COVID-19-CT-CXR: a freely accessible and weakly labeled chest X-ray and CT image collection on COVID-19 from biomedical literature

Yifan Peng[1], Yu-Xing Tang[2], Sungwon Lee[2], Yingying Zhu[2], Ronald M. Summers[2], Zhiyong Lu[1]

1. National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institutes of Health (NIH), Bethesda, MD 20894
2. Imaging Biomarkers and Computer-Aided Diagnosis Laboratory, Radiology and Imaging Sciences Department, National Institutes of Health (NIH) Clinical Center, Bethesda, MD 20892

Abstract

The latest threat to global health is the COVID-19 outbreak. Although there exist large datasets of chest X-rays (CXR) and computed tomography (CT) scans, few COVID-19 image collections are currently available due to patient privacy. At the same time, there is a rapid growth of COVID-19-relevant articles in the biomedical literature, including those that report findings on radiographs. Here, we present COVID-19-CT-CXR, a public database of COVID-19 CXR and CT images, which are automatically extracted from COVID-19-relevant articles from the PubMed Central Open Access (PMC-OA) Subset. We extracted figures, associated captions, and relevant figure descriptions in the article and separated compound figures into subfigures. Because a large portion of figures in COVID-19 articles are not CXR or CT, we designed a deep-learning model to distinguish them from other figure types and to classify them accordingly. The final database includes 1,327 CT and 263 CXR images (as of May 9, 2020) with their relevant text. To demonstrate the utility of COVID-19-CT-CXR, we conducted four case studies. (1) We show that COVID-19-CT-CXR, when used as additional training data, is able to contribute to improved deep-learning (DL) performance for the classification of COVID-19 and non-COVID-19 CT. (2) We collected CT images of influenza, another common infectious respiratory illness that may present similarly to COVID-19, and trained a baseline deep neural network to distinguish a diagnosis of COVID-19, influenza, or normal or other types of diseases on CT. (3) We trained an unsupervised one-class classifier from non-COVID-19 CXR and performed anomaly detection to detect COVID-19 CXR. (4) From text-mined captions and figure descriptions, we compared 15 clinical symptoms and 20 clinical findings of COVID-19 vs. those of influenza to demonstrate the disease differences in the scientific publications. Our database is unique, as the figures are retrieved along with relevant text with fine-grained descriptions, and it can be extended easily in the future. We believe that our work is complementary to existing resources and hope that it will contribute to medical image analysis of the COVID-19 pandemic. The dataset, code, and DL models are publicly available at https://github.com/ncbi-nlp/COVID-19-CT-CXR.

1 Introduction

The latest threat to global health is the ongoing outbreak of the COVID-19 caused by SARS-CoV-2 (Fauci et al., 2020). So far, pneumonia appears to be the most frequent and serious manifestation, and major complications, such as acute respiratory distress syndrome (ARDS), can present shortly after the onset of symptoms, contributing to the high mortality rate of COVID-19 (Chen et al., 2020b; Guan et al., 2020; Wang et al., 2020a). Chest X-rays (CXR) and chest computed tomography (CT) scans are playing a major part in the detection and monitoring of these respiratory manifestations. In some cases, CT scans have shown abnormal findings in patients prior to the development of symptoms and even before the detection of the viral RNA (Shi et al., 2020b; Xie et al., 2020).

With the shortage of specialists who have been trained to accumulate experiences with COVID-19 diagnosis, there has been a concerted move toward the adoption of artificial intelligence (AI), particularly deep-learning-based methods, in COVID-19 pandemic diagnosis and prognosis, in which well-annotated data always play a critical role (Shi et al., 2020a). Although there exist large public datasets of CXR (Irvin et al., 2019; Johnson et al., 2019; Wang et al., 2017) and CT (Irvin et al., 2019; Johnson et al., 2019; Wang et al., 2017) and CT (Yan et al., 2018), there are few collections of COVID-19 images to effectively train a deep neural network (Cohen et al., 2020; He et al., 2020; Zhang et al., 2020b). Nevertheless, we have seen a growing number of COVID-19 relevant articles in PubMed (Chen et al., 2020c; Wang et al., 2020b). In addition, there is a recent COVID-19 initiative to expand access via PubMed Central Open Access (PMC-OA) Subset to coronavirus-related publications and associated data (https://www.ncbi.nlm.nih.gov/pmc/about/covid-19-faq/). As a result, more articles ($> 10,000$ as of May 9, 2020) relevant to the COVID-19 pandemic or prior coronavirus research were added through PMC-OA with a free-reuse license for secondary analysis.

Non-textual components (e.g., figures and tables) provide key information in many scientific documents and are considered in many tasks, including search engine and knowledge base construction (Choudhury et al., 2013; Smith et al., 2018). As such, we have recently seen a growing interest in mining figures within scientific documents (Ahmed et al., 2016; Li et al., 2019; Siegel et al., 2018). In the medical domain, figures also are a topical interest because they often contain graphical images, such as CXR and CT (Lopez et al., 2013; Tsutsui and Crandall, 2017). Extracting CXR and CT from biomedical publications, however, is neither well studied nor well addressed.

For the above reasons, there is an unmet need to construct the COVID-19 image dataset from PMC-OA to allow researchers to freely access the images along with a description of the text. In this paper, we thus introduce an effective framework to construct a CXR and CT database from PMC-OA and propose a public database, termed COVID-19-CT-CXR. In contrast to previous approaches that relied solely on the manual submission of medical images to the repository, in this work, figures are automatically collected by using the integration of medical imaging and natural-language processing with limited human annotation efforts. In addition, figures in this database are partnered with text that describes these cases with details, a feature not found in other such datasets.

The framework consists of three steps. First, we extracted figures, associated captions, and relevant figure descriptions in the PMC-OA article. Such extraction is non-trivial due to the diverse layout and large volume of articles in the PMC-OA subset. Second, we separated compound figures into subfigures, as medical figures often comprise multiple image panels (Li et al., 2019; Tsutsui and Crandall, 2017). Third, we classified subfigures into CXR, CT, or others because a large portion of figures in COVID-19 articles are not CXR or CT. To this end, we designed a deep-learning model to distinguish them from other figure types and to classify them accordingly.

We further demonstrate the utility of COVID-19-CT-CXR through a series of case studies. First, using this database as additional training data, we show that existing deep neural networks can receive benefits in the task of COVID-19/non-COVID-19 classification of CT images. Second, we demonstrate that the database can be used to develop a baseline model to distinguish COVID-19, influenza, and other CT, a less-studied topic. Third, we train an unsupervised one-class classifier from non-COVID-19 CXRs and performed anomaly detection to detect COVID-19 CXRs. Fourth, we extract symptoms and clinical findings from the text, using the natural language-processing methods. The symptoms and clinical findings not only confirm the results that radiologists have found but also potentially identify other findings that may have been overlooked.

The remainder of the paper is organized as follows. Section 2 presents the material and methods to build the dataset. Section 3 contains the details of the statistics of the dataset, results of the image type classification, and the use cases. Finally, Sections 4 and 5 provide the discussion, conclusions, and recommendations for future work.

