Pneumonia Detection using Convolutional Neural Network(CNN)

By: Temesgen Tesfay

September - 6th, 2021

Introduction

Pneumonia accounts for 15% of all deaths of children under 5 years old, killing 808,694 children in 2017. Being better able to predict pneumonia quickly and accurately and at a low cost could have a large impact on healthcare related to pneumonia and on patient outcomes. In this project, I seek to use transfer learning to build a model that can accurately diagnose pneumonia based on chest X-rays.

Description of the Pneumonia Dataset

The dataset is organized into 3 folders (train, test, and validation) and contains subfolders for each image category (Pneumonia/Normal). There are 5,863 X-Ray images (JPEG) and 2 categories (Pneumonia/Normal). From the data description: "Chest X-ray images (anterior-posterior) were selected from retrospective cohorts of pediatric patients of one to five years old from Guangzhou Women and Children's Medical Center, Guangzhou. All chest X-ray imaging was performed as part of a patient's routine clinical care […] For the analysis of chest x-ray images, all chest radiographs were initially screened for quality control by removing all low-quality or unreadable scans. 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."

- **Source**

  https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia

  Description

  - The data have 1.2 GB of images, and files in .jpg containing the respective labels.
  - Classes labels, image is labeled as :
    - Pneumonia

- Normal

The dataset has 5856 images, the images are classified among train, test, and validation. 87% allotted to train set, 10% to test set, and the remaining 3% is to the validation set. There is a data imbalance between the training and test.

Exploratory Data Analysis(EDA)

The dataset contains over 5000 training images and additional 624 images for testing. Each image is stored in .jpg files The Distribution of the label instances in the training set can be seen in the bar graph below:

The distribution of dataset group between normal and pneumonia classes

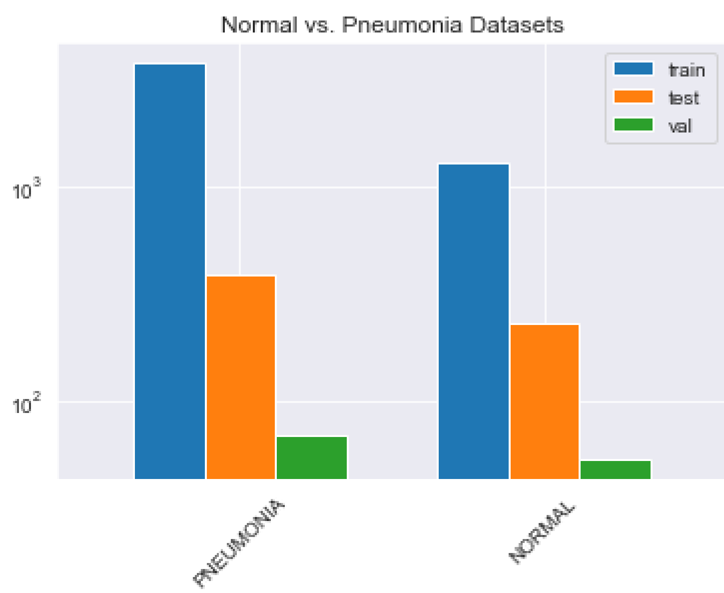

Fig 1.1 dataset proportion between normal and pneumonia

Fig 1.1 the sample size among the group has a huge difference hence, used log distribution to resize the bar graph for visualization. There is a clear indication of data imbalance.

The Distribution of Normal and Pneumonia images

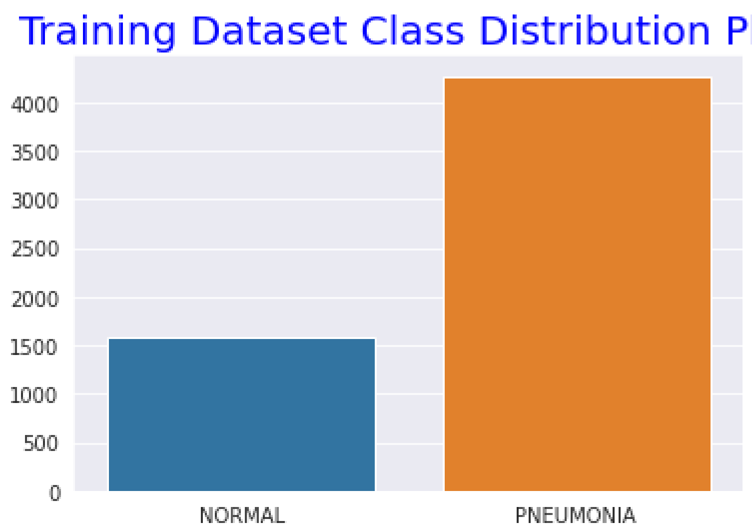

Fig 2.1 normal and pneumonia distribution

In the training dataset, we find that 75% of images show a patient with pneumonia.

