Capstone Project 2: Chest X-Ray Image Classification Using Deep Learning

Milestone Report 1

Jonathon Poage
Springboard Data Science Career Track
04/30/2019

Table of Contents

Introduction **3**

Goals of Analysis **3**

Benefit to Clients **3**

Data **4**

Raw Data 4

Processed Data 4

Feature and Target Arrays 5

Analysis **6**

Image Properties 6

Summary **10**

Next Steps **10**

Reference List **11**

Introduction

Pneumonia is the leading infectious cause of death for children worldwide. According to the World Health Organization (WHO), pneumonia caused 920,136 deaths of children under the age of 5 in 2015.[1] Pneumonia affects children everywhere, but fatalities are most common in developing areas of the world. Fatality rates could potentially be reduced if patients are diagnosed and referred for proper treatment efficiently.

The two leading causes of pneumonia are bacterial and viral pathogens. Viral pneumonia is treated with supportive care, and bacterial pneumonia requires treatment with antibiotics. Low cost treatments for pneumonia are available. The issue is that fast and accurate diagnosis of pneumonia can be challenging in facilities that have limited resources.

Diagnosis of pediatric pneumonia may involve one or more medical tests. According to the National Heart, Lung, and Blood Institute (NHLBI), the best diagnostic test involves inspecting chest radiograph (X-ray) images for signs of inflammation in the lungs.[2] This project aims to develop methods to facilitate diagnosis of pneumonia from chest X-ray images.

This projects GitHub repository can be found here.

Goals of Analysis

This project aims to develop a classification model to predict the presence or absence of pneumonia in patients, based on chest X-ray images. In addition, the model will distinguish between bacterial pneumonia and viral pneumonia. The ultimate goal of this project is to provide an automated, fast, and accurate method for interpreting chest X-ray images, in order to facilitate diagnosis of pediatric pneumonia.

Benefit to Clients

Clinicians and patients could benefit from this project, particularly in developing areas of the world. Clinicians could save time by using the results of this project to automatically interpret chest X-ray images when diagnosing patients. This could potentially greatly benefit facilities with limited resources, where clinicians may not currently have the means to quickly and accurately diagnose patients.

More importantly, the results of this project could potentially improve clinical outcomes in patients with pneumonia. Clinicians could use the results of this project to efficiently diagnose bacterial and viral pneumonia. This could ultimately facilitate rapid referral for appropriate

treatment when required. As a consequence, clinical outcomes could potentially be improved worldwide.

Data

The raw data was obtained for this project by downloading the file ZhangLabData.zip from the following website: Mendeley.[3] The relevant dataset in the downloaded file was extracted for this project, and all other content was disregarded. The raw data was then processed using Python in preparation for the analysis.

Raw Data

The raw data set contains thousands of anterior-posterior chest radiograph images from Guangzhou Women and Children's Medical Center. Kermany et al[4] describe the methods they used to collect and prepare the images in the raw data set. The chest X-ray imaging was performed on pediatric patients ranging from one to five years old as part of routine clinical care. The images were labeled based on the absence or presence of pneumonia, with a distinction between viral and bacterial pneumonia.

The raw data image files had already been split into initial train and test sets. The filename of each file provides the following information:

- Class (Normal, Bacterial Pneumonia, or Viral Pneumonia)
- Patient ID
- Image number by patient

Processed Data

The raw data was processed for this project according to the following procedure:

1. The image files in the initial train and test sets were organized into final train, validation, and test sets.
    - The image files in the initial train set were separated into the final train and validation sets by using stratified random sampling. For each image class, approximately 25% of the files in the initial train set were assigned to the final validation set. The remaining files were assigned to the final train set.
    - The final test set is identical to the initial test set.

2. A Pandas DataFrame was created for the final train set. Each row in the DataFrame represents an image.
   The columns represent the following fields:
    - image_file_base_path - The basename of the filepath.

- Image_class - The class label of the image (bacterial_pneumonia, viral_pneumonia, or normal).
- Pixel_array_custom_image_size - A NumPy array of dimensions W x H x 3 with pixel intensity values in RGB color mode. The pixel dimensions W and H are each set to 224 by default for this project.

