CAPSTONE PROJECT

Project Title: X-Ray Pneumonia Prediction

PGA-02

Abstract

The goal of the project is to predict whether a patient is diagnosed with Pneumonia disease or not. The data that is used consists of X-Ray images that are divided into 2 folders train and test. Several deep learning models are trained using train dataset & applied to test dataset in order to evaluate the model performance. These performance measures are then compared to determine which model is best in prediction of Pneumonia disease in patients.

Submitted by - Begum Zubeda

Table of Content

Capstone Project – PGA-02

X-Ray Pneumonia Prediction

1. Introduction

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 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.

The Dataset

The dataset is organized into 2 folders (train, test) and contains subfolders for each image category (Pneumonia/Normal). There are 5,863 X-Ray images (JPEG) and 2 categories (Pneumonia/Normal).

2. Normalizing Data & Image Augmentation

The X-Ray Images are converted into pixels ranging between 0-255 of shape (height, width, color channel (RGB)) so that it can be processed by the model.

```
train_datagen = ImageDataGenerator(rescale=1./255, horizontal_flip=True, zoom_range=0.2, shear_range=0.2,
                                   fill_mode='nearest')  #Divide pixels by 255(where pixel value range
test_datagen = ImageDataGenerator(rescale=1./255)
```

The process of standardizing features present in the data in a fixed range or same scale is referred to as normalization. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. In this case, Normalization of data is done by scaling image pixels. Here, each pixel is divided by 255 pixel value such that pixel values ranges between 0 and 1, because 0-255 RGB range pixel values would be too high for our models to process (given a typical learning rate).

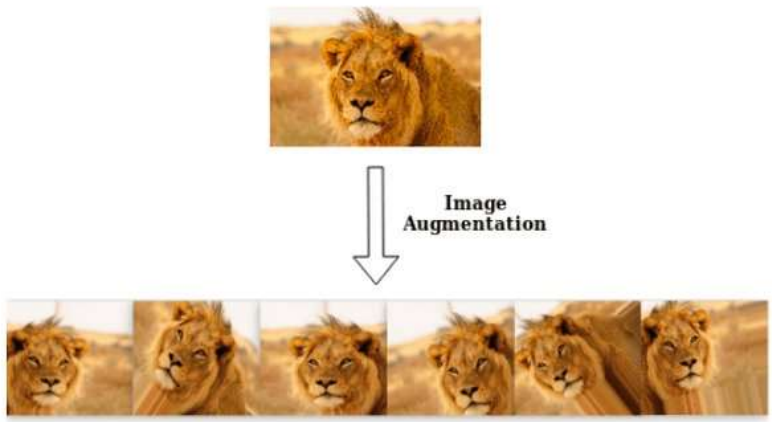
Image Augmentation is referred as a process of applying various random transformation on images, which can be applied for following reasons:

- Make the most of our few training examples.
- Helps prevent overfitting and helps the model generalize better. Overfitting happens when a model exposed to too few examples learns patterns that do not generalize to new data.

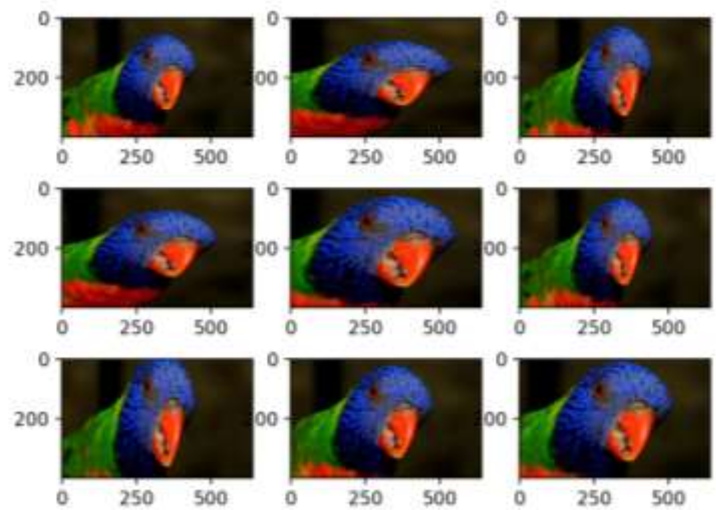
Several Image Augmentation that can be applied are:

- ‘horizontal_flip’ is for randomly flipping half of the images horizontally.

- ‘shear_range’ is for randomly applying shearing transformations.



- ‘zoom_range’ is for randomly zooming inside pictures.



- ‘fill_mode’ is the strategy used for filling in newly created pixels, which can appear after a rotation or a width/height shift.

