BMC Research Notes

DATA NOTE

Open Access

COVID19-CT-dataset: an open-access chest CT image repository of 1000+ patients with confirmed COVID-19 diagnosis

Shokouh Shakouri[1], Mohammad Amin Bakhshali[1], Parvaneh Layegh[2], Behzad Kiani[1], Farid Masoumi[1], Saeedeh Ataei Nakhaei[3] and Sayyed Mostafa Mostafavi[1*]

Abstract

Objectives: The ongoing Coronavirus disease 2019 (COVID-19) pandemic has drastically impacted the global health and economy. Computed tomography (CT) is the prime imaging modality for diagnosis of lung infections in COVID-19 patients. Data-driven and Artificial intelligence (AI)-powered solutions for automatic processing of CT images predominantly rely on large-scale, heterogeneous datasets. Owing to privacy and data availability issues, open-access and publicly available COVID-19 CT datasets are difficult to obtain, thus limiting the development of AI-enabled automatic diagnostic solutions. To tackle this problem, large CT image datasets encompassing diverse patterns of lung infections are in high demand.

Data description: In the present study, we provide an open-source repository containing 1000+ CT images of COVID-19 lung infections established by a team of board-certified radiologists. CT images were acquired from two main general university hospitals in Mashhad, Iran from March 2020 until January 2021. COVID-19 infections were ratified with matching tests including Reverse transcription polymerase chain reaction (RT-PCR) and accompanying clinical symptoms. All data are 16-bit grayscale images composed of 512 × 512 pixels and are stored in DICOM standard. Patient privacy is preserved by removing all patient-specific information from image headers. Subsequently, all images corresponding to each patient are compressed and stored in RAR format.

Keywords: Coronavirus, COVID-19, Computed tomography, Chest CT image, Lung infection, Diagnosis, Artificial intelligence, Deep learning, Clinical imaging, Radiology

Objective

Coronavirus disease 2019 (COVID-19) is an infectious, highly contagious disease with major global health implications. As of January 31, 2021, there have been 103 million confirmed infections worldwide, claiming over 2.2 million lives [1, 2]. A major hurdle in the management and control of COVID-19 is availability of timely disease screening and monitoring tests.

Computed tomography (CT) scans are routinely used in clinical practice for diagnosis, screening and management of COVID-19 worldwide. The heavy number of required scans keeps radiologists highly occupied and leaves them with limited time, which hinders the delivery of timely CT reports. Besides, there is limited access to well-trained radiologists with adequate COVID-19 imaging expertise in many underdeveloped rural regions. Collectively, these call for data-driven Artificial intelligence (AI)-powered solutions for automatic detection and quantification of COVID-19 infections.

To date, there have been numerous studies that have attempted to deploy AI-based approaches, such as deep

*Correspondence: MostafaviTM@mums.ac.ir
[1] Department of Medical Informatics, School of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran
Full list of author information is available at the end of the article

© The Author(s) 2021. This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Table 1 Overview of dataset

Label	Name of data set	File types (file extension)	Data repository and identifier (DOI or accession number)
Data set 1	Covid19-CT-dataset	RAR files (.rar)	Harvard Dataverse (https://doi.org/10.7910/DVN/6ACUZJ) [10]

convolutional neural network (CNN) models, for automatic detection and quantification of COVID-19 from CT images [3–5]. The key to success of these models is the deployment of rich datasets that encapsulate diverse patterns of lung infections. However, owing to privacy and data collection issues, CT images used in these studies are limited in size and not publicly available. This significantly impacts the development of new AI-powered solutions for more advanced diagnosis and quantification of COVID-19 infections.

Herein, we present an open-access repository of 1000 + CT images obtained since the onset of COVID-19 in March 2020, at least one order of magnitude larger than the current available datasets [6–8]. Given the diverse patterns of infection covered in this rich dataset, this can serve as the starting point for more comprehensive data-driven models. Moreover, this can also be used as an educational resource for under-trained radiologists in less developed areas around the globe.

Data description

This dataset consists of unenhanced chest CTs from 1000+ patients with confirmed COVID-19 infections. The age distribution of patients who underwent CT imaging was 47.18 ± 16.32 (mean ± standard deviation) years and age range was between 6 and 89 years. Gender distribution was 60.9% male and 39.1% female. The most prevalent self-reported coexisting conditions among patients included hypertension or coronary heart disease, diabetes, and interstitial pneumonia or emphysema (in that order). Images were obtained in the March 2020–January 2021 period, and were acquired at the point of care in an inpatient setting from patients with positive Reverse Transcription Polymerase Chain Reaction (RT-PCR) tests for COVID-19, accompanied by supporting clinical symptoms. All scans were performed with the patient in the supine position during end-inspiration. The scanning range was from the apex to lung base. CT exams were performed with a NeuViz 16-slices CT scanner machine (Neusoft medical systems) without intravenous contrast under "Helical" mode. All images are in DICOM format and consist of 16-bit grayscale images composed of 512 × 512 pixels. Slice thickness values were determined by the operator in accordance with clinical examination requirements: 1.5 or 3 mm. Patient privacy is preserved by removing all patient-specific information from image headers. Subsequently, all images corresponding to each patient are compressed and stored in RAR format. Table 1 provides an overview of the dataset.