2 Material and methods

2.1 COVID-19 relevant articles on PMC-OA

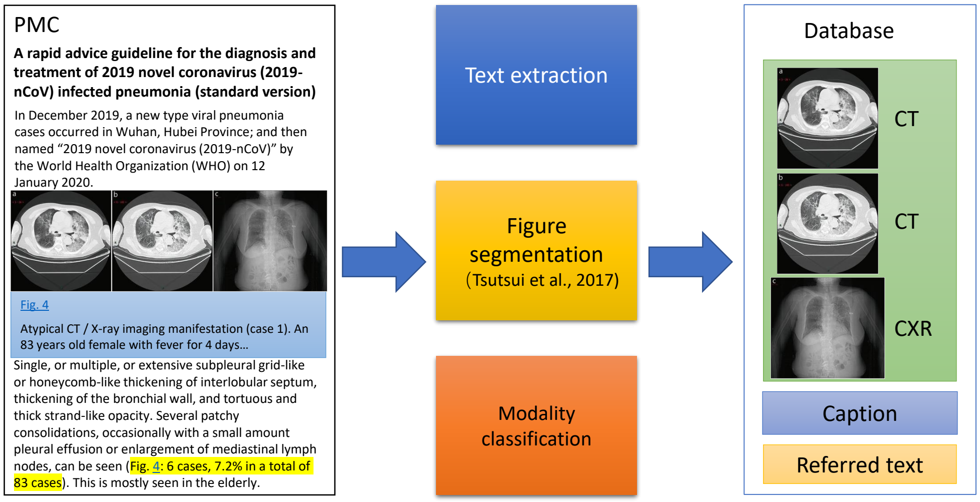
Articles in this study were collected from the PMC-OA Subset. PubMed Central® (PMC) is a free, full-text archive of biomedical and life sciences journal literature (`https://www.ncbi.nlm.nih.gov/pmc/`). PMC-OA is a well-known portion of the PMC articles under a Creative Commons license (or custom license of the Public Health Emergency COVID-19 Initiative in PMC due to the COVID pandemic) that allows for text mining, secondary analysis, and other types of reuse (`https://www.ncbi.nlm.nih.gov/pmc/about/covid-19-faq/`). In this study, we collected COVID-19 relevant articles using LitCovid (Chen et al., 2020c), a curated literature hub for tracking up-to-date scientific information about the 2019 novel coronavirus. LitCovid screens the search results of the PubMed query: `"coronavirus"[All Fields] "ncov"[All Fields] OR "cov"[All Fields] OR "2019-nCoV"[All Fields] OR "COVID-19"[All Fields] OR "SARS-CoV-2"[All Fields]`. Relevant articles are identified and curated with assistance from an automated machine-learning and text-classification algorithm. As of May 9, 2020, there were 5,381 PMC-OA articles in the collection (Table 1). The topics of articles ranged from diagnosis to treatment to case reports.

Table 1: An overview of the COVID-19 relevant articles as of May 9, 2020.

Characteristics	*n*
COVID-19 relevant articles in PMC-OA	5,381
Prevention	2,089
Mechanism	577
Diagnosis	546
Case Report	355
Transmission	354
General	238
Epidemic Forecasting	64
Others	1,158
Journals	1,145
Figures	4,407

2.2 Overview of the COVID-19-CT-CXR construction

Figure 1 shows the overview pipeline of the development. For a given PMC-OA article, we first extract figures, associated captions, and relevant figure descriptions in the PMC-OA article. Then, if figures are compound, we separate them into subfigures. We further classify the individual figures into CT, CXR, or other types of scientific images, using a deep-learning model. The final database includes figures with their types and relevant descriptions in the manuscript.



PMC

A rapid advice guideline for the diagnosis and treatment of 2019 novel coronavirus (2019-nCoV) infected pneumonia (standard version)

In December 2019, a new type viral pneumonia cases occurred in Wuhan, Hubei Province; and then named "2019 novel coronavirus (2019-nCoV)" by the World Health Organization (WHO) on 12 January 2020.

Fig. 4

Atypical CT / X-ray imaging manifestation (case 1). An 83 years old female with fever for 4 days...

Single, or multiple, or extensive subpleural grid-like or honeycomb-like thickening of interlobular septum, thickening of the bronchial wall, and tortuous and thick strand-like opacity. Several patchy consolidations, occasionally with a small amount pleural effusion or enlargement of mediastinal lymph nodes, can be seen (Fig. 4: 6 cases, 7.2% in a total of 83 cases). This is mostly seen in the elderly.

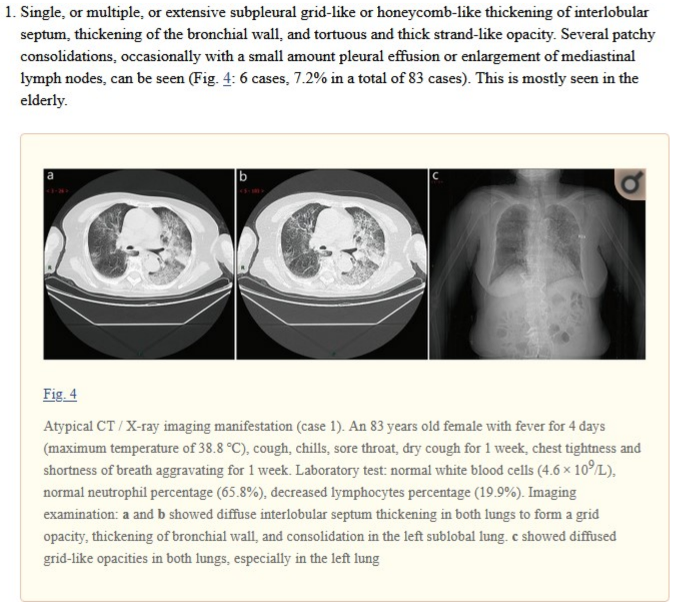
Figure 1: The overview of the pipeline to collect the images with text.

2.3 Text extraction

In this step, we identify figure captions and relevant text with the referenced figures. To facilitate the automated processing of full-text articles in PMC-OA, Comeau et al. (2019) convert PMC articles to BioC format, a data structure in XML for text sharing and processing. Each article in BioC format is encoded in UTF-8, and Unicode characters are converted to strings of ASCII characters. The article also includes section types, figures, tables, and references (Kafkas et al., 2015). In this study, we downloaded the PMC-OA articles through the RESTful web service (`https://www.ncbi.nlm.nih.gov/research/bionlp/APIs/BioC-PubMed/`). We parsed these articles to locate figures with their figure numbers and their captions. We then used the figure number and regular expressions to find where the figure is cross-referenced in the document. Figure 2 shows an example of a typical biomedical image in the article, "A rapid advice guideline for the diagnosis and treatment of 2019 novel coronavirus (2019-nCoV) infected pneumonia (standard version)" (Jin, 2020). The examples contain CXR, CT, a figure caption, and text that describes the case with rich information, such as fever, symptoms, and clinical findings.

2.4 Subfigure separation

Most of the figures in the PMC-OA articles are compound figures. A key challenge here is that one figure may have individual subfigures of the same category (e.g., four CT images) or several categories (e.g., one CXR and one CT image placed side by side). For example, Figure 2 contains a compound figure with three subfigures (Jin, 2020). Figures 2a and Figure 2b are CT images, and Figure 2c is a CXR. Notably, it is a requirement to decompose compound figures into subfigures before modality classification. In this study, we used a convolutional neural network developed by (Tsutsui and Crandall, 2017) to separate compound figures. The model was pretrained on the ImageCLEF Medical dataset with an accuracy of 85.9% (De Herrera et al., 2016).



1. Single, or multiple, or extensive subpleural grid-like or honeycomb-like thickening of interlobular septum, thickening of the bronchial wall, and tortuous and thick strand-like opacity. Several patchy consolidations, occasionally with a small amount pleural effusion or enlargement of mediastinal lymph nodes, can be seen (Fig. 4: 6 cases, 7.2% in a total of 83 cases). This is mostly seen in the elderly.

