scientific reports

Check for updates

OPEN

Biases associated with database structure for COVID-19 detection in X-ray images

Daniel Arias-Garzón[1,6], Reinel Tabares-Soto[1,2,5,6], Joshua Bernal-Salcedo[1] & Gonzalo A. Ruz[2,3,4]✉

Several artificial intelligence algorithms have been developed for COVID-19-related topics. One that has been common is the COVID-19 diagnosis using chest X-rays, where the eagerness to obtain early results has triggered the construction of a series of datasets where bias management has not been thorough from the point of view of patient information, capture conditions, class imbalance, and careless mixtures of multiple datasets. This paper analyses 19 datasets of COVID-19 chest X-ray images, identifying potential biases. Moreover, computational experiments were conducted using one of the most popular datasets in this domain, which obtains a 96.19% of classification accuracy on the complete dataset. Nevertheless, when evaluated with theethical tool Aequitas, it fails on all the metrics. Ethical tools enhanced with some distribution and image quality considerationsare the keys to developing or choosing a dataset with fewer bias issues. We aim to provide broad research on dataset problems,tools, and suggestions for future dataset developments and COVID-19 applications using chest X-ray images.

Since its first appearance in December 2019 in Wuhan, China[1], the novel coronavirus became a worldwide pandemic on March 11, 2020[2]. With its exponential growth and the early lack of an effective vaccine or treatment, health professionals focused on the disease's early diagnosis.

Reverse transcription-polymerase chain reaction (RT-PCR) is the leading diagnosis test for COVID-19. Still, the processing time is long, and the cost is high; enhancing this situation with many tests per day makes the diagnosis slow. Under the circumstances, effective separation of the patients and treatment has no support for early diagnosis[3]. Several studies have shown that chest radiograph (CXR) and computed tomography findings are typical of COVID-19-associated pneumonia[4–6]. For their lower cost compared to computed tomography, X-ray images are valuable assets for COVID-19 recognition (classification) and prognosis (Triage analysis for knowing the best treatment)[7]. Research, in particular, has focused on developing artificial intelligence (AI) models to support the diagnosis of COVID-19 using medical images[8–10].

The use of AI in diagnosing and triaging suspect COVID-19 patients can enhance the task of distinguishing COVID-19 cases from other cases, even if there is a different type of pneumonia associated. However, some claim it is possible to separate COVID-19 issues from normal ones and ones with bacterial or viral pneumonia. This potential uncertainty could be a limitation for a proper clinical application, considering that the algorithm may not be able to identify the illness, which could lead to false positives or false negative diagnoses. High-accuracy models are shown in the literature, but papers focus on obtaining high-accuracy results but do not consider possible biases present in the datasets used. As a result, some previous articles have studied the bias in datasets related to chest X-ray images in COVID-19[11–13].

In this paper, we search 46 articles that use AI to detect or triage COVID-19 on X-ray chest images, with accuracy results higher than 90.

As these papers show, COVID-19 datasets were developed in the rise of the pandemic event caused by the COVID-19 spread. Because there were no similar datasets before the sampling process and proper patient selection was rarely even implemented, most of these datasets are images that may contain or not COVID-19 of one or multiple health institutions. Previous work on COVID-19 detection[14] and the development of datasets using

[1]Departamento de Electrónica y Automatización, Universidad Autónoma de Manizales, Manizales 170001, Colombia. [2]Facultad de Ingeniería y Ciencias, Universidad Adolfo Ibáñez, 7941169 Santiago, Chile. [3]Center of Applied Ecology and Sustainability (CAPES), 8331150 Santiago, Chile. [4]Data Observatory Foundation, 7941169 Santiago, Chile. [5]Departamento de Sistemas e Informática, Universidad de Caldas, Manizales 170001, Colombia. [6]These authors contributed equally: Daniel Arias-Garzón and Reinel Tabares-Soto. ✉email: gonzalo.ruz@uai.cl

images of S.E.S Hospital de Caldas, showed, after some experiments, that the hospital tests were unsatisfactory, even if training and testing results were high. We saw a bias in terms of the device used for the image capture. An improved dataset that took into account the possible bias and tried to avoid it; has been recently developed[15].

After researching the metadata provided by some of the most used datasets, we noticed that in terms of patient characteristics and capture conditions are only provided sometimes, or they need to be better structured. Yet, these articles require further studies of the possible biases (and what are the sources of those biases) and show how to measure the bias in a dataset.

