Pneumonia Detection Report

Predicting and Localizing lung opacities in chest X-ray images of patients diagnosed with Pneumonia.

Student:

Aishwarya Saad Shariff

Course:

AI and ML

# Backstory - Pneumonia study

WHO defines clinical pneumonia simply as an acute episode of cough or difficulty breathing associated with an increased respiratory rate ([1](https://www.bing.com/search?q=WHO+Programme+for+Control+of+Acute+Respiratory+Infections.+1990.+Acute+respiratory+infections+in+children%3A+case+management+in+small+hospitals+in+developing+countries.+A+manual+for+doctors+and+other+senior+health+workers.+WHO.+Geneva,+Switzerland.+74+pp.&form=PRDLR1&src=IE11TR&pc=EUPP_DCTE)). However, pneumonia is a disease of all ages, and in adult medical wards across the developing world it is one of the most common admission diagnoses. In contrast to the industrialized world, pneumonia is found characteristically in younger adults, who have a substantial inpatient mortality of 5%–23% ([2](https://www.google.ca/search?source=hp&ei=sZeSXoHYLYzl_QblzJiYCQ&q=Aetiology%2C+outcome%2C+and+risk+factors+for+mortality+among+adults+with+acute+pneumonia+in+Kenya.&oq=Aetiology%2C+outcome%2C+and+risk+factors+for+mortality+among+adults+with+acute+pneumonia+in+Kenya.&gs_lcp=CgZwc3ktYWIQA0oJCBcSBTEyLTc2SggIGBIEMTItMVDyCVjyCWCqFmgAcAB4AIABS4gBS5IBATGYAQCgAQKgAQGqAQdnd3Mtd2l6&sclient=psy-ab&ved=0ahUKEwjB9K-ZheLoAhWMct8KHWUmBpMQ4dUDCAw&uact=5)).

Pneumonia is an acute illness in which the alveolar air spaces of the lung become inflamed and filled with fluid and white blood cells, giving rise to the appearance of consolidation on the chest radiograph. It can be caused by bacterial, viral, or parasitic infection as well as by noninfectious agents. Most severe cases of pneumonia are caused by bacteria, of which the most important are *Streptococcus pneumoniae* (pneumococcus) and *Haemophilus influenzae*.

Conventional diagnosis of pneumonia consists of two stages: first, determining the *syndrome* by history, clinical examination, and chest radiology; and second, determining the *etiology* by microbiological, serological, and molecular tests. In many cases children with clinical pneumonia and an abnormal radiograph do not satisfy the specific WHO criteria for radiological pneumonia (i.e., signs of alveolar consolidation and/or pleural effusion). Research to refine pneumonia classification, using clinical signs and a **more sophisticated radiological interpretation**, is a necessary prelude to future research on the etiology of the disease ([3](https://www.ncbi.nlm.nih.gov/pubmed/18382741))

# Introduction to Project

## Problem Statement

This project aims at building an algorithm to detect a visual signal for pneumonia in medical images. Specifically, the algorithm must automatically locate lung opacities on chest radiographs.

The main objectives would be:

1. **Predict** whether a test CXR belongs to a Pneumonia patient or not
2. **Localize** the area of lung opacity via a bounding box imposed onto the image

## Need of the study/project

* As a growing need for a faster turnaround time for diagnostic test results arise, especially in the wake of the COVID-19 pandemic, it is quintessential that there be an automated system in place to precisely diagnose known conditions like Pneumonia so that the technical expertise of a medical diagnostic professional is put to use in a more effective manner.
* The insights provided by the massive amount of data already available on Pneumonia can aid in intuitive decision making for not just this disease but other related lung conditions as well.

## Business/social opportunity

* Physicians could benefit from an automated disease prediction system in order to decrease their dependency on a radiologist without having to compromise on the integrity of the results.
* Hospitals and healthcare institutions can leverage on the automated system to avoid redundancy in workforce.
* Patients can be more informed about their health history if the image itself can be self-explanatory and can thus help them to keep track of their progress as well.