3. Step 2 was repeated for the final validation set.

4. Step 2 was repeated for the final test set.

5. The DataFrames from steps 2 - 4 were saved to pickle files in order to preserve the format of the NumPy arrays.

Table 1 shows the number of files for each class in the final train, validation, and test sets. The classes are not evenly distributed in any data set, but the imbalances are not extreme. The classes have a roughly 2:1:1 ratio in the train and validation sets, and a roughly 5:3:5 ratio in the test set.

Table 1: File counts by image class in the train, validation, and test sets. The raw data's initial train set was split into the train and validation sets for this analysis. The raw data's test set was used unaltered as the test set for this analysis.

	train	validation	test	Total
bacterial pneumonia	1904	634	242	**2780**
viral pneumonia	1009	336	148	**1493**
normal	1012	337	234	**1583**
Total	**3925**	**1307**	**624**	**5856**

The Python script used to process the raw data for this project can be found here.

Feature and Target Arrays

Feature and target arrays were prepared from data in the processed DataFrames. These arrays will be used as input for the classification models during the in-depth analysis. A separate pair of feature and target arrays was created for each of the train, validation, and test sets.

Each feature array is a NumPy array that contains the pixel intensity values in RGB color mode of every image in the designated data set. Each feature array has dimensions N x W x H x 3,

where N is the number of files in the data set, and W and H are the pixel dimensions. W and H are set to 224 by default for this project.

Each target array is a NumPy array that indicates the class for every image in the designated data set by using integer values 0 and 1. Each target array has dimensions N x 3, where N is the number of files in the data set. The rows represent image files, and the columns represent classes. A cell value of 1 indicates that the row's associated image file belongs to the column's associated class. Conversely, a cell value of 0 indicates that the row's associated image file does not belong to the column's associated class.

The Python script here contains the functions that were used for this project to load the processed data and prepare the feature and target arrays for analysis.

Analysis

The full analysis for this project will involve three parts. An analysis framework is discussed below. Note that the actual scope of the final analysis will depend on a variety of factors which will become apparent while the data is analyzed.

The first part of the analysis examined properties of the images in general. Sample images from each image class were displayed in order to visualize characteristics of typical chest X-rays for patients with bacterial pneumonia, viral pneumonia, or no pneumonia. In addition, a mean image was created for each image class by averaging over pixel intensity values. The properties of the mean images were examined to determine similarities and differences between the classes. The first part of the analysis has been completed.

The second and third parts of the analysis involve building and evaluating image classification models. See the Next Steps section for more details on the planned full analysis framework.

The Python script used to perform the first part of the analysis for this project can be found here.

Image Properties

This section describes properties of the chest X-ray images in general. Sample images are displayed in Figure 1, in order to illustrate and compare typical characteristics of chest X-ray images for each class.