All CT images were visually examined by two board-certified radiologists for the presence of COVID-19 infections. In case of a disagreement between the first two radiologists, a third more experienced radiologist rendered the final decision. CT images were identified to have a broad mixture of COVID-specific patterns of lung infections including: (i) presence of ground-glass opacities, mixed ground-glass opacities, or consolidation; (ii) presence of air bronchogram, interlobular septal thickening, or cavitation; (iii) different number of lobes affected by ground-glass or consolidative opacities; (iv) presence of fibrotic lesions; (v) presence of centri-lobular nodules; (vi) presence of a pleural effusion; (vii) presence of thoracic lymphadenopathy; (viii) presence of underlying lung disease such as tuberculosis, emphysema, or interstitial lung disease; and (ix) different distribution patterns of opacities including peripheral, central, bilateral, focal, multi-lobar and diffuse. Ground-glass opacification was defined as "hazy increased lung attenuation with preservation of bronchial and vascular margins" and consolidation was defined as "opacification with obscuration of margins of vessels and airway walls" [9].

Limitations

- A small number of images contain some form of background noise such as patient bed and/or some form of motion artifacts.
- Images were taken from only two general hospitals in Mashhad, Iran, and represent a predominantly Iranian population.

Abbreviations

COVID-19: Coronavirus disease 2019; CT: Computed tomography; AI: Artificial intelligence; RT-PCR: Reverse transcription polymerase chain reaction; DICOM: Digital imaging and communications in medicine; RAR: Roshal archive; CNN: Convolutional neural networks.

Acknowledgements

We would like to thank Mashhad University of Medical Sciences for funding this study.

Authors' contributions
PL provided access to the data for sharing and ratified image clinical diagnoses. SAN, SS and FM contributed to data collection and preparation. MAB contributed to image preprocessing and file format preparations. MAB, BK and SMM drafted the manuscript and critically revised the text. SMM was the principal investigator and project leader. All authors read and approved the final version for submission.

Funding
The study received funding from Mashhad University of Medical Sciences (Fund Number: 991315).

Availability of data and materials
The data described in this data note can be freely and openly accessed on Harvard dataverse under (https://doi.org/10.7910/DVN/6ACUZJ) [10]. Please see Table 1 for details and link to the data.

Declarations

Ethics approval and consent to participate
Regarding ethical issues, this study has been assessed by the research council of Mashhad University of Medical Sciences (Reference Number: IR.MUMS.MEDICAL.REC.1399.594). The study was approved because no identifying data have been reported.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Medical Informatics, School of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran. [2]Department of Radiology, Faculty of Medicine, Imam Reza Hospital, Mashhad University of Medical Sciences, Mashhad, Iran. [3]Nuclear Medicine Research Center, School of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran.

Received: 19 February 2021 Accepted: 29 April 2021
Published online: 12 May 2021

References

1. https://www.who.int/docs/default-source/coronaviruse/situation-reports/
2. Bergquist R, Kiani B, Manda S. First year with Covid-19: assessment and prospects. Geospat Health. 2020. https://doi.org/10.4081/gh.2020.953.
3. Islam M, Karray F, Alhajj R, Zeng J. A review on deep learning techniques for the diagnosis of novel coronavirus (covid-19). arXiv preprint,arXiv:200804815 2020
4. Shoeibi A, Khodatars M, Alizadehsani R, Ghassemi N, Jafari M, Moridian P, et al. Automated detection and forecasting of covid-19 using deep learning techniques: a review. arXiv preprint, arXiv:2007107852020
5. Shi F, Wang J, Shi J, Wu Z, Wang Q, Tang Z, et al. Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for covid-19. IEEE Rev Biomed Eng. 2020. https://doi.org/10.1109/RBME.2020.2987975.
6. Zhao J, Zhang Y, He X, Xie P. Covid-CT-dataset: a CT scan dataset about covid-19. arXiv preprint, arXiv:2003138652020
7. Yang X, He X, Zhao J, Zhang Y, Zhang S, Xie P. COVID-CT-dataset: a CT image dataset about COVID-19. arXiv preprint, arXiv:200313865 2020
8. Afshar P, Heidarian S, Enshaei N, Naderkhani F, Rafiee MJ, Oikonomou A, et al. COVID-CT-MD: COVID-19 Computed tomography (CT) scan dataset applicable in machine learning and deep learning. Sci Data. 2020. https://doi.org/10.1038/s41597-021-00900-3.
9. Hansell DM, Bankier AA, MacMahon H, McLoud TC, Muller NL, Remy J. Fleischner Society: glossary of terms for thoracic imaging. Radiology. 2008;246(3):697–722.
10. Mostafavi SM. COVID19-CT-dataset: an open-access chest CT image repository of 1000+ patients with confirmed covid-19 diagnosis. Harv Dataverse. 2021. https://doi.org/10.7910/DVN/6ACUZJ.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

BMC