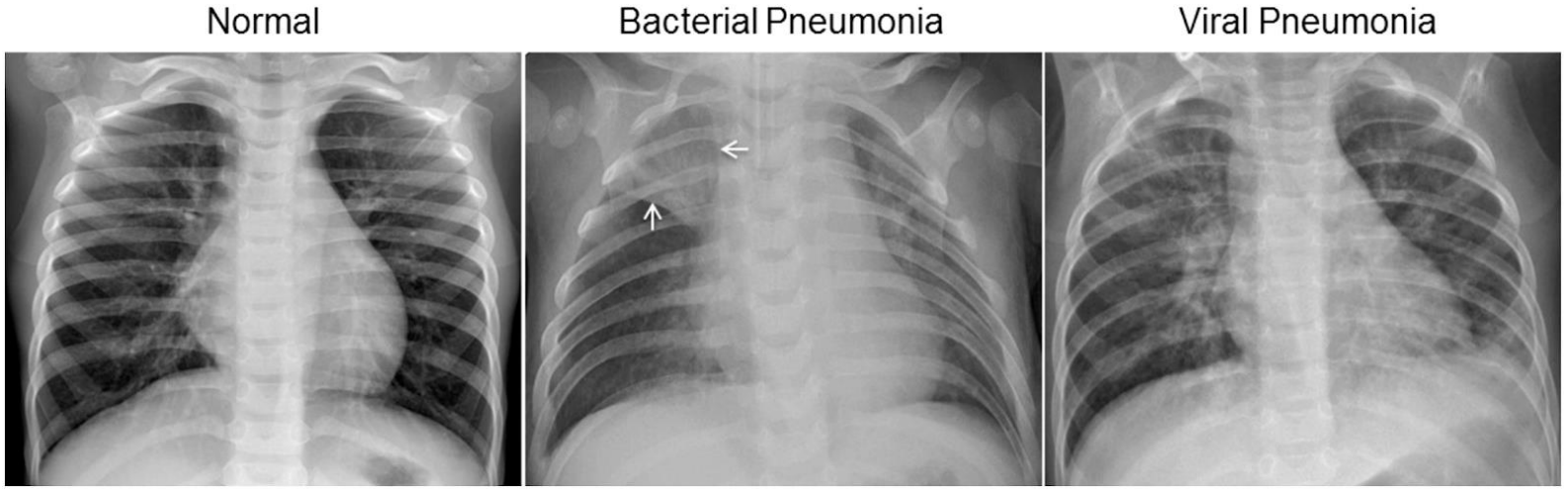
AI Model to Assist in Pneumonia Detection

CHRIS WOODS

Background



- According to the World Health Organization (WHO), pneumonia kills about 2 million children under 5 years old every year and is consistently estimated as the single leading cause of childhood mortality (Rudan et al., 2008), killing more children than HIV/AIDS, malaria, and measles combined (Adegbola, 2012)
- The WHO reports that nearly all cases (95%) of new-onset childhood clinical pneumonia occur in developing countries, particularly in Southeast Asia and Africa.
- Bacterial and viral pathogens are the two leading causes of pneumonia (Mcluckie, 2009) but require very different forms of management.
- Bacterial pneumonia requires urgent referral for immediate antibiotic treatment, while viral pneumonia is treated with supportive care.
- Therefore, accurate and timely diagnosis is a matter of life or death.

Project Goal

The purpose of this project is to determine if an AI model can be developed to assist in the diagnosis of pneumonia using common radiographs (chest x-rays).

- Success will be measured by the metric Recall, or Sensitivity. Recall answers the question, what proportion of actual positives are identified correctly. Mathematically, Recall is defined as follows:

$$\text{Recall} = \frac{TP}{TP + FN}$$

- False negatives (FN) are of critical importance here. In this context, a false negative would mean that a child who actually had bacterial pneumonia would go on undiagnosed. The ramifications of this error could be life threatening.

  "Pneumonia caused by bacterial infections poses a much greater threat to the heart than pneumonia caused by viral infections, a new study suggests. Patients in the study who were diagnosed with bacterial pneumonia had a higher risk of heart attack, stroke or death, compared with patients diagnosed with viral pneumonia, the researchers found." Link

Project Format

This project was executed in two phases:

Phase I

- Established a baseline for comparison.
- Create a model that will distinguish normal radiographs from those with pneumonia present.
- A good result in Phase I will allow us to remove normal radiographs from the analysis and focus solely on a model that can differentiate bacterial from viral pneumonia.

Phase II

- If pneumonia is present, develop a model to differentiate bacterial from viral pneumonia.

The goal of Phase I is to maximize the Recall of the pneumonia present prediction.

In Phase II, it will be maximizing the Recall of the bacterial pneumonia present prediction

Exploratory Data Analysis (EDA) and Visual Analysis

For this project the Chest X-Ray Images (Pneumonia) dataset was used. All images are represented as JPEG files

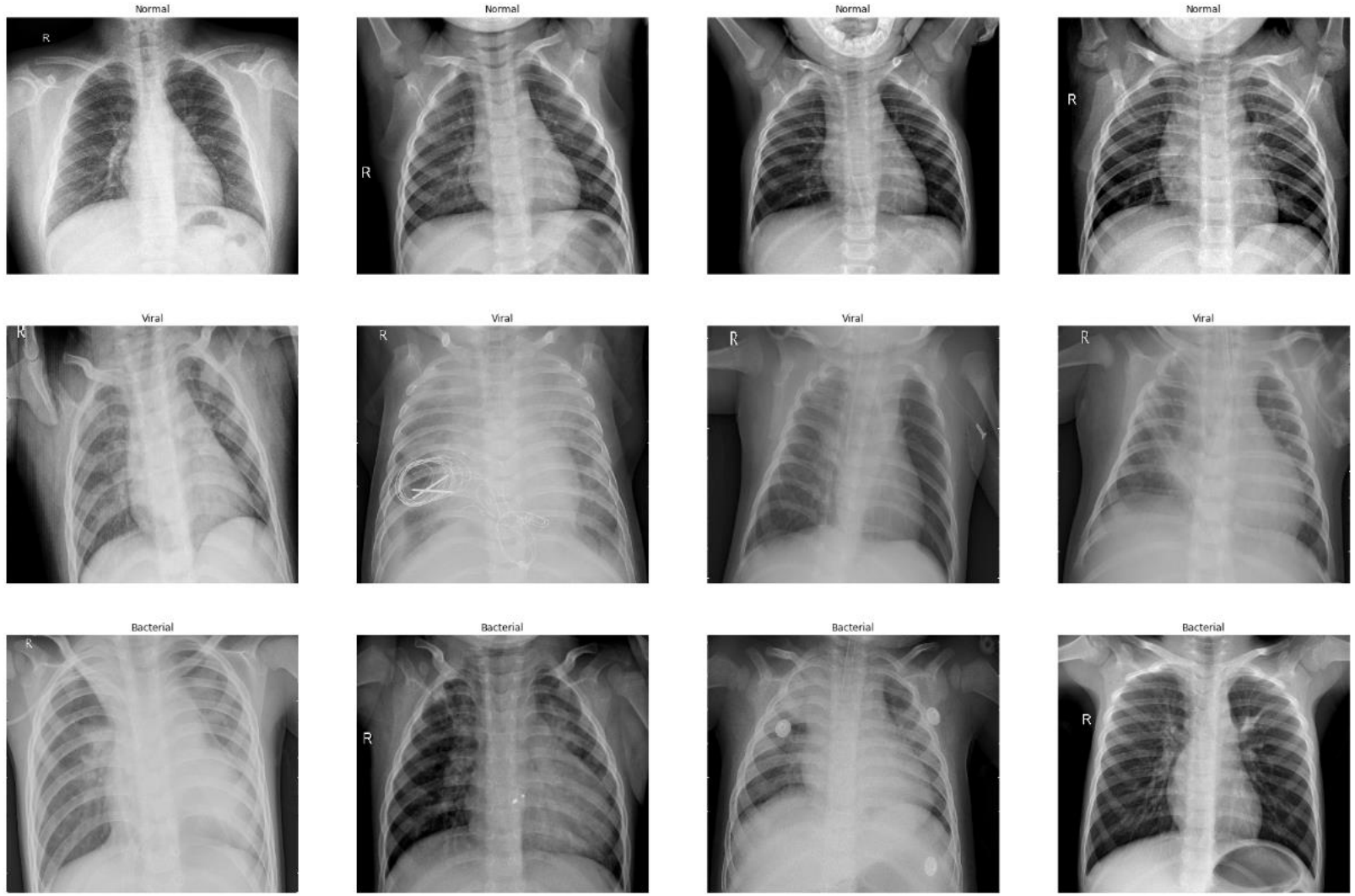
- A total of 5,840 chest X-ray images from children.
- 4,256 characterized as depicting pneumonia (2,772 bacterial and 1,493 viral)
- 1,575 normal.

