Links of Datasets

Mengyao He

Aug 2020

1 Pulmonary Chest CT Scans

Paper: Using Artificial Intelligence to Detect COVID-19 and Community-acquired Pneumonia Based on Pulmonary CT: Evaluation of the Diagnostic Accuracy

Description: The datasets were collected from six hospitals between August 2016 and February 2020. The collected dataset consisted of 4352 chest CT scans from 3322 patients.

Link: `https://github.com/bkong999/COVNet`

2 Day Level Information on Covid-19 Affected Cases

Paper: This dataset includes the daily numbers of cumulative confirmed cases, recovered cases, and death cases from January 22, 2020, to April 3, 2020.

Description: The dataset of COVID-19 has been downloaded from Kaggle. The dataset is contributed by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU). This dataset includes nationwide details of confirmed cases, recovered cases, and death cases. The presented work uses the data of China from January 22, 2020, to April 3, 2020, to predict the outcome in India in the next 22 days.

Link: `https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset`

3 Covid Chest X-ray and CT Dataset

Paper:

1. New machine learning method for image-based diagnosis of COVID-19
2. Deploying Machine and Deep Learning Models for Efficient Data-Augmented Detection of COVID-19 Infections

Description: Two dataset of chest X-ray and CT images of patients which are positive or suspected of COVID-19. In this paper, they accessed on 15 April 2020.

Link:

1. Covid-Chestxray-Dataset: `https://github.com/ieee8023/covid-chestxray-dataset`

2. COVID-CT-Dataset: A CT Scan Dataset About COVID-19: https://github.com/UCSD-AI4H/COVID-CT

4 CSSEGISandData

Paper: Machine Learning to Detect Self-Reporting of Symptoms, Testing Access, and Recovery Associated With COVID-19 on Twitter: Retrospective Big Data Infoveillance Study

Description: A total of 72,922,211 tweets were collected from March 3-20, 2020, from the Twitter public API filtered for general COVID-19-related keywords. For statistical analysis and geospatial visualization, COVID-19 cases from March 20, 2020, were obtained from the JHU GitHub CSSEGISandData file.

Link: https://github.com/CSSEGISandData/COVID-19

5 Protein Sequences of 2666 Coronaviruses

Paper: Using the spike protein feature to predict infection risk and monitor the evolutionary dynamic of coronavirus

Description: The protein sequences of 2666 coronaviruses were collected from 2019 Novel Coronavirus Resource (2019nCoVR) Database of China National Genomics Data Center (NGDC) on Jan 29, 2020. These strains had full length genomes and were isolated between 1941 and 2020, and included SARS-CoV-2 strains. The information related to these strains was summarized in Additional file 1. The 507 human-origin coronaviruses were regarded as positive samples, whereas the 2159 non-human-origin coronaviruses were regarded as negative.
Link: https://bigd.big.ac.cn/ncov

6 Real-time Data of COVID-19 and Population Mobility Data

Paper: Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions

Description: The website of Tencent news provided us with the time series data of COVID-19 by locations, including the number of confirmed cases, deaths, recovered cases, and newly diagnosed cases. Baidu migration is an open-source big data project visualizing population migration. Leveraging its Location based services system and Baidu Tianyan system, we obtained the daily migration scale index (MSI) of Beijing, Shanghai and Guangzhou, in January and February of both 2019 and 2020.

Link:

1. Real-time reporting of cases of coronavirus disease 2019 (Accessed 26 Feb 2020) https://news.qq.com/zt2020/page/feiyan.htm?ADTAG=area
2. Baidu migration (Accessed 26 Feb 2020): http://qianxi.baidu.com

7 Coivd-19 Data and Migration Data

Paper: Modeling the trend of coronavirus disease 2019 and restoration of operational capability of metropolitan medical service in China: a machine learning and mathematical model-based analysis

Description: The most recent epidemiological data based on daily COVID-19 outbreak numbers reported by the National Health Commission of China were retrieved. Migration index based on the daily number of inbound and outbound events by rail, air and road traffic, were sourced from a web-based program. The 2003 SARS epidemic data between April and June 2003 across the whole of China retrieved from an archived news-site (SOHU) was used for AI-training.

Link:

1. Situation report (in Chinese) 2020. Available online: `http://www.nhc.gov.cn/xcs/yqtb/list_gzbd.shtml`
2. Baidu qianxi (in Chinese) 2020 Available online: `https://qianxi.baidu.com/`
3. Combatting SARS (in Chinese) 2003. Available online: `http://news.sohu.com/57/26/subject206252657.shtml`

8 Dataset of SARS-CoV-2 Genome

Paper: Chaos game representation dataset of SARS-CoV-2 genome

Description: the dataset provides a chaos game representation (CGR) of SARS-CoV-2 virus nucleotide sequences. The dataset provides the CGR of 100 instances of SARS-CoV-2 virus, 11540 instances of other viruses from the Virus-Host DB dataset, and three instances of Riboviria viruses from NCBI (Betacoronavirus RaTG13, bat-SL-CoVZC45, and bat-SL-CoVZXC21).

Link: `https://data.mendeley.com/datasets/nvk5bf3m2f/1`