



OPEN

DATA DESCRIPTOR

POLCOVID: a multicenter multiclass chest X-ray database (Poland, 2020–2021)

Aleksandra Suwalska^{1,25}, Joanna Tobiasz^{1,2,25}, Wojciech Prazuch^{1,25}, Marek Socha^{1,25}, Pawel Foszner^{1,2}, Damian Piotrowski³, Katarzyna Gruszczynska⁴, Magdalena Sliwinska⁵, Jerzy Walecki⁶, Tadeusz Popiela⁷, Grzegorz Przybylski⁸, Mateusz Nowak⁹, Piotr Fiedor¹⁰, Malgorzata Pawlowska¹¹, Robert Flisiak¹², Krzysztof Simon¹³, Gabriela Zapolska¹⁴, Barbara Gizycka¹⁵, Edyta Szurowska¹⁶, for the POLCOVID Study Group*, Michal Marczyk^{1,17,25}✉, Andrzej Cieszanowski¹⁸✉ & Joanna Polanska¹

The outbreak of the SARS-CoV-2 pandemic has put healthcare systems worldwide to their limits, resulting in increased waiting time for diagnosis and required medical assistance. With chest radiographs (CXR) being one of the most common COVID-19 diagnosis methods, many artificial intelligence tools for image-based COVID-19 detection have been developed, often trained on a small number of images from COVID-19-positive patients. Thus, the need for high-quality and well-annotated CXR image databases increased. This paper introduces POLCOVID dataset, containing chest X-ray (CXR) images of patients with COVID-19 or other-type pneumonia, and healthy individuals gathered from 15 Polish hospitals. The original radiographs are accompanied by the preprocessed images limited to the lung area and the corresponding lung masks obtained with the segmentation model. Moreover, the manually created lung masks are provided for a part of POLCOVID dataset and the other four publicly available CXR image collections. POLCOVID dataset can help in pneumonia or COVID-19 diagnosis, while the set of matched images and lung masks may serve for the development of lung segmentation solutions.

¹Department of Data Science and Engineering, Silesian University of Technology, Gliwice, Poland. ²Department of Computer Graphics, Vision and Digital Systems, Silesian University of Technology, Gliwice, Poland. ³Department of Infectious Diseases and Hepatology, Medical University of Silesia, Katowice, Poland. ⁴Department of Radiology and Nuclear Medicine, Medical University of Silesia, Katowice, Poland. ⁵Department of Diagnostic Imaging, Voivodship Specialist Hospital, Wroclaw, Poland. ⁶Department of Diagnostic Radiology, Central Clinical Hospital of the Ministry of Internal Affairs and Administration, Warsaw, Poland. ⁷Department of Radiology, Jagiellonian University Medical College, Krakow, Poland. ⁸Department of Lung Diseases, Cancer and Tuberculosis, Kujawsko-Pomorskie Pulmonology Center, Bydgoszcz, Poland. ⁹Department of Radiology, Silesian Hospital, Cieszyn, Poland. ¹⁰Department of General and Transplantation Surgery, Medical University of Warsaw, Warsaw, Poland. ¹¹Department of Infectious Diseases and Hepatology, Collegium Medicum in Bydgoszcz, Nicolaus Copernicus University, Torun, Poland. ¹²Department of Infectious Diseases and Hepatology, Medical University of Bialystok, Bialystok, Poland. ¹³Department of Infectious Diseases and Hepatology, Wroclaw Medical University, Wroclaw, Poland. ¹⁴Department of Radiology, Czerniakowski Hospital, Warsaw, Poland. ¹⁵Department of Imaging Diagnostics, MEGREZ Hospital, Tychy, Poland. ¹⁶2nd Department of Radiology, Medical University of Gdansk, Gdansk, Poland. ¹⁷Yale Cancer Center, Yale School of Medicine, New Haven, CT, USA. ¹⁸Department of Radiology I, The Maria Sklodowska-Curie National Research Institute of Oncology, Warsaw, Poland. ²⁵These authors contributed equally: Aleksandra Suwalska, Joanna Tobiasz, Wojciech Prazuch, Marek Socha, Michal Marczyk. *A list of authors and their affiliations appears at the end of the paper. ✉e-mail: michal.marczyk@polsl.pl; acieszanowski@wum.edu.pl

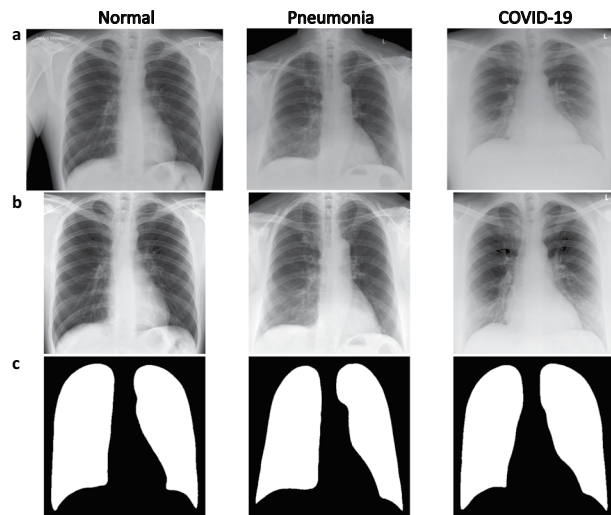


Fig. 1 Exemplary images included in the POLCOVID dataset for one representative of each diagnosis group. Original CXR images (a), preprocessed lung area images (b), and lung masks (c) of normal, pneumonia and COVID-19 cases.