Differentiating between a normal radiograph and one with pneumonia present can be quite challenging.

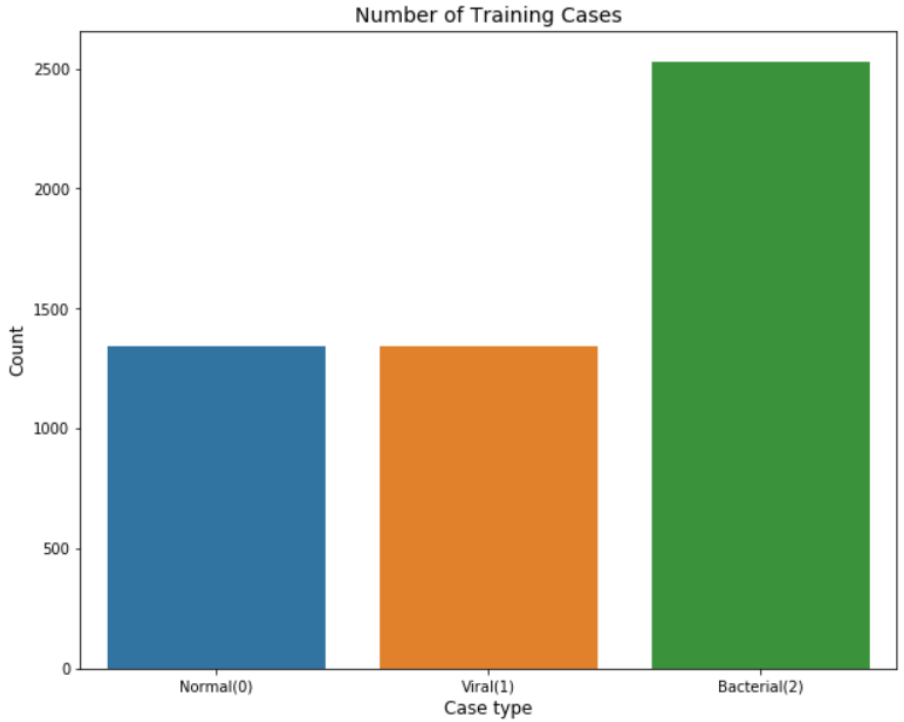
Modern medical facilities have multiple tests at their disposal in order to make a proper diagnosis.

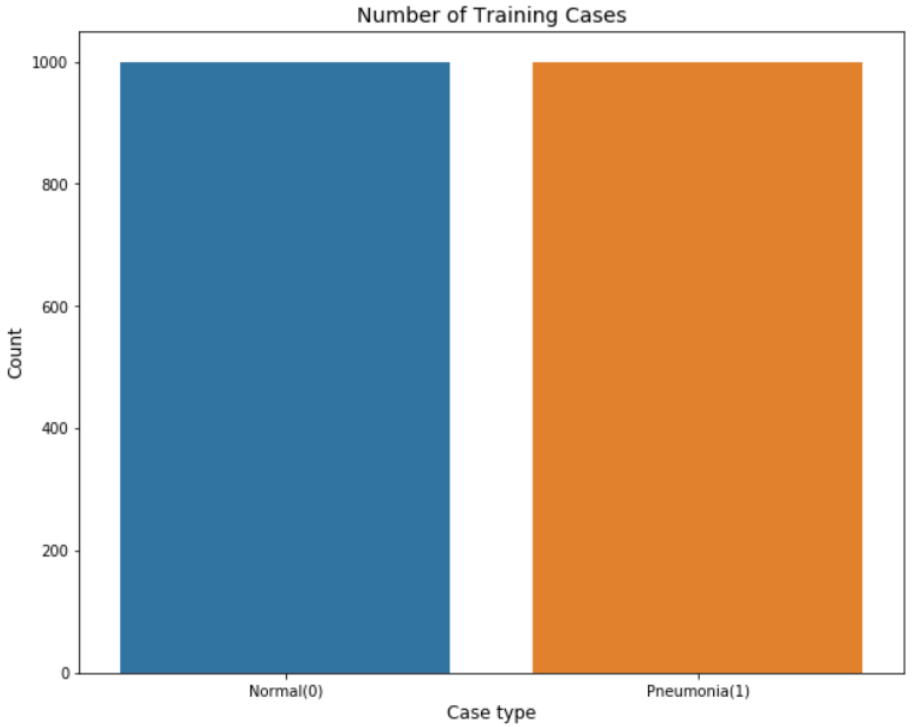
However, many third world clinics where death from pneumonia is prevalent, do not have access to the same modalities as we in the west.

