Final Project Proposal (Team 8)

1) Problem statement — this section defines the machine learning problem.

Since the Covid-19 virus acts on a person's lungs, we could theoretically distinguish whether a patient has Covid-19 based on a chest X-ray image of the patient. At the same time, our project also needs to consider common viral pneumonia, which can also cause abnormal images of people's lungs.
Therefore, our final project aimed to classify chest X-rays into 3 classes:
normal, **Covid** and **viral pneumonia**.

2) Description of data set — this part identifies the source of training data set (provide URL link to the data set if it is from Internet)

Dataset we got from https://www.kaggle.com/pranavraikokte/covid19-image-dataset.
We have a total of 251 images for training (70 viral pneumonia, 70 normal, 111 Covid) and 66 images for testing (20 viral pneumonia, 20 normal, 26 Covid). Since this dataset is relatively small, we will enlarge the data set using mirror, crop and other data augment methods.
Also, each image is not the same size. For example one is (1400, 1648, 3) and the other is (2953, 3604, 3). We also need to do data normalization.

3) Implementation plan — this section briefly describes the tentative plan for implementation, milestones and timeline.

In our final project, we want to use several ML algorithms and DL algorithms to see their performances and make a comparison. Basically, we need to do the data preprocessing first, and then train 3 different ML models for every member, and then compare these models. At last we will adapt a DL method, CNN, to see how it works.
Basically, we need to do data preprocessing first, then train 3 different ML models, one for each member, and compare the models. Finally, we will adapt a DL method, CNN, to see how good it works and do analysis.
I draw a table to show the timelines of our project.

CPE 695 Group 8 Final Project Schedule																																						
No.	STEP	Due	3/1 ~ 3/6			3/7 ~ 3/20			3/21 ~ 3/27			3/28 ~ 4/3			4/4 ~ 4/17			4/18 ~ 4/24			4/25 ~ 5/1			5/2 ~ 5/6														
			Tue	N/A	Sun	Mon	N/A	Sun	Mon	N/A	Sun	Mon	N/A	Sun	Mon	N/A	Sun	Mon	N/A	Sun	Mon	N/A	Sun	Mon	N/A	Fri													
1	Data preprocessing & Data augment	3/6	13 days																																				
2	Training **Logistic Regression** model	3/20				14 days																																	
3	Training **SVM** model	3/20				14 days																																	
4	Training **Random Forest** model	3/20				14 days																																	
5	Improve the models and Analyze & Compare	3/27							7days																														
6	Other ML method - KNN or Ensemble Method **[Optional]**	4/3										7days																											
7	Training **CNN** model	4/17													14 days																								
8	Improve the model and Analyze & Compare	4/24																7days																					
9	Other DL method - GAN **[Optional]**	5/1																			7days																	
10	Finish the report and presentation	5/6																						5 days														

4) Team members & task allocation - this section list names of all team members and defines tasks for each member.

- **Individual work:**

  A. Chengchen Zhao [10468151]:
  - Training **Random Forest** model
  - Data preprocessing: clean the data and normalization
  - Finish the report about RF model

  B. Yicong Pan [10472353]:
  - Training **SVM** model
  - Data argument: enlarge the data set with crop
  - Finish the report about SVM model

  C. Yunhan Li [10464987]:
  - Training **Logistic Regression** model
  - Data argument: enlarge the data set with mirror
  - Finish the report about Logistic Regression model

- **Team work:**
  - Training **CNN** model
  - Analyze all the ML and DL models and make a comparison
  - Train the optional ML and DL methods if successful