Background & Summary

The outbreak of the SARS-CoV-2 pandemic in 2020 has made healthcare systems worldwide face new challenges. Limited testing capacity, especially in the early phases of pandemics, shortages of adequate equipment, and overloaded hospitals were the main factors inhibiting the process of sufficient patient diagnosis and management^{1,2}. Hence, chest radiography became a crucial diagnostic tool, especially for individuals experiencing dyspnea^{3,4}. Also, patients requiring rapid treatment and support in the form of oxygenation or ventilation often were unable to wait for the RT-PCR test result. COVID-19 pandemic and the challenges it caused led to the development of many Artificial Intelligence (AI)-based tools for COVID-19 detection⁵⁻⁷. Consequently, with all the advantages of the AI-assisted diagnosis process, there appeared a great need for reliable, high-quality, and universal imaging datasets⁸.

Here, we provide two datasets used for different purposes in our studies. The first dataset was created for COVID-19 detection and includes a set of 4809 chest X-ray (CXR) images collected from COVID-19 positive and negative patients in 15 Polish hospitals. Medical doctors labelled all CXR pictures based on diagnosis as COVID-19 ($n = 1236$), other-type pneumonia ($n = 1147$), or healthy, normal lungs ($n = 2426$). Figure 1a shows the exemplary CXR images representing all groups. Some radiographs were also annotated with demographic information such as age, sex, and smoking history. The cohort is sufficiently balanced in terms of sex (1415 males, 1243 females) and heterogeneous in terms of age, ranging from 0 to 99 years. As medical centers which provided the data are in various regions of Poland, the study population is representative. As an extension to the original CXR images, we deliver their preprocessed versions limited to the lung area and the corresponding lung masks generated by our lung segmentation model. We also provide the disease subtype prediction for each patient that explains the heterogeneity within each group.

The second dataset served to build the lung segmentation model. It contains lung masks manually created by experts for 6297 chest images, including 4003 from Polish hospitals collected as the POLCOVID dataset. For those, we deliver the corresponding original CXRs. The rest of the chest images came from publicly available sources, therefore we only provide their masks.

POLCOVID dataset can serve for the generation of novel pneumonia and/or COVID-19 screening or diagnosis tools, while the set of matched images and lung masks may support the development of lung segmentation solutions.

Methods

Ethical statement. The project was approved by Institutional Review Boards (IRBs) of all collaborating medical centers (Silesian Hospital in Cieszyn, Voivodship Specialist Hospital in Wrocław, Collegium Medicum in Bydgoszcz, The Maria Skłodowska-Curie National Research Institute of Oncology in Warsaw, Medical University of Silesia in Katowice, Specialist Hospital No. 1 in Bytom, Collegium Medicum of the Jagiellonian University in Cracow, Central Clinical Hospital of the Ministry of Interior in Warsaw, Single Infectious Diseases Hospital MEGREZ Ltd. in Tychy, District Hospital in Raciborz, Kujawsko-Pomorskie Pulmonology Center in Bydgoszcz, University Clinical Hospital in Opole, Czerniakowski Hospital in Warsaw, University Clinical Center of Medical University of Gdansk, and Prognostic Specialist Clinic in Knurów). The requirement for individual patient consent was waived as we removed all identifiable patient information. We complied with all relevant ethical regulations and guidelines. The Ministry of Science and Higher Education of the Republic of Poland consented and financially supported the project for high-priority nationwide research on COVID-19 (grant no MNiSW/2/WFSN/2020), of which data collection was an integral part. The ethics approval allowed for the open publication of the data.