With the rise of AI worldwide, many algorithms are proven to contain several types of biases. AI Ethics began to grow as a study topic. Some guidelines have demonstrated effectiveness in regulating AI development and following ethical criteria[16–18]. Following this tendency, we were motivated to develop a two-stage experiment in which we can prove biases, based on medical studies that support that images change (or are affected) depending on the capture conditions and patient characteristics and then make a parity test to test the models using ethical tools and statistical processes. Thus, the contributions of this paper are, first, that we aim not only to provide a more profound argumentation with facts that the datasets have biases but also we conduct a specific experiment on the most used dataset found. Second, we use an ethical tool named Aequitas[19] that identifies biases in terms of location, sex, and age, among others, on an experiment with a 96.19

The rest of the paper is organized as follows. First, "Databases overview" Section presents the strategy for finding the databases and their corresponding articles. An overview of the database's information is given in "Bias associated to the datasets" Section. "Methodology for bias identification in the COVID-19 datasets" Section presents a classification of the types of biases for this problem. Whereas in "Methodology for bias identification in the COVID-19 datasets" section, the methodology for identifying biases present in the COVID-19 databases is described. The experiment using Cohen's dataset[20] and the results of the ethical tool are presented in "Discussion" Section. We finish with a discussion of biases and the analysis of the ethical tool in "Conclusion and recommendations" Section, and a conclusion of the study with some recommendations in "Future work" Section.

Paper search strategy

We search for papers on journal databases such as ScienceDirect, PubMed, IEEE, and Google Scholar and software repositories like GitHub or Kaggle. This search aimed to find documentation related to COVID-19 detection, classification, diagnosis, prognosis, or triage on chest X-ray images (CXR). Detection, classification, and diagnosis can be seen as a classification task where Covid cases are compared with a control group of images. Meanwhile, the triage task consists of comparing positive cases of COVID-19 to know the severity of the affectation and predict future possibilities for the patient. In terms of the excluded articles criteria, we found the following parameters.

- Articles that use another type of data different from Chest X-ray images to classify or triage (Examples CT and other applications of diagnosis validation).
- The dataset used is open. There were only two articles with private datasets and merely for showing that datasets that are not possible have a Bias validation.
- We avoid the review articles.
- In these parts, for most of the papers, we do not consider the article in which the dataset was presented.

We found 46 papers in total; 39 correspond to the classification task[8–10,14,21–55], 5 to the triage task[56–60], and two do both tasks[61,62].

Table 1 Contains the number of subjects in the database, and the metadata of each database, with this they can be associated if there are shared labels within the datasets and if the distributions of the labels are similar to be able to consider the mixture of the datasets to initially avoid biases due to irregular distributions, the references of those papers, and where to download the database. If the database is unavailable, it corresponds to a private database.

Databases overview

After the search, we obtained 21 datasets, some have parts of the others, and some do not appear in Table 1 because they are a mix of the datasets mentioned. On Mixed datasets is found COVIDx (composed by Cohen, Fig. 1, Actualmed, Qatar University, and an RSNA Pneumonia dataset for control images that appears in Table 2), QaTa (composed by SIRM, Radiopaedia, and Chest Imaging), BrixIA (which contains part of Cohen with a few changes), RSNA, Radiopaedia, and SIRM (small datasets usually used together), COVID-QU (contains QaTa, a Covid GitHub repository, Eurorad, Cohen, SIRM, Qatar University, COVID-CXNet Images, RSNA, Chest X-Ray Images (Pneumonia), and Padchest) which is the largest dataset available found with the addition that every image on the dataset has a lung segmentation mask and RYDLS(Cohen for covid, Radiopaedia for varicella and Mers and Chest X-ray 8 for normal). On the other hand, there are private datasets, which means there is no open access to download them. The datasets in this condition are Henry Ford Health System and CHUAC. Similarly, we could access other available upon request, like BrixIA (to obtain the complete dataset), AlforCOVID, and nd BIMCV. The rest of the datasets are fully open without any requirements.

We believe three main aspects characterize these datasets. First, the size of the datasets, then the type of images, and finally, but not less importantly, the metadata associated with the dataset in general. In terms of dataset size, Fig. 1 shows the number and percentage of types of images in each dataset.

COVID-19 images are validated in two ways, first, by a diagnostic test such as RT-PCR or by validation of characteristics findings by an expert. Other pathologies correspond to any pathology that is not Pneumonia or Covid. Inconclusive cases have no certainty on the diagnosis, or the radiologic report does not contain the