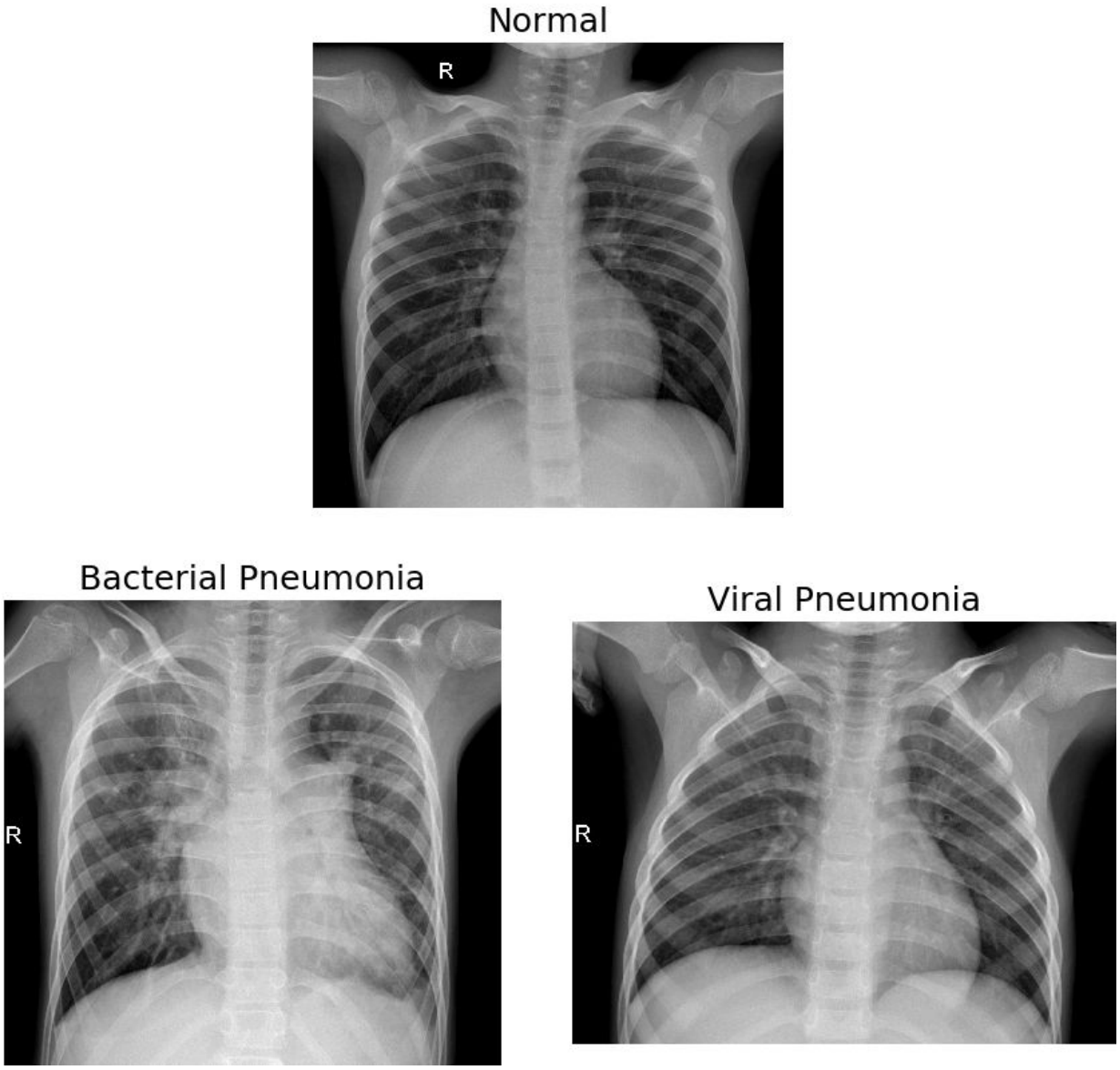

Figure 1: Sample chest X-ray images from the train set. The normal chest X-ray (top) shows clear lungs. The bacterial pneumonia (bottom left) and viral pneumonia (bottom right) chest X-rays show areas of abnormal pulmonary opacification.

The normal image in Figure 1 depicts clear lungs as a baseline for comparison. The bacterial pneumonia and viral pneumonia images exhibit abnormal pulmonary opacification. Bacterial pneumonia chest X-rays typically contain concentrated opaque areas, which indicate lobar consolidations. Viral pneumonia chest X-rays typically display more diffuse patterns of opacity. Bacterial and viral pneumonia chest X-rays may display similar patterns, and in certain cases it may be difficult to visually determine from a chest X-ray whether an infection is bacterial or viral.

The image files in the train, test, and validation sets did not all have the same pixel dimensions. The images were all resized to have identical pixel dimensions in preparation for the analysis. A mean image was then created for each image class. Each mean image was obtained by averaging the pixel intensity values over all resized images in the designated class and converting to grayscale. Figure 2 displays the mean images.

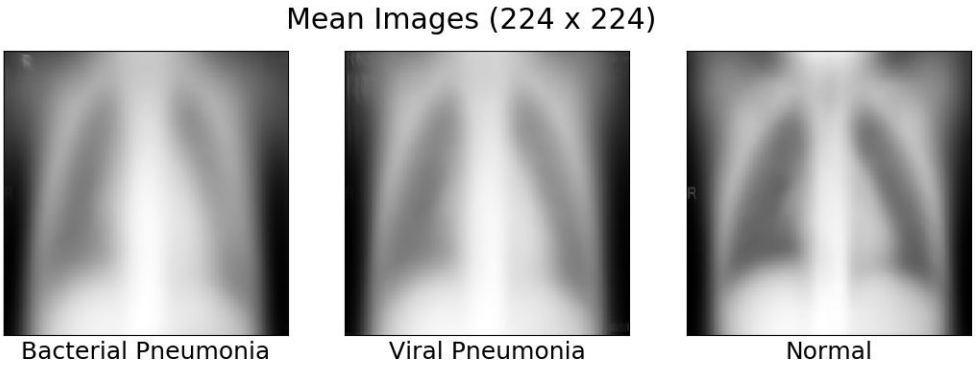

Figure 2: Mean image for each of the image classes. The image files were all resized to have pixel dimensions 224 x 224 before creating the mean images. The areas in the lungs for the normal (right) mean image are darker than the corresponding areas in the bacterial pneumonia (left) and viral pneumonia (middle) mean images.