3. Data Visualization

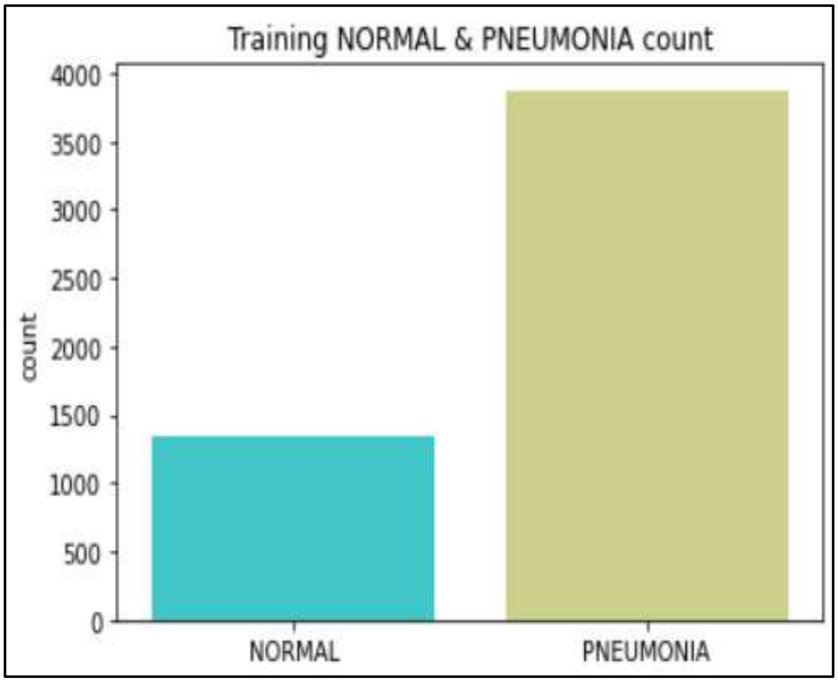


Figure 3.1

The graph on the left represents the total number of Normal and Pneumonia patients in the training dataset. We can clearly see that there are more number of Data/X-Ray Images for Pneumonia patients than Normal patients for training the model, that results in the problem of Imbalanced Data.

The graph on the right represents the total number of Normal and Pneumonia patients in the test dataset to which the trained model is going to be applied. We can clearly see that there are more number of Data/X-Ray Images for Pneumonia patients than Normal patients.

