COVID-19 Action Group on AI and Radiology
Covid19action-radiology-CXR_v1.1

Introduction

The outbreak of COVID-19 in Wuhan, China in December 2019 has rapidly spread across other countries in the world and has been declared as a global pandemic by WHO on 11th March, 2020. COVID-19 continues to have adverse effects on the health and economy of the global population and has brought immense pressure on the health care systems of the developing as well as developed countries. A critical step in the fight against COVID-19 is the early detection of the patients which not only aids in providing quicker treatment but is also useful in stopping the spread of the virus through proper isolation of the patient.

Although the reverse transcription polymerase chain reaction rt-(PCR) test of the sputum is considered as the gold standard for COVID-19 diagnosis, it is a time-consuming process and, in some cases, has led to high False Negatives. The relatively low-cost, wider availability and faster methods for sanitization of the equipment makes Chest X-ray imaging (CXR) a promising modality which is commonly being employed for treatment planning, tracking the progression of the disease and could also be useful in the detection of COVID-19.

Objectives

Our goal is to create a single repository by collating various publicly available CXR image sources and provide a standardized split for performing a five-fold cross-validation based training and a separate held out test set for evaluating different AI models. We hope that it will provide a unified platform for researchers to perform a fair comparison of different AI models. This dataset comprises images acquired from different geographical regions using different scanners and at varying resolutions.

We hope to release the baseline performance of some standard CNN architectures for the dataset in the future. We are also in the process of collecting additional CXR images from our partner hospitals which will be added to the future release versions of this dataset.

Problem Statement

The primary task is to classify a given CXR image into “COVID-19”, “Other Pneumonia” and “Non-pneumonia” classes. The “Non-pneumonia” class contains images from healthy subjects as well as subjects suffering from diseases other than pneumonia.

Additionally, the ground truth (GT) annotations of 13 radiological observations are also available for the images from the Chexpert dataset [3] (Source -3). These may be used to define an additional auxiliary task for the model or pre-training it to improve generalization.

Description of Data Sources

In Version 1.0 of this dataset, we have collated images from several publicly available sources [1-7] which are listed below.

- **Source-1** [1]: It contains CXR images collected from a hospital in Spain and primarily contains COVID-19 positive cases diagnosed using the PCR test.
- **Source-2** [2]: It contains CXR images of COVID-19 positive cases from Italy that have been publicly shared by the Italian Society of Medical and Interventional Radiology.
- **Source-3** [3]: This is a large publicly available database known as Chexpert which has been released by the Stanford Machine Learning Group. It does not contain COVID-19 cases but has CXR images of other types of pneumonia and images with other abnormalities. Note, that the Ground Truth of the training set is noisy as it was automatically extracted using a NLP tool and contains many uncertain labels (denoted by -1). We refer to [3] for further details.
- **Source-4** [4]: These images were publicly released under a Kaggle Challenge. It does not have any COVID-19 case but contains images of healthy subjects as well as subjects suffering from other types of viral and bacterial pneumonia.
- **Source-5** [5]: This dataset is currently being compiled from images from publications and other public sources. They primarily contain COVID-19 cases and some cases with other viral and bacterial pneumonia.
- **Source-6** [6]: This dataset has been publicly released recently and primarily contains COVID-19 cases and a few cases of other types of pneumonia.
- **Source-7** [7]: We have used a single CXR image of a COVID-19 case from the publication [7] with permission from the publisher.

How to download the images?

The images from Source-1, Source-2 and Source-7 have been provided with this dataset.

The remaining datasets can be downloaded from the hyperlinks provided in [3-6]. All images are available in a single folder in Source-5 and Source-6.

In case of Source-4, there are three directories for train, validation and testing. Inside each directory, there are two sub-directories containing Normal and Pneumonia images. The sub-directories for Normal and pneumonia can be merged into a single folder to easily access the images using our csv file. Also, the 8 images in the validation folder has been merged within the training set.

The images from the Chexpert dataset (Source-3) need to be properly arranged in order to use our csv files. Instead of the hierarchical file structure, used in the Chexpert dataset, we assume that all images have been arranged within a single directory (called “Train_reshape” for the Training set of Chexpert and “Val_reshape” for the Validation Set of the Chexpert dataset).