A machine learning application that could accurately diagnose pneumonia would save lives



Data Cleaning and Wrangling



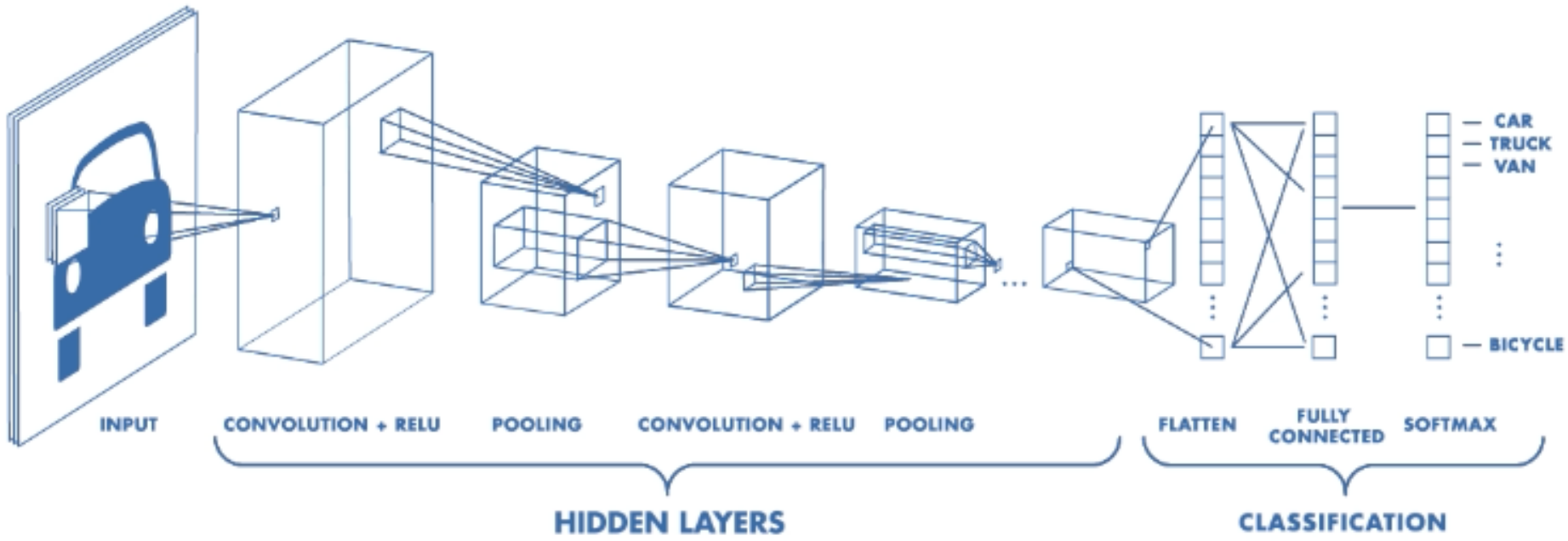


- This is a clean dataset consisting images stored as jpeg files split into training and test directories.
- Here we can see somewhat of an imbalance in the training data cases, with bacterial pneumonia represented at almost twice the rate of the viral pneumonia and normal cases.
  - 1341 Normal images
  - 1345 Viral Pneumonia images
  - 2530 Bacterial Pneumonia images

- In Phase I, to lessen the chances of overfitting and bias toward classification as bacterial pneumonia, the images in both test and training directories were redistributed and balanced.
- Now there are 1,000 normal images and 1,000 pneumonia images to train the model.

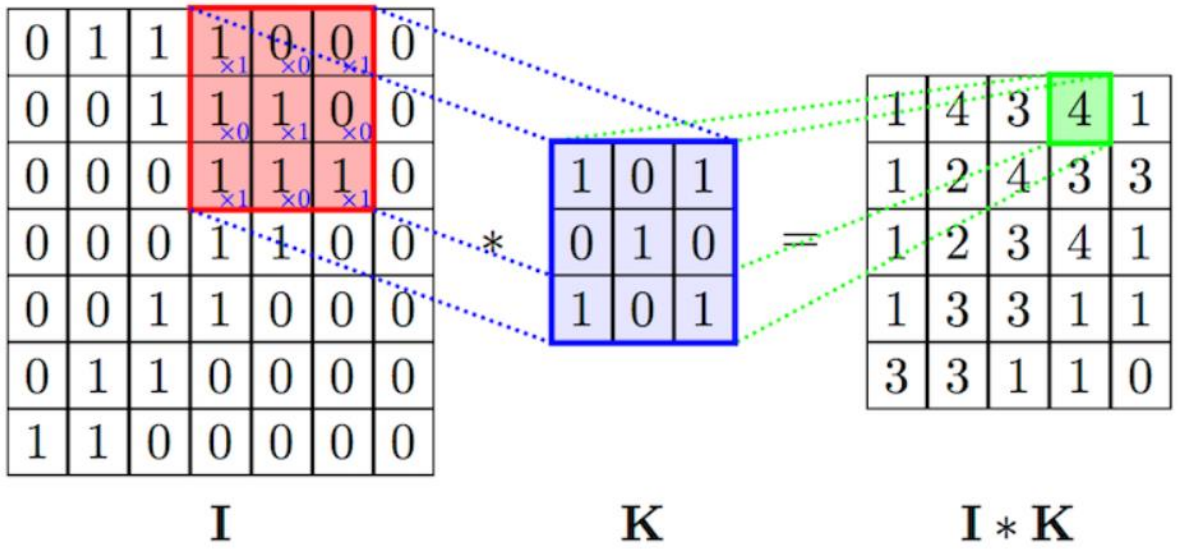
Machine Learning Model Overview

Supervised Learning Algorithm ➔ *Deep Learning*
Convolutional Neural Network (CNN or ConvNet)



Machine Learning Model Overview

- Hidden Layers /Feature Extraction – Network performs series of **convolutions** and **pooling** operations to detect features
  - A **convolution** is executed by sliding the filter over the input image. At every location matrix multiplication is performed and sums the result onto the feature map.



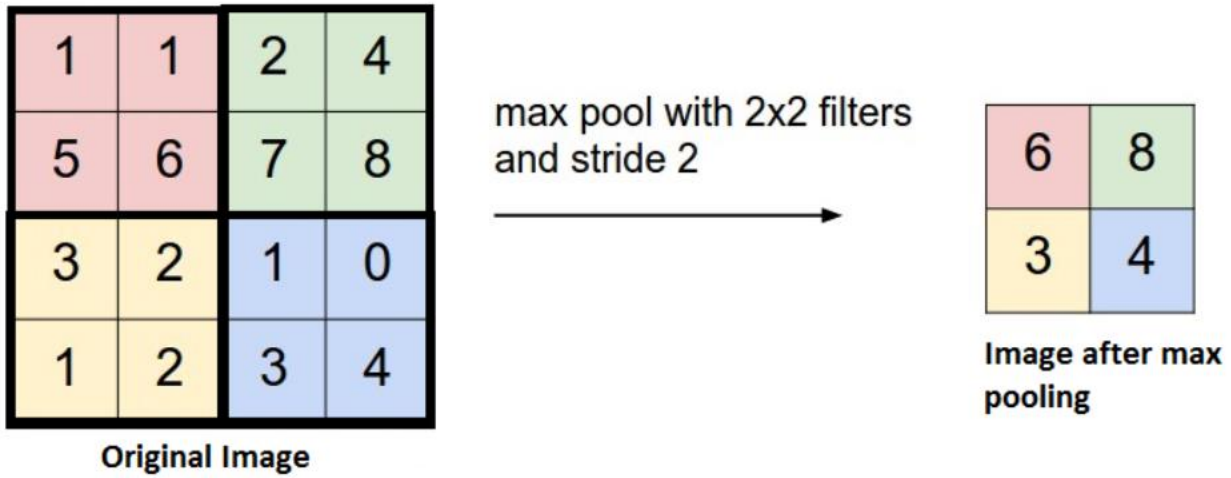
In an image, the convolutional layer detects the features that make the image unique. The convolution layers learn such complex features by building on top of each other.