# Data Report

|  |  |  |
| --- | --- | --- |
| Data – source |  | Data - collected in terms of frequency |
| The Pneumonia dataset is a subset of the original “ChestX-ray 14” dataset (an updated version of an earlier study dataset “ChestX-ray 8”) comprising of 112,120 frontal-view X-ray images of 32,717 unique patients, prepared as part of a study ([4](https://arxiv.org/abs/1705.02315)) conducted at NIH Clinical Center, Maryland, USA. |  | In total, 112,120 frontal-view X-ray images are used, of which 51,708 images contain one or more pathologies (**Pneumonia** being one among them). The remaining 60,412 images do not contain the listed 14 disease findings ([4](https://arxiv.org/abs/1705.02315)). |
| Data – collected in terms of time |  | Data – collected in terms of methodology |
| The data collected were from 32,717 unique patients whose X-ray was taken during the 1992 to 2015 timeline ([4](https://arxiv.org/abs/1705.02315)). |  | **Image**: ChestX-ray 14, X-rays images were directly extracted from a  DICOM file and resized as 1024\*1024 bitmap images without significantly losing the detail contents. Their intensity ranges were rescaled using the default window settings stored  in the DICOM header files. ([4](https://arxiv.org/abs/1705.02315))  **B-Box:** In the labeling process, 200 instances for each pathology was selected and given an image and a disease keyword, a board-certified radiologist identified only the corresponding disease  instance in the image and labeled it with a B-Box. ([4](https://arxiv.org/abs/1705.02315))  **Labeling:** Text mined 14 common disease labels from the text radiological reports via a variety of Natural Language Processing (NLP) techniques. Each radiological report would be either linked with one or more keywords or marked with ’Normal’ as the background category. ([4](https://arxiv.org/abs/1705.02315)) |

# Model Building

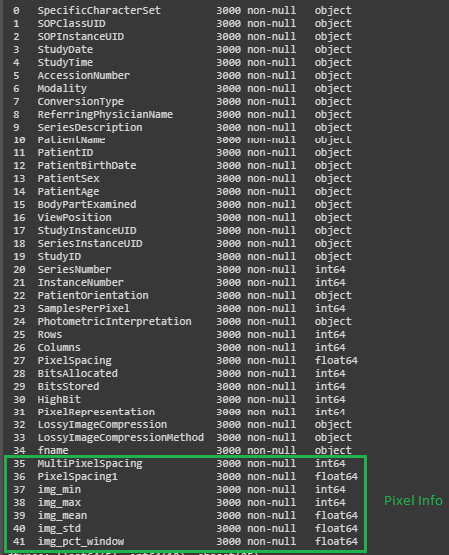
The images of Chest Radiographs were downloaded and used from [**https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/data**](https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/data)**.**

The data is structured as follows:

* **Folder 1:** Train images folder - stage\_2\_train\_images : (*26,684 images*)
* **Folder 2:** Test images folder - stage\_2\_test\_images : (*3000 images*)
  + Both the above folders have images in .dcm format
* **CSV File 1:** Images (Train) and the bounding box info - stage\_2\_train\_labels
  + {x,y,w,h} per Bbox and Target value of each image provided
    - Target values : {**1** = Pneumonia, **0** = No Pneumonia}
* **CSV File 2:** Images (Train) and their Target class - stage\_2\_detailed\_class\_info
  + Target class values :
    - { Normal, No Lung Opacity / Not Normal, Lung Opacity }

## Exploratory Data Analysis

The images of the chest radiographs being in DCM format were able to give more insights into the patient information as well as the method in which the Xray imaging was performed via the header metadata and the raw image pixel data info.



Train and Test image metadata were read and stored into separate Dataframes and the descriptive analysis for the train dataframe was performed.



The two CSV files were read and stored into 2 separate dataframes :

* CSV File 1 into ***lbls*** dataframe
* CSV File 2 into ***deta\_lbls*** dataframe

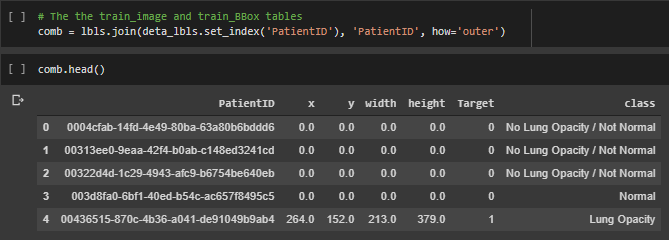
The following data analysis were done on the above two dataframes

1. The column name “patientId”was not consistent with the corresponding column name in the image metadata dataframe, so **changed that column name** to “PatientID” for both tables
2. Checked for duplicates in both dataframes. Both the dataframes had 30227 datapoints for 26684 unique patients, indicative of the fact that there were duplicate patient Ids in both dataframes. The **duplicate value of “PatientID” was dropped** from the *deta\_lbls*dataframe since all the duplicates were for that corresponding to Target class = Lung Opacity. However, for the *lbls*dataframe each row was a datapoint corresponding to a unique Bbox coordinates, so kept all the rows in it as such.
3. Checked for NaN values in both dataframes. The *lbls*dataframe had NaN values for all the Bbox coordinates whenever there was no Lung Opacity. **Treated the NaN** **values** by replacing them with 0.0 instead.
4. Checked the value count of the 3 classes within the *deta\_lbls*dataframe



There are only ~30% of the datapoints to represent the class = Lung Opacity (the class of interest to us), the rest 2/3rd of the data represents the non-Pneumonia cases, thus creating an inherent skew in the data.

