LUNG DISEASE CLASSIFICATION USING CHEST X-RAY IMAGES

Milestone report

1. Introduction

   The lungs are the centre of our respiratory system. Our body requires Oxygen to survive, and the blood is filled with oxygen in our lungs. So lungs are one of the most critical organs in our body. Malfunctioning of the lungs can seriously impair the respiratory system, and puts the life of the patient at risk. The major types of lung diseases are:

   Lung airway disease: Asthma, COPD, etc.
   Lung tissue diseases: Pulmonary Fibrosis and Sarcoidosis
   Lung circulation diseases: Pulmonary Hypertension

2. Current scenario:

   CXRs are conventionally analysed by medical experts and doctors, to diagnose the patient. Though human evaluation is very accurate, in times of emergency or while dealing with a large number of cases, there is extreme pressure on the limited number of medical staff to sieve through the huge number of images. Every case has to be carefully observed and analysed by doctors for correct diagnosis. This is a time consuming process. Moreover, human analysis is prone to errors. We need a model that can quickly filter the high criticality cases from the moderate ones, and can give doctors the opportunity to focus on critical cases.

3. Problem statement:

   We are given chest x-ray (CXR) images of patients (front view of the patient's chest). The images have been labelled as belonging to a normal patient or a diseased patient. In my project, I have collected CXR images belonging to 5 types of diseases. The task is to build a deep learning model to classify with an acceptable level of accuracy, whether the patient has any of the 5 diseases, or he/she is healthy.

4. Value to client

   Conventionally, the client in this type of project would be hospitals, clinics, radiology staff, doctors, etc. But considering the current crisis, this model has the potential to go beyond commercial purposes and help us in fighting the COVID pandemic. Deploying a model to undertake mundane tasks like filtering out obvious cases (where the patients are clearly healthy or clearly diseased), separating the critical cases from the mild ones, etc. can be carried out by a machine. This will reduce the stress on doctors, and they will have the opportunity to focus on demanding cases which need critical attention.

5. The Stakeholders

When such a project is implemented, the medical practitioners and hospitals are the stakeholders. The diagnosis and treatment of patients will depend on how well the model performs. This will impact the performance and the reputation of the medical institution. But the patients are at the highest risk here. The accuracy of the model will directly and severely affect the life and general well being of patients. If an unhealthy person is incorrectly classified as healthy, that patient will not receive treatment, leading to further complications, or death in extreme cases. This will be a disaster.

6. Data Sources and Data organization:

I have compiled the dataset from various Kaggle datasets as well as academic datasets. As medical data is protected information, it is a little difficult to find good quality CXR images with proper labelling. The same images are repeated across many datasets. I have sieved through multiple datasets and created a set of total 11276 images, with:

2530 CXR images Bacterial Pneumonia
288 CXR images COVID-19
1122 CXR images which are Normal (no disease)
5597 CXR images of Other Findings
394 CXR images of TB
1345 CXR images of Viral Pneumonia

As labelled CXR image data was not easily available, I am trying to use all images for training. As you can see, there is class imbalance in the dataset.

Later, I will create a smaller but more balanced subset of this dataset, having about 300-500 images for each category. The total number of images will be reduced to about 3000. Training will be faster, but we may overfit sooner. I will use the same training models and parameters as used for the original dataset, and compare the accuracy for both.

I have used the following datasets to compile mine:

For Normal and Pneumonia CXRs (A very popular dataset on Kaggle by Paul Mooney) - https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia -

Chest X-ray images are 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. There are 5,863 X-Ray images (JPEG) and 2 categories (Pneumonia/Normal), and pneumonia images can be further classified into Viral or Bacterial Pneumonia.

Original dataset - https://data.mendeley.com/datasets/rscbjbr9sj/2
Citation - http://www.cell.com/cell/fulltext/S0092-8674(18)30154-5

For COVID-19 CXRs -

https://github.com/agchung/Figure1-COVID-chestxray-dataset
https://www.kaggle.com/tawsifurrahman/covid19-radiography-database
https://www.kaggle.com/nabeelsajid917/covid-19-x-ray-10000-images

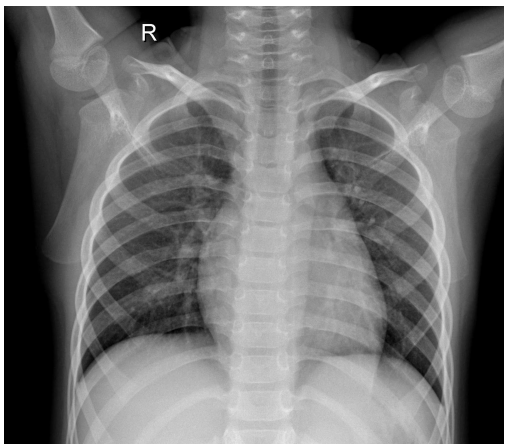
For TB CXRs -

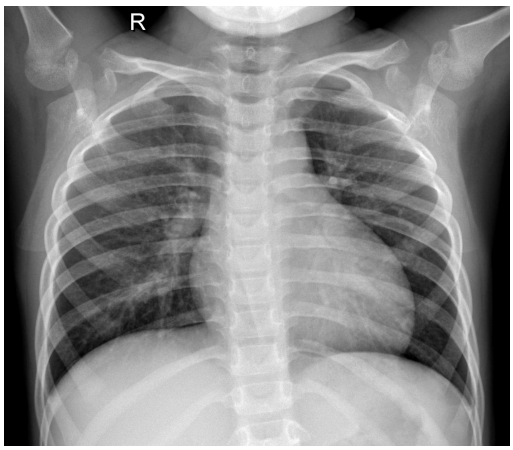
https://www.kaggle.com/kmader/pulmonary-chest-xray-abnormalities