For example, an image in the path “/CheXpert-v1.0/train/patient18936/study1/view2_lateral.jpg” is put inside “Train_reshape” directory with the image name: CheXpert-v1.0__train__patient18936__study1__view2_lateral.jpg
(The image name is formed by concatenating the entire directory path to that image in the original Chexpert dataset and separating the sub-folder names with ‘__’)

Description of the annotations

The Ground-truth annotations are provided in 4 csv files which are described below.

1. Train_Combined.csv: It comprises the images that belong to the training set. The first column of the csv file is the image name, followed by the Data Source. The third column “Partition” provides the partition index for each image which will be useful in performing a five-fold cross-validation. The partitions are stratified for each class and performed at the *patient level*. Since, in some data sources, some patients have multiple images, the number of images in each partition may vary slightly. The last three columns provide the GT for “Non-pneumonia”, “other pneumonia” and the “COVID-19” classes. Some images from Source-3 have uncertain labels which are indicated by -1.

2. Test_Combined.csv: It contains the images that belong to the test set. These images *shouldnot* be used during training *not even for* monitoring the validation accuracy. The format is similar to Train_Combined.csv with the exception that it does not have the “Partition” column.

3. Train_Source3_task2.csv: Its format is also similar to the Train_Combined.csv. The difference being that instead of the 3 classes, it provides the Ground Truth annotations for an additional set of 13 radiological findings. These GT annotations are only available for the images from Source-3. In the GT, -1 indicates that the annotation is uncertain. We refer to [3] for further details on ways to handle the uncertainty in the training labels.

4. Test_Source3_task2: It provides the test set with the GT annotations for the thirteen radiological findings for the images from Source-3. These images *shouldnot* be used for training *not even fo*r monitoring the validation accuracy.

Citation

The dataset can be cited using the IEEE dataport doi.
We also intend to release a white paper in arxiv related to this data Collection shortly.

We acknowledge the authors, contributors and administrators of all the datasets [1-7] whose efforts have made this data collection possible. The sources of these datasets should be properly cited if you use this data collection in your research.

License

About Us

Our COVID-19 Action Group on AI and Radiology (#covid19actionRadiology) group is a pro bono public health research initiative which aims to

1. Develop the COVID-19 Radiology Reporting and Data Standards (COVID-19 RADS)
2. Create a Dataport for submission of scans, clinical reports, annotations by Radiologists and Physicians to build the Database
3. Sharing of crowd annotated case reports under open source non-commercial licenses for Radiology Education and building of AI models
4. Publishing and archiving Standards, Whitepapers, Recommendations and Guidelines, and AI Model Zoo for COVID-19

For further details or queries, you can contact us at:

Email : covid19action.radiology@gmail.com
Web: https://covid19action-radiology.github.io/
Facebook: https://www.facebook.com/covid19actionRadiology
LinkedIn: https://www.linkedin.com/company/covid19action-radiology/

References

[1] https://twitter.com/ChestImaging/status/1243928581983670272

[2] https://www.sirm.org/category/senza-categoria/covid-19/

[3] Irvin, Jeremy, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund et al. "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 590-597. 2019. (link: https://stanfordmlgroup.github.io/competitions/chexpert/)

[4] https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia

[5] Joseph Paul Cohen and Paul Morrison and Lan Dao, “COVID-19 image data collection”, arXiv:2003.11597, 2020 https://github.com/ieee8023/covid-chestxray-dataset.

[6] Linda Wang, Alexander Wong, Zhong Qiu Lin, James Lee, Paul McInnis, Audrey Chung, Matt Ross, Blake VanBerlo, Ashkan Ebadi, “FIgure 1 COVID-19 Chest X-ray Dataset Initiative”, https://github.com/agchung/Figure1-COVID-chestxray-dataset

[7] Kong, Weifang, and Prachi P. Agarwal. "Chest imaging appearance of COVID-19 infection." *Radiology: Cardiothoracic Imaging* 2, no. 1 (2020): e200028. https://pubs.rsna.org/doi/full/10.1148/ryct.2020200028