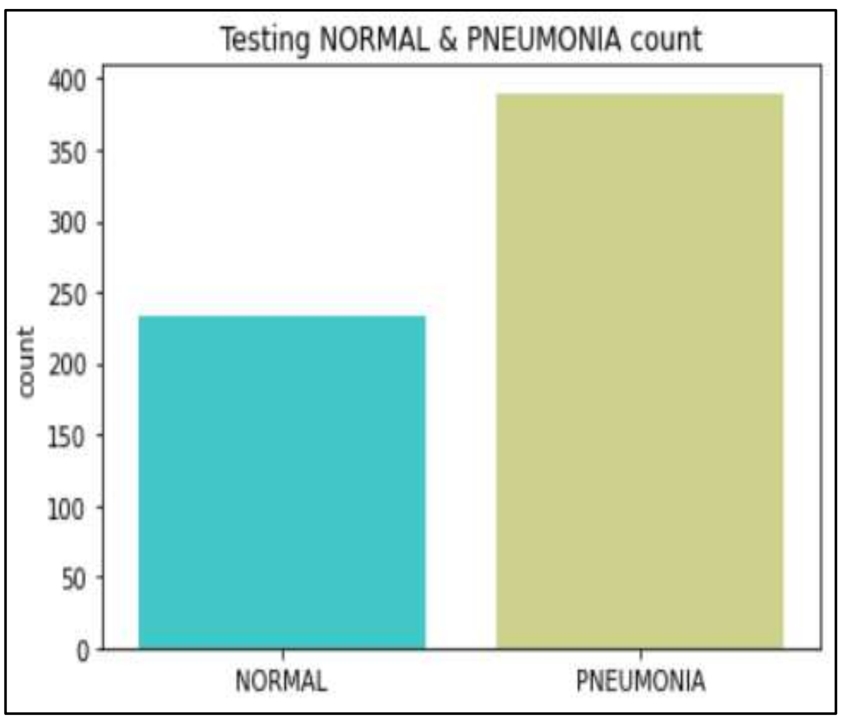


Figure 3.2

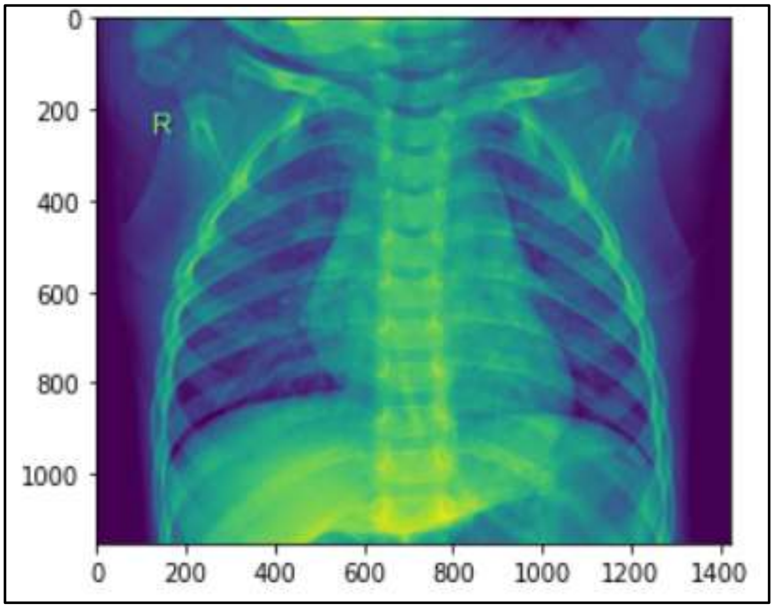


Figure 3.3

The image plot on the left represents X-ray Image for Normal patient. This normal chest X-ray depicts clear lungs without any areas of abnormal opacification in the image.

The image plot on the right represents X-ray Image for Pneumonia patient. We can see that the image manifests with a more diffuse "interstitial" pattern in both lungs.

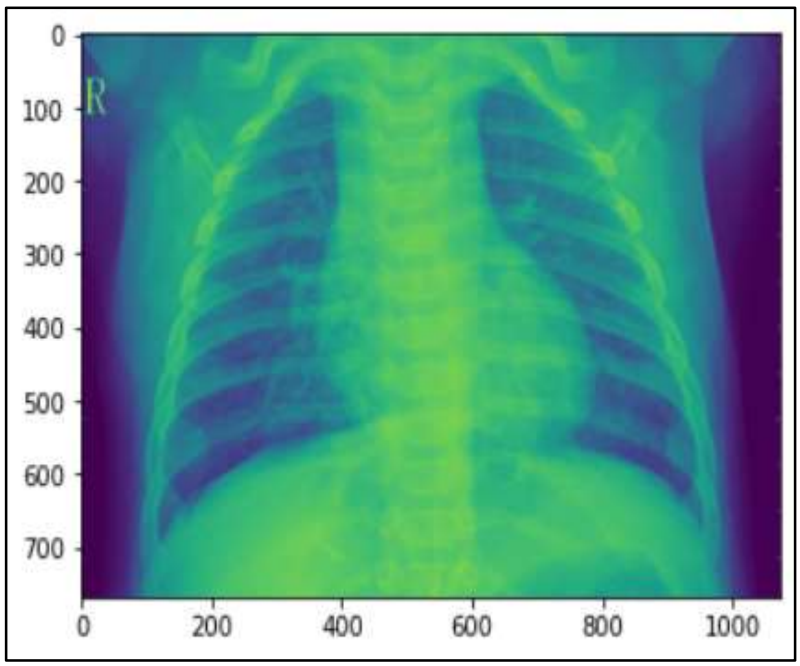


Figure 3.4

4. Handling Imbalanced Data

In a classification problem, a good model is the one that can predict or classify data well.

Imbalanced data is referred to a scenario where the dataset has unequal number of observations or data for each class or in other words difference in the number of observations or data for classes is very huge like we have seen in figure 3.1. This problem can lead to biasness for majority class and when the model is trained with such data it cannot generalize the new data well where bias refers to difference between average prediction of a model & the correct value which it is trying to predict. Therefore, when the optimization is applied while training it basically tries to reduce the misclassification error & gives more priority to the majority class data so that the total error is minimum & the accuracy is maximum, but for minority class data the performance is not so good.

In order to solve this problem we can apply various techniques like undersampling the majority class data, oversampling the minority class data and the technique we used is setting the class weights where we assign higher weights for minority class using the imbalanced ratio while training our model with train dataset, and this is the most efficient way where we can keep the data as it is without losing or adding extra data.

```
Normal:1341
Pneumonia:3875
Imbalance Ratio: 2.89

Using class_weights as:
Normal:2.89
Pneumonia:1.00
```

Figure 4.1

5. Convolutional Neural Network (CNN)