A class of pneumonia versus normal lung images

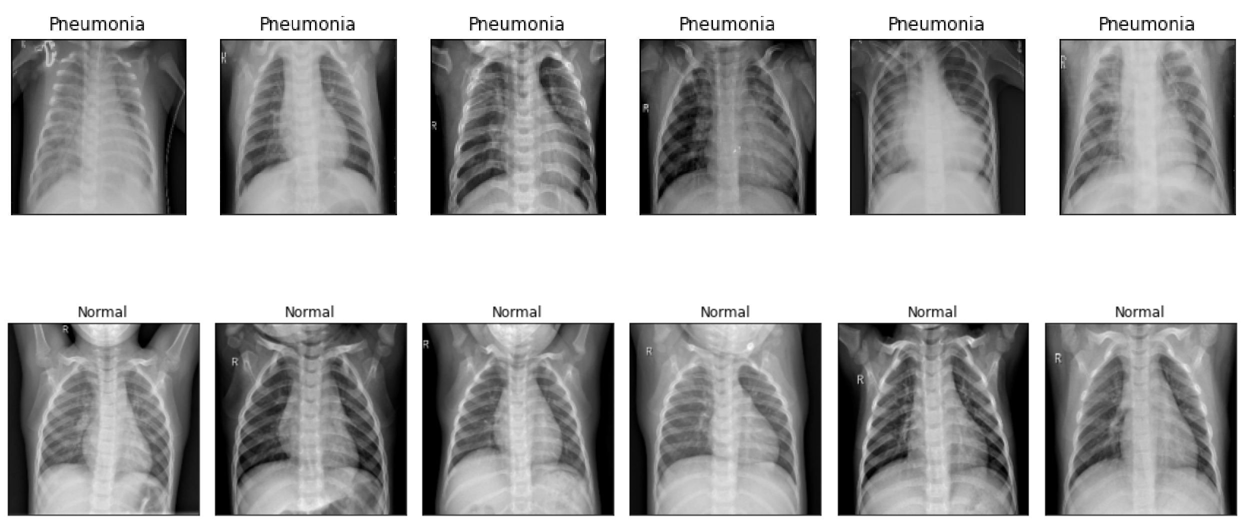

Fig 3.1 Sample images of pneumonia and normal chest-Xray image

Many Of the pneumonia-positive lungs show a wide irregular whitish area versus the normal dark brown regular pattern.

Data Augmentation

Data augmentation techniques deal with cases where the training data is limited. By applying simple techniques, images can be duplicated in ways that increase the size of the data overall but don't create exact duplicates. This can help improve model performance by improving the amount of training data and allowing a model to look at the same image in multiple different ways. For this project, I used random flip (randomly flipping the image vertically and horizontally) and random rotation (randomly rotating the image) to augment the data.

Transfer Learning Primer

Transfer learning (TL) is a research problem in machine learning (ML) that focuses on storing knowledge gained while solving one problem and applying it to a different but related problem. For example, knowledge gained while learning to recognize cars could apply when trying to recognize trucks. Practically, this is accomplished by using pre-trained models and "retraining" some layers of the model to specifically identify a new target class, while keeping the pre-trained weights frozen on other layers - thus allowing that valuable information to still be used in the new task.

Model Building

For this task, we used the ResNet50V2 model which was previously trained on a subset of ImageNet (a huge database of 14 million images manually labeled with over 22,000 categories) as part of the ImageNet Large Scale Visual Recognition Challenge. While much of the data are very different from our pneumonia data, the abstract features that are created can still be very useful for us. In this case, I froze all the layers and added a new top layer to be trained on these data and output pneumonia predictions.