The first layers detect edges, the next layers combine them to detect shapes, to following layers merge this information to infer a particular feature.

Machine Learning Model Overview

- **Pooling** layers are usually between CNN layers. The function of pooling is to continuously reduce the dimensionality to reduce the number of parameters and computation in the network.

  The most frequent type of pooling is max pooling, which takes the maximum value in each window. This decreases the feature map size while at the same time keeping the significant information.



- Classification – Fully connected layers will serve as a classifier on top of the extracted features and will assign a probability for the image being what the algorithm predicts.

Machine Learning Phase 1 - Normal vs. Pneumonia

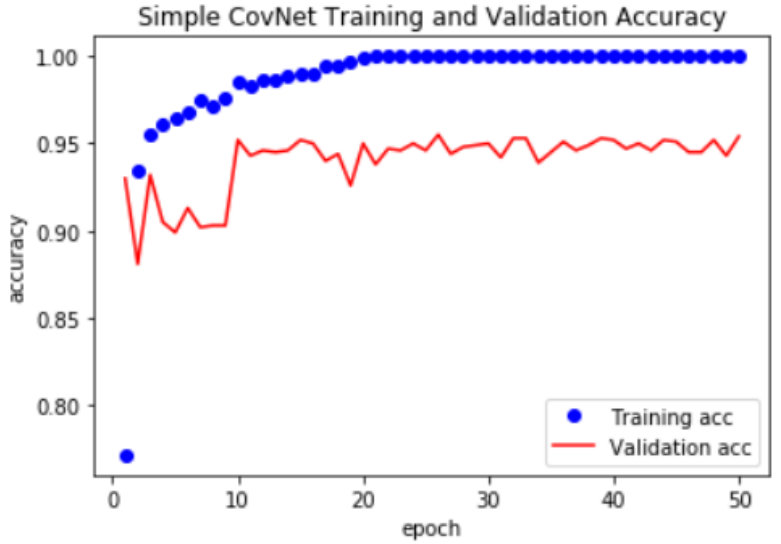
Phase I was conducted in three parts:

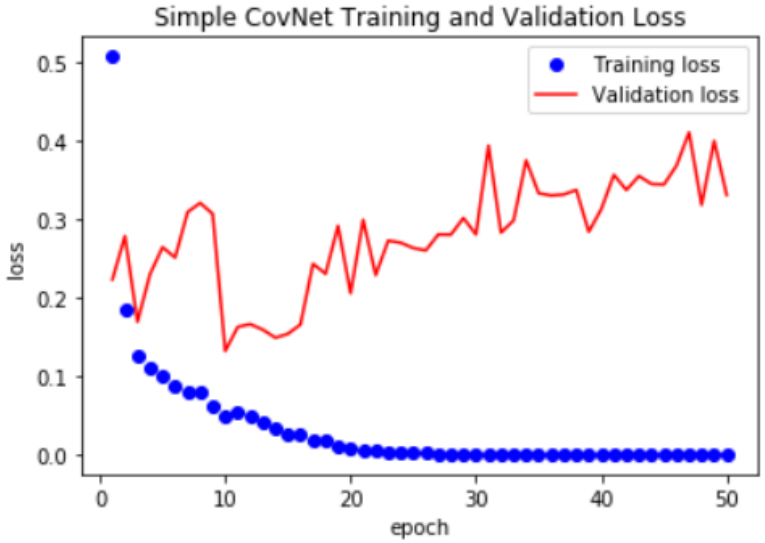
1. To establish a baseline for model performance, a small convolutional neural network (ConvNet) based on the framework suggested by Francois Chollet in his book “Deep Learning with Python” was executed.
2. A more complicated model utilizing Data Augmentation and a dropout layer.
   - **Data Augmentation** - Works by generating more training data from existing training samples, by augmenting the samples via a number of transformations that yield believable looking images. Helps expose the model to more aspects of the data and generalize better.
   - **Dropout** - randomly goes through and drops nodes. CNN nodes are unequal in their explanatory power in the training set, randomly dropping nodes improves the performance of underperforming nodes and reduces the performance of otherwise high performing nodes. Resulting in a better model overall.
3. Finally, a model utilizing Transfer Learning.
   - **Transfer Learning** – Use the learning from the convolutional base of the VGG16 model and rework the fully connected layer where classification occurs according to our model.

Machine Learning Phase 1 - Normal vs. Pneumonia

1 - Baseline Convnet Accuracy and Loss

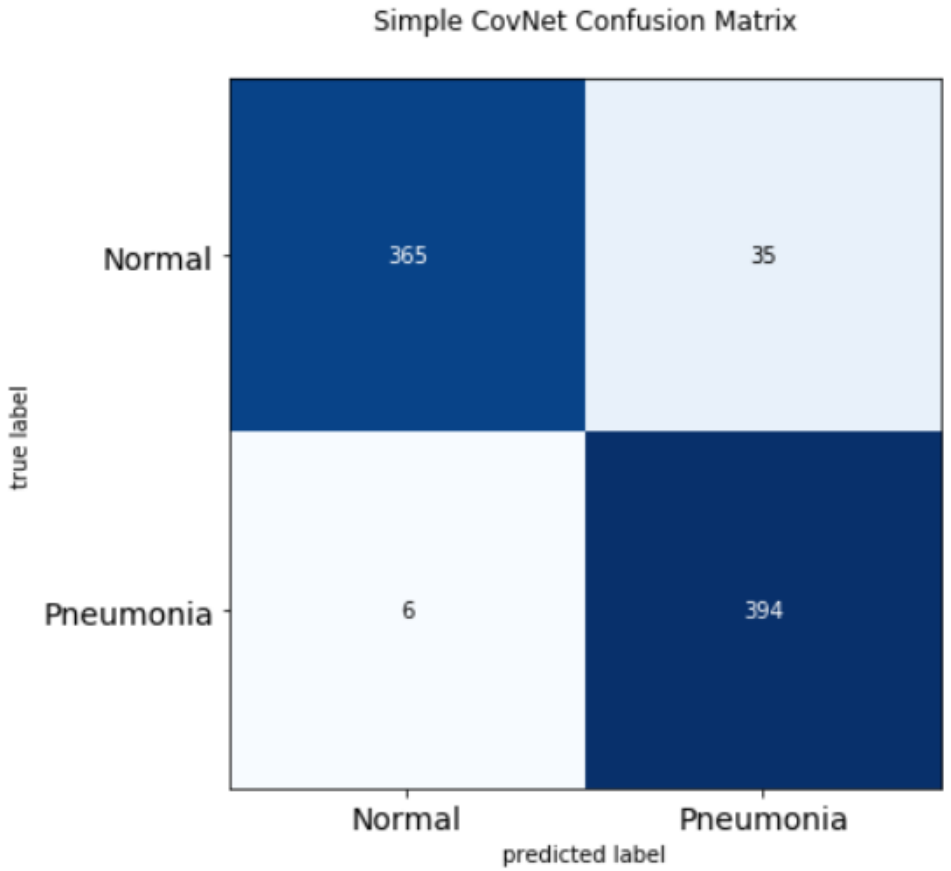
- The model performs well on the data it is trained with and not so well on the data it has not seen before.
- After the 10th epoch, the model is over optimizing on the training data, **overfitting**.
- This causes the model to learn representations that are specific to the training data and do not generalize to data outside of the training data set.
- Even with overfitting, it is still a reasonably good model. Validation accuracy is close to 95%, and validation loss below 0.4.