Database	Subjects	Metadata	References	Download page
Cohen[20]	332	Sex-(Image Amount), Age-(Included), Location-(Europe>Others), Dates-(Included), Others-(ICU admission-Survival model).	8–10,24–31,55 32–38,41,42,44,45,52–54	Cohen
BIMCV[63]	4899 positive 5242 negative	Sex-(Patient amount), Age-(Included), Location-(Spain) Dates-(Included),Others-(Classification test).	14,22,43,46,47,62	BIMCV-COVID19
Cancer Image Archive[64]	105	Sex-(Image/patient amount), Age-(Included), Location-(EE.UU) Dates-(Not included),Others-(ICU admit-Mortality).	43,46	Cancer Image Archive
ML Hannover[65]	71	Sex-(Image amount), Age-(Not included), Location-(Germany) Dates-(Not included),Others-(ICU admission offset-Death offset).	43	ML Hannover
BrixIA[57]	2351	Sex-(Image amount), Age-(Included), Location-(Italy) Device-(Included),Dates-(Included).	23,48,57,62	BrixIA
Actualmed COVID-19 chest X-rays[66]	216	Sex-(Not included), Age-(Not included) Location-(Spain), Dates-(Not included).	41,42,48,49	Actualmed
RYDLS-20[50]	Not mentioned	Sex-(Not included), Age-(Not included) Location-(Not specified), Dates-(Not included).	50,53	RYDLS-20
COVID-19 Radiography Database (Qatar university)[67]	Not mentioned	Sex-(Not included), Age-(Not included) Location-(Not specified), Dates-(Not included).	40–42,45,46,48,49	Qatar university
SIRM[68]	65	Sex-(Patient amount), Age-(Included) Location-(Italy),Dates-(Not included), Other-(Some symthoms).	10,60,69	SIRM
CHUAC dataset[51]	Privated not validated	Sex-(Private not validated),Age-(Private not validated),Location-(Spain).	51	Not available for downloading
Radiopaedia[70]	16	Sex-(Patient amount), Age-(Included) Location-(Not mentioned), Dates-(Not mentioned).	10,69	Radiopaedia
Eurorad[71]	41	Sex-(Patient amount), Age-(Included) Location-(Europa, America, Asia and Oceania),Dates-(Included).	10	Eurorad
AlforCOVID[72]	Not mentioned	Sex-(Image amount), Age-(Included) Location-(Not specified), Dates-(Not mentioned), Other-(ICU admision, Death and Prognosis).	48	AlforCOVID
Chest Imaging[73]	50	Sex-(Patient amount), Age-(Included) Location-(Spain), Dates-(Not specified)	10,60,69	Chest Imaging COVID-19 CXR Spain
BSTI[74]	40	Sex-(Patient amound), Age-(Included) Location-(UK),Dates-(Included)	10	BSTI
Henry Ford Health System[21]	2060	Sex-(Patient amound), Age-(Included) Location-(EE.UU),Dates-(Included)	21	Not available for downloading
Figure 1 Covid-19 Chest X ray dataset[75]	48	Sex-(Image amount), Age-(Inlcuded) Location-(Not specified), Dates-(Not included)	40–42,46,48,49,60	Figure 1
Covid-QU[76]	Not mentioned	Sex-(Not included), Age-(Not included) Location-(Not specified), Dates-(Not included), Others-(Lungs segmentation mask)	61	Covid-QU

Table 1. Databases found with the metadata provided, references, and download information.

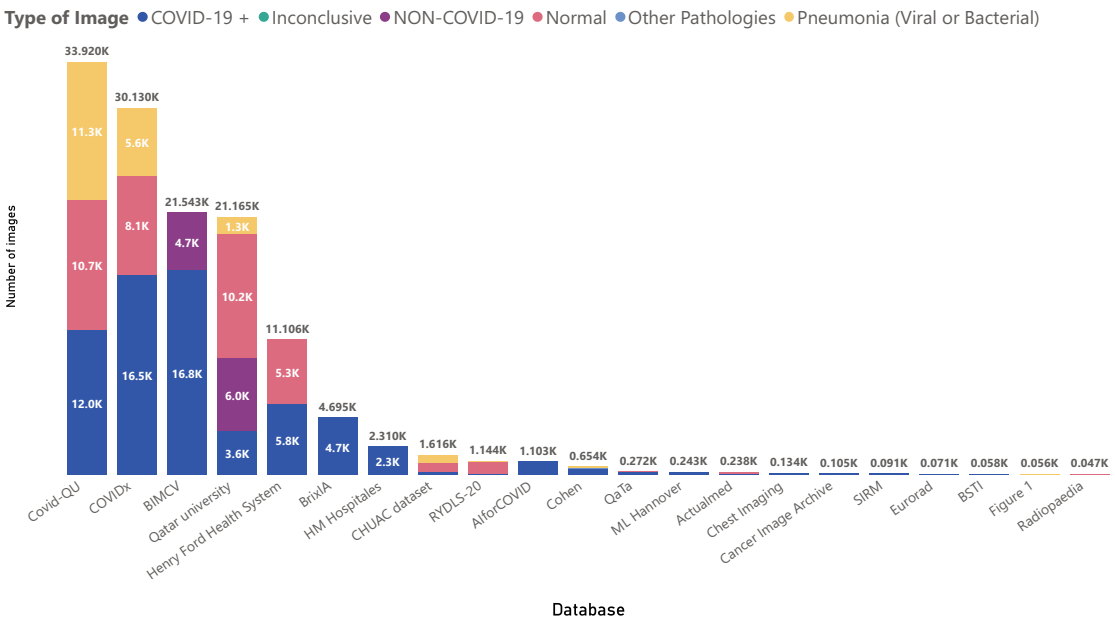
associated diagnostic. NON-COVID-19 is a control category that groups images with diverse pathologies and healthy patients. This is mainly used in binary classification so that the models could detect a COVID diagnosis among many possible images, not only of other types of pneumonia or Normal images.

In terms of the content of the datasets, this refers to the image format (normally DICOM or PNG). Image view for area coverage in classification anteroposterior (AP), posteroanterior (PA), or similar views is the leading used, and the lateral test is usually avoided.