The base model of CNN architecture was created using ResNet50V2 Keras applications, and its parameters are as follows:

- Image resolution input shape = (height = 220, width, 220, and 3 channels = (220,220,3)
- Training set size: 5100
- Validation set size: 124
- Test set size: 624
- Batch size = 32

The model's Training and Validation accuracies and losses plots:

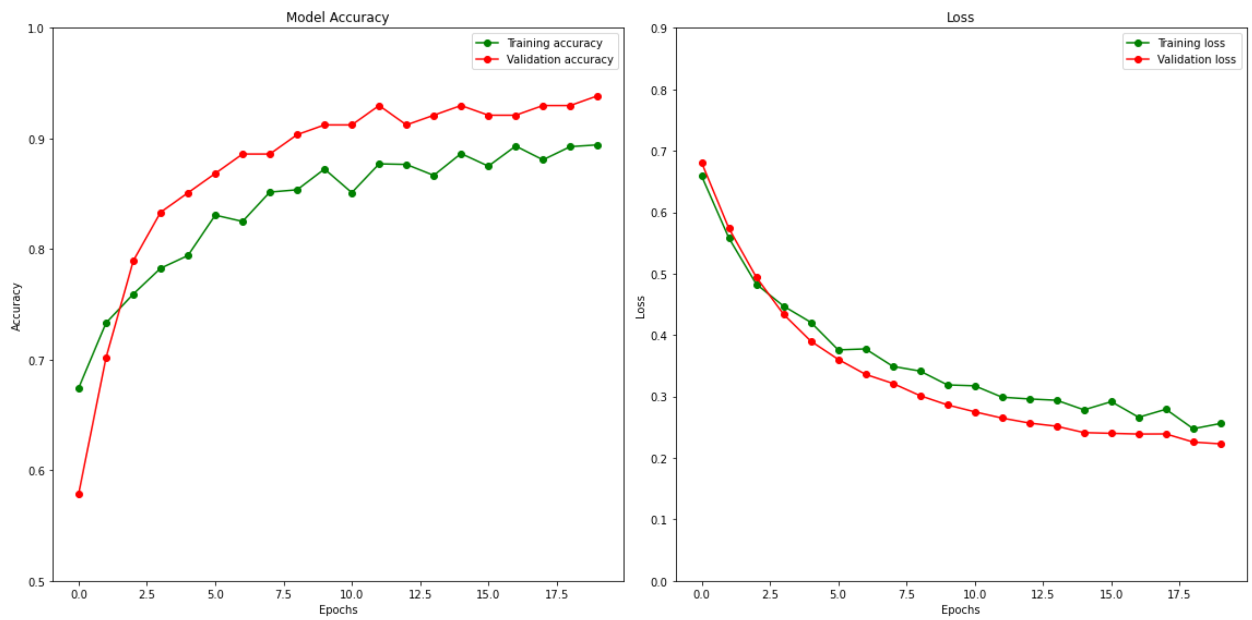

Fig 4.1 Model Accuracy vs Loss Distribution

Fig 4.1 shows that the model outcomes begin to plateau around 12 epochs and stop at 17 epochs. Strangely, we find that our model performs better on the validation data than the training data. This may be due to the size of the validation set but would require a further inquiry to determine the exact cause of this issue.

Modeling Result : ROC CURVE (Receiver Operating Characteristic Curve)

