Evolving Systems (2024) 15:635–640
https://doi.org/10.1007/s12530-023-09511-2

ORIGINAL PAPER

# A large multiclass dataset of CT scans for COVID-19 identification

Eduardo Soares[1] · Plamen Angelov[1] · Sarah Biaso[2] · Marcelo Cury[3] · Daniel Abe[2]

Received: 20 October 2022 / Accepted: 24 May 2023 / Published online: 27 June 2023


**Abstract**
The infection by SARS-CoV-2 which causes the COVID-19 disease has spread widely over the whole world since the beginning of 2020. Following the epidemic which started in Wuhan, China on January 30, 2020 the World Health Organization (WHO) declared a global health emergency and a pandemic. In this paper, we describe a publicly available multiclass CT scan dataset for SARS-CoV-2 infection identification. Which currently contains 4173 CT-scans of 210 different patients, out of which 2168 correspond to 80 patients infected with SARS-CoV-2 and confirmed by RT-PCR. These data have been collected in the Public Hospital of the Government Employees of Sao Paulo and the Metropolitan Hospital of Lapa, both in Sao Paulo – Brazil. The aim of this data set is to encourage the research and development of artificial intelligent methods that are able to identify SARS-CoV-2 or other diseases through the analysis of CT scans. As a baseline result for this data set, we used the recently introduced eXplainable Deep Learning approach (xDNN), which is a transparent deep learning approach that allows users to inspect the decisions of the network.

**Keywords** CT-scans · COVID-19 detection · Machine learning · Explainable AI

## 1 Introduction

Coronavirus disease 2019 (COVID-19) is an infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The disease was first identified in December 2019 in Wuhan, the capital of China's Hubei province, and has since spread worldwide (Andersen et al. 2020). On January 30, 2020 the World Health Organization (WHO) declared a global health emergency [1]. Common symptoms of COVID-19 include fever, cough, and shortness of breath (Guan et al. 2020; Xu et al. 2020). While the majority of cases result in mild symptoms, some progress to viral pneumonia. By 7 August 2020, over 19 million officially confirmed cases were reported in practically every corner of the Earth with 717,687 officially reported deaths documented (Dong et al. 2020).

✉ Eduardo Soares
e.almeidasoares@lancaster.ac.uk

[1] School of Computing and Communications, Lancaster University, Lancaster, LA1 4WA, UK

[2] Public Hospital of the Government Employees of Sao Paulo, 01532-001 Sao Paulo, BR, Brazil

[3] Paulista School of Medicine, Federal University of Sao Paulo, 04021-001 Sao Paulo, BR, Brazil

As the first countries explore deconfinement strategies (Cousins 2020; Salathé et al. 2020) after a long period of quarantine, the death toll of COVID-19 keeps rising, specially in US, UK, and Brazil (Dong et al. 2020). New technologies and strategies have emerged in order to support healthcare systems during this pandemic time (Hu et al. 2020; Ting et al. 2020). As early as March 2020, Chinese hospitals used artificial intelligence (AI)-assisted computed tomography (CT) imaging analysis to screen COVID-19 cases and streamline diagnosis (Jin et al. 2020).

In this work, we build a large multiclass dataset of CT scans for SARS-CoV-2 infection identification. The dataset is built upon on the recently introduced dataset (Soares et al. 2020). The proposed dataset contains 4173 CT-scans of 210 different patients which are divided into 3 different classes (healthy, COVID-19, and other pulmonary diseases). These data have been collected from real patients in hospitals from Sao Paulo, Brazil. The aim of this dataset is to encourage the research and development of artificial intelligence (AI) methods that are able to identify if a person is infected by SARS-CoV-2 through the analysis of his/her CT scans.

An open-source dataset for COVID-19 identification through CT scans has been proposed by Yang et al. (2020), however, the data collected for this dataset has been acquired from scientific journals and may not provide the necessary

quality to train an algorithm for complex applications as such. Moreover, other authors as Santa Cruz et al. (2021); Mangal et al. (2020); Pham (2021); Signoroni et al. (2021) provided open-source datasets and solutions based on X-ray scans which are not detailed as CT scans.

As a baseline result for the new dataset based on CT scans, we consider the eXplainable Deep Learning approach (xDNN) (Angelov and Soares 2020). As the explanation of AI systems is essential to medical applications, we used the xDNN approach as baseline for this application. XDNN is a prototype-based approach that allows users to audit the decisions of the network through its similarity mechanism. XDNN obtained an $F1$ score of 97.31%, which is higher than traditional deep learning approaches such as ResNet.

