Chest X-Ray- Pneumonia Image Classification

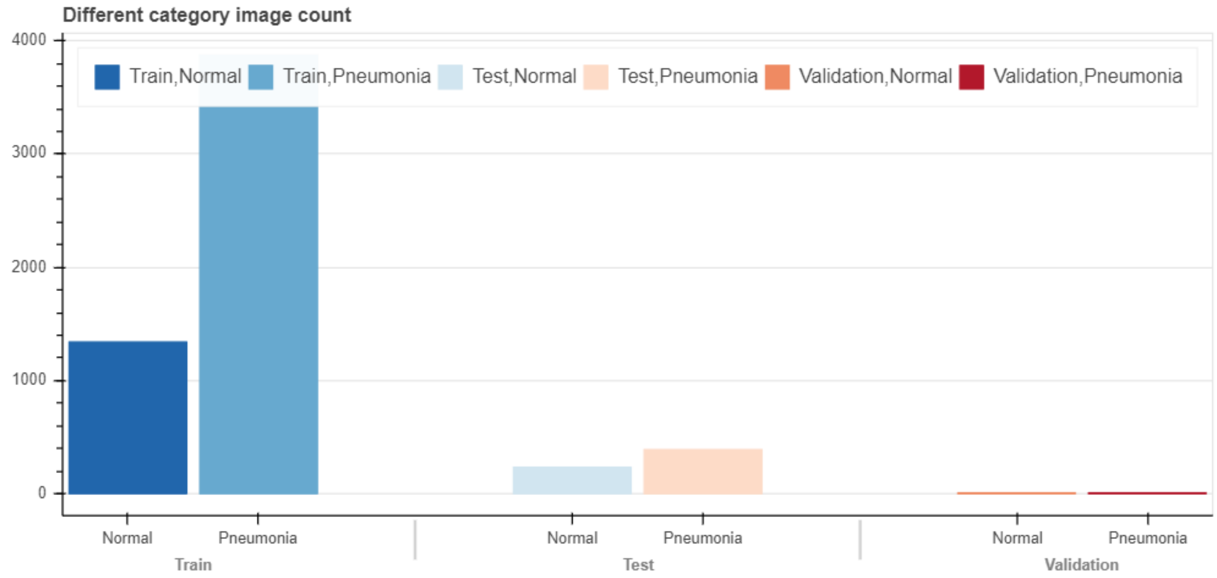
By Lahari Kuchibhotla, Stuti Sanghavi, Sanjana Ramankandath and Tanmayi Varanasi

Introduction

Pneumonia is an infection that impacts the lungs. It causes difficulty in breathing, fever, chills, etc. by filling the air sacs in the lungs with fluid or pus. Chest X-ray, blood tests and various other exams can be used to determine the presence of this infection.Fast detection and diagnosis is very essential for Pneumonia.

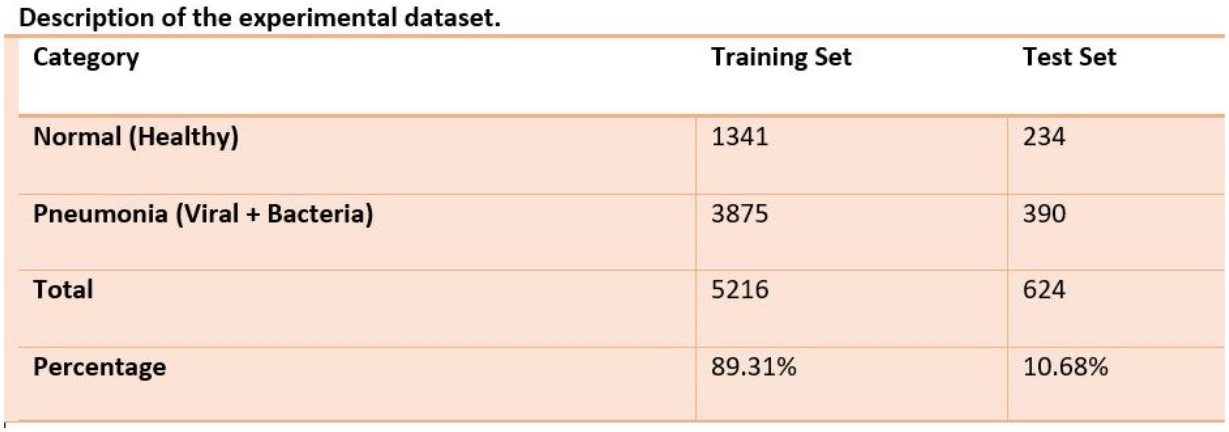
Basic Information about Dataset

The X-rays were collected from patients at the Guangzhou Women and Children's Medical Center. The X-ray images were screened and any low quality or unreadable scans were removed. The dataset is split into 3 categories: training, test and validation. And within these 3 categories, we have images for normal chest x-rays and pneumonia chest x-rays. Both bacterial and viral pneumonia were considered as a single category and together classified as pneumonia infected. The dataset used in this study did not include any case of viral and bacterial co-infection together. All chest X-ray images were taken during the routine clinical care of the patients. There are overall 5,863 X-ray images which can be categorized as either normal or pneumonia.



Link to the dataset : https://data.mendeley.com/datasets/rscbjbr9sj/2

In the training dataset we have more images labelled as Pneumonia than Normal as seen below:



Description of the experimental dataset.

Category	Training Set	Test Set
Normal (Healthy)	1341	234
Pneumonia (Viral + Bacteria)	3875	390
Total	5216	624
Percentage	89.31%	10.68%

This implies that the dataset is imbalanced since there are more X-rays for Pneumonia as compared to the Normal X-rays. We address this issue by implementing the data augmentation technique.

Goal of project

The final goal of the project is to build a model, which can accurately classify the X-ray images with a minimum accuracy of 90%.

Proposed Methodology

1. **Image Preprocessing and Augmentation:**

   Depending on the pre trained networks we use for the project, the data will have to be preprocessed and adequate resizing and normalization of the images would have to be performed. Also, for the neural networks to train properly we would require an adequate amount of data. Data augmentation will help solve this problem by utilizing existing data more efficiently. It aids in increasing the size of the existing training dataset and helps the model not to overfit this dataset.

2. **Using Convolutional Neural Networks:**

   The major advantage of using CNN is that it can detect the relevant features from the image without any human supervision. A series of such convolution and pooling operations can be performed on the input image, which can then be followed by a single or multiple fully connected layers. The output layer depends on the operations being performed. For multiclass classification, the output layer majorly used will be a softmax layer.

3. **Transfer learning using pre-trained neural networks:**

   Due to the lack of a sufficient dataset, training a deep learning model for medical diagnosis related problems is not feasible and is computationally expensive. The results achieved would also not be as expected. Hence, using pre-trained deep learning models, which were previously trained on datasets like ImageNet can be used for the model to produce better results and be cost effective at the same time.

4. **Evaluating performance metrics for Classification:**

   After the completion of training the model, we use the test data to validate our findings to see how accurate the classification is. The performance can be validated by looking at metrics like accuracy, recall, precision, F1, area under the curve (AUC) score etc.