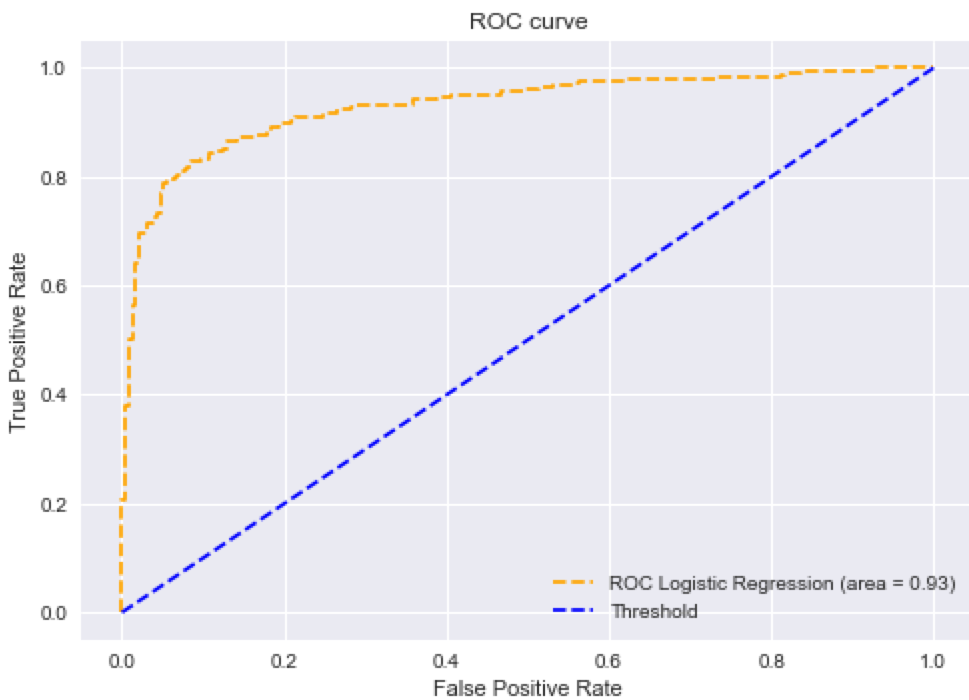

Fig 5.1 Receiver Operating Characteristic Curve (ROC Curve)

After completing modeling and optimization, the final Transfer Learning model generated predicted probabilities that resulted in the ROC curve shown above in Fig 5.1. This curve shows an AUC score of .93, which is .43 greater than random guessing which would generate an AUC score of .5.

However, the AUC score gives us an idea of how well the model performs independent of a threshold, meaning that if we knew nothing about how this model would be used we'd have some idea of how it performs generally. However, I do know how this model will be used, and it's clear that, given the deadly impact of false negatives, recall is far more important than precision here. So, the next step would be to choose the appropriate threshold for the model to optimize recall for this business case.

The focus of this project is to minimize false negatives. Therefore, the recall will be prioritized in evaluating model performance. However, the impact of the high false positives (200 out of ~624) that result from the confusion matrix. Those poor patients would be subjected to more tests, and there would be a lot more expense involved.

