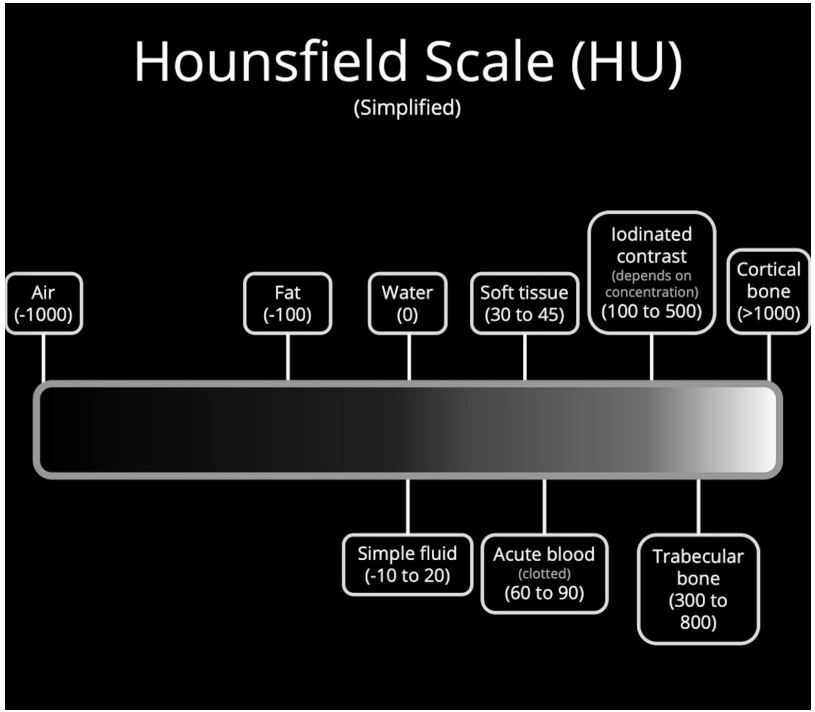


Normal vs. Pneumonia vs. ...

	F1 Score (95% CI)
Radiologist 1	0.383 (0.309, 0.453)
Radiologist 2	0.356 (0.282, 0.428)
Radiologist 3	0.365 (0.291, 0.435)
Radiologist 4	0.442 (0.390, 0.492)
Radiologist Avg.	0.387 (0.330, 0.442)
CheXNet	0.435 (0.387, 0.481)

Source: https://arxiv.org/pdf/1711.05225.pdf

Background: Medical Imaging



Hounsfield Scale
Scale based on absorption/attenuation coefficient which is a measure of physical density of tissue.

- 0: Pure Water
- -1000: Air

DICOM (Digital Imaging and Communications in Medicine)
Medical imaging standard for X-Rays and other CT scans

Much research is being conducted in using HU thresholds for diagnosis.

Data

- Chest X-Ray Kaggle Data Set

- ~2GB, 7135 photos

- Some Initial Hurdles:
  - RBG, Mode P, Mode L ➔ Mode L
  - Vectorization
  - Proportional vs. even import size
  - “Creation” of development set

set	train	test	val
COVID	460	106	10
NORMAL	1341	234	8
PNEUMONIA	3875	390	8
TUBERCULOSIS	650	41	12
TOTAL	**6326**	**771**	**38**

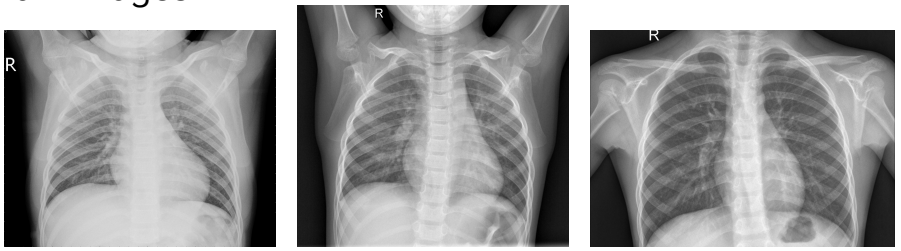
```
X_train.shape, X_train[1].shape
```

```
((650, 40000), (40000,))
```

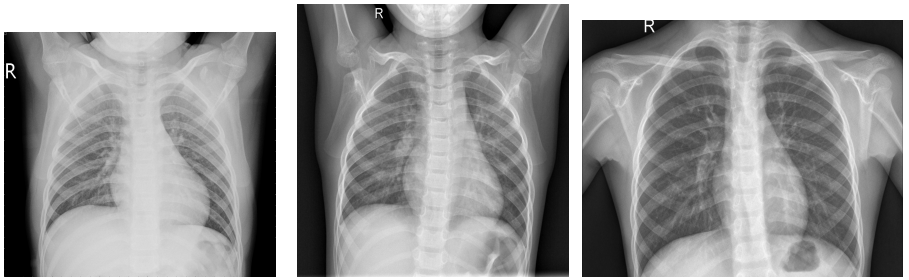
Image Pre-Processing

- Files are provided in JPG format
  - 1000x1000 Resolution
  - Images vary in aspect ratio
- Using Pillow image processing library
- Using Pillow's Luminance conversion to grey scale
- Subsampling to 200x200

Raw Images



Converted To Luminance



Numpy Arrays

Re-scaling and Normalization

- Images start out on a calibrated scale in the DICOM format
- All features are on the same data scale
- Pillow image conversion is converting these to a [0,1] grayscale value
- Mean of the existing data is .4949, Standard deviation is .249.
- This isn't quite a Normal or Gaussian distribution, but it is close.
- Experimental runs with SVM show that re-normalizing does not yield any improvement in the models.

Therefore: We have decided to not do additional normalization

Algorithms & Initial Results

- Multi Layer Perceptron
  - Parameters: Hidden layer Values = (5,2), alpha = 1, activation = identity, solver = lbfgs
    - Default Model: 95.2%
    - Optimal Model (5 Fold Cross Validated GridSearch): 95.9%
- Naive Bayes
  - Parameters: alpha = 3001
    - Default Model: 82%
    - Optimal Model: 82%
- KNN
  - Parameters: K = 5
    - Default Model: 95.8%
    - Optimal Model (5 Fold Cross Validated GridSearch): 96.1%
- SVM
  - Parameters: C=10, gamma= 0.0001, kernel= 'rbf'
    - Default Model: 96.5%
    - Optimal Model (5 Fold Cross Validated GridSearch): **97.9%**

Next Steps

Ensemble Learning

Expand into TB and COVID - Viral Pneumonia vs. COVID difficulty

Other pre-processing steps - Blurring, Sharpening

Q&A