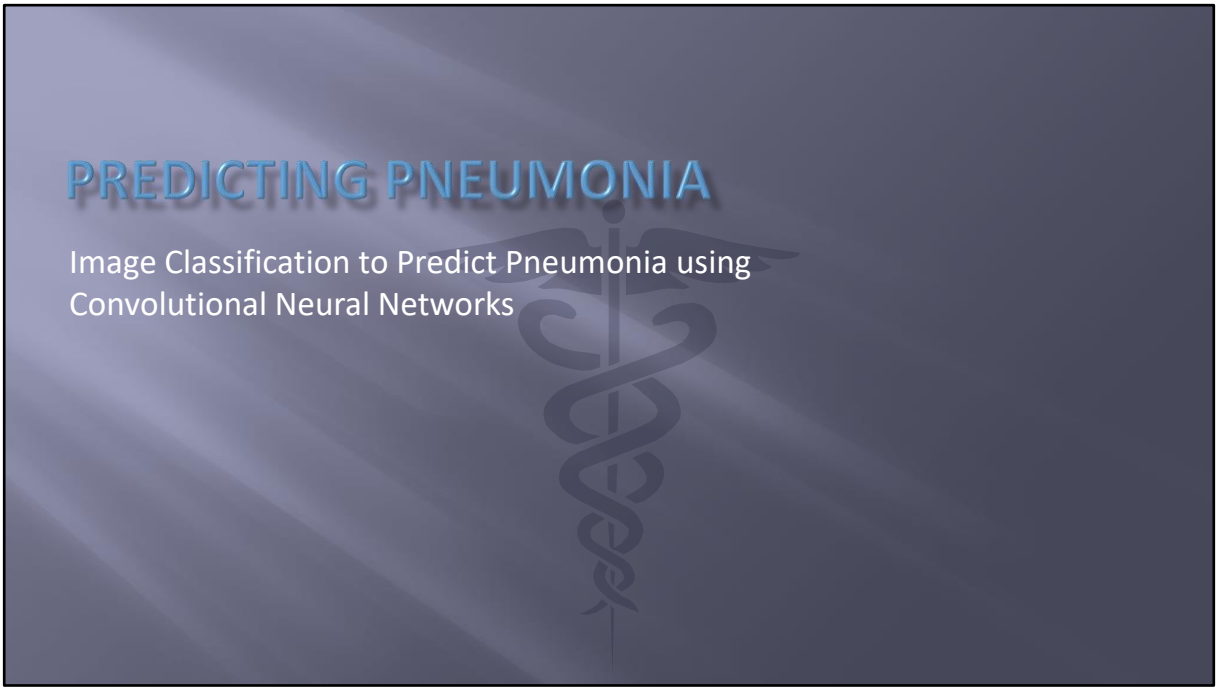
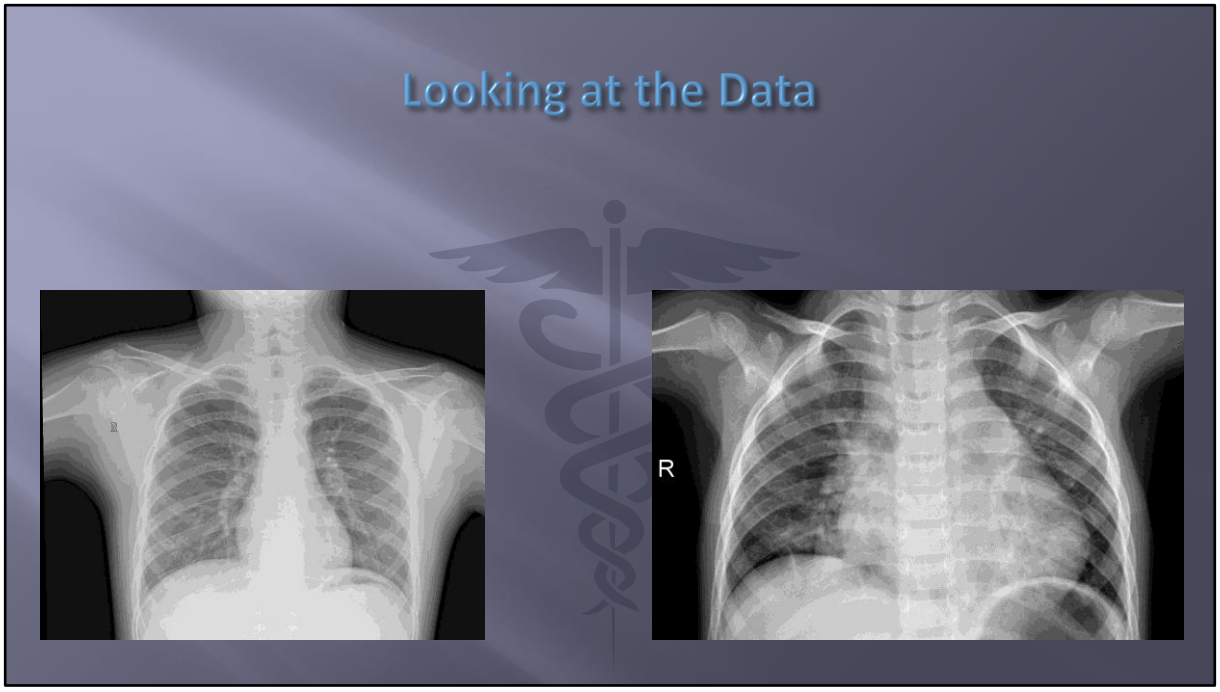


For this project, I assumed the role of a Data Scientist working with a medical research firm. The goal is to train a machine learning model to classify whether a patient has pneumonia or not, given images of their chest x-rays.

