Detecting Pneumonia from X-ray Images

Ayush Dewan
Computer Science Dept.
New Jersey Institute of Technology
Newark, NJ, USA
ad2293@njit.edu

Dejoe Varghese John
Computer Science Dept.
New Jersey Institute of Technology
Newark, NJ, USA
dj324@njit.edu

Kiran Mayee Bellam
Computer Science Dept.
New Jersey Institute of Technology
Newark, NJ, USA
kb634@njit.edu

Salma Shaik
Computer Science Dept.
New Jersey Institute of Technology
Newark, NJ, USA
ss4942@njit.edu

***Abstract*—Pneumonia is an infection of the lungs affecting primarily the small air sacs known as alveoli. The Global Burden of Disease Study 2015 found lower respiratory infec tions(LRIs) were the leading infectious cause of death and the fifth leading cause of death overall. Pneumonia caused 55% of LRI deaths in all ages (1.5 million deaths). Symptoms typically include some combination of productive or dry cough, chest pain, fever and difficulty breathing. Risk factors include cystic fibrosis, chronic obstructive pulmonary disease (COPD), asthma, diabetes, heart failure, a history of smoking, a poor ability to cough such as following a stroke and a weak immune system. Diagnosis is often based on symptoms and physical examination.**

***Index Terms*—pneumonia, chest X-ray, Discrete Wavelet Transform, Bilateral Filter, Convolutional Neural Networks**

I. Introduction

Chest X-ray, blood tests, and culture of the sputum may help confirm the diagnosis. The chest X-ray are examined by a radiologist through manual examination. The examination can be error-prone due to the complexity involved in pneumonia and their properties. One of the primary steps involved is data collection. During data collection, the dataset is trained and tested and is categorized into pneumonia and normal. The X-ray images are key to developing a classification system. In this phase, the dataset obtained will be passed through an Exploratory Data Analysis before building the classification model.

II. Proposed Methodology

A. Image Acquisition

Finding and gathering pertinent data for the research was the aim of the data collection phase, which was the first part of this investigation. Facts are essential for effectively training a machine learning model. In order to guarantee that the machine learning model produces the intended results, an enormous volume of data is required during the training phase. Numerous sources, including internet databases and hospitals, are available for data acquisition. The data used in this study came from an online platform that lets users construct models in a web-based data science environment as well as search for and exchange data sets. Two-dimensional X-ray pictures of both normal and pneumonia-infected chests were obtained from several datasets during the data collection phase.

B. Dataset Imbalance

The splits in the dataset contain some significant inconsistencies, particularly with the validation set, which contains only 16 images.

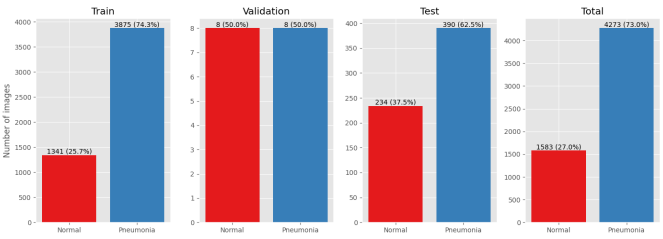

Fig. 1. Class distribution of the original dataset

We implemented a utility method to resample and resplit the whole dataset using an 80-10-10 ratio and save them into a new folder with the same structure as the original dataset. The overall class balance is preserved in each set. This is the new distribution of the dataset.

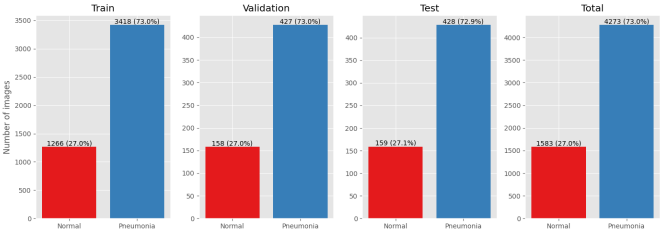

Fig. 2. Class distribution of dataset after resampling and resplitting

C. Image Augmentation

This technique is commonly used to expand the dataset size. The ImageDataGenerator from Keras lets you augment images by providing various techniques like standardization, rotation,

shifts, flips, brightness, etc. Multiple data generators with different augmentation parameters help increase training data diversity, help prevent overfitting, and improve the model's generalization ability.

