Chest X-Ray Images for Pneumonia Documentation

1. Introduction of Data

This data set consists of 5840 x-ray images. These images include two different categorical labels, “Normal” or “Pneumonia”.

2. Problem to Solve

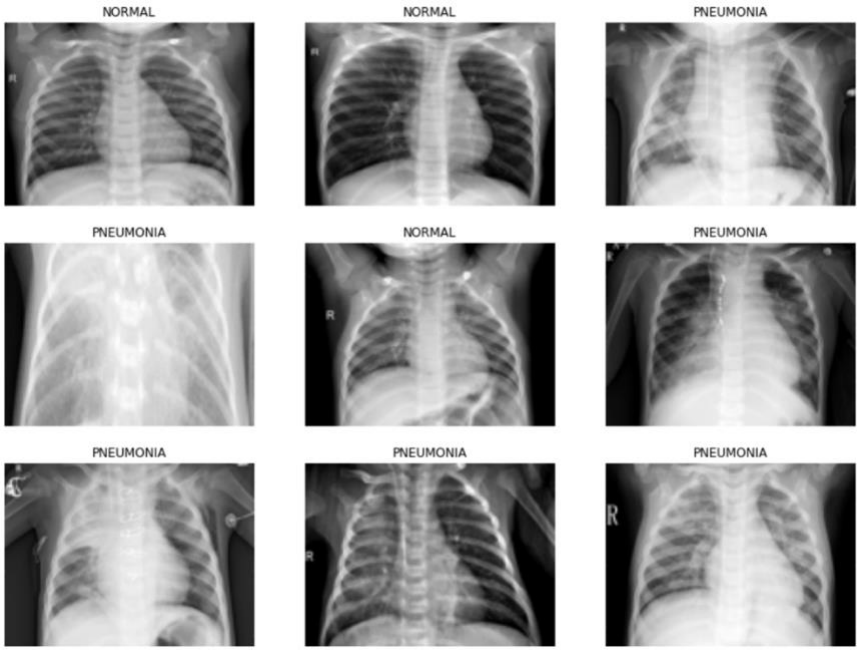
To accurately predict whether a patient has pneumonia, chest x-rays are analyzed to look for inflammation in the lungs. The model we create must have very few misclassified x-rays for the health of the patient. We also must look closely at our recall score. While add accuracy, precision, recall, and f1 scores will be evaluated, this is a medical model, so we must make sure we have as few false negatives as possible.

3. Data Cleaning

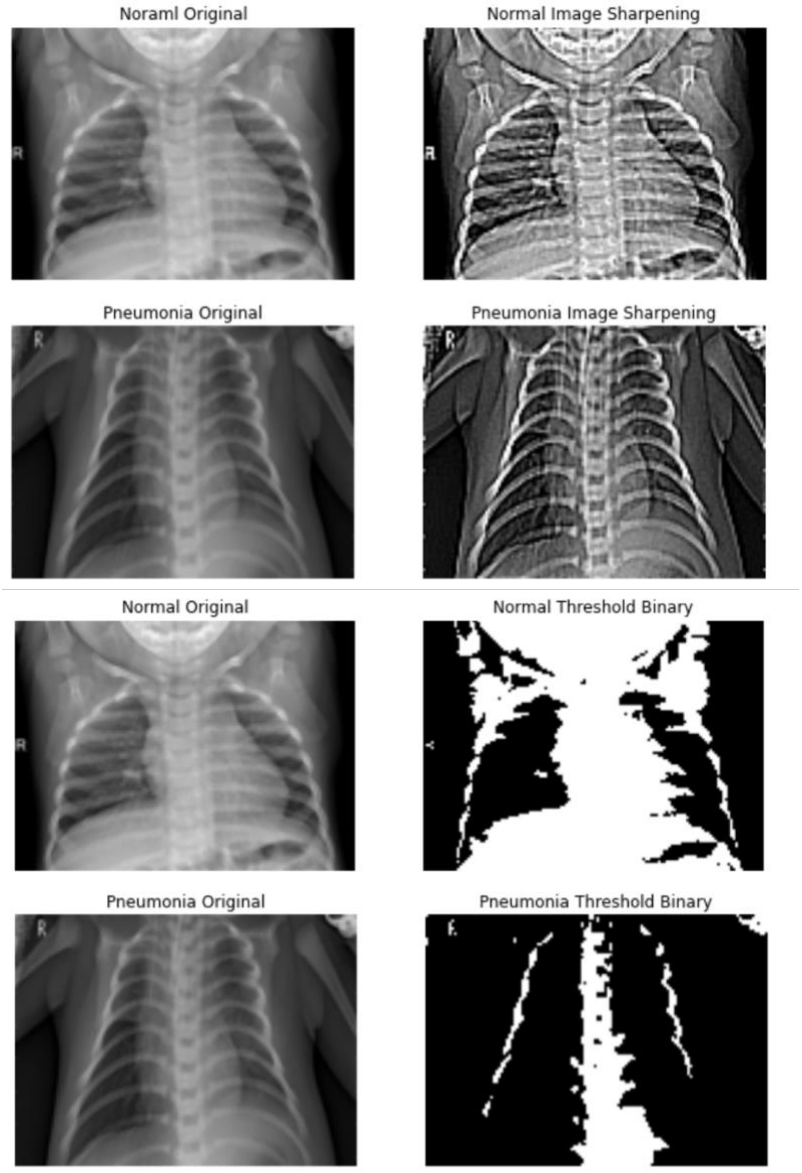
Because this data set was images, there was not much cleaning necessary. The images were all imported, as well as their labels. Then all the images were resized to 132 X 97. This was the average size of the images.