Hospital	Hospital ID	Number of images			
		NORMAL	PNEUMONIA	COVID-19	TOTAL
Department of Radiology, Silesian Hospital, Cieszyn	1	889	162	2	1053
Voivodship Specialist Hospital, Wroclaw	2	333	234	349	916
Department of Infectious Diseases and Hepatology, Collegium Medicum in Bydgoszcz	3	0	11	80	91
Department of Imaging Diagnostics, The Maria Skłodowska-Curie National Research Institute of Oncology, Warsaw	4	742	180	1	923
Faculty of Medical Sciences, Medical University of Silesia, Katowice	5	1	1	0	2
Specialist Hospital No. 1, Bytom	6	95	49	21	165
Collegium Medicum of the Jagiellonian University, Cracow	7	51	25	268	344
Central Clinical Hospital of the Ministry of Interior in Warsaw	8	3	0	151	154
Department of Imaging Diagnostics, Single Infectious Diseases Hospital MEGREZ Ltd., Tychy	9	19	17	33	69
District Hospital, Raciborz	10	0	0	10	10
Kujawsko-Pomorskie Pulmonology Center, Bydgoszcz	11	93	159	20	272
University Clinical Hospital, Opole	12	3	3	0	6
Czerniakowski Hospital, Warsaw	13	0	0	114	114
University Clinical Center, Medical University of Gdansk	14	18	22	170	210
Prognostic Specialist Clinic, Knurów	15	179	284	17	480

Table 1. Numbers of CXR images provided by each medical center with regard to the diagnosis.

Data source. Fifteen medical centers from seven regions of Poland participated in the data acquisition. At each hospital, patients were diagnosed with COVID-19 or other types of pneumonia based on radiological findings or labeled as normal otherwise. COVID-19 was confirmed radiologically in all COVID-19 positive cases. This diagnosis was moreover supported with an RT-PCR test. All COVID-19 positive patients required medical assistance, although they might have developed various symptoms. The centers uploaded the data in the time range from August 7th, 2020, to April 7th, 2021. Hence, no Omicron SARS-CoV-2 variant-infected patients participated in the study, as the first reports of this variant appeared in November 2021⁹. The summary of the number of CXR images provided by each medical center is presented in Table 1 with regard to diagnosis.

Imaging. CXR images were collected using various devices and parameters due to differences in equipment between medical centers. All radiographs were performed in a frontal projection.

Data collection. We created a web service dedicated to medical centers participating in the project to provide the data in a secure manner. Registered users from the POLCOVID Study Group uploaded radiographs annotated with a medical diagnosis. When available, medical centers attached a more detailed patient description including demographic and clinical information such as sex, age, and smoking history. X-ray images were stored in the Digital Imaging and Communication in Medicine (DICOM)¹⁰ or JPEG formats, depending on the uploader. Exemplary CXR images representing COVID-19, pneumonia, and normal patients are presented in Fig. 1a.

Data preparation. We applied the U-Net neural network to segment the lung area from the standardized and contrast-enhanced CXR images¹¹. For lung segmentation model training and testing, we used 6297 CXR images referred to as the lung segmentation dataset. Out of those pictures, 4003 radiographs were a part of our POLCOVID dataset. The remaining 2294 CXRs came from the publicly available collections: the National Institute of Health – Clinical Center database¹² (1124 CXRs), Shenzhen No.3 Hospital, Shenzhen, China¹³ (662 CXRs), the tuberculosis control program of the Department of Health and Human Services of Montgomery County, USA¹³ (138 CXRs), and Guangzhou Women and Children’s Medical Center, Guangzhou, China¹⁴ (370 CXRs). We converted the original CXRs to TIFF format and we scaled the intensity values to range 0–1. Experts manually annotated each CXR picture with a lung mask. We randomly divided the CXRs into the training (n = 5247), validation (n = 500), and test (n = 550) subsets. A detailed summary of subsets regarding the image source is presented in Table 2.

During the model generation, the sigmoid (for the last convolutional layer) and the Scaled Exponential Linear Unit (SELU) (for all remaining layers) served as activation functions, the Sorensen-Dice coefficient (SDC) as a similarity measure for the loss function, and the adaptive learning rate method ADAM¹⁵ as the optimization algorithm. With the model-generated masks, we limited the standardized image to the lung area – the region of interest (ROI), further resized to 512 × 512 pixels with the original aspect ratio. Prazuch *et al.*¹⁶ precisely described the lung segmentation procedure.

For all the POLCOVID CXRs, we deliver resized ROI images and model-generated lung masks adjusted to the ROI dimensions. Exemplary ROI images and lung masks representing COVID-19, pneumonia, and normal patients are presented in Fig. 1b,c. As a separate data subset, we also provide all manually annotated lung masks and the original POLCOVID CXRs used to generate the lung segmentation model.

Demographic summary. The patient sex is well-balanced in the normal (554 males, 583 females) and COVID-19 (492 males, 437 females) groups and in the whole cohort (1415 males, 1243 females). In the

Source	Subset			TOTAL
	Training	Validation	Testing	
POLCOVID	3403	300	300	4003
National Institute of Health – Clinical Center ¹²	904	20	200	1124
Shenzhen No.3 Hospital, Shenzhen, China ¹³	525	137	0	662
Department of Health and Human Services of Montgomery County, USA ¹³	115	23	0	138
Guangzhou Women and Children’s Medical Center, Guangzhou, China ¹⁴	300	20	50	370
TOTAL	5247	500	550	6297

Table 2. Numbers of CXR images used for the lung segmentation model training, with regard to the data source and subset.

