Team Project Proposal

Jessie Zheng, Zhou Fang, Joseph Amegatcher, Nhia Lee, Tom Flaherty

**Description of data source and web link(s).**

The data is taken from a Kaggle dataset <https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia?>

Which is originally made available by Mendeley as [**Large Dataset of Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images**](https://data.mendeley.com/datasets/rscbjbr9sj/2)

*“The dataset is organized into 3 folders (train, test, val) and contains subfolders for each image category (Pneumonia/Normal). There are 5,863 X-Ray images (JPEG) and 2 categories (Pneumonia/Normal).*

*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 patients’ 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.” --source Kaggle*

**Size/# of records of the dataset or files.**

*There are 5,863 X-Ray images.*

**# of attributes of the dataset and description of each attribute.**

Since we have images as input for our model, we do not have a set number of attributes.

**Some general statistics of the dataset.**

Training Dataset

* Normal: 1341 images
* Bacterial pneumonia: 2530 images
* Virus pneumonia: 1345 images

Validation Dataset:

* Normal: 8 images
* Bacterial pneumonia: 8 images
* Virus pneumonia: 0 images

Test Dataset:

* Normal: 234 images
* Bacterial pneumonia: 242 images
* Virus pneumonia: 148 images

**Tools/methods your team (plan to) use in your study.**

We will be using AWS’s deep learning machine image with Jupyter Notebook. As the inputs are all images, CNN would definitely be our choice of method for this study.

**Exactly what problems/questions your team plans to predict/study.**

Classify normal, bacterial and viral pneumonia with certain accuracy           