**Table 1** This table demonstrates the number of patients considered to compose the dataset. In this case, we considered data from 80 patients infected by SARS-CoV-2, out of which 41 were male and 39 were female. We also considered data from 80 patients presenting other pulmonary diseases such as lung cancer, bronchitis, etc. The dataset is also composed of CT scans that do not present any pulmonary disease, These data refer to data of 50 patients

| Condition | Patients | CT-Scans | Average CT-Scans per patient |
|---|---|---|---|
| Healthy | 50 | 758 | 15 |
| COVID-19 | 80 | 2168 | 27 |
| Other pulmonary diseases | 80 | 1247 | 16 |
| TOTAL | 210 | 4173 | 20 |

## 2 Methods

The proposed dataset is composed of 4173 CT-scans of 210 different patients which are divided into: 80 patients infected by SARS-CoV-2; 80 patients with other pulmonary diseases as non-COVID pneumonia, bronchitis, and lung cancer; and 50 patients with healthy lung conditions. The data was collected from March 15 to June 15 2020 in the Public Hospital of the Government Employees of Sao Paulo, and the Metropolitan Hospital of Lapa, Sao Paulo – Brazil. The following demographic data have been collected during the clinical analysis of each patient:

- Sex
- Age
- Number of days since the 1st symptoms
- Comorbidities
- Hypertension
- Diabetes
- Chronic obstructive pulmonary disease (COPD)
- Obesity
- Pulmonary involvement > 50%
- Outcome

Table 1 details the patient's considered in this study.

The inclusion criteria for this study are listed as follows:

- Patients with a positive new coronavirus nucleic acid antibody and confirmed by the CDC;
- Patients who underwent thin-section CT;
- Age>= 18;
- Presence of lung infection in CT images.

The median duration from the onset of the illness to CT scan was 5 days, ranging from 1 to 14 days. The CT protocol was as follows: 120 kV; automatic tube current (180 mA-400 mA); iterative reconstruction; 64 mm detector; rotation time, 0.35 sec; slice thickness, 5 mm; collimation, 0.625 mm; pitch, 1.5; matrix, $512 \times 512$; and breath hold at full inspiration. The reconstruction kernel used is set as "lung smooth with a thickness of 1 mm and an interval of 0.8 mm". During reading, the lung window (with window wiDecision Treeh 1200 HU and window level-600 HU) was used. Figure 2 illustrates some examples of CT scans found in the dataset.

## 3 Data records

The database can be downloaded from Synapse (https://www.synapse.org/#!Synapse:syn22174850), and it has been presented in two formats: PNG and CSV, where PNG represents the CT scans files and CVS are the demographic data. Figure 1 illustrates the data distribution for the patients infected by SARS-CoV-2 and considered in this study.

The data types of the demographic data variables considered in this study are depicted below:

- Sex (Boolean)
- Age (Integer)
- Number of days since the 1st symptoms (Integer)
- Comorbidities (Boolean)
- Hypertension (Boolean)
- Diabetes (Boolean)
- Chronic obstructive pulmonary disease (COPD) (Boolean)
- Obesity (Boolean)
- Pulmonary involvement > 50% (Boolean)
- Outcome (Boolean)

Figure 2 illustrates different examples of data available in the proposed dataset.

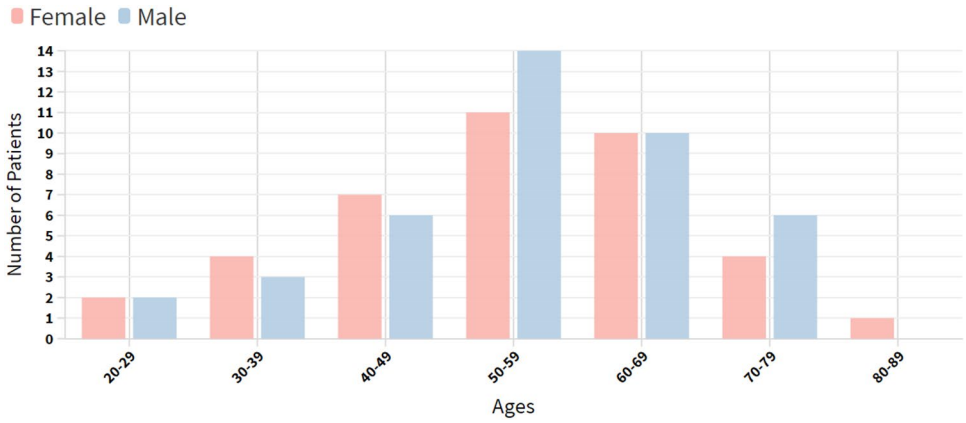


**Fig. 1** The study considered data of 80 different patients (41 male and 39 female patients). The data revealed that the major of the patients are 50–59 years old