	NORMAL	PNEUMONIA	COVID-19	ALL
	(n/% of N)	(n/% of N)	(n/% of N)	(N/% of N.ALL)
All	2426/50.45%	1147/23.85%	1236/25.70%	N.ALL = 4809/100%
SEX				
Male	554/39.15%	369/26.08%	492/34.77%	1415/29.42%
Female	583/46.90%	223/17.94%	437/35.16%	1243/25.85%
No information	1289/59.93%	555/25.80%	307/14.27%	2151/44.73%
SMOKING STATUS				
Non-smoker	104/17.45%	183/30.70%	309/51.85%	596/12.39%
Smoker	55/24.23%	85/37.44%	87/38.33%	227/4.72%
No information	2267/56.87%	879/22.05%	840/21.07%	3986/82.89%

Table 3. The summary of sex and smoking status in the whole cohort and in diagnosis groups.

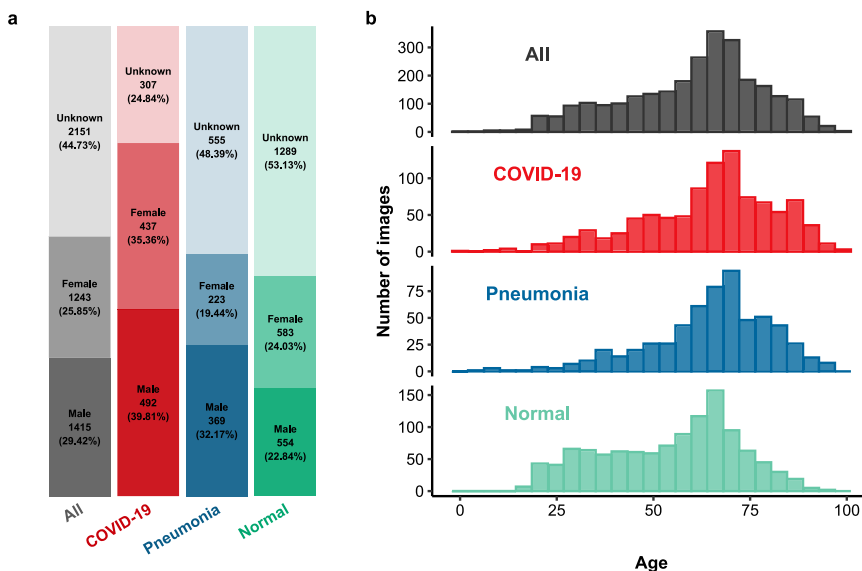


Fig. 2 Demographic summary of the cohort. Proportions of sexes in diagnosis groups and in total are accompanied by numbers of images and numbers of missing records (a). Age distributions in diagnosis groups and in total (b).

pneumonia group, male patients are overrepresented (369 males, 223 females). The summary of sex and smoking status in total and regarding diagnosis is presented in Table 3. Proportions of sexes in diagnosis groups and in the whole cohort are presented in Fig. 2a.

The dataset is highly heterogeneous in terms of patient age, ranging from 0 to 99 years, with a mean and median equal to 60.24 and 63 years, respectively. Age distributions differ significantly between the patient groups (Kruskal–Wallis one-way analysis of variance p -value $< 10^{-6}$). The median age of COVID-19 and pneumonia patients is equal (67 years) with a similar range. In the normal group, the median age is lower with a smaller range compared to other patients. Conover post-hoc tests showed significant differences in age distribution only

		Min.	1 st quartile	Median	Mean \pm SD	3 rd quartile	Max.	#Missing
Age	All	0	49	63	60.24 \pm 17.83	72	99	2086
	NORMAL	17	40	58	54.32 \pm 17.57	67	96	1273
	PNEUMONIA	4	57	67	64.82 \pm 15.86	76	96	545
	COVID-19	0	54	67	64.45 \pm 17.27	76	99	268
Pack-years	All	1	11.5	25	27.94 \pm 20.18	39.5	114	68
	NORMAL	1	10	13.5	17.66 \pm 13.83	23.75	60	18
	PNEUMONIA	5	20.75	33.5	36.48 \pm 22.72	41.5	114	27
	COVID-19	2	10	20	26.02 \pm 17.66	32.5	80	23

Table 4. The summary of age and pack-years status in the whole cohort and in diagnosis groups.

