

COMS 4771 COVID Kaggle Competition (Spring 2020)

Due: May 9, 2020 at 11:59 PM

This is a major class project that will allow you to apply methods you've learned about this semester, explore new algorithms, and work on a concrete project relevant to the current coronavirus epidemic. You will have over a month to work on this project, which will be a classification challenge hosted on Kaggle. This project is to be done **alone**. In this competition, you will attempt to diagnose patients with COVID-19, viral pneumonia, and bacterial pneumonia from images of chest X-rays. Your goal is to develop a multiclass classifier that achieves good weighted categorization accuracy on a set of unseen examples. All data is available on Kaggle and you will submit predictions through Kaggle as well. This document is a general outline of the project, but more specific guidelines can be found on the Kaggle site. You can access the competition at <https://www.kaggle.com/t/4104cd43b8814618bb03e8e479393869>. The public website is <https://www.kaggle.com/c/4771-sp20-covid>.

1 Competition Details

Welcome to the COMS 4771 Spring 2020 COVID diagnosis challenge. Given the present impact of the coronavirus on our lives, it made sense to create a challenge with some real-world relevance to the current crisis. In this competition, you will attempt to diagnose patients with COVID-19, viral pneumonia, and bacterial pneumonia from images of chest X-rays. Your goal is to develop a multiclass classifier that achieves good weighted categorization accuracy on a set of unseen examples.

Early in the epidemic, **physicians were actually diagnosing cases of coronavirus using X-Ray and CT images**. You can help them by developing more accurate methods for diagnosing COVID-19 from chest X-Rays. Since COVID X-Rays are frequently confused with ordinary pneumonia, you will be asked to perform multi-class classification, distinguishing patients with COVID-19 from those who have viral and bacterial pneumonia or who are healthy.

The training data (available on Kaggle) includes 1127 chest xrays drawn from several different sources (of varying size and quality) and a set of multiclass labels indicating whether each patient was healthy or diagnosed with bacterial pneumonia, viral pneumonia, or COVID-19. The test data includes 484 images without labels for which you will provide a diagnosis.

1.1 Kaggle Ranking

Your grade will correspond to your public and private leaderboard rankings. The public leaderboard on Kaggle includes a **minimum benchmark score (about 57% weighted accuracy)** which is the minimum weighted accuracy you must achieve to get partial credit. To get full credit, you will need to achieve a **80% weighted accuracy** on the public leaderboard. Your grade on this portion will increase linearly from 0 to 100 over this margin. We will also award extra credit to the top 5% of

teams on the private leaderboard, which is calculated separately from the public leaderboard but uses data drawn from the same distribution. This will be evaluated after the competition has ended.

1.2 Rules

Please see the rules and restrictions on the Kaggle site. The most notable rule, which goes without saying, is that you **cannot cheat**. That means you cannot try to find the dataset online and submit the true labels (or train on the dataset if you find it). You should submit all submissions under your UNI, i.e. your "team name" should be your UNI. Failure to comply with this may result in your disqualification. The competition will close on Saturday, May 9.

1.3 Writeup

In addition to the online competition, you should submit a PDF writeup on Gradescope describing your approach to the competition, details of your final submission, and lessons learned during the process. *This is as important as your ranking on the Kaggle competition.* Your writeup is a narrative describing the process you took exploring the data, investigating different algorithms, and assessing your results as if you were aiming to deploy this algorithm in a clinical setting. In your writeup, you should address at least (but not limited to) the following:

1. What methods/algorithms did you try? How well did they work? Why did you choose one over another?
2. What pre-processing steps did you apply? How did they affect your overall performance?
3. Perform a detailed error analysis of your best classifiers on a holdout validation set. You may wish to plot a confusion matrix or compare different metrics for multi-class classification like F1 score, precision, and recall. Given this error analysis, what issues would present themselves if your method were applied in a clinical setting?
4. Compare the *interpretability* of the different methods you tried. Can you gain any insight into what features your classifier is using to distinguish between different classes? Could this be useful to a clinician combating COVID-19?
5. What bugs or issues did you encounter while working on this projects? What slowed you down? How would approach this differently next time?

You should submit your code on Courseworks as usual.

1.4 What to submit?

Here are the items you are expected to submit for this competition.

1. Classification predictions for the competition on Kaggle (submitted as CSV files, see more details on Kaggle)
2. A detailed writeup of your experimental design and methods (submitted as a PDF on Gradescope)
3. Any code you write for the Kaggle competition (on Courseworks)