一、本周研究内容

对 COVID-19 CXR 数据集进行了处理，将图片的详细信息等收集在一个 CSV 文件 **dataset.csv** 里，并把 254 个 COVID-19 CXR 图片按照约定的格式全部重新命名处理。

二、项目实施当前状态

COVID-19 CXR 数据集已经完成了流程的制定，规范了文件的命名，下一步等暑期科研的同学们一起做合作完成对数据集的扩充。

三、本周成果

1. 图片重新命名：

所有图片处理后的文件名称对应文件为 ‘dataset.csv’ 。该 CSV 一共包含 finding、modality、filename、url, paper_name 和 new_filename。其中 finding 值均为 COVID-19，modality 均为 X-ray, filename 为图片原始名字，url 为该 CXR 的来源(论文链接或网站链接)，**paper_name** 为该论文的名字，**new_filename** 为根据新规则命名的图片新名字。

paper_name 如下图 1 所示，为包含 CXR 图片的论文名称。

		dataset
1		**paper_name**
2		Importation and human-to-human transmission of a novel coronavirus in Vietnam
3		Importation and human-to-human transmission of a novel coronavirus in Vietnam
4		Importation and human-to-human transmission of a novel coronavirus in Vietnam
5		Importation and human-to-human transmission of a novel coronavirus in Vietnam
6		A locally transmitted case of SARS-CoV-2 infection in Taiwan
7		A locally transmitted case of SARS-CoV-2 infection in Taiwan
8		Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study
9		Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study
10		First imported case of 2019 novel coronavirus in Canada, presenting as mild pneumonia
11		Evolution of CT Manifestations in a Patient Recovered from 2019 Novel Coronavirus (2019-nCoV) Pneumonia in Wuhan, China
12		First Case of 2019 Novel Coronavirus in the United States
13		First Case of 2019 Novel Coronavirus in the United States
14		First Case of 2019 Novel Coronavirus in the United States
15		First Case of 2019 Novel Coronavirus in the United States
16		First Case of 2019 Novel Coronavirus in the United States
17		First Case of 2019 Novel Coronavirus in the United States
18		First Case of 2019 Novel Coronavirus in the United States
19		Imaging Profile of the COVID-19 Infection: Radiologic Findings and Literature Review
20		Imaging Profile of the COVID-19 Infection: Radiologic Findings and Literature Review
21		Imaging Profile of the COVID-19 Infection: Radiologic Findings and Literature Review
22		Imaging Profile of the COVID-19 Infection: Radiologic Findings and Literature Review
23		Chest Imaging Appearance of COVID-19 Infection
24		Case of the Index Patient Who Caused Tertiary Transmission of Coronavirus Disease 2019 in Korea: the Application of Lopinavir/Ritonavir for the Treatment of COVID-19 Pneumonia Monitored by Quantitative RT-PCR
25		Case of the Index Patient Who Caused Tertiary Transmission of Coronavirus Disease 2019 in Korea: the Application of Lopinavir/Ritonavir for the Treatment of COVID-19 Pneumonia Monitored by Quantitative RT-PCR
26		Case of the Index Patient Who Caused Tertiary Transmission of Coronavirus Disease 2019 in Korea: the Application of Lopinavir/Ritonavir for the Treatment of COVID-19 Pneumonia Monitored by Quantitative RT-PCR
27		Coronavirus Disease 2019 (COVID-19): A Perspective from China
28		First case of Coronavirus Disease 2019 (COVID-19) pneumonia in Taiwan
29		First case of Coronavirus Disease 2019 (COVID-19) pneumonia in Taiwan
30		First case of Coronavirus Disease 2019 (COVID-19) pneumonia in Taiwan
31		First case of Coronavirus Disease 2019 (COVID-19) pneumonia in Taiwan