Machine Learning Phase 1 - Normal vs. Pneumonia

1- Baseline Convnet Confusion Matrix and Classification



	precision	recall	f1-score	support
0	0.98	0.91	0.95	400
1	0.92	0.98	0.95	400
micro avg	0.95	0.95	0.95	800
macro avg	0.95	0.95	0.95	800
weighted avg	0.95	0.95	0.95	800

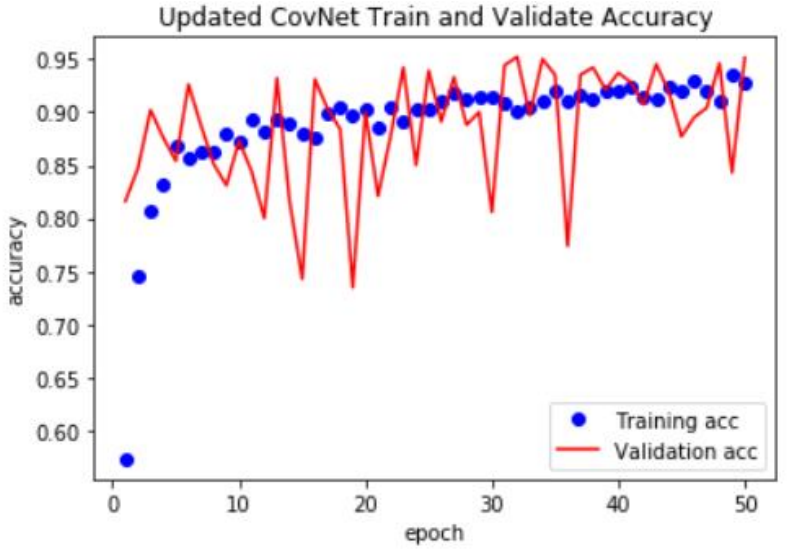
Good Results for a first pass at the data!

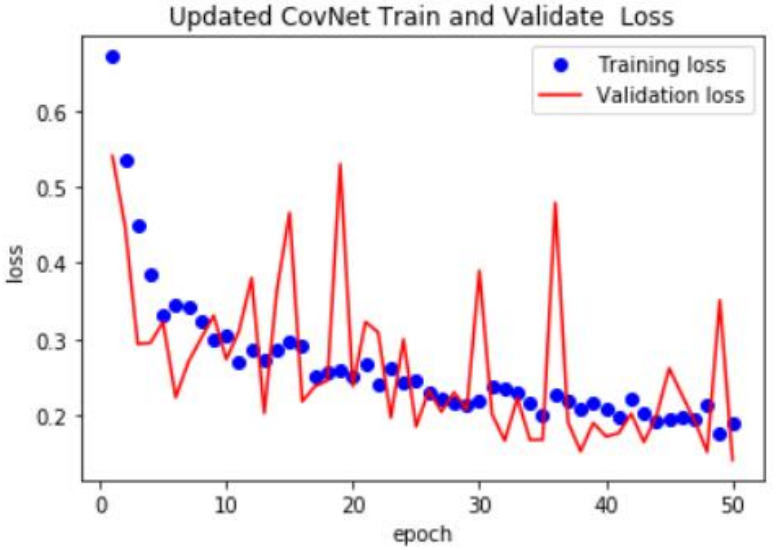
- Good Recall with the true pneumonia radiographs, 98.5%. Or to put another way, 1.5 % who truly had pneumonia would be classified as normal with this model.
- Reasonably good Recall for normal radiographs, 91%. With this model 9% of children without pneumonia would get diagnosed with pneumonia.

Machine Learning Phase 1 - Normal vs. Pneumonia

2 - ConvNet with Data Augmentation and Dropout Layer

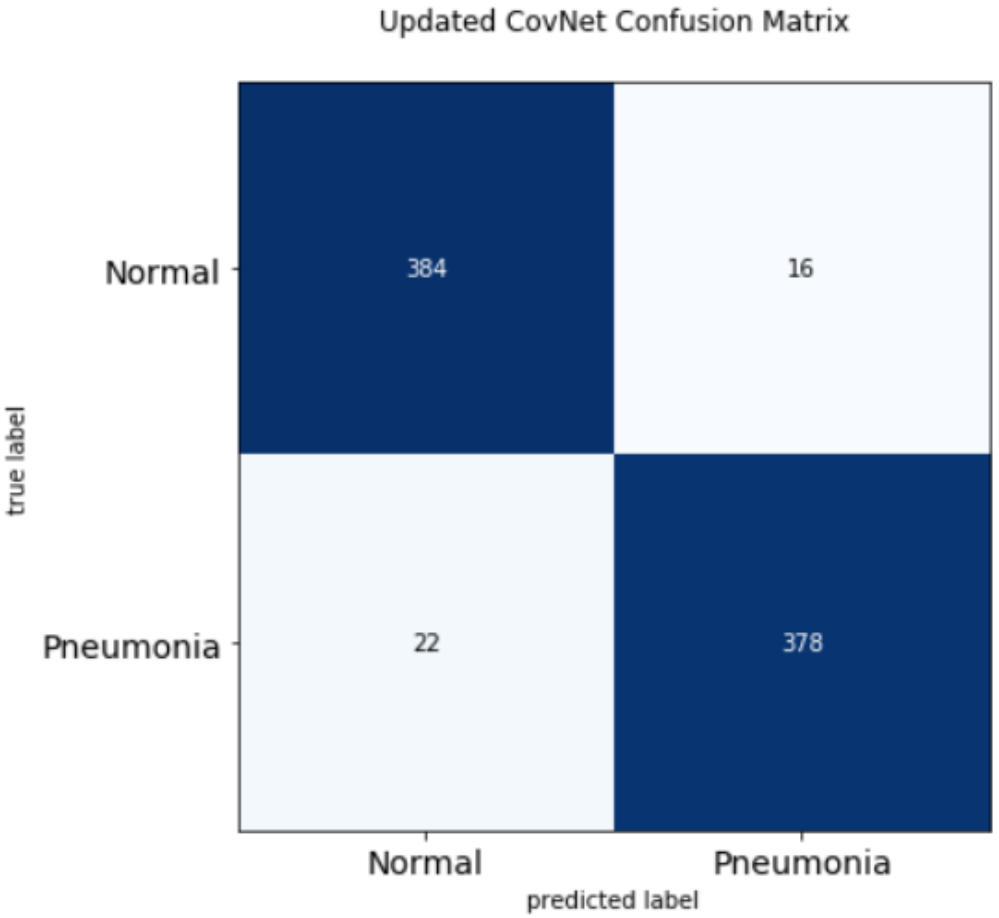
- no longer overfitting: the training curves are closely tracking the validation.
- Validation accuracy is similar to the first model.
- Validation loss is much improved at ~0.2.