**Fig. 2** **a** A 27-year-old male patient presented with fever and headache for 2 days. CT scans do not show the presence of any pulmonary disease. The RT-PCR test revealed negative for SARS-CoV-2 infection. **b** A 63-year-old woman patient presented shortness of breath and cough for 4 days. CT scan shows the presence of subpulmonic pleural effusion. The RT-PCR test revealed negative for SARS-CoV-2. **c** A 31-year-old woman presented fever, dry cough, shortness of breath for 4 days. CT scan revealed multifocal bilateral consolidation with ground-glass opacities with typical distribution. The RT-PCR tested positve for SARS-CoV-2

## 4 Technical validation

In order to validate our data in this section we report the results by different classification approaches. The following metrics have been used to evaluate the classification of the CT scans in different classes:

$$Accuracy\,(\%) = \frac{TP + TN}{TP + FP + TN + FN} \times 100, \tag{1}$$

Precision:

$$Precision\,(\%) = \frac{TP}{TP + FP} \times 100, \tag{2}$$

Recall:

$$Recall\,(\%) = \frac{TP}{TP + FN} \times 100, \tag{3}$$

$F1$ Score:

$$F1\ Score\,(\%) = 2 \times \frac{Precision \times Recall}{Precision + Recall} \times 100, \tag{4}$$

where $TP$, $FP$, $TN$, $FN$ denote true and false, negative and positive respectively.

The area under the curve, $AUC$, is defined through the $TP$ rate and $FN$ rate.

In this section, we report the results obtained by the xDNN classification approach (Angelov and Soares 2020; Soares et al. 2019) when applied to the proposed SARS-CoV-2 CT scan data set. We divided the dataset into 80% for training purposes and 20% for validation purposes. The division has been made in terms of patients; therefore, we separated data of 168 patients for training and data for 42 patients for validation. Results presented in Table 2 compare the performance of the xDNN algorithm with other state-of-the-art approaches, including ResNet, GoogleNet, VGG-16, AlexNet, Decision Tree, and AdaBoost.

The xDNN (Angelov and Soares 2020, 2020) classifier provided highly interpretable results (Angelov et al. 2021) that may be helpful for specialists (medical doctors). Rules generated by the identified prototypes are illustrated by Figs. 3 and 4, respectively. xDNN identified data of 18 patients with COVID-19 as prototypes and data of 11 patients non-infected as prototypes. The training time

**Table 2** Results considering different methods for the COVID-19 identification

| MethodMetric | Accuracy (%) | Precision (%) | Recall (%) | Specificity (%) | F1 Score (%) | AUC (%) |
|---|---|---|---|---|---|---|
| xDNN | 97.38 | 99.16 | 95.53 | 96.42 | 97.31 | 97.36 |
| ResNet | 94.96 | 93.00 | 97.15 | 94.36 | 95.03 | 94.98 |
| GoogleNet | 91.73 | 90.20 | 93.50 | 90.17 | 91.82 | 91.79 |
| VGG-16 | 94.96 | 94.02 | 95.43 | 94.51 | 94.97 | 94.96 |
| AlexNet | 93.75 | 94.98 | 92.28 | 92.32 | 93.61 | 93.68 |
| Decision Tree | 79.44 | 76.81 | 83.13 | 77.16 | 79.84 | 79.51 |
| AdaBoost | 95.16 | 93.63 | 96.71 | 94.98 | 95.14 | 95.19 |

for the xDNN algorithm (Angelov and Soares 2020) was only 11.82 s for all images (an average of 5 milliseconds per image. On the other hand, the traditional deep learning approach may take hours for the same task and usually requires hardware accelerators such as GPUs and once trained is not flexible to new data. We have to stress that xDNN does not require full re-training if new data is presented (Angelov and Soares 2021)—it keeps all prototypes identified so far and may add new ones if the data pattern requires that Soares et al. (2020, 2019).

