Assignment 4
Pulmonary Disease Detection

- Deadline: 2021/12/23 23:59

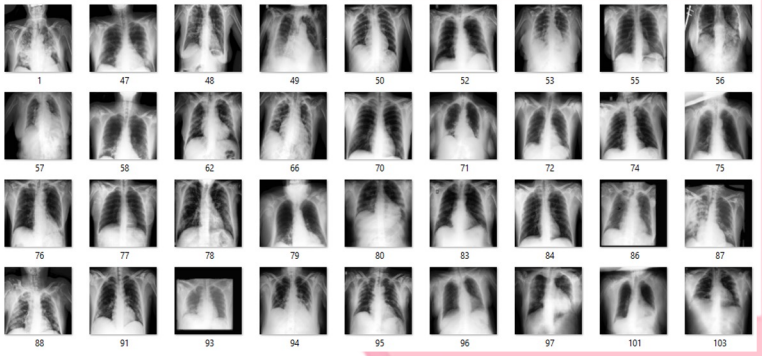
Goal

- **Build a convolutional neural network to predict the pulmonary disease of patients from their chest X-ray (CXR) images.**
- **You need to deal with limited and imbalanced data.**

Dataset: MIMIC-CXR

- MIMIC-CXR dataset is a large publicly available dataset of chest X-ray (CXR) images.
- The dataset contains 377,110 JPG format images and structured labels derived from the 227,827 free-text radiology reports associated with these images.
- In this assignment, we randomly extract 12,500 images for you.

Dataset

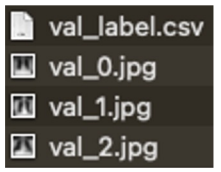


Please keep the data confidential!

- There are **10,000** images in the training set, 1,000 images in the validation set and 1,500 in the testing set.
- 2 demographic features (age and gender) are provided for all patients.
- **7** pulmonary disease labels for patients in training and validation set.
- All CXR images has been processed with histogram equalization and down-sampled to **128 by 128.**

Dataset

- You can find the data on elearn platform.
- Training and validation data are in the folder name as “public”.
- Testing data are in the folder name as “private”.
- The file name of each image contains an index and the split.
- Labels and demographic features for training and validation data are provided in csv files.
- The index of each row corresponds to the file name of the image.
- You can refer to the data loading example.



	Atelectasis	Cardiomegaly	Edema	Lung Opacity	No Finding	Pleural Effusion	Support Devices	Age	Gender
0	0	0	0	0	1	0	0	4	0

Basic-prediction (50%)

- It is a binary classification problem.
- 0: negative; 1: positive
- Use the training data to predict whether the patient has “Edema”.
- In this part, any existing model architectures are prohibited, but functions for constructing models are allowed. You should design you own model architecture.
- You will get all 50 points if your prediction achieves **f1-score greater than 0.6** in the testing set.

	Edema
0	0
1	0
2	0
3	1
4	0

Basic-report (10%)

- Describe how you design your own model architecture.
- Describe how you choose hyperparameters (eg. optimizer, learning rate).
- Describe difficulties you encountered.
- Summarize your implementation.
- No more than 1 page.
- Don’t screenshot any codes or copy paste.

Bonus (10%)

- If you use the demographic data such as age and gender to improve the model performance and f1-score achieve **0.65**, you will get extra 10 points as bonus.
- Illustrate how you combine image data with demographic data in the report and submit an extra prediction for bonus part.

Value	Age	Gender
0	0-20	Male
1	20-40	Female
2	40-60	
3	60-80	
4	80-	

Advanced-prediction (35%)

- It is a **multi-label** classification problem.
- Use the training data to predict 7 labels.
- In this part, any state-of-the-art model architectures are allowed.
- Points in this part will be given based on the **weighted f1-score** of your prediction in testing set.

	Atelectasis	Cardiomegaly	Edema	Lung Opacity	No Finding	Pleural Effusion	Support Devices
0	0	0	0	0	1	0	0
1	0	0	0	0	1	0	0
2	0	0	1	0	0	0	0
3	0	0	1	0	0	0	1
4	0	0	0	0	1	0	0

Advanced-report (5%)

- Describe how you design or choose your own model architecture.
- Describe how you choose loss function and optimizer.
- Describe difficulties you encountered.
- Summarize your implementation.
- No more than 1 page.
- Don’t screenshot any codes or copy paste.
- Write it after the basic part in the same report file.

Submission format-prediction

- All predictions should be binary (0 or 1), you can choose your own threshold or I will round it up if your predictions is not in binary form.
- The index of your prediction should match the index of images.
- The format of prediction is provided.
- For basic part, submit your prediction in a single csv file name as [Student_ID]_basic_prediction.csv

  [Student_ID]_bonus_prediction.csv (if you implement)
- For advanced part, submit your prediction in a single csv file name as [Student_ID]_advanced_prediction.csv.

Submission format-report, code

- Name your report as [Student_ID]_report.pdf.
- No more than 2 pages (1 for basic part, 1 for advanced part). If you implement bonus part, an extra page is acceptable.
- Zip all your codes in 1 zip file, and name as [Student_ID]_code.zip
- Put all your stuff into one zip file name as [Student_ID].zip, including basic prediction, bonus prediction, advance prediction, code, and report.

Notification

- It is an individual assignment.
- We do not provide any template, but we provide a sample code for loading images.
- You can use any framework as you like, such as tensorflow, or pytorch. Colab is recommended.
- 0 point will be given in the following conditions.

  Use SOTA model in the basic part.

  late submission.

  Incorrect prediction format.

  Incorrect file name.

  Plagiarism.

Contact & supplements

- **Ryan Wang ryanatnthu114@gmail.com**
- **Don’t ask me for debugging.**
- **Be polite!**
- **Tensorflow example:**
  https://www.tensorflow.org/tutorials/images/classification
- **Pytorch example:**
  https://pytorch.org/tutorials/beginner/blitz/cifar10_tutorial.html