图 1：paper_name 为包含 CXR 图片的论文

new_filename 如下图 2 所示，为新的 CXR 图片的名称。

new_filename
P1_ImportationAndHuman-to-human_20200128_Fig1a.jpeg
P1_ImportationAndHuman-to-human_20200128_Fig1b.jpeg
P1_ImportationAndHuman-to-human_20200128_Fig1c.jpeg
P1_ImportationAndHuman-to-human_20200128_Fig1d.jpeg
P2_ALocallyTransmitted_20200212_Fig1a.jpeg
P2_ALocallyTransmitted_20200212_Fig1b.jpeg
P3_EpidemiologicalAndClinical_20200130_Case2a.jpg
P3_EpidemiologicalAndClinical_20200130_Case2b.jpg
P4_FirstImportedCase_20200229_Fig1.jpeg
P5_EvolutionofCT_20200207_Figa.jpeg
P6_FirstCaseof_20200305_Fig1pa.jpeg
P6_FirstCaseof_20200305_Fig1l.jpeg
P6_FirstCaseof_20200305_Fig3pa.jpeg
P6_FirstCaseof_20200305_Fig3l.jpeg
P6_FirstCaseof_20200305_Fig4.jpeg
P6_FirstCaseof_20200305_Fig5pa.jpeg
P6_FirstCaseof_20200305_Fig5l.jpeg
P7_ImagingProfileof_20200213_Fig2.jpeg
P7_ImagingProfileof_20200213_Fig5day0.jpeg
P7_ImagingProfileof_20200213_Fig5day4.jpeg
P7_ImagingProfileof_20200213_Fig5day7.jpeg
P8_ImagingProfileof_20200213_Fig5day7.jpg
P9_CaseOfThe_20200214_Fig1a.jpg
P9_CaseOfThe_20200214_Fig1b.jpg
P9_CaseOfThe_20200214_Fig1c.jpg
P10_CoronavirusDisease2019_20200221_Fig3.jpg

图 2：paper_name 为新的 CXR 图片名称

2. CXR 数据集收集 Pipeline 更新整理：

2.1 在暑期和其他本科同学一起收集 CXR 图片时，我将每天负责维护一个 CSV 文档 dataset.csv，共享给所有参与数据集搜集的同学们，然后大家分别从以下渠道：

- 预印本：包括但不限于 medRxiv, bioRxiv
- 期刊论文：包括但不限于 Lacent, Radiology
- 网站：包括但不限 SIRM, radiopaedia

去搜集 CXR 图片，然后将该图片的来源等信息填写到 CSV 文档里，我每天对文档进行核查，确保来源不重复。

2.2 图片信息填写 CSV 文档：

- 当找到了一个新 COVID CXR 后，添加图片来源 url 到 CSV 文件的 URL 栏；
- 如果是来源于论文的话，还需将论文的名称添加到 CSV 文件的 paper_name 栏；

- 来源是论文：按照 ***论文 ID+ 论文 title 前 3 个单词+时间（数字）+图片编号（数字）*** 的格式给该图片命名，例如 **P2_ALocallyTransmitted_20200212_Fig1a.jpeg**，并且添加到 CSV 文件的 new_filename 栏。
- 来源是网站：按照 ***网站名称 + 图片编号***（参考网站 URL 给它的命名方式）来给图片命名，例如 网站 url 为 https://radiopaedia.org/cases/covid-19-pneumonia-7，那么该图片名称为 **radiopaedia_covid19Pneumonia7.jpg**。

3. 发现的问题：

在 5 月 14 日 Webinar 上第四位分享的 speaker Jong Chul Ye 指出了 COVIDX 数据集的问题，指出它分的 category 选择有问题，不应该简单讲一些病毒肺炎归为 non-covid19。