Machine Learning Phase 1 - Normal vs. Pneumonia

2 - ConvNet with Data Augmentation and Dropout Layer Confusion Matrix and Classification Report



	precision	recall	f1-score	support
0	0.95	0.96	0.95	400
1	0.96	0.94	0.95	400
micro avg	0.95	0.95	0.95	800
macro avg	0.95	0.95	0.95	800
weighted avg	0.95	0.95	0.95	800

Mixed Results for a second model

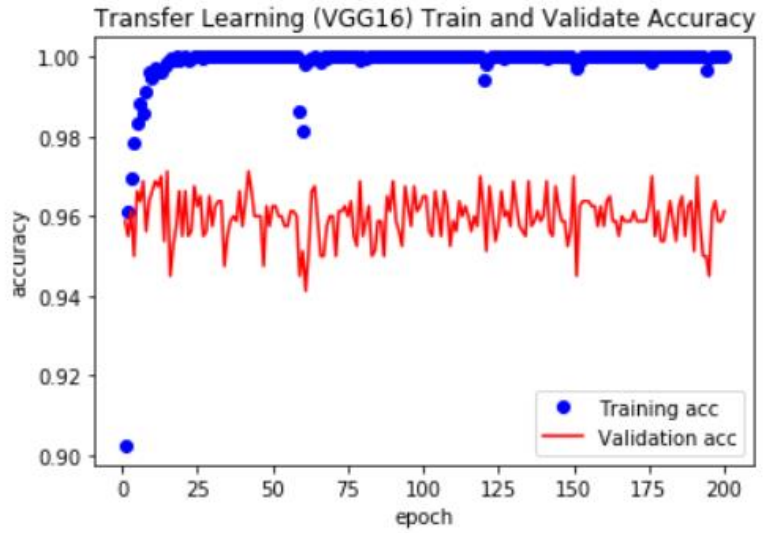
- Reduced Recall with the true pneumonia radiographs, 94.5%. ~ 4% lower than the first model
- Better Recall for normal radiographs, 96%. In this model only 4% of children without pneumonia would get diagnosed with pneumonia.

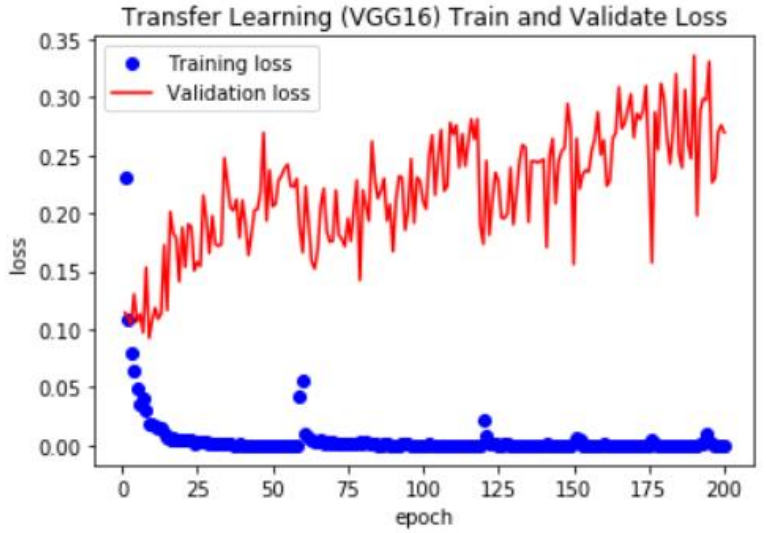
Machine Learning Phase 1 - Normal vs. Pneumonia

3 – Transfer Learning ConvNet

Very Good Model!

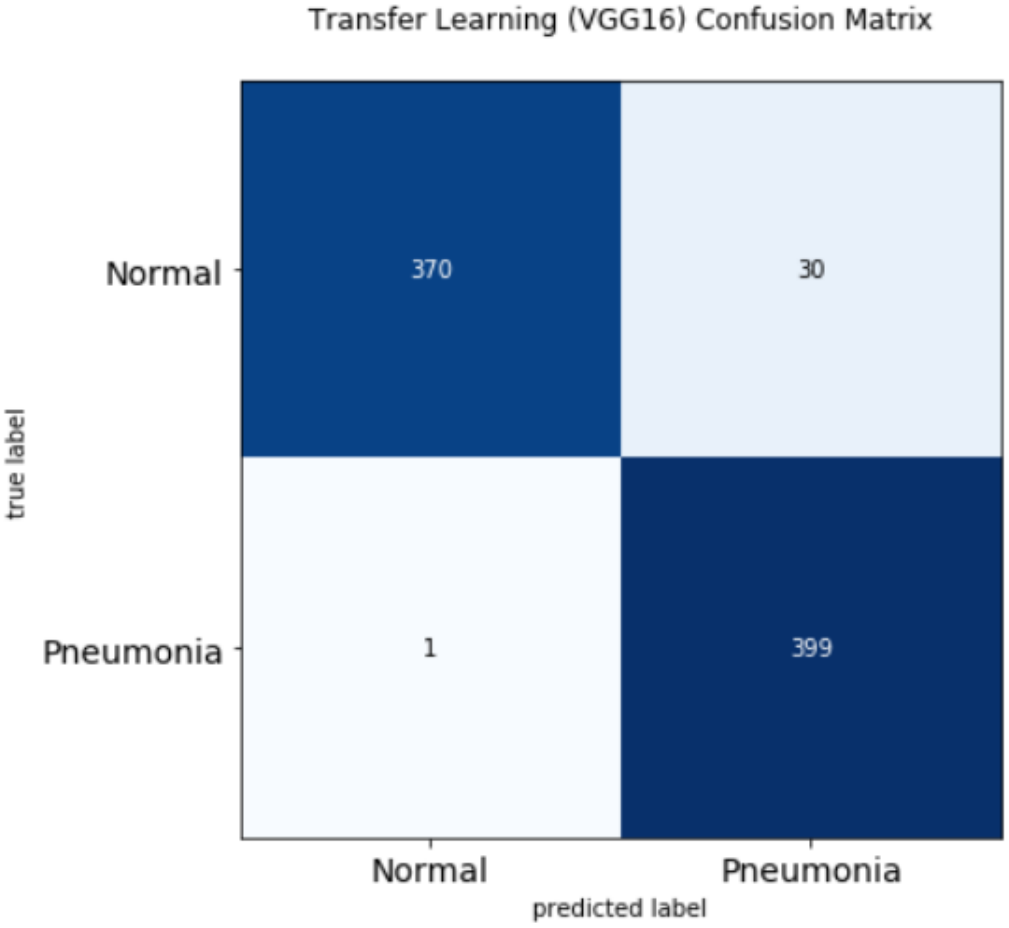
- Validation Accuracy hovering around 96%.
- Validation Loss seems to diverge from training loss but still in the excellent range < 0.30.