For CXRs of numerous other lung diseases -
https://www.kaggle.com/nih-chest-xrays/sample (It is a randomly selected subset of the full NIH Chest X-Ray dataset (42 GB))

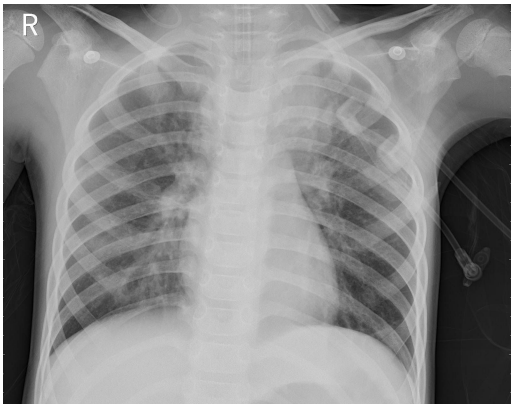
7. Image samples:

NORMAL:

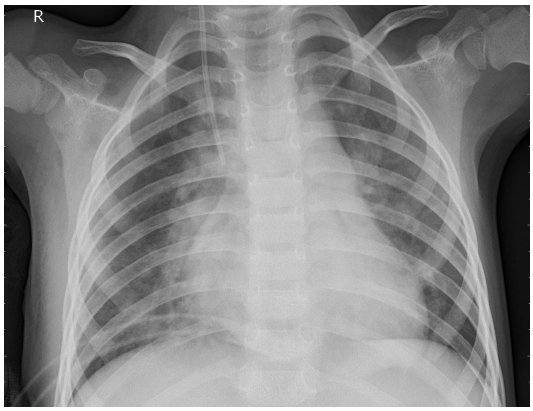


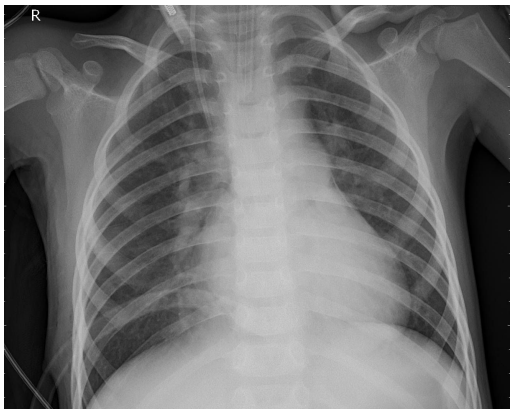


VIRAL PNEUMONIA:



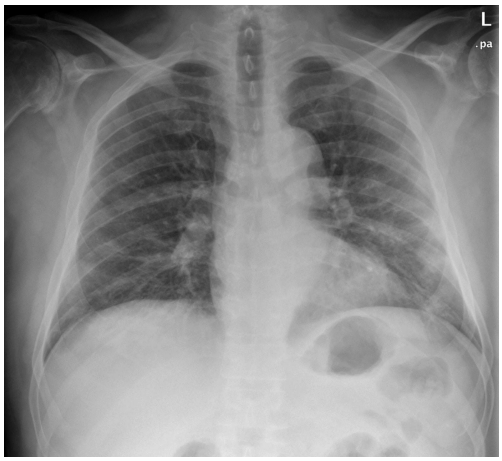
BACTERIAL PNEUMONIA:

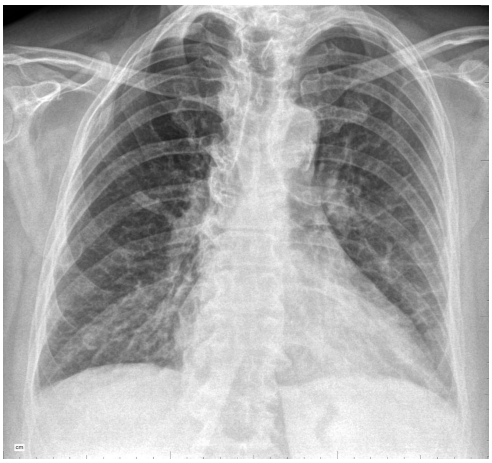




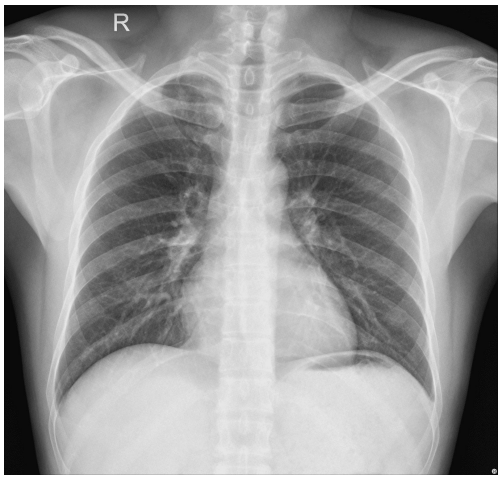
Taking a look at the images with naked eye shows that bacterial pneumonia typically exhibits a focal consolidation, .i.e. a region of aggregated “material”, whereas viral pneumonia (right) manifests with a more diffuse ‘‘interstitial’’ pattern in both lungs.