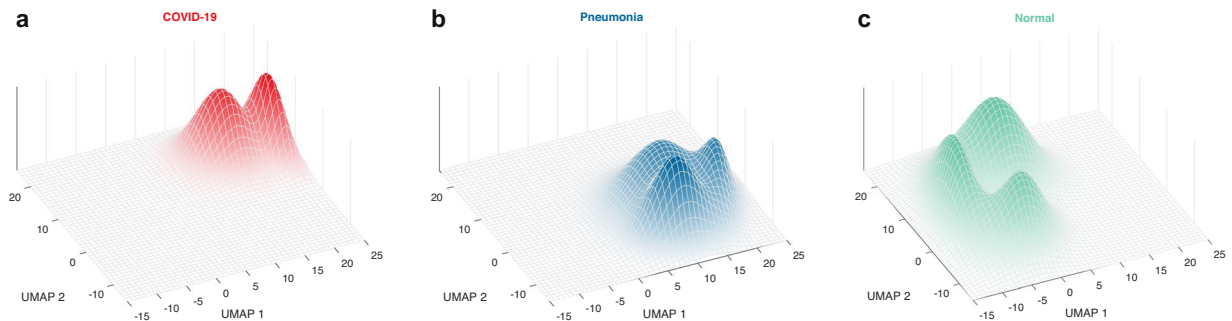


Fig. 3 Two-dimensional Gaussian mixture model (2D GMM) fitted on the results of nUMAP feature extraction for each diagnosis category. COVID-19 (a), other-type pneumonia (b), and normals (c).

in the normal group compared to the remaining two (both p -values $< 10^{-6}$). For COVID-19 versus pneumonia comparison, the corrected p -value equaled 0.77. The summary of age and pack-years in total and regarding diagnosis is presented in Table 4. Age distributions in diagnosis groups and in the whole cohort are presented in Fig. 2b.

Medical centers failed to provide additional information (sex, age, smoking status) concerning many patients. The completeness of data is the poorest for the normal group (53.13%, 52.47%, and 93.45% of missing records for sex, age, and smoking status, respectively) and the highest for COVID-19 patients (24.84%, 21.68%, and 67.96% of missing records for sex, age, and smoking status, respectively).

Disease subtype prediction. We used the nUMAP method from Suwalska *et al.*¹⁷ to predict the disease subtype, as described in Prazuch *et al.*¹⁶ For this step, we extended our POLCOVID dataset with two publicly available chest CXRs databases: COVIDx¹⁸ ($n = 15403$) and AIforCovid¹⁹ ($n = 1105$). The nUMAP approach involves the neural network serving as a feature extractor. It takes CXR images with clinical information as an input and provides a numerical data matrix with features' values per image as an output of the final fully connected layer. We applied the standard UMAP algorithm with the cosine distance metrics on the numerical feature vectors to visualize the data in the two-dimensional space. This projection served for fitting the two-dimensional Gaussian mixture model (2D GMM) with the modified expectation-maximization (EM) algorithm, as explained in Marczyk²⁰. We obtained three mixture model components per diagnosis category (COVID-19, pneumonia, and normal), each representing a different disease subtype. The first subtypes correspond to the typical representatives of each group (denoted as C1, P1, and N1, respectively). The second subtypes contain mild cases (C2, P2, and N2, respectively). The third subtypes (C3, P3, N3) show the smallest differences between the groups and represent the atypical cases. The results of 2D GMM fitting to the nUMAP embedding are shown in Fig. 3.

Data Records

POLCOVID image collection. All CXR images are available in de-identified form on Synapse (www.synapse.org/#!Synapse:syn50877085; <https://doi.org/10.7303/syn50877085>)²¹ and CIRCA COVID-19 CXR/CT-based diagnosis web service (<https://covid.aei.polsl.pl>). Once registered, the user may download the radiographs (in the DICOM or JPEG format), preprocessed images, and lung masks of a group of interest.

The image files are divided into three parts: original CXRs, preprocessed CXRs, and lung masks. For each of those, the files are organized in three individual ZIP files, one per diagnosis category (COVID-19, other-type pneumonia, or normal). The ZIP files are named according to the following pattern: Polcovid_<record_type>_<diagnosis_category>.zip. The organization and content of provided files is shown in Table 5.

POLCOVID metadata structure. The metadata files are available for registered users on Synapse (www.synapse.org/#!Synapse:syn50877085; <https://doi.org/10.7303/syn50877085>)²¹ and CIRCA COVID-19 CXR/CT-based diagnosis web service in the form of the Microsoft Excel spreadsheet for all files and CSV files for

Record type	Diagnosis category	Filename	#Files
Metadata	All	metadata.xlsx	1
	COVID-19	metadata_COVID.csv	1
	Pneumonia	metadata_PNEUMONIA.csv	1
	Normal	metadata_NORMAL.csv	1
Original CXRs	COVID-19	Polcovid_data_covid.zip	1236
	Pneumonia	Polcovid_data_pneumonia.zip	1147
	Normal	Polcovid_data_normal.zip	2426
Preprocessed CXRs	COVID-19	Polcovid_data_processed_covid.zip	1236
	Pneumonia	Polcovid_data_processed_pneumonia.zip	1147
	Normal	Polcovid_data_processed_normal.zip	2426
Lung masks	COVID-19	Polcovid_data_masks_covid.zip	1236
	Pneumonia	Polcovid_data_masks_pneumonia.zip	1147
	Normal	Polcovid_data_masks_normal.zip	2426

Table 5. File organization of the POLCOVID dataset.