Machine Learning Phase 1 - Normal vs. Pneumonia

2 – Transfer Learning Confusion Matrix and Classification Report



	precision	recall	f1-score	support
0	1.00	0.93	0.96	400
1	0.93	1.00	0.96	400
micro avg	0.96	0.96	0.96	800
macro avg	0.96	0.96	0.96	800
weighted avg	0.96	0.96	0.96	800

- Excellent Recall with the true pneumonia radiographs, 99.8%.
- With this model only 0.2%, or 1 of the 400 validation cases, who truly had pneumonia would be classified as normal with this model.

- Small loss in the Recall for normal radiographs, 92.5%. 7.5% of children without pneumonia would get diagnosed with pneumonia if chest x-rays were the only available modality for diagnosis.

Machine Learning Phase 1 - Normal vs. Pneumonia

The resulting high-accuracy models suggests that this AI system has the potential to effectively learn from increasingly complicated images with a high degree of generalization using a relatively small repository of data

Machine Learning Phase 2 – Bacterial Pneumonia vs. Viral Pneumonia

Phase II:

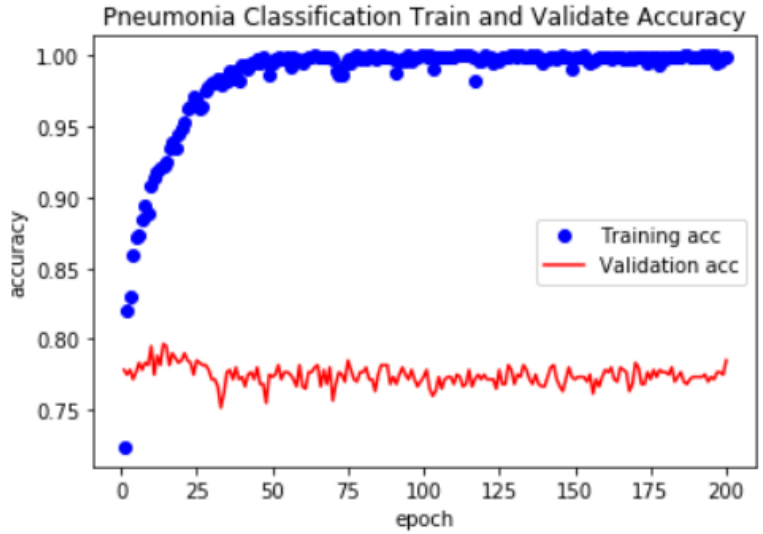
1. Because best results achieved with the VGG16 transfer learning model, we will use it again.
2. The only difference is that we will be using bacterial and viral pneumonia images, and no normal images

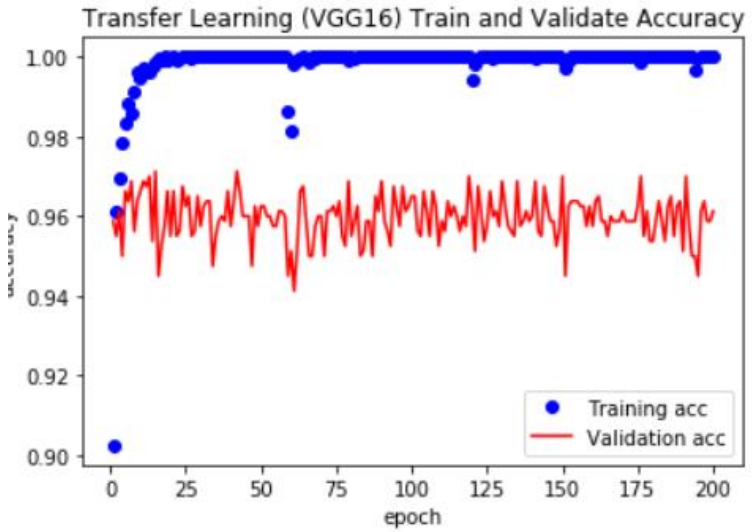
Machine Learning Phase 2 – Bacterial Pneumonia vs. Viral Pneumonia

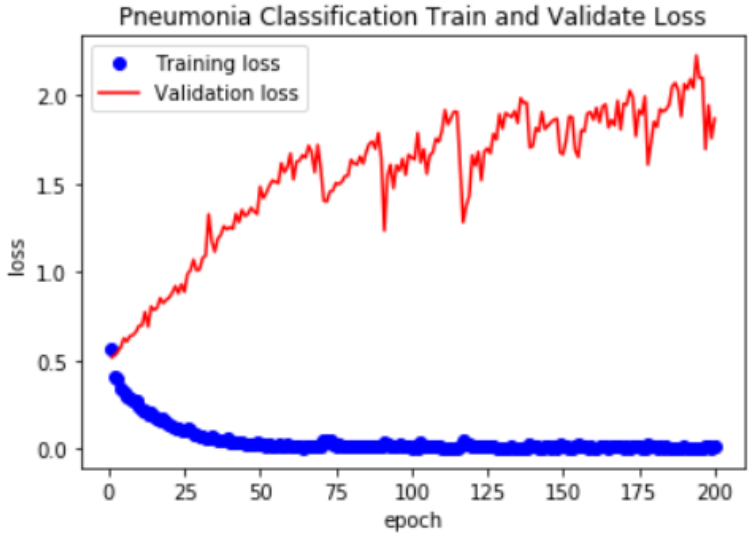
Transfer Learning ConvNet Comparison

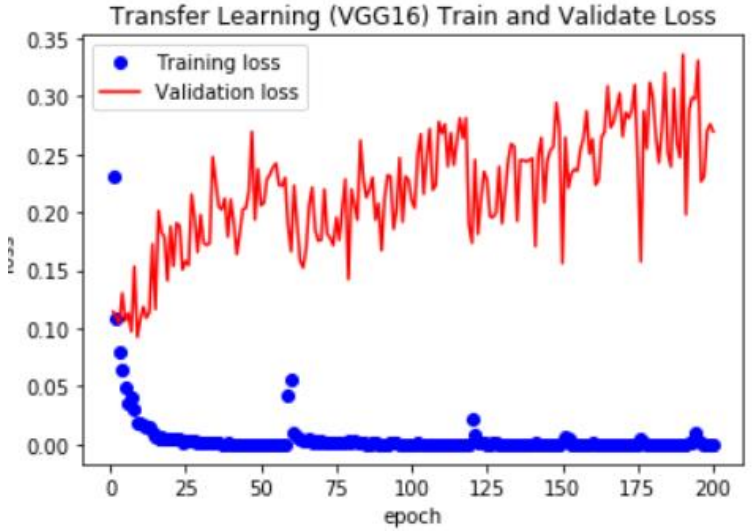
Detecting the Difference Between Bacterial and Viral Pneumonia Much More Challenging

- Accuracy slips from 96.1% in the previous model where we were comparing normal x-ray to those with pneumonia (both viral and bacterial), to 78.5%.