Fig. 4

Atypical CT / X-ray imaging manifestation (case 1). An 83 years old female with fever for 4 days (maximum temperature of 38.8 °C), cough, chills, sore throat, dry cough for 1 week, chest tightness and shortness of breath aggravating for 1 week. Laboratory test: normal white blood cells (4.6 × 10^9/L), normal neutrophil percentage (65.8%), decreased lymphocytes percentage (19.9%). Imaging examination: **a** and **b** showed diffuse interlobular septum thickening in both lungs to form a grid opacity, thickening of bronchial wall, and consolidation in the left sublobal lung. **c** showed diffused grid-like opacities in both lungs, especially in the left lung

Figure 2: Examples of CT and CXR that are positive for COVID-19. The figures are from the article, "A rapid advice guideline for the diagnosis and treatment of 2019 novel coronavirus (2019-nCoV) infected pneumonia (standard version)" (Jin, 2020).

We applied the model on the figures obtained in previous steps and filtered the subfigures with a size smaller than 224 x 224 pixels. We consider that subfigures with fewer pixels might be deformed, and most state-of-the-art neural networks in image analysis, such as Inception-v3 (Szegedy et al., 2016) and DenseNet (Iandola et al., 2014), require an input size of 224 or larger.

2.5 Image modality classification

A large portion of figures in the PMC-OA articles are not CXR or CT images. To distinguish them from other types of scientific figures, we designed a scientific figure classifier that was trained on a newly created dataset (`https://github.com/ncbi-nlp/COVID-19-CT-CXR`). Table 2 shows the breakdown of the figures by their category in the training and test set. This dataset consists of 2,700 figures in three categories: CXR, CT, and Other scientific figure types. A total of 500 CXRs are randomly picked from the NIH Chest X-ray (Wang et al., 2017), and 500 CT images are randomly picked from DeepLesion (Yan et al., 2018). Other scientific figures are randomly picked from DocFigure (Jobin et al., 2019). The original DocFigure annotated figures of 28 categories, such as Heat map, Bar plots, and Histogram. Here, we combined these categories into one for simplicity of training the classifier. In addition, we curated 1,200 figures from PMC-OA, using the annotation tool developed by Tang et al. (2020).

Our framework uses DenseNet121 to classify image types (Huang et al., 2016). The weights (or parameters) were pretrained on ImageNet (Russakovsky et al., 2015). We replaced the last classification layer with a fully connected layer with a softmax operation that outputs the approximate probability that an input image is a CXR, CT, or other scientific figure type. All images were resized to 224 x 224 pixels. The hyperparameters include a learning rate of 0.0001, a batch size of 16, and 50 training epochs. All experiments were conducted on a server with an NVIDIA V100 128G GPU from the NIH HPC Biowulf cluster (`http://hpc.nih.gov`). We implemented the framework using the Keras deep-learning library with

Table 2: Summary of the dataset for image modality classification.

Modality	Training	Test
CXR		
NIH Chest X-ray (Wang et al., 2017)	399	101
PMC-OA	38	7
CT		
DeepLesion (Yan et al., 2018)	415	85
PMC-OA	225	21
Other scientific document figures		
DocFigure (Jobin et al., 2019)	386	114
PMC-OA	737	172
Total	2,200	500

TensorFlow backend (https://www.tensorflow.org/guide/keras).

2.6 Qualification and statistical analysis

The performance metrics include the area under the receiver operating characteristic curve (AUC), sensitivity, specificity (recall), precision (positive predictive value), and F1 score. For the classification problem, we chose the label with the highest probability when required in computing the metrics. Each of the models was trained and tested five times, using the same parameters, training, and testing images each time. The validation set was randomly selected from 10% of the training set. Fisher's exact test was used to determine whether there are nonrandom associations between COVID-19 and influenza's symptoms and clinical findings (Fisher, 1922). We conduct above statistical analysis using numpy, scipy, matplotlib, and scikit-learn built on Python.

3 Results

3.1 COVID-19-CT-CXR characteristics

Table 3 shows the breakdown of the figures by modality. We obtained 1,327 CT images and 263 CXR text-mined labeled as positive for COVID-19 from 1,831 PMC-OA articles. These images have different sizes. The minimum, maximum, and average heights are 224, 2,703, and 387.5 pixels, respectively. The minimum, maximum, and average widths are 224, 1,961, and 472.4, respectively. For each article, we also include major elements, such as DOI, title, journal, and publication date for reference. Figure 3 shows the cumulative numbers of articles and figures on a weekly basis.

3.2 Image modality classification

Table 4 shows the performance of the model to classify image modality. The macro average F-score is 0.996. The F-score was 0.993 0.004 for CT, 1.000 0.000 for CXR, and 0.998 0.001 for other scientific figure types.

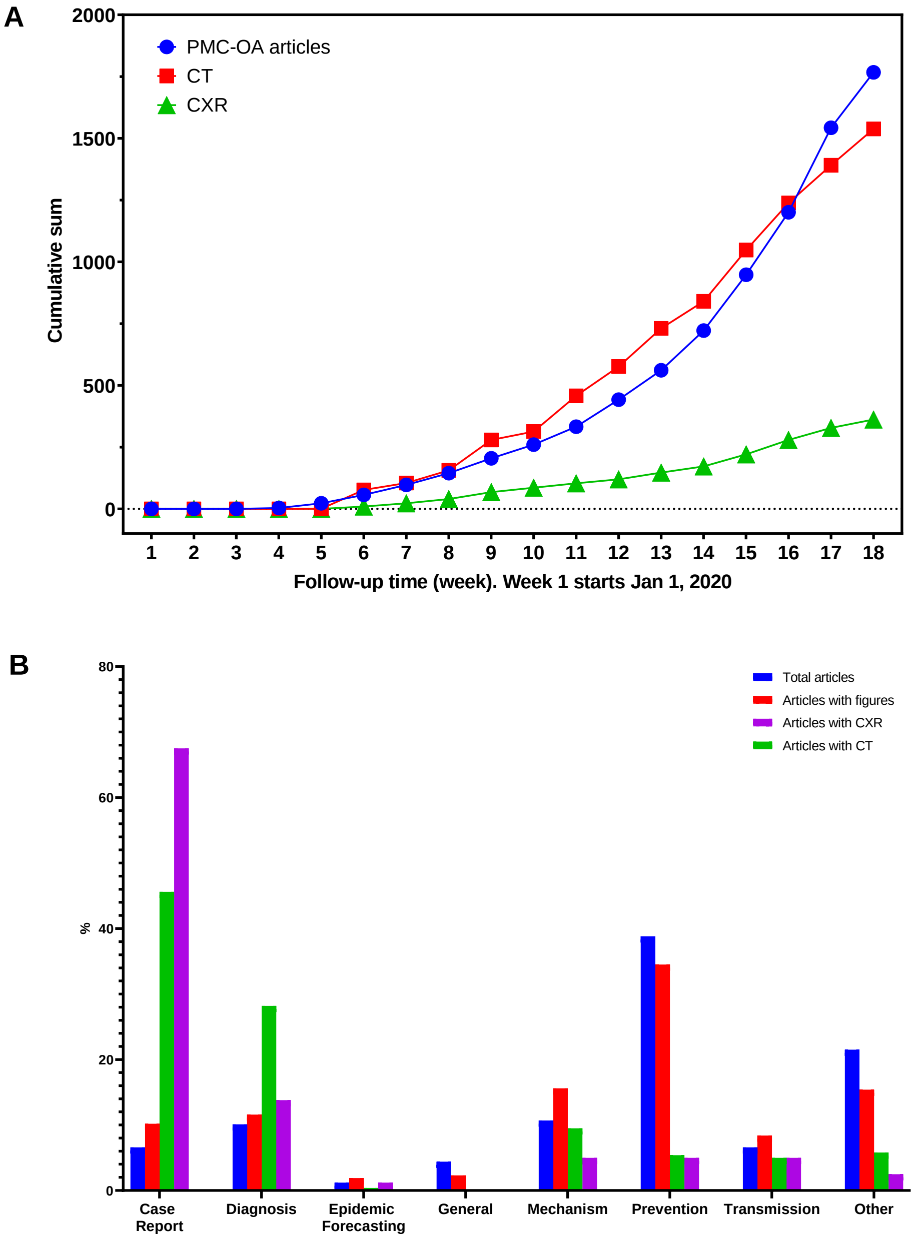


Figure 3: Characteristics of the COVID-19-CT-CXR. (A) The rapid growth of the number of COVID-19-relevant articles, CT, and CXR in PMC-OA from January 1, 2020 (Week 1). (B) The distribution of categories in COVID-19-relevant PMC-OA articles and articles with figures, CT, and CXR.