Variable Name	Definition
<i>origin</i>	Name of the dataset.
<i>filename</i>	Anonymized unique file name of the following structure: Anonymous_<hospital_id>_<patient_id>_<class_id>.<file_format>.
<i>patient_id</i>	Anonymized patient identifier, unique for patients examined in the same medical center, ranging from 1 to the number of patients.
<i>hospital</i>	Name of the medical center where the image was created (in Polish).
<i>hospital_eng</i>	Name of the medical center where the image was created (translated to English).
<i>hospital_id</i>	Unique hospital identifier ranging from 1 to 15.
<i>sex</i>	Patient sex.
<i>age</i>	Patient age in years.
<i>smoke</i>	Smoking status: “Yes” for smokers, “No” for non-smokers.
<i>smoke_packyears</i>	Number of pack-years for smokers.
<i>class</i>	Diagnosis: “COVID-19” for COVID-19, “PNEUMONIA” for types of pneumonia other than COVID-19-related, and “NORMAL” for the remaining cases.
<i>class_id</i>	Class identifier: 1 - normal, 2 - pneumonia, 3 - COVID-19.
<i>quality</i>	Image quality category: “Good” - sufficient quality, “Bad” - insufficient quality. The criteria for quality assessment are described in the Technical Validation section.
<i>subtype</i>	Subtype label: “C1”, “C2”, “C3” for COVID-19; “P1”, “P2”, “P3” for pneumonia other than COVID-19-related; “N1”, “N2”, “N3” for the remaining cases.
<i>set</i>	Set to which the image was included in Prazuch <i>et al.</i> ¹⁶ : “train” – training set, “hold-out test” – testing set.

Table 6. Definition of variables included in the POLCOVID metadata file.

each group separately (Table 5). They contain patient demographic and clinical data, group and subtype labels, information regarding the hospital of data collection, and image quality category. Table 6 defines the variables included in the metadata.

Lung segmentation image collection. We provide the manually created lung masks in the PNG format for all CXRs images used to generate the lung segmentation model (radiographs delivered by the POLCOVID Study Group and collected from the publicly available databases). Moreover, the original de-identified versions of POLCOVID CXR images used for the manual mask annotation are also available. Registered users may download the POLCOVID unprocessed images as the TIFF files and all lung masks in the PNG format for each data source separately from Synapse (www.synapse.org/#!Synapse:syn50877085; <https://doi.org/10.7303/syn50877085>)²¹ and CIRCA COVID-19 CXR/CT-based diagnosis web service.

The files are divided into two parts: original CXRs and manually created lung masks. The lung masks are organized in three individual ZIP files, one per source collection. The organization and content of provided files is shown in Table 7.

Lung segmentation metadata. The metadata for radiographs used to create the lung segmentation model are available in the CSV file for registered users from Synapse (www.synapse.org/#!Synapse:syn50877085; <https://doi.org/10.7303/syn50877085>)²¹ and the CIRCA COVID-19 CXR/CT-based diagnosis web service. They contain information regarding the data source, the file names consistent with those used by data providers, and the subset

Record type	Source collection	Filename	#Files
Metadata	All	metadata_segmentation.csv	1
Original CXRs	POLCOVID	POLCOVID.zip	4003
Manually created lung masks	POLCOVID	POLCOVID_masks.zip	4003
	National Institute of Health – Clinical Center	NIH.zip	1124
	Shenzhen No.3 Hospital, Shenzhen, China	SHENZHEN.zip	662
	Department of Health and Human Services of Montgomery County, USA	DHHS.zip	138
	Guangzhou Women and Children's Medical Center, Guangzhou, China	GUANGZHOU.zip	370

Table 7. File organization of the lung segmentation dataset.

Variable Name	Definition
<i>source</i>	Name of dataset
<i>source_id</i>	Dataset abbreviation: “POLCOVID” for the POLCOVID dataset; “NIH” for National Institute of Health – Clinical Center ¹² ; “SHENZHEN” for Shenzhen No.3 Hospital, Shenzhen, China ¹³ ; “DHHS” for Department of Health and Human Services of Montgomery County, USA ¹³ ; “GUANGZHOU” for Guangzhou Women and Children's Medical Center, Guangzhou, China ¹⁴ .
<i>filename</i>	Anonymized unique file name: for POLCOVID Anonymus_<hospital_id>_<patient_id>_<class_id>.<file_format>; for the remaining datasets the name of the file given by the data provider.
<i>set</i>	Set to which the image was included during the generation of the lung segmentation model: “train” – training set, “validation” – validation set, “hold-out test” – testing set.