We have used:

1) rescale - This scales each pixel value to the range [0,1]
2) rotation_range, width_shift_range, height_shift_range, shear_range, zoom_range, brightness_range are parameters fine-tuned to generate different augmented images.
3) fill_mode is the strategy used to fill the newly created pixels that appear after augmenting the image.

D. Model Training

We implemented ResNeXt50 architecture and trained a model with a high accuracy for detecting pneumonia.

1) The architecture begins with two standard convolution layers and is followed by max-pooling.
2) Multiple blocks of group convolution layers are interleaved with max-pooling layers.
3) Each block of grouped convolution increases the filter number and reduces the spatial dimensions of the feature maps using convolutions and max-pooling.
4) The architecture finishes with fully connected (dense) layers for classification.
5) ReLU activation functions are piecewise linear functions that output the input directly if it is positive or zero if it is negative. This is used throughout the network.
6) To fix the overfitting issue that we faced when we initially trained the model, we have utilized -
    a) L2 regularization - which adds a penalty to lose function, which encourages the smaller weights by adding the L2 norm square of the weights, which is then scaled by a regularization parameter.
    b) Dropout - randomly disables a few neurons during training to prevent overfitting.

E. Model Evaluation

After the training procedure is finished, the goal of this phase is to assess the trained model's correctness. A distinct subset of the dataset, making up 20% of the total, is put aside for testing in order to mitigate the problem of biased data.

Classification performance criteria are used to evaluate the model's performance; these criteria are especially useful when dealing with labels for categorical data. The findings in this study were classified as either normal chest X-rays or pneumonia.

Evaluation criteria like accuracy score and precision are used throughout the model testing phase. The True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) indices are used to obtain insights into the model's performance.

As performance metrics, recall, accuracy, precision, and F1-score were chosen.

Precision - Precision evaluates the fraction of correctly classified instances among those classified as positives. It is defined as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

where **TP** represents True Positives, and **FP** represents False Positives.

Recall - Recall, Sensitivity or True Positive Rate is the ratio of true positive predictions to the total number of actual positive instances in the dataset. It quantifies the model's ability to correctly identify all positive instances. It is given by:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

where **TP** represents True Positives, and **FN** represents False Negatives.

F1 Score- The F1 Score is the harmonic mean of Precision and Recall, providing a balance between them. It is given by:

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

Accuracy- The Accuracy is the harmonic mean of Precision and Recall, providing a balance between them. It is given by:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{4}$$

where **TP** represents True Positives, TN represents True Negatives, FP represents False Positives and **FN** represents False Negatives.

III. Experimental Result

A. Model Generated Parameters

The most crucial factor influencing the model's complexity, training time, and generalization ability is the total number of parameters in a neural network. The trained model has 4,488,834 parameters.

Total params: 4488834 (17.12 MB)
Trainable params: 4488834 (17.12 MB)
Non-trainable params: 0 (0.00 Byte)

B. Epoch Results

Training Accuracy - This is the proportion of training samples that are correctly classified. In this model, the increasing training accuracy means that the model has become accurate at classifying the training data. **Validation Accuracy** - This is the proportion of validation samples that are correctly classified. In this model, the increasing validation accuracy means the model performs well on new, unseen data.

Training Loss - This is the average error between the model's predictions and actual labels on the training data. Decreasing training loss shows that the model has improved predictions over several epochs.

Validation Loss - Average error between the model's predictions and actual labels on the validation dataset. Decreasing validation loss shows that the model has been generalizing the new data correctly in the validation dataset.