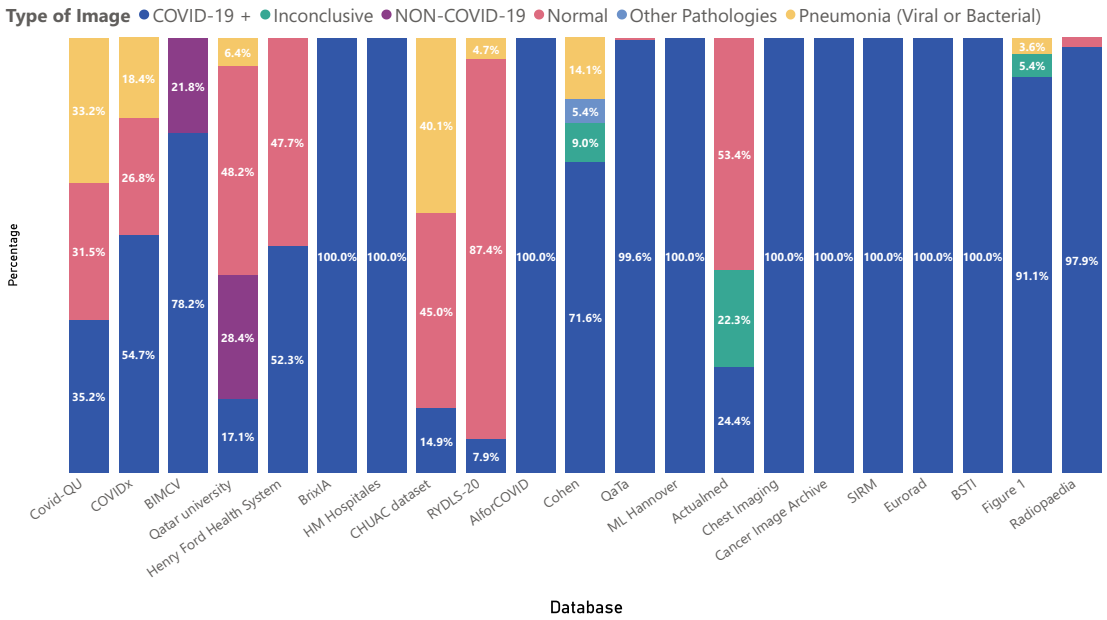
There are some peculiar distributions on some metadata for the datasets. Figure 2 shows the atypical distribution for Age, and Fig. 3 shows the Cohen Study date distribution that is different from what it should be because the information came from various sources. There are problems with the proper label of the metadata. For more details regarding the metadata of the datasets, Table 1, Figs. 1 and 2 of supplementary material show more complete information on the metadata of each COVID database.

For these datasets, it is important to consider how damaging it can be to mix positive and negative COVID-19 databases. Table 2 shows the databases used for control or to associate with other diseases, with the same information as provided in Table 1.

As Table 2 Shows that there are eight control image datasets used. Most of these datasets were used for Pneumonia tasks, such as detection and severity or triage tasks. Still, there are also datasets for lung segmentation and simply significant X-ray recollection of different pathologies. There are no private datasets among them, and



a) Number of Images per database



b) Percentage of Images per database

Figure 1. Number and Percentage of images according to a specific pathology on each dataset.

with a request, we found CheXpert and JSRT. The rest are open, but the RSNA Pneumonia Detection Challenge dataset used in many experiments and as part of the COVIDx dataset is a subset of Chest X-ray 8 and Chest X-ray 8 in its last version is also called Chest X-ray 14. Also, Chest X-ray Images (Pneumonia) is a project mixed with Optical Coherence Tomographies so that it can also be found as "Large Dataset of Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images"[85]. As on positive ones, Fig. 4 show the percentage and quantity of images associated with specific pathologies.

In terms of Control databases, the metadata associated is sometimes more organized. Still, we also found strange features such as Chest X-Ray images (Pneumonia) and age features that are not specified, but the CXRs are from a pediatric hospital. For more details on Control datasets metadata, follow Table 2 and Fig. 3 of the supplementary material.