Balanced one-way ANalysis Of VAriance (ANOVA) (McHugh 2011) was used to compare the results provided by the classification methods. The null hypothesis is that the mean results provided by the methods are the same. A cutoff value $p$ less than 0.05 suggests that the accuracy of at least one of the algorithms is significantly different from the others. A $p = 4.38e - 22$ was obtained and, therefore, the mean accuracy of the algorithms is not all the same; the null hypothesis was rejected.

The Tukey Honestly Significant Difference (HSD) test (McHugh 2011) was performed to compare pairs of classifiers over accuracy. Table 3 shows the results of the Tuckey HSD test for a 95% confidence interval for the true difference of the means.

If the $p - adj < 0.05$ then the null hypothesis is rejected and the difference between the methods is statistically significant. As shown in Table 3 the proposed xDNN has results statistically different from 4 traditional approaches, including well-known deep learning approaches as GoogleNet, VGG-16, and AlexNet.

Through the xDNN method we generated (extracted from the data) linguistic *IF...THEN* rules which involve actual images of both cases (COVID-19 and NO COVID-19) as illustrated in Figs. 3 and 4. Such transparent rules can be

**Table 3** Tukey Test Results

| Method 1 | Method 2 | Meandiff | p-adj | Lower | Upper | Reject |
|---|---|---|---|---|---|---|
| xDNN | Resnet | −2.28 | 0.068 | −4.6604 | 0.1004 | False |
| xDNN | GoogleNet | −5.6583 | 0.001 | −8.0387 | −3.278 | True |
| xDNN | Vgg16 | −2.385 | 0.0493 | −4.7654 | −0.0046 | True |
| xDNN | Alexnet | −3.7567 | 0.001 | −6.137 | −1.3763 | True |
| xDNN | Decision Tree | −17.8783 | 0.001 | −20.2587 | −15.498 | True |
| xDNN | Adaboost | −2.0583 | 0.1272 | −4.4387 | 0.322 | False |
| Resnet | GoogleNet | −3.3783 | 0.0015 | −5.7587 | −0.998 | True |
| Resnet | Vgg16 | −0.105 | 0.9 | −2.4854 | 2.2754 | False |
| Resnet | Alexnet | −1.4767 | 0.4709 | −3.857 | 0.9037 | False |
| Resnet | Decision Tree | −15.5983 | 0.001 | −17.9787 | −13.218 | True |
| Resnet | Adaboost | 0.2217 | 0.9 | −2.1587 | 2.602 | False |
| GoogleNet | Vgg16 | 3.2733 | 0.0023 | 0.893 | 5.6537 | True |
| GoogleNet | Alexnet | 1.9017 | 0.1912 | −0.4787 | 4.282 | False |
| GoogleNet | Decision Tree | −12.22 | 0.001 | −14.6004 | −9.8396 | True |
| GoogleNet | Adaboost | 3.6 | 0.001 | 1.2196 | 5.9804 | True |
| Vgg16 | Alexnet | −1.3717 | 0.5491 | −3.752 | 1.0087 | False |
| Vgg16 | Decision Tree | −15.4933 | 0.001 | −17.8737 | −13.113 | True |
| Vgg16 | Adaboost | 0.3267 | 0.9 | −2.0537 | 2.707 | False |
| Alexnet | Decision Tree | −14.1217 | 0.001 | −16.502 | −11.7413 | True |
| Alexnet | Adaboost | 1.6983 | 0.3061 | −0.682 | 4.0787 | False |
| Decision Tree | Adaboost | 15.82 | 0.001 | 13.4396 | 18.2004 | True |

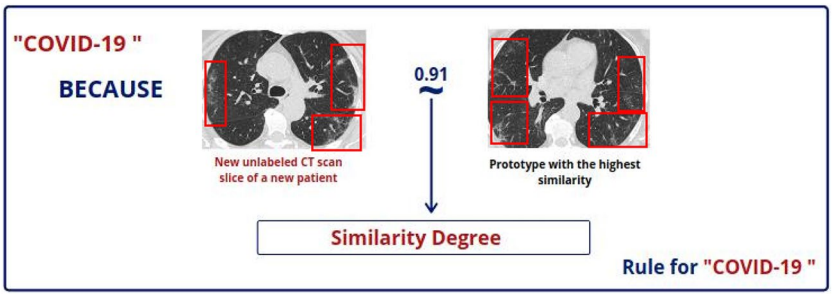


**Fig. 3** Final rule given by xDNN classifier for the COVID-19 identification. Differently, from typical deep neural networks, xDNN provides highly interpretable rules which can be visualised and used by human experts for the early evaluation of patients suspected of COVID-19 infection. The classification is done based on the similarity of the unlabeled CT scan slice to the identified prototypes