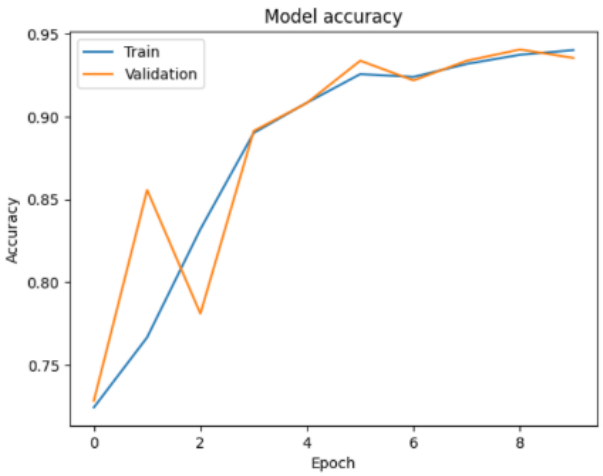

Fig. 3. Model Accuracy over Epochs graph

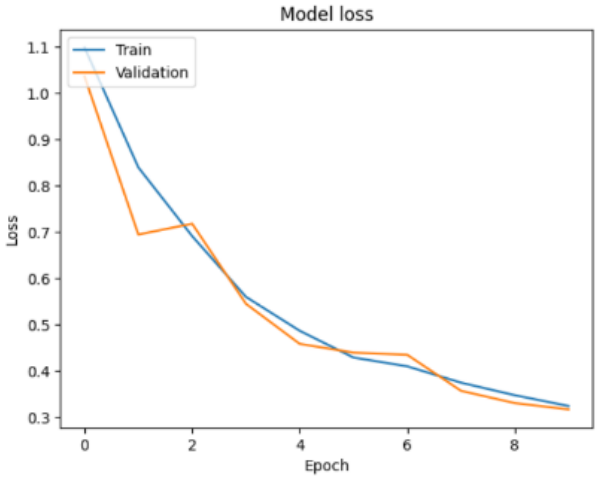

Fig. 4. Fig 2 Model Loss over Epochs graph

Both training and validation loses decrease over epochs which shows that the model is being trained and is generalizing properly.
The discrepancy between training and validation metrics in epochs 1 and 2 could indicate overfitting, but the subsequent epochs are corrected.

C. Classification Report

	precision	recall	f1-score	support
NORMAL	0.31	0.37	0.34	158
PNEUMONIA	0.75	0.70	0.72	428
accuracy			0.61	586
macro avg	0.53	0.53	0.53	586
weighted avg	0.63	0.61	0.62	586

Fig. 5. Classification Report of ResNeXt50 model

Precision: Measures the accuracy of positive predictions.

- NORMAL class = 0.31 which means of all samples predicted as NORMAL, only 31% are correct.
- PNEUMONIA class = 0.75 which means of all samples predicted as PNEUMONIA, only 75% are correct.

Recall: Measures the proportion of actual positives that were corrected identified.

- NORMAL class = 0.37 which means 37% of all the actual NORMAL samples were correctly identified
- PNEUMONIA class = 0.70 which means 70% of all the actual PNEUMONIA samples were correctly classified.

F1-score: Mean of precision and recall.

- NORMAL class = 0.34
- PNEUMONIA class = 0.72

D. Confusion Matrix

```
[[ 59  99]
[130 298]]
```

Fig. 6. Confusion Matrix of ResNeXt50 model

The confusion matrix provides a summary of correct and incorrect classifications.

For NORMAL class:

- True Negatives: 59 samples were correctly classified as NORMAL
- False Positives: 99 samples that were actually NORMAL were incorrectly classified as PNEUMONIA

For PNEUMONIA class:

- False Negatives: 130 samples that were actually PNEUMONIA were incorrectly classified as NORMAL
- True Positives: 298 samples were correctly classified as PNEUMONIA.

Conclusion

- The model is better at identifying PNEUMONIA cases compared to NORMAL cases which is shown by higher precision, recall and F1-score for the PNEUMONIA class.
- Future work would incluse investigation into misclassified samples which will provide insights into the model's weaknesses and potential areas of improvement.
- Class imbalance needs to be addressed which may improve the model's performance especially with NORMAL class cases.