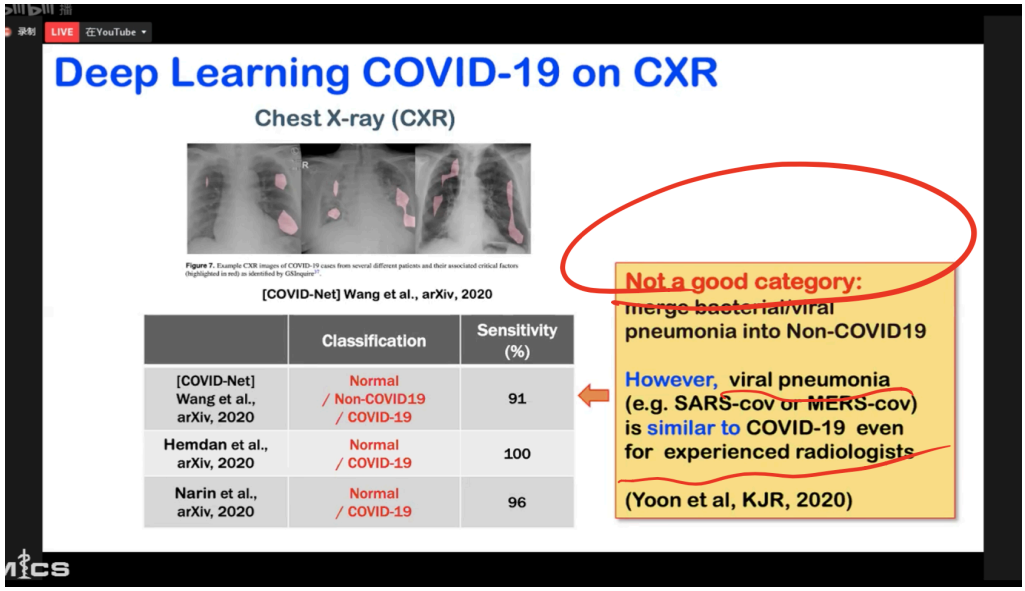

图 3：5.14 MICCAI Webniar Jong Chul Ye

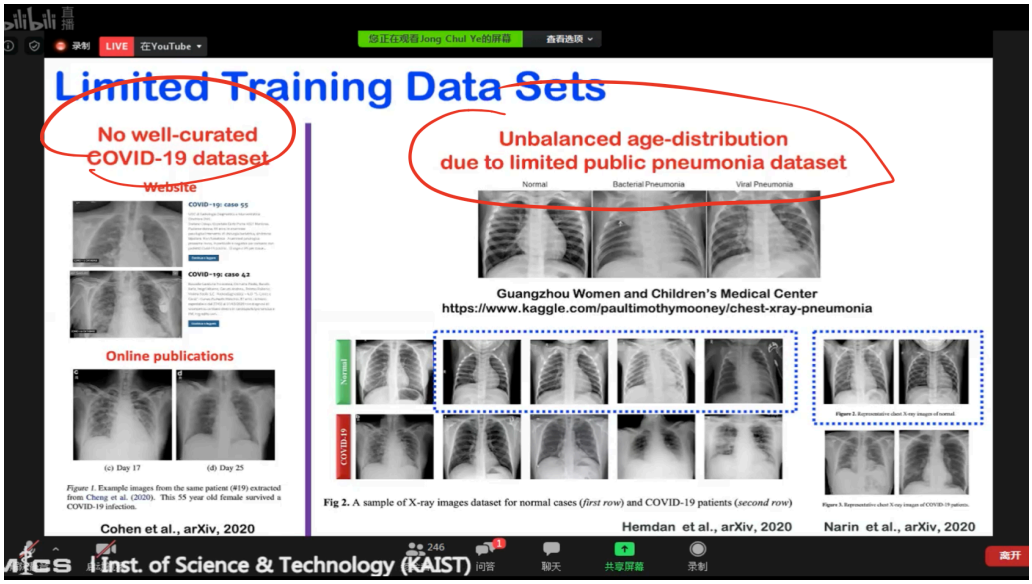

图 4：5.14 MICCAI Webniar Jong Chul Ye

我去查看了 Jong Chui Ye 的这篇 paper 以及去核实 COVIDX 的 paper, 发现其实 COVIDX 是 Normal，Pneumonia 和 COVID-19 的三分类，不存在 Jong Chui Ye 说到的将病毒肺炎分类到 non-covid19 的问题。但是在看了 Webniar 后，我觉得可以讨论一下我们的数据集的 category 该如何选择的问题，是 Normal, Pneumonia 和 COVID-19 三分类，还是 COVID-19 和 CAP 二分类，或其他。

四、下周计划

下周继续”COVID-19 Twitter Mining and Text Classification”项目，预计下周完成 twitter 数据标注 2000 个，并且建立经典机器学习 model baseline: KNN, SVM 对已经打标的。