The differences between the normal mean image and each of the pneumonia mean images are visually apparent. The areas in the lung regions of the normal mean image are darker than the corresponding areas in bacterial and viral pneumonia mean images. The differences between the bacterial pneumonia mean image and the viral pneumonia mean image are difficult to detect visually.

The mean images are quantitatively compared by using their pixel intensity distributions. Figure 3 displays the intensity histograms for the mean images.

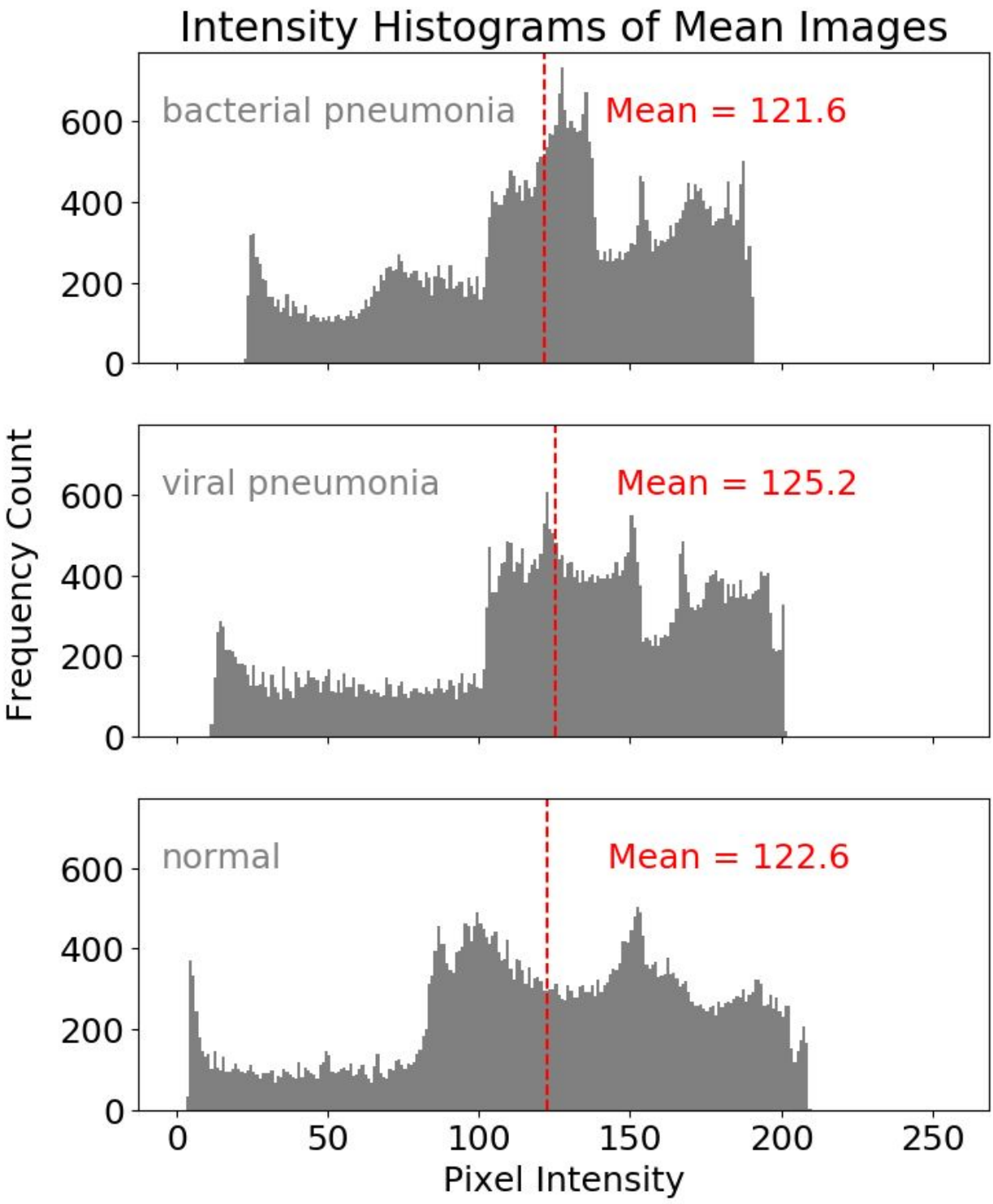

Figure 3: Intensity histograms for the mean images. Pixel intensity values indicate the brightness of the pixels, with 0 representing black and 255 representing white. The mean pixel intensity values of the mean images are indicated by dashed red lines.

Figure 3 shows similarities and differences between the pixel intensity distributions of the mean images. All three distributions are left skewed, with a bulk of the pixel intensity values in the moderate to high intensity region. This indicates that there are more light grey and grey pixels in the mean images than there are dark grey pixels. The distributions have peaks at slightly different locations, indicating that there are subtle differences in the predominant pixel intensity values for each of the mean images.

The distribution for the normal mean image has the largest range, with moderate peaks spread out over the moderate to high intensity regions. The distribution for the viral pneumonia mean image has a smaller range and stronger, more concentrated peaks than the normal mean image's distribution. The distribution for the bacterial pneumonia mean image has the smallest range and the strongest peaks in the moderate to high intensity regions. The histograms indicate that the normal mean image pixel intensities are more dispersed, ranging in color from near black to light grey. The viral and bacterial mean image pixel intensities are more concentrated towards moderate values, with a larger proportion of grey and light grey pixels compared to the normal mean image.

The mean images have similar mean pixel intensity values, ranging from 121.6 to 125.2. It might seem intuitive to expect the mean pixel intensity values to be higher for the bacterial and pneumonia mean images, considering that these images typically contain abnormal pulmonary opacities. The means are likely similar for this analysis because the patients' bodies had slightly different sizes, shapes, and orientations in the images. For example, some images may contain more area with dark background relative to others.

