Sprint 3- Status Check In 1

Chang Liu
chl769@gatech.edu

1 PROJECT INTRODUCTION

Over the last two years, Covid19 has quickly spread worldwide, resulting in the Covid-19 pandemic which causes thousands and thousands of infections and death.

Chest X-ray images have proven to be powerful to monitor various lung diseases and have been used to monitor Covid 19 disease. In recent years, there is a rapid development of machine learning models. Some of which have achieved human-like performance in various classification tasks.

In this project, the aim is to leverage machine learning for Covid 19 classification and display the result in an App/Web page. The App/Web page will generate results based on patients' chest X-rays in real-time.

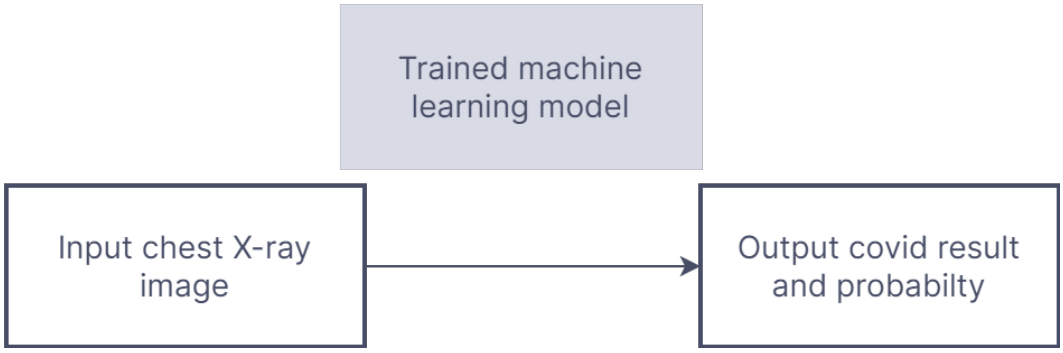

Figure 1—Architecture Diagram

2 ACCCOMPLISHIMENTS

Based on the previously proposed timeline, I have been working on the identifying the dataset and data preprocessing.

Data cleaning and preprocessing: 10/17 - 10/30

2.1 Download dataset

-After some search, I have downloaded the covid chest x-ray data set from the Kaggle challenge. It is the largest covid 19 dataset which constist of 3616 covid

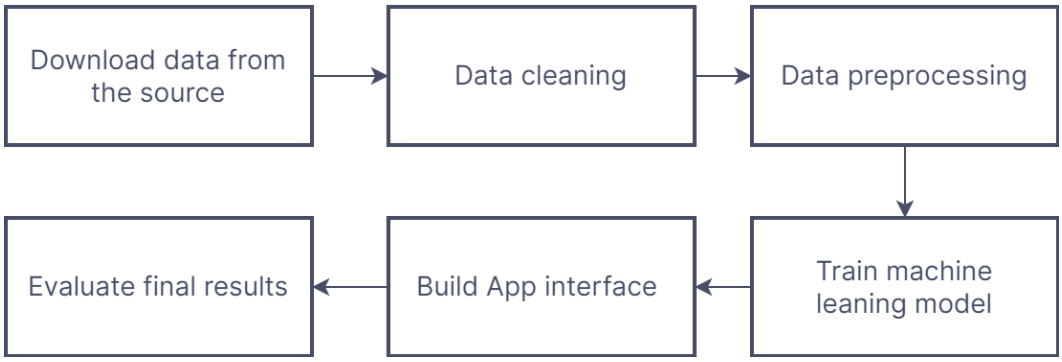

Figure 2—Project Task

images and 10701 normal chest x-ray images.

Covid chest X-ray data has been downloaded from this source:

https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database

2.2 Visualize the data

I have loaded the images and visualized them in a jupyter notebook. It looks like patients with covid have a more 'foggy' lung image in x-ray.

Figure 3—Normal image

Figure 4—Covid image

2.3 Data preprocessing

After some literature search, I decide to use 2 different methods for image preprocessing.

1. Just the baseline approach apart from size reduction.

2. Use histogram equalization to see if we can somehow increase the range of the pixel intensity and potentially a higher contrast.

2.4 Evaluation benchmark

Based on the discussion with my mentor, I have established several benchmarks for model evaluations.

Currenlty, there are three benchmark that I am consider using:

1. Accuracy for classification is greater than 0.5.

2. AUC for classification is greater than 0.5.

3. Also, because we don't want to miss any covid patients, I think specificity greater than 0.50 can be another benchmark.

3 CHALLENGES

1. Currently, I don't have any major roadblocks. One thing that I am little concerned is that the size of the data is a little bit large and my computer might run out of memory. I think I might start with a subset of the large dataset and try to fine-tune the model and select some best models.

2. Building the application might be difficult since I don't have much web development experience. I have been looking into some tutorials on FAST API and hopefully, I will be able to implement it after I train the model.

4 SPRINT PLANS

In the next 2/3 weeks, I am planning to train the machine learning model for covid 19 classification. I am planning to start with some traditional machine-learning models. For example, support vector machine, xgboost, etc.

I will resize the image matrix to a long vector and then consider each pixel as a feature. I will try different models and select the model with best performance.