Thresholding

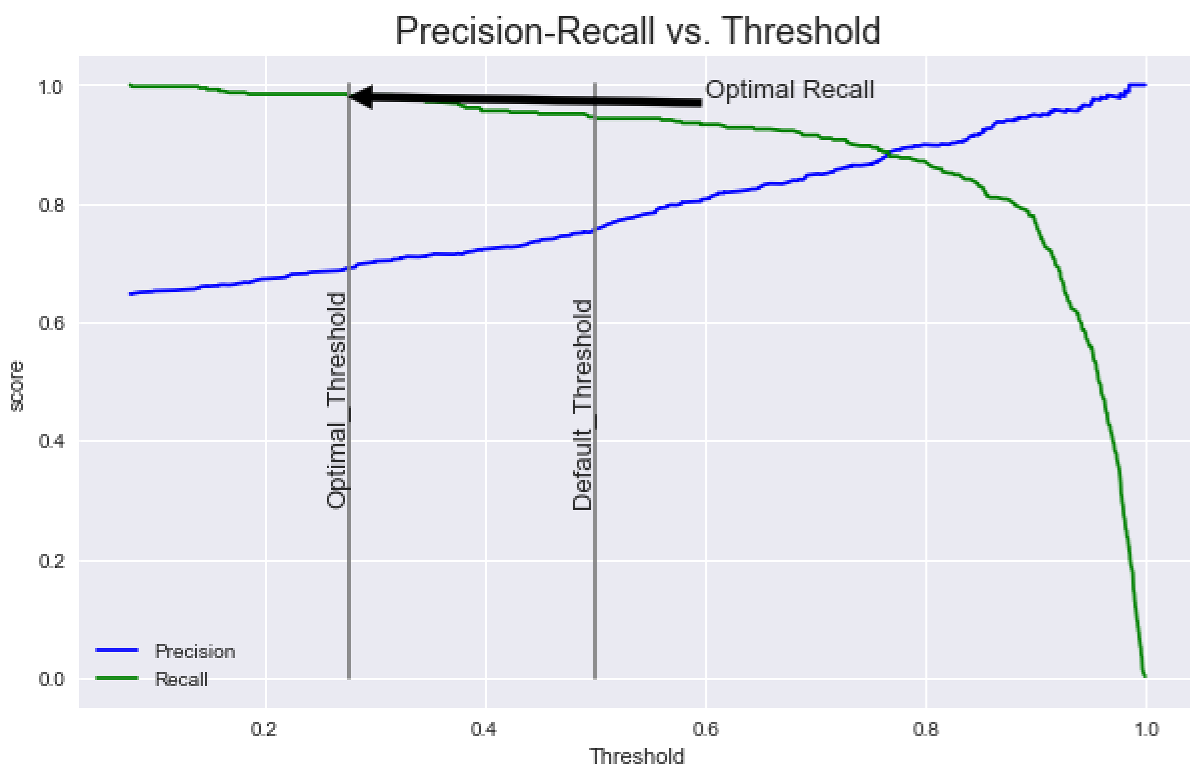

Fig 5.2 Precision and Recall by Threshold

At the default threshold of .5, as used by sklearn - the recall that our model gives is ~.78 However, the choice of threshold is based on the goal of the project and the decision on the recall and precision trade-off. I deemed that having a <98% recall would render the model unusable as the effect of that many false negatives would be too deadly for the patients. As such, to get recall > 98%, the choice was made to set the threshold to 0.25. This threshold does dramatically improve recall but is also associated with a corresponding sacrifice to precision which can be seen in Fig 7.1 and 7.2 below.

Confusion Matrix resulted from the threshold set at 0.1

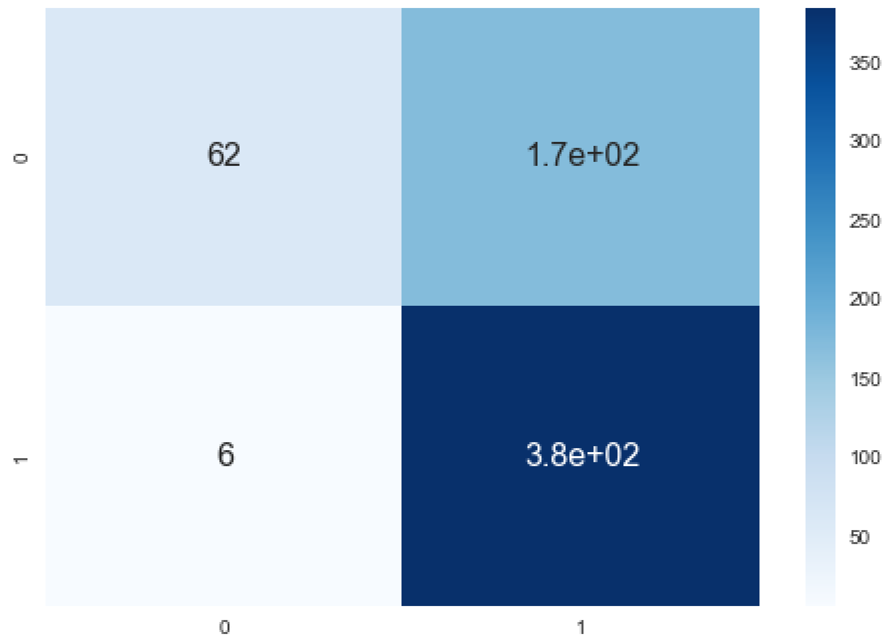

Fig 6.1 confusion matrix using the final model with a probability threshold set at 0.1

	precision	recall	f1-score	support
0	0.91	0.26	0.41	234
1	0.69	0.98	0.81	390
accuracy			0.71	624
macro avg	0.80	0.62	0.61	624
weighted avg	0.77	0.71	0.66	624

Fig 6.2 classification report using the final model with a probability threshold set at 0.1

This sacrifice to precision is not insignificant. We have 170 false positives out of ~624 patients. Those poor patients would be subjected to more tests, and there would be a lot more expense involved. As such, it would be important to speak with stakeholders who would be using this technology to ensure that having a high recall was worth this sacrifice in precision or if they would prefer to have the threshold adjusted.

Conclusion

In this project, transfer learning was used to predict pneumonia in images of chest X-Rays. The model generated was able to perform classification with a recall of ~.98 and precision of .69. It would be worthwhile to speak with stakeholders to better understand the business need and whether or not a higher recall would be worth the associated sacrifice in precision.

Insufficient data size was a significant obstacle here, though the issue was minimized using data augmentation. However, in the future having more labeled data could be very useful to improve the predictive power of the model.