Table 3: Summary of the COVID-19-CT-CXR dataset

Characteristics	*n*
PMC-OA articles with figures	1,831
Subfigures	10,650
CXR	263
CT	1,327
Others	9,060

Table 4: The performance of image type classification. The test set is the combination of NIH Chest X-ray, DeepLesion, DocFigure, and PMC-OA.

Metrics	CT	CXR	Other scientific figures	*Macro Avg*
Precision	0.989 0.004	1.000 0.000	0.999 0.001	0.996 0.002
Recall/Sensitivity	0.998 0.004	1.000 0.000	0.996 0.001	0.998 0.002
Specificity	0.997 0.001	1.000 0.000	0.999 0.002	0.999 0.001
F-score	0.993 0.004	1.000 0.000	0.998 0.001	0.997 0.002

3.3 Use cases

To demonstrate the utility of COVID-19-CT-CXR, we conducted four case studies. (1) We combined COVID-19-CT-CXR with previously curated data at `https://github.com/UCSD-AI4H/COVID-CT` (He et al., 2020) and trained a deep neural network to perform the classification of COVID-19 and non-COVID-19 CT. (2) We collected CT of influenza, using a similar method, and trained a deep neural network to distinguish among the diagnoses of COVID-19, influenza, and normal or other types of diseases on CT. (3) We trained an unsupervised one-class learning model, using only non-COVID-19 CXR to perform anomaly detection, to detect COVID-19 CXR. (4) We extracted 15 clinical symptoms and 26 clinical findings from the captions and relevant descriptions. We then compared their frequencies to those described in articles on influenza, another common infectious respiratory illness that may present similarly to COVID-19.

3.3.1 Classification of COVID-19 and non-COVID-19 on CT

In the context of the COVID-19 pandemic, it is important to separate patients likely to be infected with COVID-19 from other non-COVID-19 patients. As it is time-consuming for specialists to both accumulate experiences and read a large volume of CT scans to diagnose COVID-19, many studies use machine learning to separate COVID-19 patients from non-COVID-19 patients (Chen et al., 2020a; He et al., 2020; Jin et al., 2020; Wang et al., 2020c; Zheng et al., 2020). In this work, we hypothesize that our creation of additional training data from existing articles can improve the performance of the system and reduce the effort of manual image annotation. To test this hypothesis, we compared the performance of deep neural networks trained on the existing benchmark (He et al., 2020) and COVID-19-CT-CXR (Supplementary Table S1). For a fair comparison, we added additional training examples only in the training set and used the same test set as described in (He et al., 2020).

In this experiment, DenseNet121 was pre-trained on ImageNet, fine-tuned, and evaluated on the training and test sets. We then replaced the last classification layer with a single neuron with sigmoid that outputs the approximate probability that an input image is COVID-19 or non-COVID-19. Other experimental settings are the same as that of training the image modality classifier. Figure 4 shows that the model significantly outperforms the baseline when PMC-OA CT figures were added for training. Specifically, we achieved the highest performance of 0.891 0.012 in AUC, 0.780 0.074 in recall, 0.816 0.053 in precision, and 0.792 0.015 in *F*-score (Supplementary Table S2).

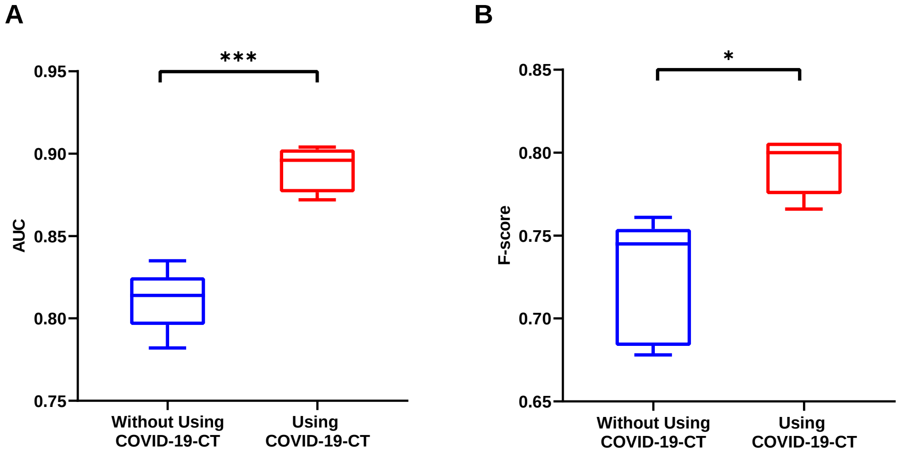


Figure 4: Comparison of AUC and *F*-score by models fine-tuned with and without using additional COVID-19 CT extracted from PMC-OA. *: $P \leq 0.05$; ***: $P \leq 0.001$ (t-test).

3.3.2 Classification of COVID-19, influenza, and other types of disease on CT

As the COVID-19 outbreak continues to evolve, there is an increasing number of studies that compare COVID-19 with other viral pneumonias, such as influenza (Luo et al., 2020). Distinguishing patients infected by COVID-19 and influenza is important for public health measures because the current treatment guidelines are different (Kimberlin, 2018). This task is non-trivial because both viruses have a similar radiological presentation. To assist clinicians at triage, several studies have proposed to use deep learning to distinguish COVID-19 from influenza and no-infection with 3D CT scans (Xu et al., 2020). In this paper, we aim to establish a baseline model to distinguish COVID-19 from influenza on single CT figures. To collect CT figures with influenza, we searched the PMC using the query "`(Influenza[Title] OR (flu[Title] AND pneumonia[Title]) AND open access[Filter]`" and extracted the most recent 10,000 PMC-OA articles. We used the same method to extract CT and its caption and relevant text from the articles (called Influenza-CT). Taken together, we construct a dataset with 983 CT for training and 242 CT for testing (Supplementary Table S3).

To obtain the baseline model, we use the same model and experimental settings as described in the "Image modality classification" section. Figure 5 shows the performance of the deep-learning model by its receiver operating characteristic (ROC) curves. The AUC was 0.855 0.012 for COVID-19 detection and 0.889 0.014 for influenza detection. Supplementary Table S4 shows more detail for the results. We achieved the highest precision (0.845 0.026) for COVID-19 detection and high recall (0.711 0.053) for influenza detection.

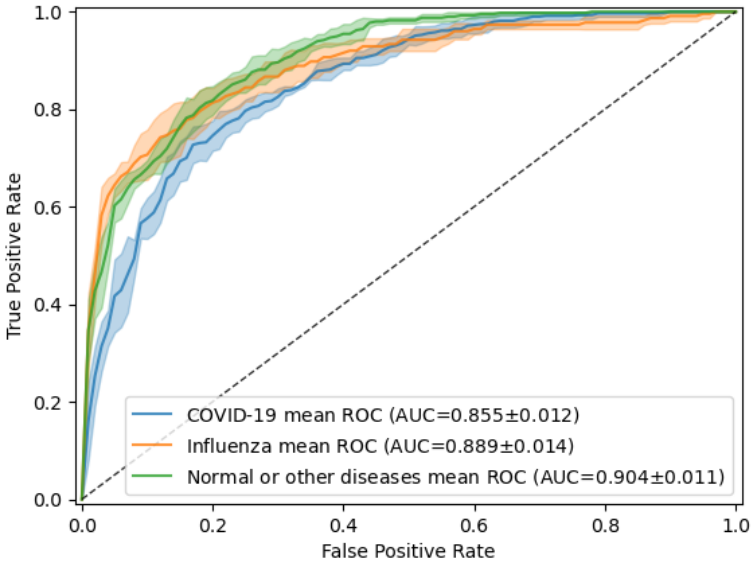


Figure 5: Receiver operating characteristic (ROC) curves of the classification of COVID-19, influenza, and normal or other types of diseases in CT. The model was trained and tested 5 times, using the same training and testing images each time. The mean ROC curve is shown together with its standard deviation (shaded area).