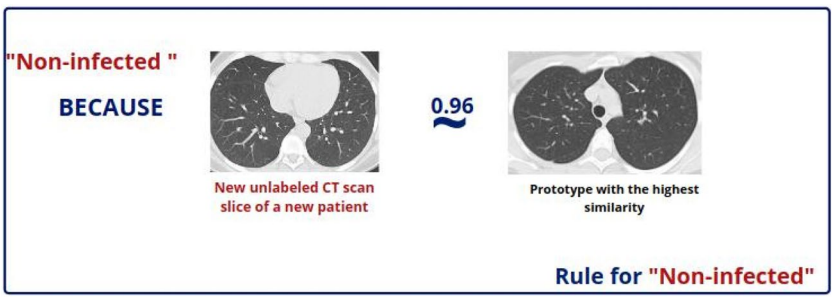


**Fig. 4** Non-SARS-CoV-2 final rule given by the proposed eXplainable Deep Learning classifier

used in a clear decision-making process for early diagnostics for COVID-19 infection. Rapid detection with high sensitivity of viral infection may allow better control of the viral spread. Early diagnosis of COVID-19 is crucial for disease treatment and control.

## 5 Conclusion

In the context of a pandemic and the urgency to contain the crisis, research has increased exponentially in order to alleviate the healthcare systems burden (Cohen et al. 2020). However, many prediction models for diagnosis and prognosis of COVID-19 infection are at high risk of bias and model overfitting as well as poorly reported, their alleged performance being likely optimistic. In order to prevent premature implementation in hospitals, tools must be robustly evaluated along several practical tests. Indeed, while some AI-assisted tools might be powerful, they do not replace clinical judgment and their diagnostic performance cannot be assessed or claimed without a proper clinical trial.

Moreover, The lack of a public database made it difficult to conduct large-scale robust evaluations. This small number of samples prevents proper cohort selection which is a limitation of this study and exposes our evaluation to sample bias. In this study, we present a database which is composed of 4173 CT-scans of 210 different patients, out of which 2168 correspond to 80 patients infected with SARS-CoV-2 and confirmed by RT-PCR. These data have been collected at the Public Hospital of the Government Employees of Sao Paulo and the Metropolitan Hospital of Lapa, Sao Paulo, Brazil. Sao Paulo is now one of the global epicenters of the COVID-19 disease.

As a baseline result for the proposed dataset, we used an explainable deep learning approach. The xDNN classifier presented an $F1$ score of 97.31% for the proposed task. Moreover, the xDNN approach provided insights into the decision-making process which is helpful to support specialists in the diagnosis of the disease. This is of great importance for medical specialists to understand and diagnose COVID-19 at its early stages via computer tomography. The proposed dataset is available https://www.synapse.org/#!Synapse:syn22174850 and xDNN (Angelov and Soares 2020) code is available at https://github.com/Plamen-Eduardo/xDNN-SARS-CoV-2-CT-Scan.

**Data availability** The data that support the findings of this study are openly available in Synapse at https://www.synapse.org/#!Synapse:syn22174850 and a small version of it in Kaggle at https://www.kaggle.com/datasets/plameneduardo/sarscov2-ctscan-dataset.

**Code availability** We provided the code used in this research at https://github.com/Plamen-Eduardo/xDNN-SARS-CoV-2-CT-Scan. Other codes are available upon request to the corresponding author.