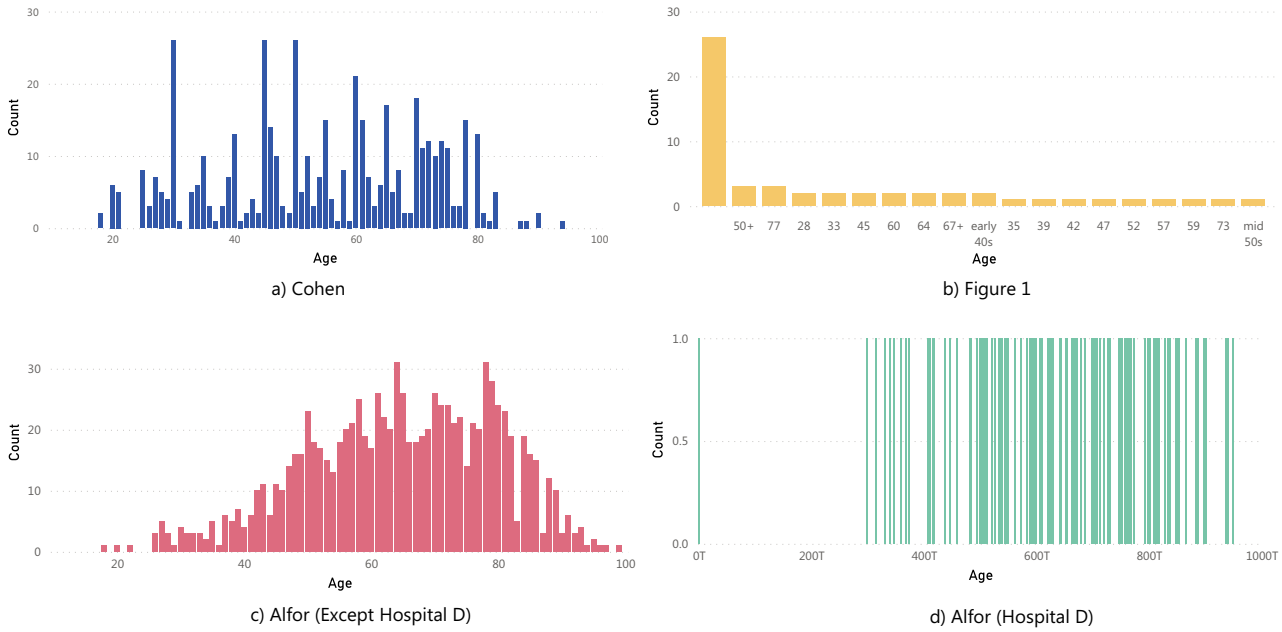


Figure 2. Metadata of some datasets related to the age of the patient with peculiar distributions.

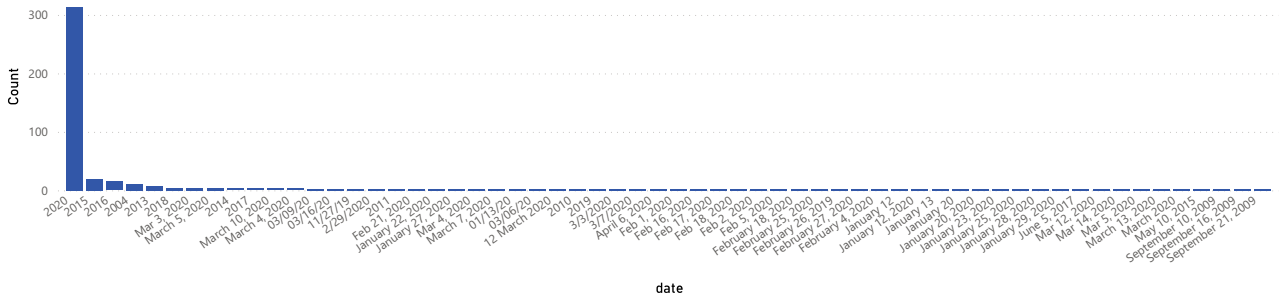
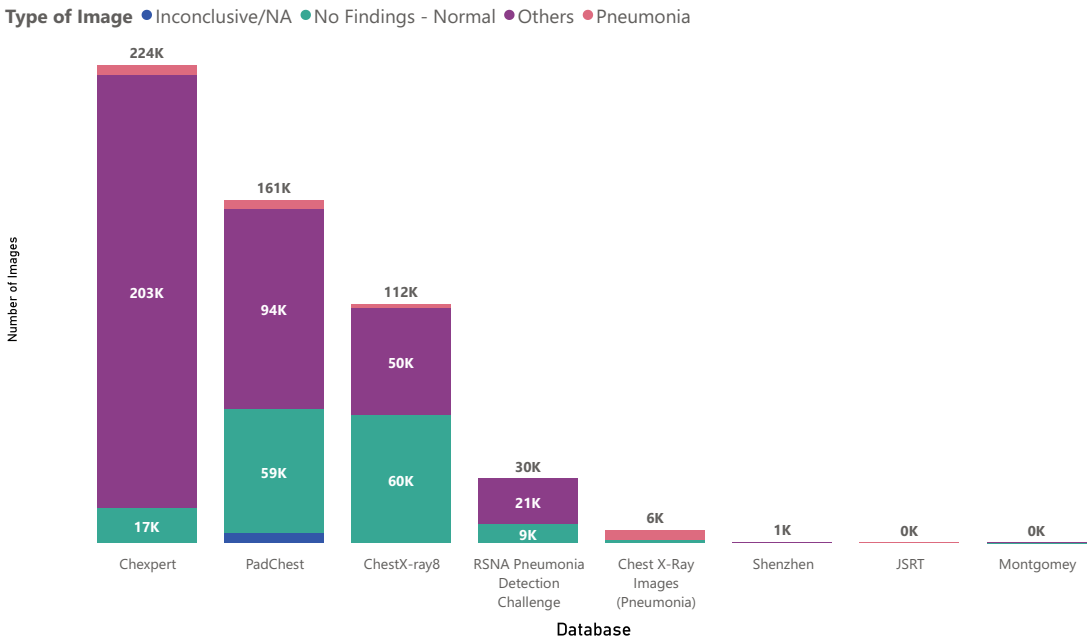