3.3.3 Anomaly detection of COVID-19 in CXR using one-class learning

As they lack annotated COVID-19 CXR for training powerful deep-learning classifiers, unsupervised and semi-supervised approaches are highly desired for automated COVID-19 diagnosis. The presence of COVID-19 can be considered a novel anomaly in CXR for the NIH Chest X-ray dataset, in which no COVID-19 cases are available. In this experiment, we performed anomaly detection (Chandola et al., 2009; Zhang et al., 2020a) to detect COVID-19 CXR. We trained a one-class classifier, using only non-COVID-19 CXR, and used this classifier to distinguish COVID-19 CXR from non-COVID-19 CXR. The non-COVID-19 images were a subset extracted from the NIH Chest X-ray dataset by combining 14 abnormalities and a no-finding category. The detailed numbers of training and testing CXR are shown in Supplementary Table S5. We adopted the generative adversarial one-class learning approach from (Tang et al., 2019). Figure 6 shows the performance of the unsupervised one-class learning by its ROC curves. Supplementary Table S6 shows more detail for the results. Our model achieved 0.828 0.019 in AUC, 0.767 0.020 in precision, 0.772 0.017 in recall, and 0.769 0.018 in *F*-score for COVID-19 anomaly detection.

3.3.4 Extraction of clinical symptoms and findings using text-mining

In this case, we extracted clinical symptoms or signs from the figure captions and relevant text that describes the case. A total of 15 symptoms or signs were collected from (Guan et al., 2020) and the CDC website (https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html), including chest pain, constipation, cough, diarrhea, dizziness, dyspnea, fatigue, fever, headache, myalgia, proteinuria, runny nose, sputum production, throat pain, and vomiting.

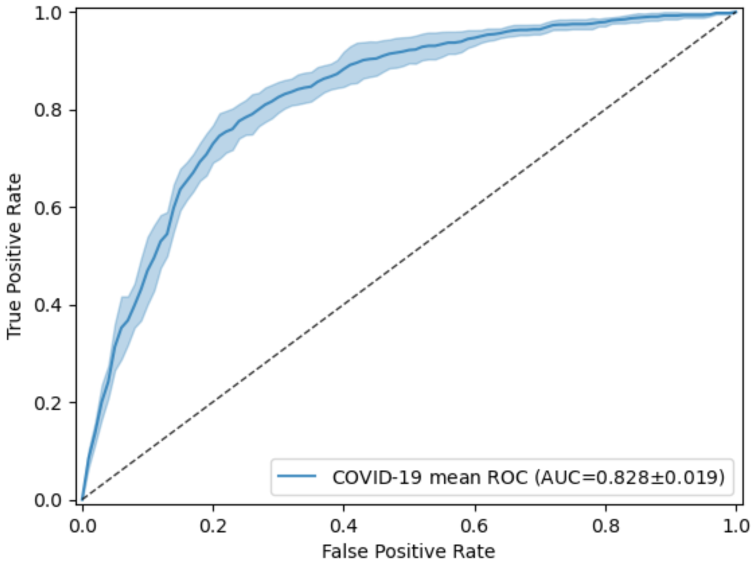


Figure 6: Receiver operating characteristic (ROC) curves of the classification of COVID-19 anomaly detection in CXR. The model was trained and tested 5 times, using the same training and testing images each time. The mean ROC curve is shown together with its standard deviation (shaded area).

Extracting these symptoms from text is a challenging task because their mentions in the text can be positive or negative. For example, "fever" is negative in the sentence, "She experienced headache and pharyngalgia but no fever on 29 January." To discriminate between positive and negative mentions, we applied our previously developed tool, NegBio, on the figure caption and referred text (Peng et al., 2018). In short, NegBio utilizes patterns in universal dependencies to identify the scope of triggers that are indicative of negation; thus, it is highly accurate for detecting negative symptom mentions. Figure 7A shows the proportion of symptoms for COVID-19 and influenza. The most common symptoms are fever, cough, dyspnea, and myalgia.

We then extracted the radiographic findings from the figure caption and text. The findings (and their synonyms) are based on 20 common thoracic disease types, which are expanded from NIH Chest X-ray 14 labels (Wang et al., 2017). Figure 7B shows the 20 findings in both COVID-19 and influenza datasets. Both illnesses can result in lung opacity, pneumonia, and consolidation. COVID-19 more likely results in ground-glass opacification (GGO), while influenza more likely results in infiltration than does COVID-19 (Fisher's exact test, $p < 0.0001$).

4 Discussion

In this abrupt outbreak of SARS-CoV-2, the demand for chest radiographs and CT scans is growing rapidly, but there is a shortage of experienced specialists, radiologists, and researchers. Further, we are still new to this virus and have yet to discover the full radiologic features and prognosis of this disease. The tremendous increase in the number of patients has led to a substantial increase of COVID-19-related PMC-OA articles over the past few months (Figur 3A), especially in the case report and diagnosis-relevant articles (Figure 3B).