1. Since both the dataframes had data of interest to us, I **combined both the two dataframes into one**, with each row within the combined dataframe representing a Bbox on an image indicating a Lung Opacity.



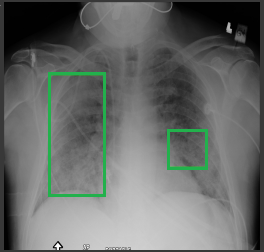
All Train and Test images were converted from DCM to PNG format, since conversion into any conventional image formats would provide more flexibility while training. The PNG files were all stored into separate Train and Test folders namely stage\_2\_train\_PNG and stage\_2\_test\_PNG

Train images were further divided into 80% Training and 20% Validation images before training them. Based on the above 80-20 split the combined dataframe was also split and the Train and Validation image details were stored into separate csv files by the names TRAIN\_CSV.csv and VALID\_CSV.csv, within the working directory.

## Modeling

The following are the logical steps through which the model has been build:

* **define a DataGenerator class that accepts a Sequence as input**
  + Train data can be generated from this class by passing in the TRAIN\_CSV.csv file as input
  + Validation data can be generated from this class by passing in the VALID\_CSV.csv file as input
    - Input (to the model) images are read from image file path provided in csv
    - Output (from the model) masks coordinates are read from the {x,y,w,h} columns of csv
      * based on the BATCH\_SIZE (pre-defined), batch\_images are created after reading each of the input images
      * based on the BATCH\_SIZE (pre-defined), batch\_masks are created after reading each of the output masks
* **define a custom Callback class to monitor the validation dice\_coefficient at the end of each Epoch while training**
  + define the method *on\_epoch\_end(self, epoch, logs=None)* to make a model prediction on every batch of images from the validation DataGenerator, and to calculate the Dice Coefficient at the end of each Epoch based on the above predictions made and their corresponding ground truths. The log of the dice coefficient is calculated and is stored as “val\_dice” for monitoring the training process
* **define create\_model method** (inspiration for this model selection is from [5](https://arxiv.org/pdf/1803.02758.pdf))
  + MobileNetV2 is used as base model for feature extraction (Encoder)
    - input image shape = (224, 224, 3)
    - weights initialized using imagenet stats
  + SkipNet architecture is used to define the decoding methodology (Decoder)
    - output mask shape = (28, 28, 1)
* **define loss method**
  + a combination of dice coefficient loss and binary cross-entropy loss is used
* **define optimizer**
  + Adam is used as the optimizer of choice
* **define ModelCheckpoint**
  + *ModelCheckpoint("model-{val\_dice:.2f}.h5", monitor="val\_dice", verbose=1, save\_best\_only=True, save\_weights\_only=True, mode="max")*
* **define EarlyStopping**
  + *EarlyStopping(monitor="val\_dice", patience=PATIENCE, mode="max")*
* **define ReduceLRonPlateau**
  + *ReduceLROnPlateau(monitor="val\_dice", factor=0.2, patience=5, min\_lr=1e-6, verbose=1, mode="max")*
* **Create, compile and fit model**
* **Evaluate model over the validation dataset**
* **Predict output for a Test image**

******

Results

The model was trained for the below mentioned settings:

ALPHA = 1.0

GRID\_SIZE = 28

IMAGE\_SIZE = 224

image\_width = 1024

image\_height = 1024

WEIGHTS = "model-0.89.h5"

EPOCHS = 4

BATCH\_SIZE = 8

PATIENCE = 15

MULTI\_PROCESSING = False

THREADS = 1

The validation accuracy achieved is **98.06%**

Recommendations

The accuracy achieved suggests that the model could be slightly overfitting

* choosing Resnet50 as the base model might help improve this

The inherent skew in the dataset (i.e. Class 1 being underrepresented) could be affecting the accuracy of the overall model

* image augmentations like zoom, lighting, vertical rotation, horizontal rotation and warping might serve helpful to offset this problem

The images being very noisy by nature could be affecting the efficiency of the training to a good extend

* most libraries aren’t flexible to work with medical imaging formats such as DCM and thus conversion to more familiar image formats such as PNG is inevitable. This leads to further reduce the quality of images. Using a more versatile library like Fastai would be well suited in such occasions

The size of the images being large (1024\*1024) and the dataset being this large causes the training time to be unreasonably long.