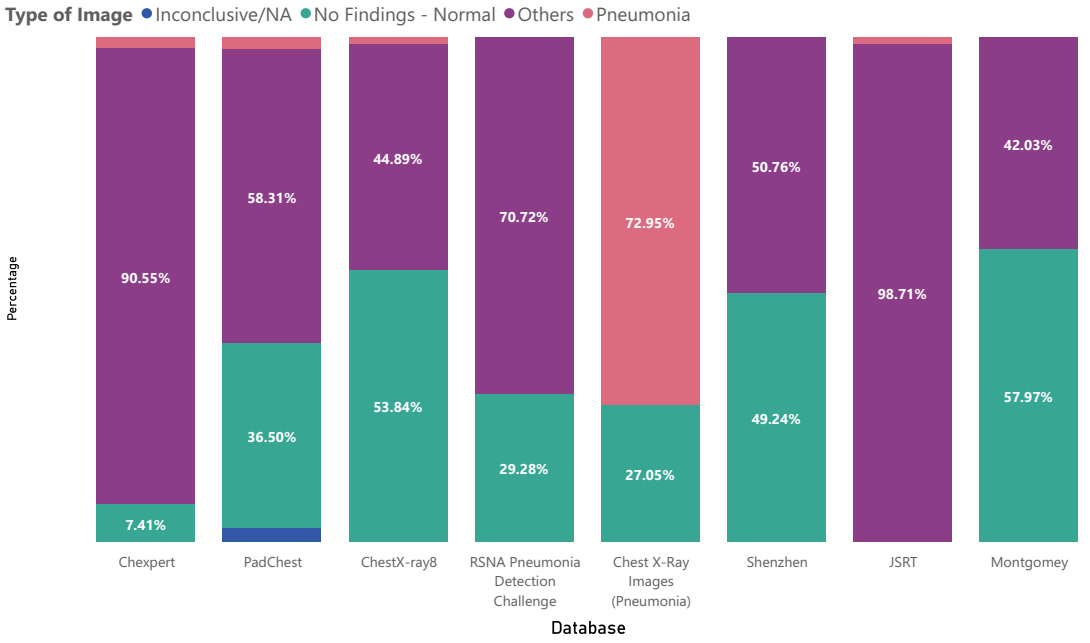
Figure 3. Cohen dataset metadata related to the date of the study (showing inconsistency on this data label).

Database	Subjects	Metadata	References	Download page
ChestX-ray 8[77]	30805	Sex-(Images amount), Age-(Included) Location-(No mentioned), Dates-(2017-2020), Others-(Findings)	10,23,33,44,47,54	Link
CheXpert[78]	64540	Sex-(Images amount), Age-(Included), Location-(EE.UU) Dates-(2002-2017), Others-(Findings)	10,35,43,54,62,79	Link
PadChest[80]	67625	Sex-(Images amount), Age-(Included), Location-(Spain) Dates-(2009-2017), Others-(Findings-symptoms)	14,22,43	Link
Chest X-Ray Images(Pneumonia)[81]	5856	Sex-(Not mentioned), Age-(Children) Location-(China), Dates-(Not mentioned)	8,9,25,27–29,31–34,36–40,45,52,54,55,58	Link
RSNA PneumoniaDetection Challenge[82]	Not mentioned	Sex-(Not mentioned), Age-(Not mentioned) Location-(China), Dates-(Not mentioned)	8,30,40–42,46,48,49,52,53,56	Link
JSRT[83]	Not mentioned	Sex-(Images amount), Age-(Included) Location-(Japan-EE.UU), Dates-(Not mentioned), Others-(Diagnosis)	28,53	Link
Montgomey[84]	Not mentioned	Sex-(Images amount), Age-(Included), Location-(EE.UU) Device-(Included), Dates-(Not mentioned), Others-(Findings)	8,53	Link
Shenzhen[84]	Not mentioned	Sex-(Images amount), Age-(Included), Location-(China) Deviced-(Included), Dates-(2012), Others-(Findings)	8,53	Link

Table 2. Control databases found with the number of papers that use it, references, and where to find the database.



a) Number of Images per database



b) Percentage of Images per database

Figure 4. Number and Percentage of images according to a specific pathology on each Control dataset.

Bias associated to the datasets

Bias in AI can come from many sources, and it is essential to develop equitable systems, generating the lowest bias and relying on ethics tools in AI to prevent models from being racist or sexist, generating problems of discrimination and assumptions that can generate a decrease in the performance of the model, which results in a lack of reliability for the use of these tools. In terms of medical images, many factors can affect the performance of the model generation bias. We want to group by bias associated with COVID-19 Detection or Triage in CXR in four possible groups. Some articles before have shown the possible bias risk in this dataset[86].