Table 8. Definition of variables included in the lung segmentation metadata file.

to which we assigned an image in the model generation process. Table 8 defines the variables included in the metadata.

Technical Validation

Anonymization. We carefully de-identified all radiographs. We deleted all identifiable metadata stored in DICOM objects and manually reviewed all image data. All personal information on radiographs was also removed.

Data quality control. We curated the database based on the DICOM headers when available. We visually inspected every X-ray image and removed all radiographs with lateral projections, incomplete lung regions, and improperly saved or stored. We reviewed the clinical data for consistency and filled in the missing demographic fields if an uploader provided the lacking information elsewhere.

Moreover, we further investigated the image quality. We selected very low-resolution images characterized by lung area smaller than 300 pixels in height or width. We also identified the radiographs whose quality prevents proper lung segmentation, leaving one or both lungs mostly or entirely undetected. We characterized the segmentation quality by the score defined as the mean value of four lung mask properties: eccentricity, orientation, area, and solidity, as explained in Prazuch *et al.*¹⁶. The lung segmentation quality score was normalized to range from 0 to 1. We identified poor-quality images with outlying quality scores with the outlier detection method dedicated to skewed data²².

The lung segmentation model performed satisfactorily with SDC equal to 94.86% and 93.36% for the validation and testing datasets, respectively. We moreover visually inspected the obtained lung masks to ensure the high quality of the segmentation process.

Usage Notes

Only registered users are permitted to download the data from Synapse repository (www.synapse.org/#!/Synapse:syn50877085; <https://doi.org/10.7303/syn50877085>)²¹. However, anyone can view the project and its documentation. Similarly, at the CIRCA COVID-19 CXR/CT-based diagnosis web service, the user is required to register and provide the name, institution, e-mail address, and the purpose of data usage. We recommend IrfanView software for previewing CXRs images converted to TIFF format. For any publication using these data, the authors must cite this original paper. The data are available under the CC-BY license.

Code availability

The code used for generating preprocessed images and lung masks from the original CXR images is available on GitHub (<https://github.com/ZAEDPolSI/PolCovid>).

Received: 16 February 2023; Accepted: 11 May 2023;

Published online: 02 June 2023

References

- Haldane, V. *et al.* Health systems resilience in managing the COVID-19 pandemic: lessons from 28 countries. *Nat Med* **27**, 964–980 (2021).
- Vandenberg, O., Martiny, D., Rochas, O., van Belkum, A. & Kozlakidis, Z. Considerations for diagnostic COVID-19 tests. *Nat Rev Microbiol* **19**, 171–183 (2021).
- Cleverley, J., Piper, J. & Jones, M. M. The role of chest radiography in confirming covid-19 pneumonia. *BMJ* **370**, m2426 (2020).
- Wong, H. Y. F. *et al.* Frequency and distribution of chest radiographic findings in patients positive for COVID-19. *Radiology* **296**, E72–E78 (2020).
- Wang, L., Lin, Z. Q. & Wong, A. COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. *Sci Rep* **10**, 19549 (2020).
- Minaee, S., Kafieh, R., Sonka, M., Yazdani, S. & Soufi, G. J. Deep-COVID: Predicting COVID-19 from chest X-ray images using deep transfer learning. *Med Image Anal* **65**, 101794 (2020).
- Ismael, A. M. & Şengür, A. Deep learning approaches for COVID-19 detection based on chest X-ray images. *Expert Syst Appl* **164**, 114054 (2021).
- Santa Cruz, B. G., Bossa, M. N., Sölter, J. & Husch, A. D. Public covid-19 x-ray datasets and their impact on model bias—a systematic review of a significant problem. *Med Image Anal* **74**, 102225 (2021).
- World Health Organization. *Tracking SARS-CoV-2 variants*. <https://www.who.int/activities/tracking-SARS-CoV-2-variants>
- DICOM. In *NEMA PS3/ISO 12052 Digital Imaging and Communications in Medicine (DICOM) Standard*, (NEMA, Rosslyn, VA, 2020).
- Ronneberger, O., Fischer, P. & Brox, T. in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science* Vol. 9351 (ed. Navab, N., Hornegger, J., Wells, W. & Frangi, A.) U-Net: Convolutional networks for biomedical image segmentation (Springer, Cham, 2015).
- Wang, X. *et al.* in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases (IEEE, 2017).
- Jaeger, S. *et al.* Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. *Quant Imaging Med Surg* **4**, 475 (2014).
- Kermany, D. S. *et al.* Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* **172**, 1122–1131 (2018).
- Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. in *3rd International Conference on Learning Representations (ICLR, 2015)*.
- Prazuch, W. *et al.* CIRCA: comprehensible online system in support of chest X-rays-based COVID-19 diagnosis. Preprint at <https://arxiv.org/abs/2210.05440> (2022).
- Suwalska, A. *et al.* in *Recent Advances in Computational Oncology and Personalized Medicine* (ed. Krukiewicz, K., Marczyk, M. & Polanska, J.) Ch. 12 (Silesian University of Technology, 2022).
- Chowdhury, M. E. *et al.* Can AI help in screening viral and COVID-19 pneumonia? *IEEE Access* **8**, 132665–132676 (2020).
- Soda, P. *et al.* AIforCOVID: Predicting the clinical outcomes in patients with COVID-19 applying AI to chest-X-rays. An Italian multicentre study. *Med Image Anal* **74**, 102216 (2021).
- Marczyk, M. Mixture modeling of 2-D gel electrophoresis spots enhances the performance of spot detection. *IEEE Trans. Nanobioscience* **16**, 91–99 (2017).
- Suwalska, A. *et al.* POLCOVID – a multicenter multiclass chest X-ray database (Poland, 2020–2021). *Synapse* <https://doi.org/10.7303/syn50877085> (2023).
- Hubert, M. & Van der Veeken, S. Outlier detection for skewed data. *Journal of Chemometrics* **22**, 235–246 (2008).