## Declarations

**Conflict of interest** The authors declare no competing interests.



## References

Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF (2020) The proximal origin of sars-cov-2. Nat Med 26(4):450–452

Angelov P, Soares E (2020) Towards explainable deep neural networks (xDNN). Neural Netw 130:185–194

Angelov P, Soares E (2021) Detecting and learning from unknown by extremely weak supervision: exploratory classifier (xclass). Neural Comput Appl 33(22):15145–15157

Angelov PP, Soares EA, Jiang R, Arnold NI, Atkinson PM (2021) Explainable artificial intelligence: an analytical review. WIREs Data Min Knowl Discovery 11(5):e1424. https://doi.org/10.1002/widm.1424

Angelov P, Soares E (2020) Towards deep machine reasoning: a prototype-based deep neural network with decision tree inference. In: 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 2092–2099. https://doi.org/10.1109/SMC42975.2020.9282812

Cohen JP, Dao L, Roth K, Morrison P, Bengio Y, Abbasi AF, Shen B, Mahsa HK, Ghassemi M, Li H, et al (2020) Predicting covid-19 pneumonia severity on chest x-ray with deep learning. Cureus **12**(7)

Cousins S (2020) New Zealand eliminates covid-19. The Lancet 395(10235):1474

Dong E, Du H, Gardner L (2020) An interactive web-based dashboard to track covid-19 in real time. The Lancet infectious diseases

Guan Wj, Ni Zy, Hu Y, Liang Wh, Ou Cq, He Jx, Liu L, Shan H, Lei Cl, Hui DS et al (2020) Clinical characteristics of coronavirus disease 2019 in china. N Engl J Med 382(18):1708–1720

Hu Z, Ge Q, Li S, Jin L, Xiong M (2020) Artificial intelligence forecasting of Covid-19 in China. arXiv preprint arXiv:2002.07112 (01-20)

Jin YH, Cai L, Cheng ZS, Cheng H, Deng T, Fan YP, Fang C, Huang D, Huang LQ, Huang Q et al (2020) A rapid advice guideline for the diagnosis and treatment of 2019 novel coronavirus (2019-ncov) infected pneumonia (standard version). Mil Med Res 7(1):4

Mangal A, Kalia S, Rajgopal H, Rangarajan K, Namboodiri V, Banerjee S, Arora C (2020) Covidaid: Covid-19 detection using chest x-ray. arXiv preprint arXiv:2004.09803

McHugh ML (2011) Multiple comparison analysis testing in anova. Biochem Med 21(3):203–209

Pham TD (2021) Classification of covid-19 chest X-rays with deep learning: new models or fine tuning? Health Inf Sci Syst 9:1–11

Salathé M, Althaus CL, Neher R, Stringhini S, Hodcroft E, Fellay J, Zwahlen M, Senti G, Battegay M, Wilder-Smith A et al (2020) Covid-19 epidemic in Switzerland: on the importance of testing, contact tracing and isolation. Swiss Med Wkly 150(11–12):w20225

Santa Cruz BG, Bossa MN, Sölter J, Husch AD (2021) Public covid-19 x-ray datasets and their impact on model bias-a systematic review of a significant problem. Med Image Anal 74:102225

Signoroni A, Savardi M, Benini S, Adami N, Leonardi R, Gibellini P, Vaccher F, Ravanelli M, Borghesi A, Maroldi R et al (2021) Bs-net: Learning covid-19 pneumonia severity on a large chest x-ray dataset. Med Image Anal 71:102046

Soares E, Angelov P, Biaso S, Froes MH, Abe DK (2020) Sars-cov-2 ct-scan dataset: a large dataset of real patients ct scans for sars-cov-2 identification. medRxiv 1(1)

Soares E, Angelov P, Costa B, Castro M (2019) Actively semi-supervised deep rule-based classifier applied to adverse driving scenarios. In: 2019 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. https://doi.org/10.1109/IJCNN.2019.8851842

Soares E, Angelov P, Filev D, Costa B, Castro M, Nageshrao S (2019) Explainable density-based approach for self-driving actions classification. In: 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA), pp 469–474. https://doi.org/10.1109/ICMLA.2019.00087

Ting DSW, Carin L, Dzau V, Wong TY (2020) Digital technology and covid-19. Nat Med 26(4):459–461

World Health Organization Declares Global Emergency (2020) A review of the 2019 novel coronavirus (covid-19). Int J Surg 76:71–76. https://doi.org/10.1016/j.ijsu.2020.02.034

Xu Z, Shi L, Wang Y, Zhang J, Huang L, Zhang C, Liu S, Zhao P, Liu H, Zhu L et al (2020) Pathological findings of covid-19 associated with acute respiratory distress syndrome. Lancet Respir Med 8(4):420–422

Yang X, He X, Zhao J, Zhang Y, Zhang S, Xie P (2020) Covid-ct-dataset: a ct scan dataset about covid-19. arXiv preprint arXiv:2003.13865, 14