Overall, the image classes show many similarities, but ultimately they show suitable differences for a comparison. The remainder of the analysis will apply deep learning in order to exploit the differences between the classes and build an image classification model.

Summary

All steps for this project up through the first part of the analysis have been completed. The raw data has been obtained and processed in preparation for the full analysis. The general image properties were examined in order to evaluate typical characteristics of chest X-rays for each image class. The remainder of the analysis will be completed after the submission of this milestone report.

The next steps for the analysis are described in the following section.

Next Steps

The second part of the analysis will apply deep learning techniques to develop a set of multi-class classification models. The models will predict the probabilities that a chest X-ray

image indicates the presence of bacterial pneumonia, the presence of viral pneumonia, or the absence of pneumonia. The second part of the analysis is in progress.

The third part of the analysis will evaluate the predictive performance of the models on the test set. The predictive performances of the models will be evaluated and compared based on various scoring metrics, such as accuracy, recall, and ROC AUC. The practical significance of the best performing model will be assessed based on its ability to predict the presence or absence of pneumonia as well as its ability to distinguish between bacterial and viral pneumonia.

Time permitting, the analysis may include visualizations of occlusion maps. Occlusion maps could be overlayed over sample images to visually indicate the regions that contributed the highest importance to the deep learning algorithm.

The final report for this project will contain a summary of the results of the full analysis. The final report will describe the practical significance of the results with respect to the goals of the project, and provide actionable recommendations.

Reference List

1. Pneumonia. World Health Organization Web site. https://www.who.int/news-room/fact-sheets/detail/pneumonia. Published November 7, 2016. Accessed April 30, 2019.

2. Pneumonia. National Heart, Lung, and Blood Institute Web site. https://www.nhlbi.nih.gov/health-topics/pneumonia. Accessed April 30, 2019.

3. Kermany D, Zhang K, Goldbaum M. Large dataset of labeled optical coherence tomography (OCT) and chest X-ray images. Mendeley Web site. https://data.mendeley.com/datasets/rscbjbr9sj/3. Published June 1, 2018. Accessed April 13, 2019.

4. Kermany DS, Goldbaum M, Cai W, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*. 2018;172(5):1122-1131. doi:10.1016/j.cell.2018.02.010.