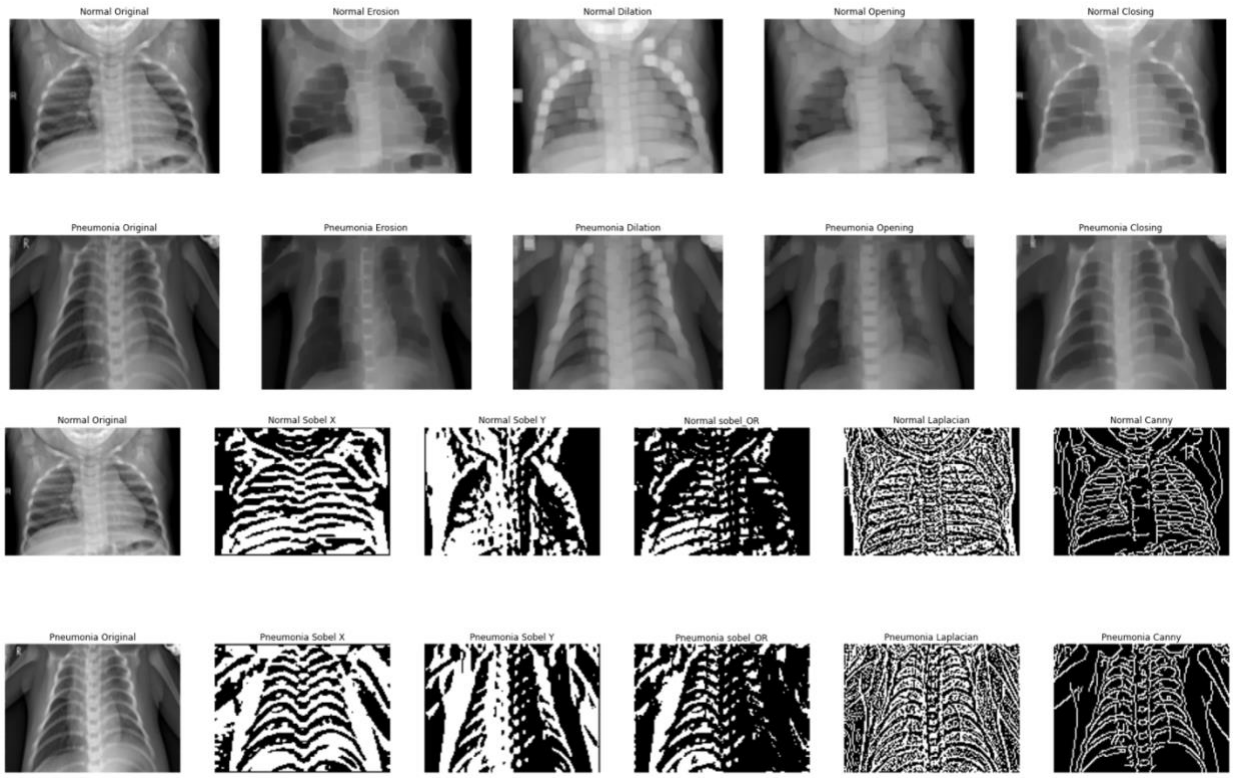
4. Exploratory Data Analysis

During the exploratory data analysis phase, I first explored the two different labels of the images. I broke down the dataset and learned that roughly 73% of the images were labeled with pneumonia. Then, I looked at the images for each label. Here were the results:



I then began to use image manipulation to see the effects on a normal and pneumonia image. Here are the results:





While there were not specific findings, we know that pneumonia patients have inflamed lungs, which may become a factor in the images during certain manipulations.

5. Modeling Approach

I used two different techniques in my modeling approach. First, I used a PCA on the image features. I then used four different Sklearn models to make predictions on the image labels. Those models were a Linear Regression, Decision Tree, Random Forest, and XGBoost. While it was a great start, none of these models' products the best results. I then transitioned to using Tensor Flow to create Convolutional Neutral Networks. I created two models that were not hyperparameter tuned to got very similar results. Then, I used the Keras Tuner to hyperparameter tune my Convolutional Neutral Network. It took some time for the tuner to run, but the best model from the tuner produced the best results overall.

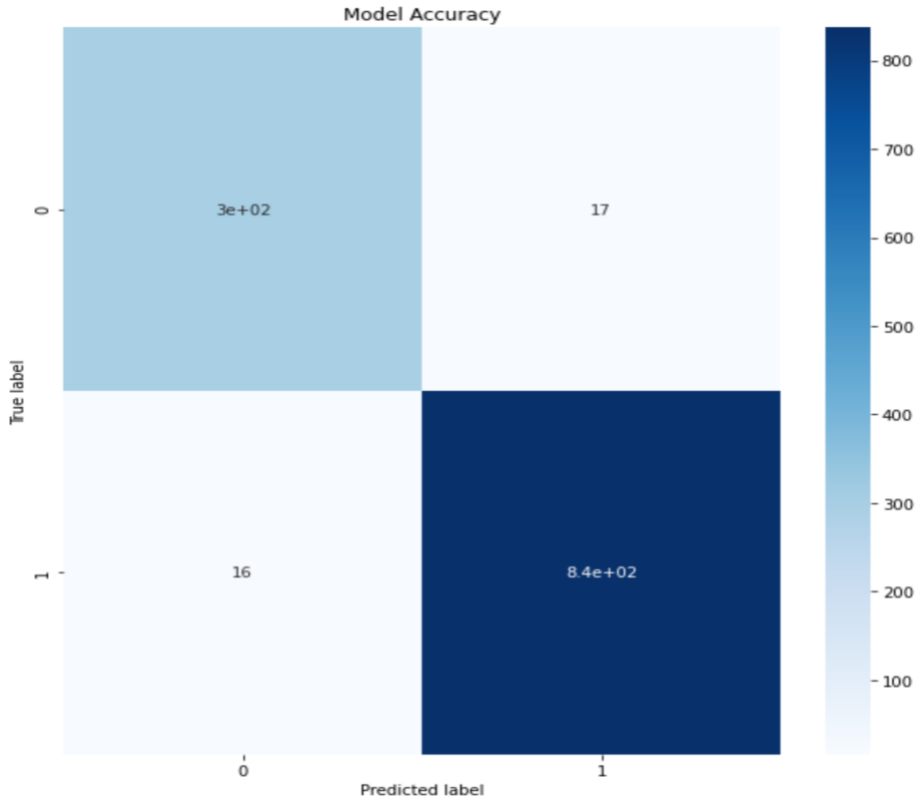
6. Findings

The results from my Sklearn models were not wonderful. It appears that approach was not the best option. The Tensorflow models performed better as a whole. Here were the final results:

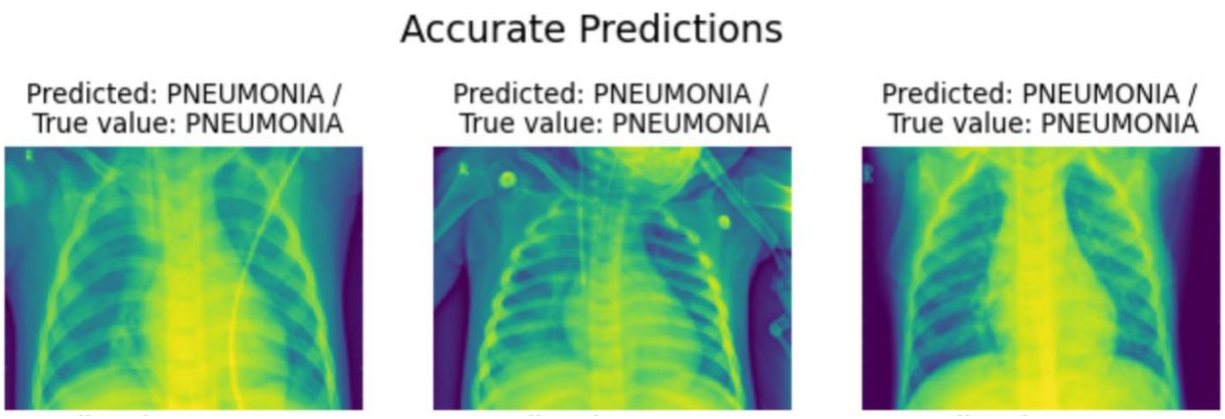
	Model	Accuracy Score	Precision Score	Recall Score	F1 Score
0	Logistic Regession	0.893836	0.969112	0.882767	0.923926
1	Decision Tree Classifier	0.720890	0.730131	0.980070	0.836837
2	Random Forest Classifier	0.730308	0.730308	1.000000	0.844137
3	XGBClassifier	0.881849	0.867420	0.989449	0.924425
4	Convolutional Neural Network 1	0.956336	0.957763	0.983587	0.970503
5	Convolutional Neural Network 2	0.954623	0.955581	0.983587	0.969382
6	Hyperparameter Tuned Convolutional Neural Network	0.971747	0.980094	0.981243	0.980668

The Hyperparameter tuned Convolutional Neural Network outperformed the other models in Accuracy, Precision, and F1 scores. Recall scores were very close and after looking at the confusion matrix comparison Hyperparameter tuned Convolutional Neural Network was still the best model. After the hyperparameter tuning, her was the best model selected:

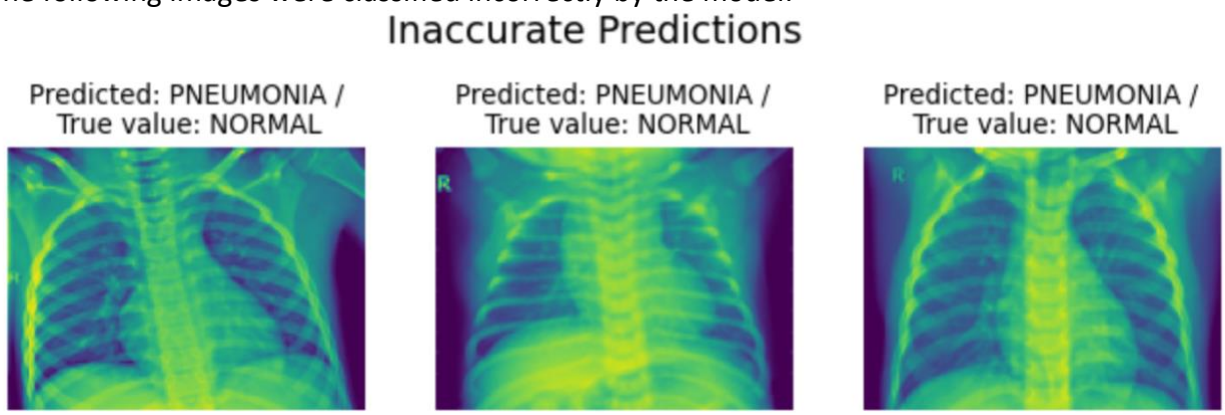
The Confusion Matrix for the prediction results are as follows:



The following images were classified correctly by the model:



The following images were classified incorrectly by the model:



7. Use of Findings

The model can be used to predict the presents of pneumonia in a patient's chest x-ray with 97% accuracy and few misclassifications.

8. Further Research

With more chest x-rays and labels this model can become more accurate over time. Also, if the misclassified images are examined based on color saturation, the model can become more accurate. The better the accuracy the more helpful the model will be come to the health care industry.