Sprint 5- Status Check In 3

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.

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.

2 ACCCOMPLISHIMENTS

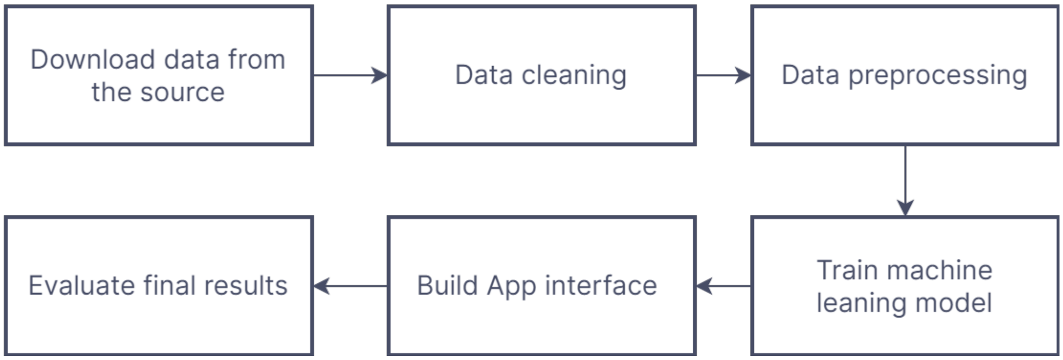

Figure 1—Project Task

Based on the previously proposed timeline, I have finished data cleaning, data preprocessing and training machine learning model in the past weeks.

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

Machine learning model training 11/1 - 11/27

Web development and report writing 11/28 - 12/7

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 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. (See previous sprint for more details)

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. (See previous sprint for more details)

2.4 Machine learning training

I train the model with image data and support vector machine achieves the best performance in terms of AUC.

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

Benchmarks evaluated on test set (randomly selected 60 covid positive and 60 normal patients, the testing set is never seen by the model during training):

1. Accuracy for classification: 0.941.

2. AUC for classification: 0.979. See the model ROC curve below. Dot line indicates 0.5 ROC(random guessing).

3. Specificity: 0.933.

4. Sensitivity: 0.950.

I also look at sensitivity and specificity at different thresholds of probability. Here are their values on 2 extreme scenarios:

When sensitivity is 1, which means that we do not miss any covid patients, the model has a specificity of 0.617. This means that when the model is able to identify all the covid patients, it can correctly identify a majority of the benign patients as well. I think this will be more applicable in real-life application since we want to contain the virus and don't want to miss any covid patients.

When specificity is 1, which means that all the healthy people are identified as not having covid, the model has a sensitivity of 0.783 which means that it is able to identify most of the patients with covid. I think this is more applicable when the detection tool is limited or expensive.

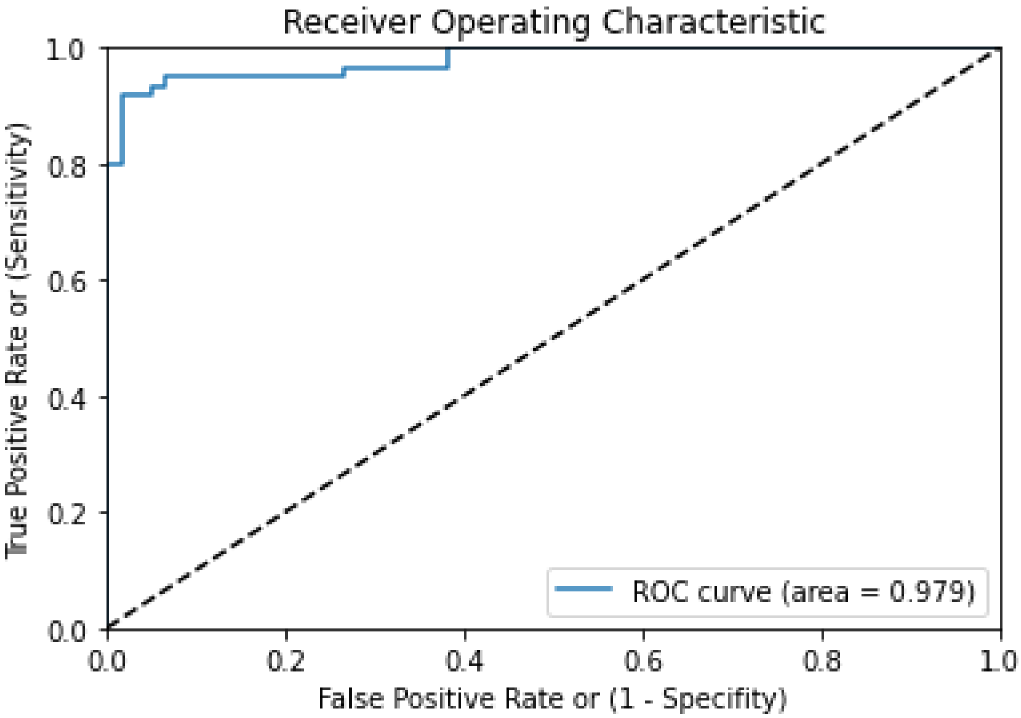

Figure 2—Roc curve of the model

3 CHALLENGES

Currently, I don't have any major roadblocks. I am working on integrating my code into web/app development. I have been looking into some tutorials on FAST API and other web/app development materials.

4 SPRINT PLANS

In the next week, I will keep looking into some online materials for web/app development. I will be working on integrating the existing model into a web/app interface. After that, I will work on the final report of the project.