* Like mentioned earlier using a more versatile library could be a solution here

Refernces

**Attribute Info (Image Metadata):**

|  |  |
| --- | --- |
| **Attribute Name** | **Description** |
| *Specific Character Set* | ISO\_IR 100 for Latin alphabet No. 1 |
| *SOP Class UID* | Unique Identifier |
| *SOP Instance UID* | Unique Identifier (UI) |
| *StudyDate* | Date the Study started. |
| *StudyTime* | Time the Study started. |
| *AccessionNumber* | A RIS generated number that identifies the order  for the Study. |
| *Modality* | Type of equipment that originally acquired the  data used to create the images in this Series. |
| *ConversionType* | Describes the kind of image conversion. WSD Workstation |
| *Referring Physician's Name* | Name of the Patient's referring physician |
| *Series Description* | Description of the Series |
| *Patient's Name* | Patient's full name. |
| *Patient ID* | Primary identifier for the Patient. |
| *Patient's Birth Date* | Birth date of the Patient. |
| *Patient's Sex* | Sex of the named Patient. |
| *Patient's Age* | Age of the Patient. |
| *Body Part Examined* | Text description of the part of the body examined. |
| *View Position* | Radiographic view of the image relative to the  real-world patient orientation |
| *Study Instance UID* | Unique identifier of the Study containing the  referenced Instances. |
| *Series Instance UID* | Unique identifier of the Series containing the  referenced Instances. |
| *Study ID* | User or equipment generated Study identifier. |
| *Series Number* | A number that identifies this Series. |
| *Instance Number* | A number that identifies this image. |
| *Patient Orientation* | Patient direction of the rows and columns of the image. |
| *Samples per Pixel* | Samples per Pixel (0028,0002) is the number of  separate planes in this image. For monochrome  (gray scale) and palette color images value = 1. For RGB and other three vector color models this attribute’s value = 3. |
| *Photometric Interpretation* | The value of Photometric Interpretation (0028,0004)  specifies the intended interpretation of the image  pixel data. |
| *Rows* | Number of rows in the image. |
| *Columns* | Number of columns in the image. |
| *Pixel Spacing* | Physical distance between the centers of each  two-dimensional pixel, specified by two numeric values. |
| *Bits Allocated* | Number of bits allocated for each pixel sample. |
| *Bits Stored* | Number of bits stored for each pixel sample. |
| *High Bit* | Most significant bit for pixel sample data. |
| *Pixel Representation* | Data representation of the pixel samples. |
| *Lossy Image Compression* | Specifies whether an Image has undergone lossy  compression (at a point in its lifetime). |
| *Lossy Image Compression Method* | A label for the lossy compression method(s) that have  been applied to this image. |
| *Pixel Data* | A data stream of the pixel samples that comprise the  Image. |

1. [WHO Programme for Control of Acute Respiratory Infections. 1990. Acute respiratory infections in children: case management in small hospitals in developing countries. A manual for doctors and other senior health workers. WHO. Geneva, Switzerland. 74 pp.](#_Backstory_-_Pneumonia)
2. [Aetiology, outcome, and risk factors for mortality among adults with acute pneumonia in Kenya. Scott JA, Hall AJ, Muyodi C, Lowe B, Ross M, Chohan B, Mandaliya K, Getambu E, Gleeson F, Drobniewski F, Marsh K, Lancet. 2000 Apr 8; 355(9211):1225-30.](#_Backstory_-_Pneumonia)
3. [Scott JA, Brooks WA, Peiris JS, Holtzman D, Mulholland EK. Pneumonia research to reduce childhood mortality in the developing world. J Clin Invest. 2008 Apr;118(4):1291-300. doi: 10.1172/JCI33947. PMID: 18382741; PMCID: PMC2276784.](#_Backstory_-_Pneumonia)
4. [Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, Ronald M. Summers. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly- Supervised Classification and Localization of Common Thorax Diseases, IEEE CVPR, pp. 3462-3471,2017](#_Data_-_collected)
5. <https://arxiv.org/pdf/1803.02758.pdf>

RTSEG: REAL-TIME SEMANTIC SEGMENTATION COMPARATIVE STUDY Mennatullah Siam∗ , Mostafa Gamal∗ , Moemen Abdel-Razek∗ , Senthil Yogamani, Martin Jagersand mennatul@ualberta.ca, senthil.yogamani@valeo.com University of Alberta, Valeo Vision Systems, Cairo University