Bias associated with patient information. People come with different characteristics, and the physical characteristics mainly cause images to look different and that pathologies express differently. Hence, in these cases, we think there are four main reasons that a dataset could have a bias taking into account information of the patients.

- *Sex* The sex can affect CXR mainly because breast tissue in some women is opaque in parts of the images, so datasets with a high number of women compared to men could generate some bias in COVID-19 early stages because the images are initially a little more opaque.
- *Age* The age affects these images in two terms. First, patients with more images are usually old, so the age distribution is generally 60–80 years old. Also, pediatric X-ray chests are not only of the chest (in the image also appears other parts of the patient's body). Hence, the images differ from the others, and the quantity is considerably less. Age affects the opaque of the images in terms of bone density. Some older people have x-ray findings for some habits like smoking.
- *Patients distribution* The label of patients is super important because in the moment of data splitting for training and testing, using the same patients on both datasets could generate that the model identifies the patients but not the pathology (overfitting).
- *Demographic characteristics* The characteristics of the general population change in every community of the planet, so datasets with many countries in which there are more images from one hospital/city/country than the rest generate that the patient's characteristics, evolution, and devices used to capture the images change. These changes can cause the model to be biased to recognize images coming from a particular location.

For this group of bias, Fig. 5 shows the comparison between a child image from the Chest X-Ray Images (Pneumonia), specifically person_97_virus_180.jpeg, and two adult images (nejmc2001573_f1a.jpeg and 7EF28E12-F628-4BEC-A8C5-E6277C2E4F60.png) of different sex of the Cohen dataset that correspond both to COVID-19 cases.

Bias associated to capture conditions. The way an image is captured is significant. It can vary much in different devices, hospitals, and countries, so we think three main factors on capturing affect the resulting images, thus, generating bias that may influence the models that use them.

- The device used: many types of devices with different X-ray frequencies affect an image's shades. Also, portable devices do not have the same quality as normal ones. The unification of this factor is difficult, making the models in many cases biased because the devices used are not the same in COVID-19 images and on the Control dataset, enhanced by the fact that even in the same hospital, if a COVID-19 patient is in the ICU the device must be portable, meanwhile for control cases, the device used is the normal one because there is no limitation on movement.
- Good image capture: even though trained personnel usually take images, some of these could get captured wrong. For example, if the picture is taken on expiration instead of inspiration, the image will be more opaque. If the patient moves in the capture, it could be in a position not suitable for the diagnosis.
- Cables or tubes: if a patient is in the ICU and has ECG cables, nasopharyngeal intubation, and invasive pressure devices, among others, for classifying severity or even COVID-19 from other diseases, the models could learn to identify cables instead of the disease.

In this case, Fig. 6 shows examples of different capture conditions on the Cohen dataset. It is found that 000005-5-a.jpg is bluer and zoomed than the others, 000012.jpg is yellow, 11547_2020_1202_Fig1_HTML-a.png has labels, and ajr.20.23034.pdf-003.png is whiter than the others and with cables. Also, we found the images 108115246579239728.dcm and 46529543479051320.dcm that are DICOM images from BrixIA that use Siemens and Agfa devices, respectively and for c) and d) sub-S03562_ses-E07248_run-2_bp-chest_vp-ap_cr.png from the BIMCV dataset of a positive case, a 90-year-old male in its original 16 bits PNG format and fixed by changing it with python cv2.IMREAD_GRAYSCALE format.

Bias associated with unbalanced datasets. Unbalanced datasets are standard in AI problems; in this context, it is not the exception. In particular, COVID-19-positive images are low in comparison with other CXR images. Using it how it is presented may cause the system not to find a remarkable disease pattern. There are two prominent cases to solve this issue. We may balance to the less quantity class, in these cases mainly COVID-

a) Child Image with viral Pneumonia b) 60 years old male Image with COVID-19 c) 52 years old female Image with COVID-19

Figure 5. Examples of bias associated with the patient's information.

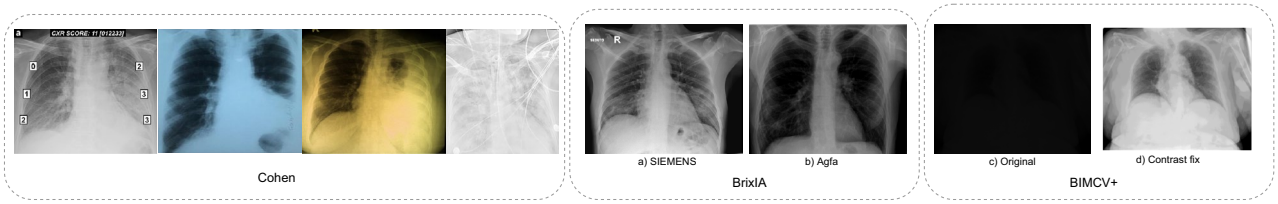


Figure 6. Examples of bias associated to capture conditions in different datasets.

