CSCE 633
Homework 5: Designing and disseminating ML for a real-world problem

---

Instructions for homework timeline and submission
Please submit on eCampus a **single pdf** file containing your solutions.
a) Form teams of 5 classmates! Please try to have **exactly 5 members** in your team. Please email us your team names and UIN with a private message on Piazza by **Monday 11/9**.
b) Work with your teammates in this project.
c) With your team, you will present your work in class: **Friday 11/20** (half of the teams) and **Monday 11/23** (the other half of the teams) during class time (**4.15-5.05pm CT**). Most of the work should be ready by the class presentations, including the elevator pitch video (question (e)) and the e-poster (question (f)). The material for question (g) will be obtained during in-class presentations.
d) The final report isuue on **December 5, 2020 @ 11.59pm**. Please create a zip file with the following: (i) final report; (ii) e-poster; (iii) csv file with decisions on test (optional).
e) Please start early :)

The goal is to build machine learning models to perform COVID-19 diagnosis from chest X-ray images. The dataset comes from COVID-19 image data collection and is uploaded under *Homework5* folder in the Google Drive. The training images are in the train folder, while testing images are in the test folder. The corresponding labels and additional information can be found at *train.csv* and *test.csv* files, that include:

- filename: the filename of the image
- gender: the patient's gender
- age: the patient's age
- location: the patient's location
- covid: the COVID-19 diagnosis, 1: positive, 0: negative **(outcome)**

For the following, you can use *any publicly available toolboxes*.

(a) (1 point) Image pre-processing. The images that are provided in the data are of different sizes. Center crop the images to a square shape (i.e., image length equals image width) and resize the cropped image into a fixed size.

(b.i) (2 points) Visual feature extraction. Extract image features, which will be used to predict the COVID-19 diagnosis. Please describe the features and provide a brief justification on how these features might work well for the outcome of interest.
Note: Please find here some useful toolboxes for extracting features from images: Gabor feature extraction, Computer Vision Feature Extraction Toolbox for Image Classification, Techniques to Extract Features from Image Data using Python, HoG feature extraction.

(b.ii) (2 points) Feature exploration. Provide visualizations of the features with respect to the outcome (e.g., overlaying histograms, scatter plots), and quantify associations between the features and the outcome.
Note: You can compute the Fisher's score, conditional information entropy, gini index, or any other type of metric that indicates associations between features and categorical outcome.

(b.iii) (2 points) Feature selection. Using the features that you have designed, explore two different feature selection methods of your choice. One method should be part of the **Filter** category and the other should be part of the **Wrapper** category. Using a simple classifier (e.g., SVM, logistic regression), plot the classification performance using a 5-fold cross-validation on the training data against the number of features for both feature selection methods. Compare and contrast between the two (e.g., in terms of performance and computation time).

loss / Acc

(b.iv) (1 point) Ensemble learning. Using the features that you have designed, employ the Adaboost method to estimate the COVID-19 diagnosis. Report results using a 5-fold cross-validation on the training data. Compare the result from Adaboost with the ones from feature selection.

(c) (2 points) Improving performance: Use any type of machine learning algorithm or feature design to improve the performance of your system. Describe your proposed approach and report the performance using a 5-fold cross-validation on the training data.
Note: You can also use information related to a patient's age, gender, and location.
Note: You can find a library of datasets and pre-trained models of chest X-ray images at: `https://github.com/mlmed/torchxrayvision`.

(d) (2 Bonus points) Select the **two** models that you believe that work best and are the most *generalizable*. Apply these models on the testing data. Provide the decisions in a csv file that includes three columns (in the provided order): test filename, model 1 decision, model 2 decision. The three teams that achieve the best performance will get 2 bonus points.
Note: Please do not provide more than two models.

(e) (2 points + 1 Bonus point) Elevator pitch video. Prepare a **45 sec** video to describe your work and results. In your elevator pitch, you have to attract the interest of your audience, so that they want to listen to more details on your presentation. You can use any type of visuals that you would like. Please upload your video on any type of social media (e.g., YouTube, TikTok) and provide the url in your final report. You will be also sending us the url of the video before the day that you are presenting in class, so that we can show it everyone! We will send you a reminder regarding this.
Note: The elevator pitch will be presented in beginning of the class on 11/20 and 11/23. The best videos, as voted by the rest of the class, will get 1 bonus point.
Note: You can use a private link if you would like.

(f) (2 points) E-poster. Create an e-poster presentation of your work. The e-poster will give the main gist of your work, including the problem statement, your methodology, and the main results from your experiments. **Add visuals to your poster so that people understand the main concepts.** Do not make your e-poster too crowded, since you want other people to be able to see through the screen projection. You can find here the link to prepare your poster presentation `https://www.youtube.com/watch?v=1RwJbhkCA58&feature=youtu.be`.
Note: Each team will be assigned to a Zoom link. Zoom links will be available to everyone in the class. The teams that are not presenting in the current day will log in to discuss your e-poster presentation. All members of the team need to be present during the presentation.

(g) (1 point) Reporting other teams' work. During the day that your team is not presenting, your will go around the posters of the teams that are presenting and report the main findings of the other teams. In the final report, provide a brief description of the work and main results from **8 other teams**.
Note: You can distribute the work among your team members.