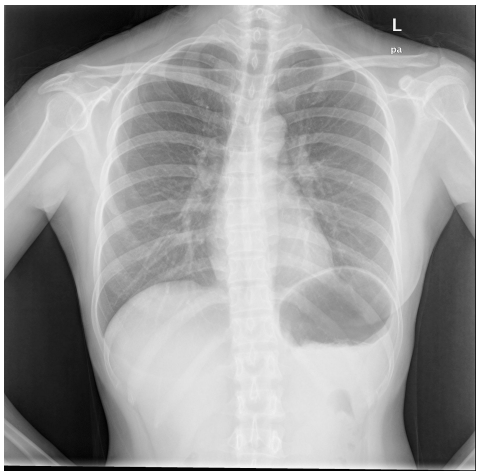
COVID-19:



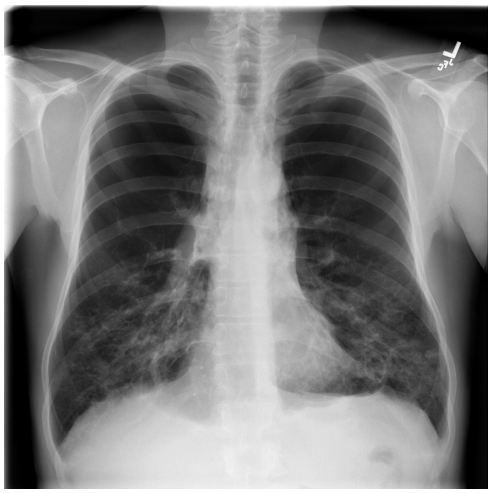


TUBERCULOSIS:





OTHER FINDINGS:



8. Preliminary Observations with naked eye

   We do not possess radiology knowledge to analyse these images and correctly diagnose the disease. But it is visible that ‘Normal’ CXRs are more “clear”, without any consolidated material or fluids obstructing the path of X-Rays. Diseases CXRs have more whitish or grayish regions spread throughout the lung area.
   For example, bacterial pneumonia typically exhibits a focal consolidation, whereas viral pneumonia manifests with a more diffuse ‘‘interstitial’’ pattern in both lungs.
   Likewise, CXRs of COVID-19 also show opacity inside the lung area, representing development of or aggregation of fluid-like substances.

9. Basics of Chest X-Ray Radiology:

   Regions of different opacity in the CXR images can be broadly divided into the following categories:

   a. Black: These areas offer no resistance to X-rays - mostly air and empty space
   b. Dark Grey (very slightly opaque): fat and subcutaneous tissue

c. Light Grey (mildly opaque): soft tissue (heart, blood vessels, etc.)
d. Off white: Bones
e. Bright white: metal implants (buttons, heart stents, nuts, etc.)

10. Broad methodology to be used:

    a. Collecting data and going through data to ensure integrity and sanity.
    b. Organising data into separate folders for each category.
    c. Uploading data to google drive/kaggle to work with Deep Learning models.
    d. Performance evaluation and reporting of model accuracy.
    e. Further training and improvement.

11. Packages and libraries to be used:

    Keras (based on tensorflow) is a powerful and highly user-friendly library for image processing. It has support for data augmentation, auto-categorising from folders, etc. We will try to use this library for modelling.
    The “fastai” library developed by Jeremy Howard and the fast.ai team is also very user friendly and suited for beginners. It has easy-to-use inbuilt methods for all major tasks.
    We will use a combination of the above 2 libraries to build our model

12. Initial training without any pre-processing:

    I organised the data into 6 separate classes for each category, and fit a basic ‘resnet34’ model, training for 4 epochs. No image preprocessing or data augmentation was used. This is to get an initial idea about the dataset and our model performance. I obtained the following results:

epoch	train_loss	valid_loss	accuracy	error_rate	time
0	0.740564	0.447471	0.849667	0.150333	26:42
1	0.485865	0.331255	0.866962	0.133038	09:05
2	0.375516	0.339502	0.859867	0.140133	09:04
3	0.337040	0.300028	0.875388	0.124612	09:06

    After training for 4 epochs on a set of 11276 images, with high class imbalance, we down to 12% error rate. The training loss is still more than validation loss, thus we are not yet overfitting.

13. Next steps:

    a. Training for more epochs to reduce bias

b. Creating a more balanced dataset and comparing performance with original dataset
c. Analysing accuracy, precision, recall and its impact