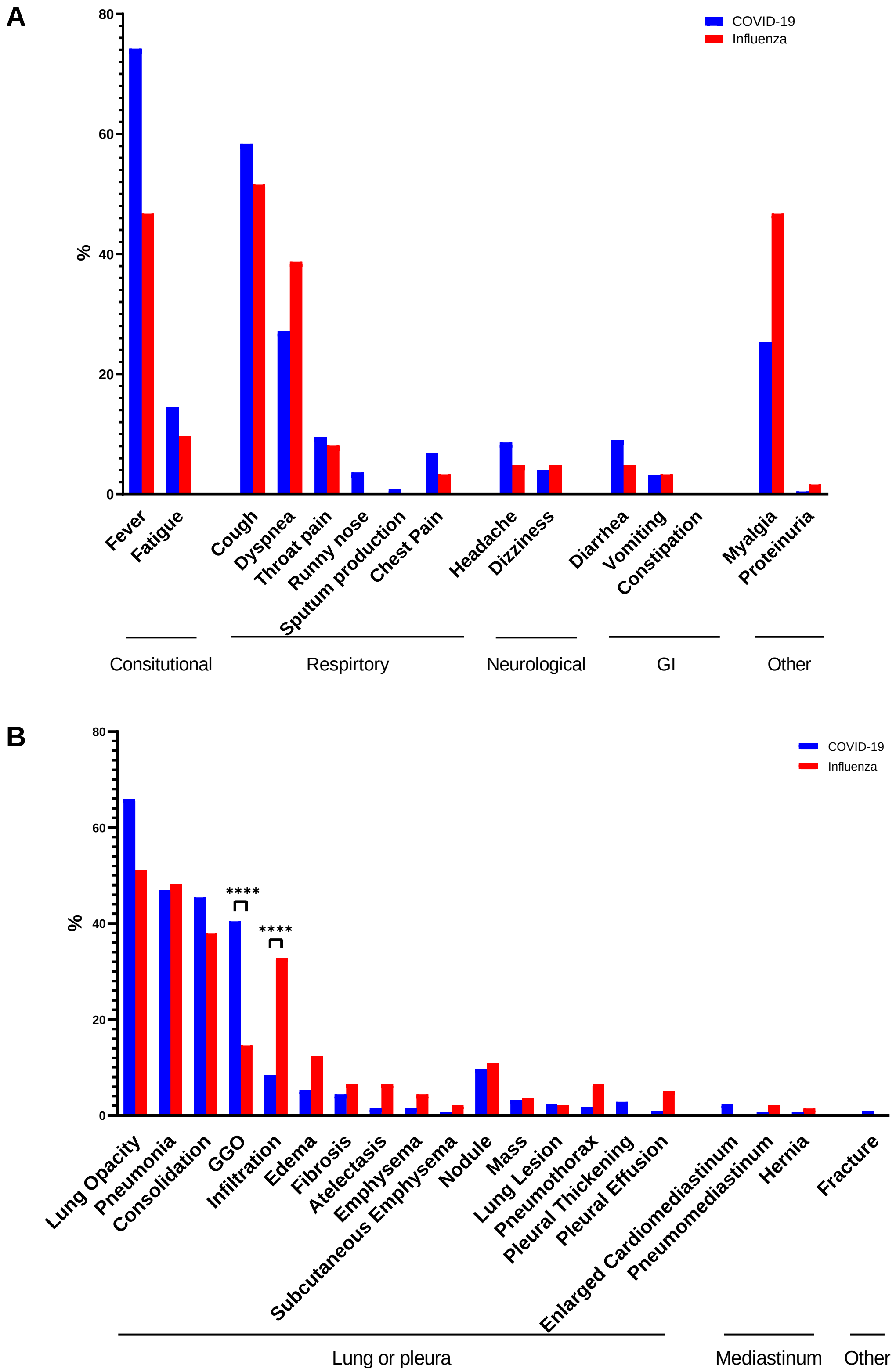


Figure 7: The frequencies of (A) 15 symptoms and (B) 20 clinical findings text mined from the figure captions and relevant text from the collection of COVID-19- and influenza-relevant articles. ****: $p \leq 0.0001$ (Fisher exact test).

These articles contain rich chest radiographs and CT images that are helpful for scientists and clinicians in describing COVID-19 cases. Thus, it is important to analyze these images and text to construct a large-scale database. By using the quickly increasing dataset, AI methods can help to find significant features of COVID-19 and speed up the clinical workload. Among others, deep learning is undoubtedly a powerful approach in dealing with a pandemic outbreak of COVID-19.

Although deep learning has shown promise in diagnosing/screening COVID-19, using CT, it remains difficult to collect large-scale labeled imaging data, especially in the public domain. In this work, we present a set of repeatable techniques to rapidly build a CT and CXR dataset of COVID-19 from PMC-OA COVID-19-relevant articles. The strength of the study lies in its multidisciplinary integration of medical imagining and natural-language processing. It provides a new way to annotate large-scale medical images required by deep-learning models.

An additional strength includes a highly accurate model for image type classification. As a large portion of figures in the PMC-OA articles are not CXR or CT images, we provided a model to classify these two types from other scientific figure types. Our model achieved both high precision and high recall (Table 4).

To assess the hypothesis that deep neural network training on this additional dataset enables us to diagnose COVID-19 with almost no hand-labeled data, we conducted several experiments. First, we showed that this additional data enable significant performance gains to classify COVID-19 versus non-COVID-19 lung infection on CT (Figure 4 and Supplementary Table S2). For our own system, we show that our baseline performance compares favorably to the results in (He et al., 2020). Then, we added more automatically labeled training data and achieved the highest performance of 0.891 0.012 in AUC. The comparison shows that, with additional data, both precision and recall substantially improve (7.4% and 6.6%, respectively). This observation indicates that additional COVID-19 CT helps to not only find more but also to restrict the positive predictions to those with the highest certainty in the model.

In a more challenging scenario, we built a baseline system to distinguish COVID-19, influenza, and no-infection CT, which is a more clinically interesting but also more challenging task. We observed that we could achieve high AUCs for both COVID-19 and influenza detection. The recall of COVID-19 detection and the precision of influenza, however, are low (0.597 0.030 and 0.609 0.033, respectively). Although several studies have tackled this problem (Xu et al., 2020), to the best of our knowledge, there is no publicly available benchmarking. Although our work only scratches the surface of the classification of COVID-19, influenza, and normal or other types of diseases, we hope that it sheds light on the development of generalizable deep-learning models that can assist frontline radiologists.

In addition, we presented a one-class learning model for anomaly detection of COVID-19 in CXR by learning only from non-COVID-19 radiographs. Compared to the CT-based method, the one-class model achieves comparable performance, showing great potential in discriminating COVID-19 from CXR. The performance of our model, however, is worse than that of Zhang et al. (2020a), suggesting that this weakly labeled dataset should be used as additional training data obtained without additional annotation cost from existing entries in curated databases.

The unique characteristic of our database is that figures are retrieved along with relevant text that describes these cases in detail. Thus, text mining can be applied to extract additional information that confirms the existing results and potentially identifies other findings that may have been overlooked. As proof of this concept, we extracted clinical symptoms and findings from the text. We found that the most common symptoms of COVID-19 were fever and cough (Figure 7A), which are consistent with the clinical characteristics in (Zhang et al., 2020b). Other common symptoms include dyspnea (shortness of breath), fatigue, and throat pain. These symptoms are consistent with those reported by the CDC. When comparing the frequencies of these 20 clinical findings to those described in articles on influenza, Figure 7 shows that both conditions cause lung opacity, pneumonia, and consolidation. Further, GGO appears more frequently for COVID-19, whereas "infiltration" appears more frequently for influenza. This is because radiologists use the term *GGO* to describe most COVID-19 findings. In addition, the influenza articles are older than are the COVID articles, and, according to Fleischner Society recommendations, the use of the term *infiltrate* remains controversial, and it is recommended that it no longer be used in reports (Bueno et al., 2018).

In terms of limitations, first, the subfigure segmentation model needs to be improved. In this study, we applied a deep-learning model that was pretrained on an ImageCLEF Medical dataset to this task (Tsutsui and Crandall, 2017). Although this model is robust to variations in background color and spaces between subfigures, it sometimes fails to recognize similar subfigures that are aligned very closely. Unfortunately, these cases appear more frequently in our study than in others (e.g., several CT images are placed in a grid). Other errors occur when the model incorrectly treated the spine as spaces in the anteroposterior (AP) chest X-ray and split the large figure into two subfigures. In the future, the figure synthesis approach should be applied to augment the training datasets. Another limitation is that this work extracted only the passage that contains the referred figure. Sometimes, the case is not described in this passage. In the future, we plan to text mine the associated case description in the full text.

5 Conclusions

We have developed a framework for rapidly constructing a CXR/CT database from PMC full-text articles. Our database is unique, as figures are retrieved along with relevant text that describes these cases in detail, and it can be extended easily in the future. Hence, the work is complementary to existing resources. Applications of this database show that our creation of additional training data from existing articles improves the system performance on COVID-19 vs. non-COVID-19 classification in CT and CXR. We hope that the public dataset can facilitate deep-learning model development, educate medical students and residents, help to evaluate findings reported by radiologists, and provide additional insights for COVID-19 diagnosis. With an ongoing commitment to data sharing, we anticipate increasingly adding CXR and CT images to be made available as well in the coming months. The code that extracts the text from PMC, segments subfigures, and classifies image modality is openly available at https://github.com/ncbi-nlp/COVID-19-CT-CXR.

