Assessing the Feasibility of Diagnosis of Pneumonia using Chest X-Ray Images

Group - 3ml
Krishnakanta Maity ||*RajatGaur*||*SaikatPatra*

June 4, 2022

Abstract

Pneumonia is one of the widely found diseases among people especially children. Recent studies suggest that globally there are over 1400 cases of this lung disease per 100,000 children, or 1 case per 71 children [6]. This disease is caused by bacterial infection in the lungs. One challenge which is usually faced in the medical domain is early diagnosis which becomes critical in the treatment process. Our study aims to develop a computer aided diagnosis system which will assist to guide the clinicians. Through our study, we would be applying deep learning concepts and will be applying modelling techniques for the detection and diagnosis of Pneumonia.
As the timeline progresses, we will come up with the detailed modelling progress based on the relevant literature post which we will present our conclusions along with the results of our analysis.

1 Introduction

Pneumonia is a disease that adversely affects the lungs of an individual and this inflammatory condition primarily affects the small air sacs in the lungs called *alveoli*. Certain symptoms which emerge in this condition include dry cough, fever, chest pain and the person faces difficulty in breathing.
It is usually caused by the infection with viruses and bacteria. Available data suggests that there were 2.5 million deaths from pneumonia in 2019 and pneumonia is the single largest infectious cause of death in children worldwide. Among many other reasons, one of the probable reasons is non diagnosis of pneumonia in the very early phases.
Our study aims to analyze the chest x ray images and to come up with a modelling classification problem to assess the feasibility of detecting the disease so that proper diagnosis can be made available to the patients at an adequate time.

Our study is an effort in order to assist doctors and to leave a meaningful impact in the medical domain.

2 Proposed methodology

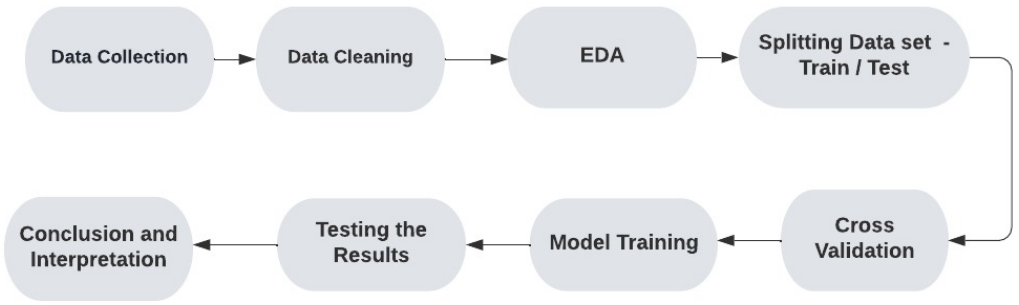

Figure 1: Timeline of this project

The above visualization summarizes our proposed methodology and a brief of each of those points has been mentioned below,

- **Data Collection** : We have explored the sources to collect the relevant data required for the analysis. The total dataset is the updated version 3 which has a size of 8 GB.

- **Data Preprocessing**: This step will include cleaning the dataset and figuring out the structure of the features as well as response variables.

- **EDA**: In this step, further exploration of the dataset will be done and any kind of anomaly in the dataset will be handled according such as missing values and dropping of the redundant features.

- **Splitting Dataset**: In order to build up the model, we will split the dataset and will dedicatedly keep training and test dataset separate to analyze the accuracy results during later stages of the project.

- **Cross Validation**: For having a greater efficiency of the model we are proposing cross validation of the training dataset so that our model accuracy would be improved

for unknown data points as well.

- **Model Training**: Based on the literature we will explore the modelling techniques that can be used for successfully training the dataset.
- **Testing**: After model training, we will test the efficacy and performance of our model by feeding the test dataset which we kept separately to analyze the model efficiency.
- **Inference**: Based on the results, we will provide the concluding remarks consisting of the performance of the model and its efficiency.
- **Presentation**: Finally we will present our findings and will seek the feedback of sir and batchmates.

3 Work plan and time line

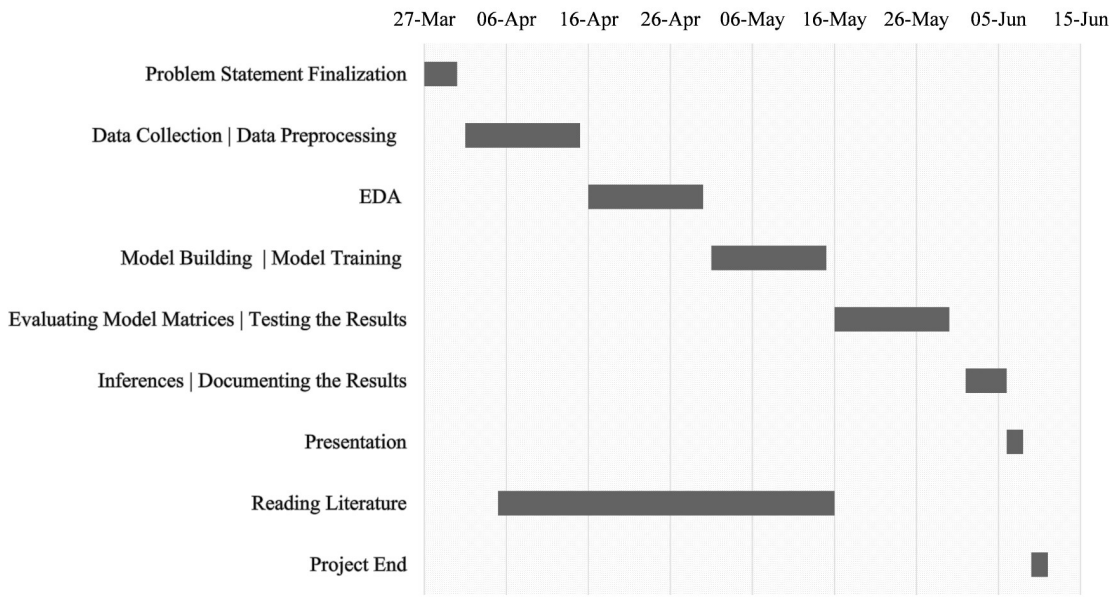

Figure 2: Project methodology

4 Work plan division in your group

1. Literature Reading - Saikat Patra (assisted by other 2 members)

2. Data Processing and EDA – Krishnakanta and Rajat
3. Model Training - 3 pilot batch divisions by each of us
4. Model Testing - Post application of the model on the 3 batches, the model will be tested on the testing dataset
5. Interpretation - Collaboratively we will assist each other to come up to the conclusions.
6. Presentation - All 3 of us

References

[1] Bernhard E Boser, Isabelle M Guyon, and Vladimir Naumovich Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pages 144–152. ACM, 1992.

[2] Leo Breiman. Random forests. *Machine Learning*, 45(1):5 – 32, 2001.

[3] Daniel Kermany, Kang Zhang, Michael Goldbaum, et al. Labeled optical coherence tomography (oct) and chest x-ray images for classification. *Mendeley data*, 2(2), 2018.

[4] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. The MIT Press, 2nd edition, 2018.

[5] Frank Rosenblatt. Principles of neurodynamics: Perceptrons and the theory of brain mechanisms. Technical report, Cornell Aeronautical Laboratory, 1961.

[6] UNICEF. Pneumonia. `https://data.unicef.org/topic/child-health/pneumonia/`, April, 2021.