Acknowledgements

This work was supported in part by the National Science Centre, Poland grant no MNiSW/2/WFSN/2020, and Silesian University of Technology grant no. 02/070/BK_22/0033 for Support and Development of Research Potential. Calculations were carried out using GeCONil infrastructure funded by NCBiR project no. POIG.02.03.01-24-099/13. Additionally, AS and WP are holders of the European Union scholarship through the European Social Fund, grant POWR.03.05.00-00-Z305, and JT is the holder of a European Union scholarship through the European Social Fund, grant no. POWR.03.02.00-00-I029. In memory of a good friend, we would like to thank and dedicate this work to Dr Franciszek Binczyk, whose great contribution and support made it possible to conduct the project and publish this dataset.

Author contributions

A.S., J.T., W.P., M.S. and M.M. contributed equally. J.P., M.M. and A.C. conceived the idea of the study. P.F. created the web service and database and gave technical support. J.J., K.G., M.S.I., J.W., T.P., G.P., M.N., P.Fi., M.P., R.F., K.S., G.Z., B.G., E.S., A.C. and POLCOVID Study Group collected the clinical and imaging data. W.P., M.S. and A.S. de-identified the data. M.M., W.P., M.S., A.S. and J.T. investigated image quality and data consistency. M.M., W.P., M.S., A.S. and J.T. performed data curation and validation. A.S. and J.T. prepared metadata. M.M. and J.T. prepared figures. J.T. prepared data summaries and wrote the manuscript. All authors reviewed and contributed to the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to M.M. or A.C.

Reprints and permissions information is available at www.nature.com/reprints.

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



Open Access 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 license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license 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 license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023

for the POLCOVID Study Group Investigators

Joanna Polanska¹, Michal Marczyk^{1,17}, Wojciech Prazuch¹, Aleksandra Suwalska¹, Marek Socha¹, Pawel Foszner^{1,2}, Joanna Tobiasz^{1,2}, Mateusz Nowak⁹, Piotr Fiedor¹⁰, Andrzej Cieszanowski¹⁸, Agnieszka Oronowicz-Jaskowiak¹⁹, Bogumil Golebiewski¹⁹, Krzysztof Simon¹³, Magdalena Sliwinska⁵, Mateusz Rataj⁵, Przemyslaw Chmielarz⁵, Adrianna Tur²⁰, Grzegorz Drabik²⁰, Tadeusz Popiela⁷, Justyna Kozub⁷, Grzegorz Przybylski⁸, Anna Kozanecka⁸, Edyta Szurowska¹⁶, Sebastian Hildebrandt²¹, Katarzyna Krutul-Walenciej²¹, Jan Baron⁴, Katarzyna Gruszczynska⁴, Jerzy Jaroszewicz³, Damian Piotrowski³, Jerzy Walecki⁶, Piotr Wasilewski⁶, Samuel Mazur⁶, Robert Flisiak¹², Gabriela Zapolska¹⁴, Krzysztof Klaude¹⁴, Katarzyna Rataj¹⁴, Bogumil Golebiewski¹⁴, Malgorzata Pawlowska¹¹, Piotr Rabiko¹¹, Pawel Rajewski¹¹, Barbara Gizycka¹⁵, Piotr Blewaska²², Katarzyna Sznajder²³ & Robert Plesniak²⁴

¹⁹Department of Imaging Diagnostics, The Maria Skłodowska-Curie National Research Institute of Oncology, Warsaw, Poland. ²⁰Prognostic Specialist Clinic, Knurów, Poland. ²¹Central Clinical Hospital, Medical University of Gdansk, Gdansk, Poland. ²²District Hospital, Raciborz, Poland. ²³University Clinical Hospital, Opole, Poland. ²⁴University of Rzeszów, Medical Center, Lancut, Poland.