Acknowledgments

This research was supported in part by the Intramural Research Programs of the National Library of Medicine (NLM) and National Institutes of Health (NIH) Clinical Center. It also was supported by NLM under Award No. K99LM013001. This work utilized the computational resources of the NIH HPC Biowulf cluster (http://hpc.nih.gov).

References

Z. Ahmed, S. Zeeshan, and T. Dandekar. Mining biomedical images towards valuable information retrieval in biomedical and life sciences. *Database : the journal of biological databases and curation*, 2016, 2016. ISSN 1758-0463. doi: 10.1093/database/baw118.

J. Bueno, L. Landeras, and J. H. Chung. Updated fleischner society guidelines for managing incidental pulmonary nodules: Common questions and challenging scenarios. *Radiographics : a review publication of the Radiological Society of North America, Inc*, 38:1337–1350, 2018. ISSN 1527-1323. doi: 10.1148/rg.2018180017.

V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection. *ACM Computing Surveys*, 41(3):1–58, July 2009. doi: 10.1145/1541880.1541882.

J. Chen, L. Wu, J. Zhang, L. Zhang, D. Gong, Y. Zhao, S. Hu, Y. Wang, X. Hu, B. Zheng, K. Zhang, H. Wu, Z. Dong, Y. Xu, Y. Zhu, X. Chen, L. Yu, and H. Yu. Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography: a prospective study. *medrxiv preprint*, Feb. 2020a. doi: 10.1101/2020.02.25.20021568.

N. Chen, M. Zhou, X. Dong, J. Qu, F. Gong, Y. Han, Y. Qiu, J. Wang, Y. Liu, Y. Wei, J. Xia, T. Yu, X. Zhang, and L. Zhang. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in wuhan, china: a descriptive study. *Lancet (London, England)*, 395:507–513, Feb. 2020b. ISSN 1474-547X. doi: 10.1016/S0140-6736(20)30211-7.

Q. Chen, A. Allot, and Z. Lu. Keep up with the latest coronavirus research. *Nature*, 579:193, Mar. 2020c. ISSN 1476-4687. doi: 10.1038/d41586-020-00694-1.

S. R. Choudhury, S. Tuarob, P. Mitra, L. Rokach, A. Kirk, S. Szep, D. Pellegrino, S. Jones, and C. L. Giles. A figure search engine architecture for a chemistry digital library. In *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries - JCDL '13*, pages 369–370. ACM Press, 2013. doi: 10.1145/2467696.2467757.

J. P. Cohen, P. Morrison, and L. Dao. Covid-19 image data collection. *arxiv preprint*, 2020.

D. C. Comeau, C.-H. Wei, R. Islamaj Doğan, and Z. Lu. Pmc text mining subset in bioc: about three million full-text articles and growing. *Bioinformatics (Oxford, England)*, 35:3533–3535, Sept. 2019. ISSN 1367-4811. doi: 10.1093/bioinformatics/btz070.

A. G. S. De Herrera, S. Bromuri, R. Schaer, and H. Müller. Overview of the medical tasks in imageclef 2016. *CLEF Working Notes. Evora, Portugal*, 2016.

A. S. Fauci, H. C. Lane, and R. R. Redfield. Covid-19 - navigating the uncharted. *The New England journal of medicine*, 382:1268–1269, Mar. 2020. ISSN 1533-4406. doi: 10.1056/NEJMe2002387.

R. A. Fisher. On the interpretation of χ 2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85(1):87, jan 1922. doi: 10.2307/2340521.

W.-J. Guan, Z.-Y. Ni, Y. Hu, W.-H. Liang, C.-Q. Ou, J.-X. He, L. Liu, H. Shan, C.-L. Lei, D. S. C. Hui, B. Du, L.-J. Li, G. Zeng, K.-Y. Yuen, R.-C. Chen, C.-L. Tang, T. Wang, P.-Y. Chen, J. Xiang, S.-Y. Li, J.-L. Wang, Z.-J. Liang, Y.-X. Peng, L. Wei, Y. Liu, Y.-H. Hu, P. Peng, J.-M. Wang, J.-Y. Liu, Z. Chen, G. Li, Z.-J. Zheng, S.-Q. Qiu, J. Luo, C.-J. Ye, S.-Y. Zhu, N.-S. Zhong, and C. M. T. E. G. for Covid-19. Clinical characteristics of coronavirus disease 2019 in china. *The New England journal of medicine*, 382: 1708–1720, Apr. 2020. ISSN 1533-4406. doi: 10.1056/NEJMoa2002032.

X. He, X. Yang, S. Zhang, J. Zhao, Y. Zhang, E. Xing, and P. Xie. Sample-efficient deep learning for COVID-19 diagnosis based on CT scans. *medrxiv preprint*, Apr. 2020. doi: 10.1101/2020.04.13.20063941.

G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. *arxiv preprint*, 2016.

F. Iandola, M. Moskewicz, S. Karayev, R. Girshick, T. Darrell, and K. Keutzer. Densenet: Implementing efficient convnet descriptor pyramids. *arxiv preprint*, 2014.

J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, and H. Marklund. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597, 2019.

C. Jin, W. Chen, Y. Cao, Z. Xu, X. Zhang, L. Deng, C. Zheng, J. Zhou, H. Shi, and J. Feng. Development and evaluation of an AI system for COVID-19 diagnosis. *medrxiv preprint*, mar 2020. doi: 10.1101/2020.03.20.20039834.

Y.-H. e. a. Jin. A rapid advice guideline for the diagnosis and treatment of 2019 novel coronavirus (2019-ncov) infected pneumonia (standard version). *Military Medical Research*, 7:4, Feb. 2020. ISSN 2054-9369. doi: 10.1186/s40779-020-0233-6.

K. V. Jobin, A. Mondal, and C. V. Jawahar. DocFigure: A dataset for scientific document figure classification. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, pages 74–79. IEEE, Sept. 2019. doi: 10.1109/icdarw.2019.00018.

A. E. W. Johnson, T. J. Pollard, N. R. Greenbaum, M. P. Lungren, C. ying Deng, Y. Peng, Z. Lu, R. G. Mark, S. J. Berkowitz, and S. Horng. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv preprint*, 2019.

Ş. Kafkas, X. Pi, N. Marinos, F. Talo', A. Morrison, and J. R. McEntyre. Section level search functionality in europe pmc. *Journal of biomedical semantics*, 6:7, 2015. ISSN 2041-1480. doi: 10.1186/s13326-015-0003-7.

D. Kimberlin. *Red book : 2018-2021 report of the Committee on Infectious Diseases*. American Academy of Pediatrics, Elk Grove Village, IL, 2018. ISBN 9781610021463.

P. Li, X. Jiang, and H. Shatkay. Figure and caption extraction from biomedical documents. *Bioinformatics (Oxford, England)*, 35:4381–4388, Nov. 2019. ISSN 1367-4811. doi: 10.1093/bioinformatics/btz228.

L. D. Lopez, J. Yu, C. Arighi, C. O. Tudor, M. Torii, H. Huang, K. Vijay-Shanker, and C. Wu. A framework for biomedical figure segmentation towards image-based document retrieval. *BMC systems biology*, 7 Suppl 4:S8, 2013. ISSN 1752-0509. doi: 10.1186/1752-0509-7-S4-S8.

Y. Luo, X. Yuan, Y. Xue, L. Mao, Q. Lin, G. Tang, H. Song, W. Liu, H. Hou, F. Wang, and Z. Sun. Using a diagnostic model based on routine laboratory tests to distinguish patients infected with sars-cov-2 from those infected with influenza virus. *International journal of infectious diseases : IJID : official publication of the International Society for Infectious Diseases*, 95:436–440, May 2020. ISSN 1878-3511. doi: 10.1016/j.ijid.2020.04.078.