What Is Pneumonia?

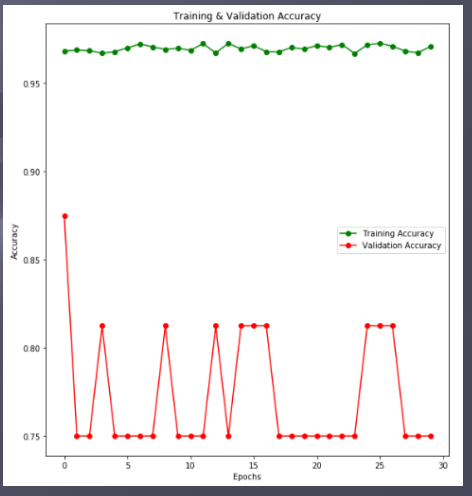
- Pneumonia is an inflammatory infection that primarily affects small air sacks in the lungs known as alveoli. In more severe cases, these sacks may fill with fluid.
- Symptoms include coughing, chest pain, fever, and labored breathing.
- Each year, ~450 million people are infected globally and about 4 million of those die.
- The 20th century brought antibiotics and vaccines, which greatly increase survival rates, however pneumonia remains the leading cause of death in developing nations, as well as infants, seniors, and the chronically ill.



The dataset used contains images of healthy lungs, like those on the left, and infected lungs, like those in the right image.
The goal is to train a convolutional neural network to accurately identify patients that have pneumonia. The goal in this instance is more of a proof of concept rather than aiming for pure accuracy, due to the size of the dataset.

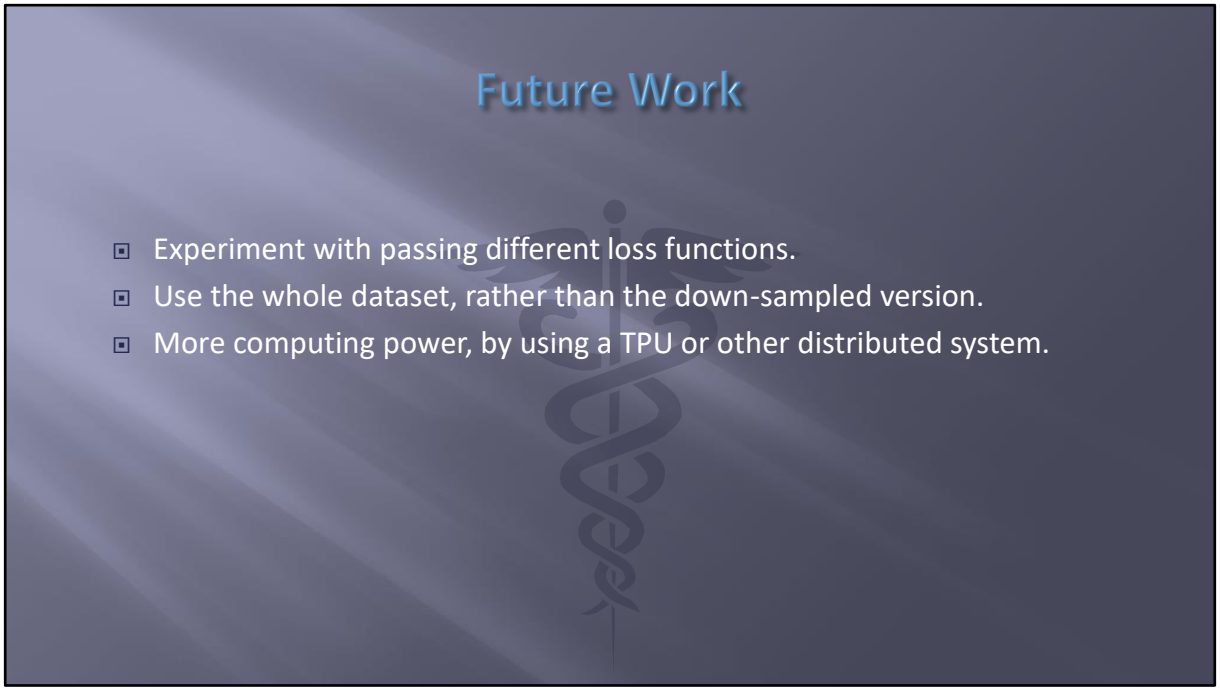
Model Performance

	Recall	Accuracy
Baseline CNN	42.39365	88.94290
CNN with Dropout	23.68644	91.18589
CNN with dropout, batch normalization, and learning rate reduction on plateau	28.10528	91.18589



The learning rate reduction on plateau/batch normalization with dropout model performed the best. It was the most stable of the models trained, but had slightly higher recall than the dropout model without batch normalization or LRReduction on plateau but it also boasted higher accuracy.
It was also pretty consistent through the epochs, which is likely due to the batch normalization. The dropout assisted in preventing overfitting, and the learning rate reduction on plateau kept the validation accuracy high through all the epochs.

On the right is a visualization of the training and validation accuracy of the final model.



The third model I made performed the best, but there is still some work to be done to manage recall. In the future, I'd like to train another model and pass a different loss function, like a softmax classifier to see if there are improvements to be had over my chosen binary crossentropy classifier.
To better train the model, it would be advantageous to use the entire dataset rather than the down-sampled set used in this notebook. With more images, the model would be trained significantly better.
Related to the above point of using the whole dataset, I'd need to make use of a Tensor Processing Unit or other distributed system to handle that amount of data. My local machine is quite powerful, but still an unweildy tool for a task that large.

Appendix

- Additional work and full in-depth analysis of the data can be found in the Predicting_Pneumonia.ipynb file of this repo: https://github.com/RS291/Mod4_Pneumonia_Project