Problem

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. 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. This project aims to develop a classification model to predict the presence or absence of pneumonia in patients, based on chest radiograph (X-ray) images. The model will distinguish between bacterial pneumonia and viral pneumonia.

Clients and How They Benefit

Clinicians and patients could benefit from this project, particularly in developing areas of the world. This project aims to produce an automated, fast, and accurate method for interpreting chest X-ray images. 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 adequate time 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

Data will be obtained by downloading the file ZhangLabData.zip from the following website: Mendeley. The relevant dataset in the downloaded file will be extracted for analysis, and all other content will be disregarded.

The dataset contains thousands of anterior-posterior chest radiograph images from Guangzhou Women and Children's Medical Center. 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. Kermany et al., (2018) describe the methods used to collect and prepare the images in the dataset.

The image files have already been split into 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

Analysis Strategy

An analysis framework is proposed 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 collected and explored.

The first part of the analysis will apply machine learning techniques to develop a multi-class classification model. The model will predict the absence of pneumonia, the presence of bacterial pneumonia, or the presence of viral pneumonia based on an image. This analysis will utilize transfer learning in order to train a convolutional neural network (CNN) on a relatively small dataset. The analysis will use TensorFlow to adapt the pre-trained Inception V3 Keras model and retrain it on the training dataset. The model will take an array of image pixel values as input and produce a classification as output.

The second part of the analysis will evaluate the predictive performance of the CNN on a test dataset. The predictive performance will be evaluated based on various scoring metrics, such as accuracy, recall, and ROC AUC. The practical significance of the 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.

Deliverables

The results of this analysis will be presented in a report and a PowerPoint presentation.

References

Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C. S., Liang, H., Baxter, S. L., … Zhang, K. (2018). Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell, 172*, 1122-1131.