Y. Peng, X. Wang, L. Lu, M. Bagheri, R. Summers, and Z. Lu. Negbio: a high-performance tool for negation and uncertainty detection in radiology reports. *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, 2017:188–196, 2018. ISSN 2153-4063. URL https://arxiv.org/abs/1712.05898.

O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

F. Shi, J. Wang, J. Shi, Z. Wu, Q. Wang, Z. Tang, K. He, Y. Shi, and D. Shen. Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for covid-19. *IEEE reviews in biomedical engineering*, Apr. 2020a. ISSN 1941-1189. doi: 10.1109/RBME.2020.2987975.

H. Shi, X. Han, N. Jiang, Y. Cao, O. Alwalid, J. Gu, Y. Fan, and C. Zheng. Radiological findings from 81 patients with covid-19 pneumonia in wuhan, china: a descriptive study. *The Lancet. Infectious diseases*, 20:425–434, Apr. 2020b. ISSN 1474-4457. doi: 10.1016/S1473-3099(20)30086-4.

N. Siegel, N. Lourie, R. Power, and W. Ammar. Extracting scientific figures with distantly supervised neural networks. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, pages 223–232. ACM, may 2018. doi: 10.1145/3197026.3197040.

C. L. Smith, J. A. Blake, J. A. Kadin, J. E. Richardson, C. J. Bult, and M. G. D. Group. Mouse genome database (mgd)-2018: knowledgebase for the laboratory mouse. *Nucleic acids research*, 46:D836–D842, Jan. 2018. ISSN 1362-4962. doi: 10.1093/nar/gkx1006.

C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016.

Y.-X. Tang, Y.-B. Tang, M. Han, J. Xiao, and R. M. Summers. Abnormal Chest X-Ray Identification With Generative Adversarial One-Class Classifier. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 1358–1361, Venice, Italy, Apr. 2019. IEEE. ISBN 978-1-5386-3641-1. doi: 10.1109/ISBI.2019.8759442.

Y.-X. Tang, Y.-B. Tang, Y. Peng, K. Yan, M. Bagheri, B. A. Redd, C. J. Brandon, Z. Lu, M. Han, J. Xiao, and R. M. Summers. Automated abnormality classification of chest radiographs using deep convolutional neural networks. *NPJ digital medicine*, 3:70, 2020. ISSN 2398-6352. doi: 10.1038/s41746-020-0273-z.

S. Tsutsui and D. Crandall. A data driven approach for compound figure separation using convolutional neural networks. *arxiv preprint*, 2017.

D. Wang, B. Hu, C. Hu, F. Zhu, X. Liu, J. Zhang, B. Wang, H. Xiang, Z. Cheng, Y. Xiong, Y. Zhao, Y. Li, X. Wang, and Z. Peng. Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus-infected pneumonia in wuhan, china. *JAMA*, Feb. 2020a. ISSN 1538-3598. doi: 10.1001/jama.2020.1585.

L. L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Eide, K. Funk, R. Kinney, Z. Liu, W. Merrill, P. Mooney, D. Murdick, D. Rishi, J. Sheehan, Z. Shen, B. Stilson, A. D. Wade, K. Wang, C. Wilhelm, B. Xie, D. Raymond, D. S. Weld, O. Etzioni, and S. Kohlmeier. Cord-19: The covid-19 open research dataset. *arxiv preprint*, 2020b.

S. Wang, B. Kang, J. Ma, X. Zeng, M. Xiao, J. Guo, M. Cai, J. Yang, Y. Li, X. Meng, and B. Xu. A deep learning algorithm using CT images to screen for corona virus disease (COVID-19). *medrxiv preprint*, feb 2020c. doi: 10.1101/2020.02.14.20023028.

X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers. ChestX-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2097–2106. IEEE, July 2017. doi: 10.1109/cvpr.2017.369.

X. Xie, Z. Zhong, W. Zhao, C. Zheng, F. Wang, and J. Liu. Chest ct for typical 2019-ncov pneumonia: Relationship to negative rt-pcr testing. *Radiology*, page 200343, Feb. 2020. ISSN 1527-1315. doi: 10.1148/radiol.2020200343.

X. Xu, X. Jiang, C. Ma, P. Du, X. Li, S. Lv, L. Yu, Y. Chen, J. Su, G. Lang, Y. Li, H. Zhao, K. Xu, L. Ruan, and W. Wu. Deep learning system to screen coronavirus disease 2019 pneumonia. *arxiv preprint*, 2020.

K. Yan, X. Wang, L. Lu, and R. M. Summers. DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *Journal of medical imaging (Bellingham, Wash.)*, 5:036501, July 2018. ISSN 2329-4302. doi: 10.1117/1.JMI.5.3.036501.

J. Zhang, Y. Xie, Y. Li, C. Shen, and Y. Xia. Covid-19 screening on chest x-ray images using deep learning based anomaly detection. *arxiv preprint*, 2020a.

K. Zhang, X. Liu, J. Shen, Z. Li, Y. Sang, X. Wu, Y. Zha, W. Liang, C. Wang, K. Wang, L. Ye, M. Gao, Z. Zhou, L. Li, J. Wang, Z. Yang, H. Cai, J. Xu, L. Yang, W. Cai, W. Xu, S. Wu, and et al. Clinically applicable ai system for accurate diagnosis, quantitative measurements, and prognosis of covid-19 pneumonia using computed tomography. *Cell*, May 2020b. ISSN 1097-4172. doi: 10.1016/j.cell.2020.04.045.

C. Zheng, X. Deng, Q. Fu, Q. Zhou, J. Feng, H. Ma, W. Liu, and X. Wang. Deep learning-based detection for COVID-19 from chest CT using weak label. *medrxiv preprint*, Mar. 2020. doi: 10.1101/2020.03.12.20027185.

Supplementary Material

Table S1: Summary of the dataset for classification of COVID-19 and non-COVID-19 CT.

Dataset		COVID-19	Non-COVID-19
Training	(He et al., 2020)	251	292
	COVID-19-CT	542	67
Test	(He et al., 2020)	98	105

Table S2: Performance metrics for classification of COVID-19 and non-COVID-19 CT.

Metrics	Without using COVID-19-CT	Using COVID-19-CT
AUC	0.811 0.017	0.891 0.012
Precision	0.742 0.029	0.816 0.053
Recall/Sensitivity	0.714 0.083	0.780 0.074
Specificity	0.764 0.059	0.827 0.073
F-score	0.724 0.034	0.792 0.015

Table S3: Summary of the dataset for classification of COVID-19, influenza, and others in CT.

Dataset	COVID-19	Influenza	Normal or other diseases
Training	488	177	318
Test	118	45	79

Table S4: Performance metrics for classification of COVID-19, influenza, and normal or other types of diseases in CT.

Metrics	COVID-19	Influenza	Normal or other diseases	*Macro Avg*
AUC	0.855 0.012	0.889 0.014	0.904 0.011	0.879 0.010
Precision	0.845 0.026	0.609 0.033	0.642 0.021	0.699 0.019
Recall/Sensitivity	0.597 0.030	0.711 0.053	0.861 0.033	0.723 0.022
Specificity	0.895 0.024	0.895 0.013	0.767 0.025	0.852 0.009
F-score	0.699 0.018	0.655 0.034	0.735 0.015	0.696 0.018

Table S5: Summary of dataset used for anomaly detection of COVID-19 in CXR in unsupervised one-class classification.

Dataset	COVID-19	Non-COVID-19
Training	0	37,829
Test	184	184

Table S6: Anomaly detection performance of COVID-19 vs. non-COVID-19 using unsupervised one-class learning.

Metrics	COVID-19 vs Non-COVID-19
AUC	0.828 0.019
Precision	0.767 0.020
Recall/Sensitivity	0.772 0.017
Specificity	0.765 0.023
F-score	0.769 0.018