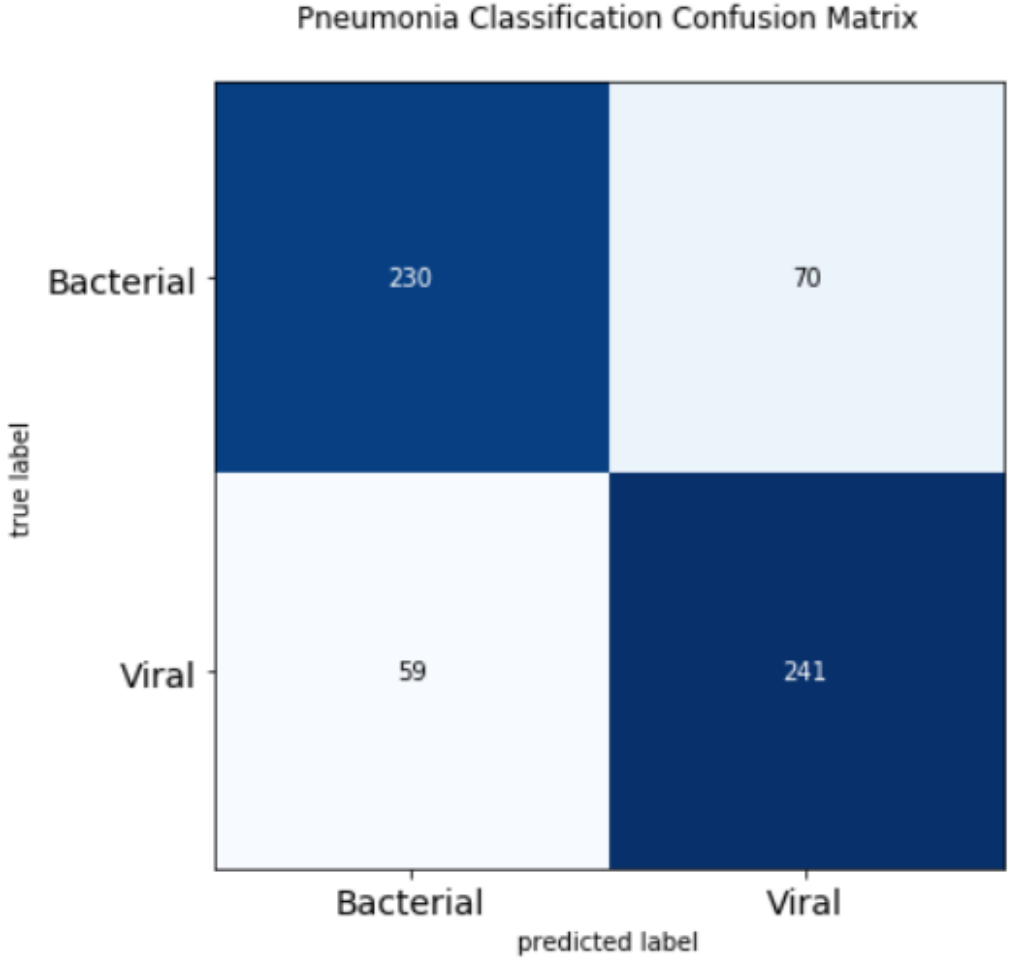






Machine Learning Phase 2 – Bacterial Pneumonia vs. Viral Pneumonia

Transfer Learning – Bacterial vs. Viral Pneumonia Confusion Matrix and Classification Report



	precision	recall	f1-score	support
0	0.80	0.77	0.78	300
1	0.77	0.80	0.79	300
micro avg	0.79	0.79	0.79	600
macro avg	0.79	0.79	0.78	600
weighted avg	0.79	0.79	0.78	600

Due to the potential consequences of misdiagnosing bacterial pneumonia, our primary measure of interest is Bacterial (0) Recall.

- Recall for both Bacterial and Viral Pneumonia does not perform as well as the model for Normal vs. Pneumonia images.

Conclusion

***The purpose of this project is to determine if an AI model can be developed to assist in the diagnosis of pneumonia using common radiographs* (chest x-rays).**

In differentiating normal radiographs from those where pneumonia was present (Phase I), all three models showed excellent results. The table below summarizes the findings of this report and compares them to the findings of the original research. "Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning"

Normal vs. Pneumonia Comparison

Model	Metric		
	Recall	Accuracy	Specificity
Original Study (TL Inception-v3 ConvNet)	93.2%	92.8%	90.1%
Simple ConvNet	98.5%	94.9%	91.3%
Data Augmentation and Dropout ConvNet	94.5%	95.3%	96.0%
Transfer Learning VGG16 ConvNet	99.8%	96.1%	92.5%

Conclusion

Phase II of the project, differentiating between bacterial and viral pneumonia, the results were not as impressive with a Recall of 76.7%.

Bacterial vs Viral Comparison

Model	Metric		
	Recall	Accuracy	Specificity
Original Study (TL Inception-v3 ConvNet)	88.6%	90.7%	90.9%
Transfer Learning VGG16 ConvNet	76.7%	78.5%	80.3%

Possible explanation for the reduced Recall compared to Phase I:

- ❑ Radiographically, the difference between bacterial and viral pneumonia is not that great.
  - Several studies have suggested that bacterial pneumonia cannot be differentiated from non-bacterial pneumonia on the basis of the chest radiograph. Virkki R, Juven T, Rikalainen H, et al Differentiation of bacterial and viral pneumonia in children Thorax 2002;57:438-441.
  - This study also found that in 30% of cases there was evidence of a mixed viral/bacterial infection.

Finally, the last sentence in the of the paper the researchers state ***"It is evident that all children with radiologically confirmed pneumonia should be treated with antibiotics because, in clinical practice, it is virtually impossible to distinguish exclusively between viral pneumonia and bacterial pneumonia."***

Recommendations for Future Research

In order to achieve better model performance when classifying viral vs. bacterial pneumonia, the following suggestions are offered:

- Increase the number of images to train the model. More images make models that predict more accurately.
- Come up with a way to systematically crop the portions of the x-ray that have nothing to do with pneumonia diagnosis. This way the model could focus on only the most important features.
- And finally, the images used in this project may have some inaccuracies. The diagnoses for the images were then graded by two expert physicians before being cleared for training the AI system. In order to account for any grading errors, the evaluation set was also checked by a third expert. X-rays were the only modality used to classify the images. In the study referenced above, "in 30% of cases there was evidence of a mixed viral/bacterial infection.". Image classification needs to be based and confirmed using multiple tests, not just x-rays.