19 type that is not suitable because the models will get a small number of images that could affect the model's generalization power. The other alternative is to use data augmentation for models to have more information. Data augmentation in medical images is quite risky, mainly because some data augmentation processes, such as adding Gaussian noise, may affect the shades on the images and could generate bias on radiological findings associated to a specific pathology different from COVID-19. So the main problem of this approach is the lack of formal clinical validation of the resulting images.

Bias associated with mixing datasets. Mixing datasets is a standard practice in AI. When only a few COVID-19 images are available (the most common case), mixing with other datasets is helpful because this would be an approach to avoid the unbalanced problem. At the same time, it helps with variability during training allowing the trained model to potentially generalize better. Nevertheless, a common mistake in COVID-19 dataset fusion, mainly on classification tasks, is that by mixing many datasets of COVID-19 and using as Control images another dataset, there will be many variable characteristics on the side of the COVID-19 images. Still, they will be significantly different from the Control images so in the end, most of the time the model will be making differentiation on the dataset used instead of the content of the image. Therefore, in this case, the bias is associated with some past discrimination. There may be people from different places in the COVID-19 dataset in comparison to the Control dataset, and there could be different age or sex distributions.

Methodology for bias identification in the COVID-19 datasets

In this work, we will analyze biases in the context of COVID-19 classification or diagnosis; other tasks, such as triage, are not considered. Also, we are not assuming that the datasets are deliberately biased, but in terms of the information provided, it is probable to make mistakes that could lead to bias if we were to use it. Finally, if we do not find any data or information about the attributes or a particular attribute, we say it has a bias. For example, suppose the age distribution is not known. In that case, the control dataset could match the COVID-19 dataset. Also, with the patient distribution, even for another task such as triage, if we do not know which images correspond to a specific patient, it is possible to have images of a patient in train and test folds, introducing a bias in the model evaluation.

Figure 7 shows the suggested bias of each COVID-19 dataset, in which we group the bias described before into seven types of bias. Figure 4 of supplementary material shows three of the four groups described in the previous section.

To generate Fig. 7, we took these conditions and clarified each bias label. First, in terms of sex, if one label has more than 70.

Radiopedia is also a particular case among some labels because a study or patient has 31 images apparently from different people. Still, they are all grouped in one Age, Sex, and other information about these images is not provided. Thus, we put this dataset on Age, Sex, and patient distribution bias. Finally, mixed datasets such as RYLDS and Qatar University have unclear demographic information, so they are enclosed in this type of

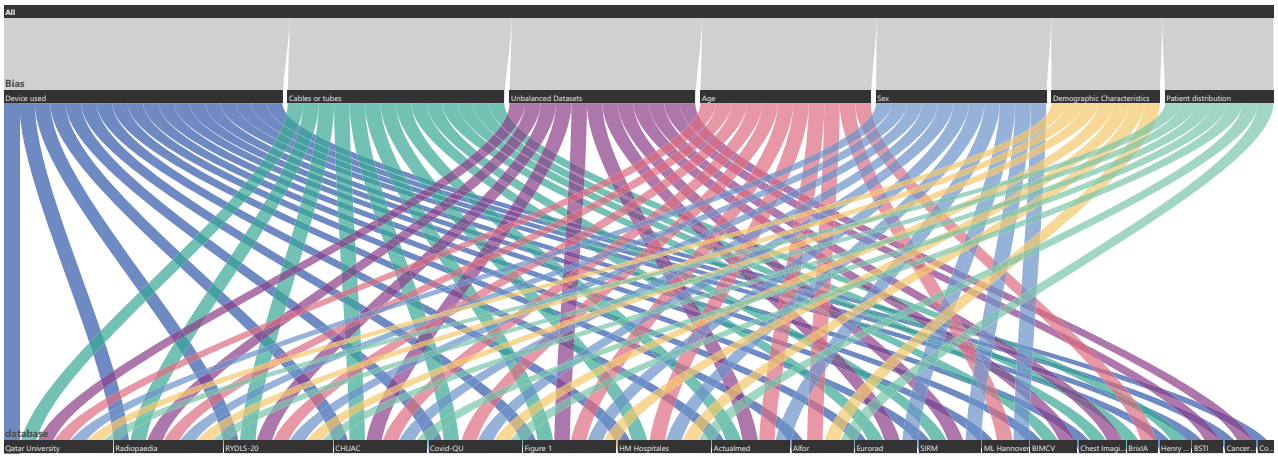


Figure 7. Bias for each COVID-19 dataset.

bias. Also, Alfor enters this category because it classifies entities as letters without giving details on the hospital's locations. Figure 1 is a dataset of images provided by people, so it is difficult to find the place of the images. For the Cables or tubes label, we enclosed the datasets that do not inform if the images belong to a UCI patient. However, if we do not consider these labels in the classification task in the other datasets, it is also a bias in the dataset. Still, we can eliminate those images or perform some cleaning to use those images. Meanwhile, if we do not see each image, it is impossible to identify if a picture contains cables or tubes.

Last but not least, for unbalanced data, if the COVID-19 data is too small or too big, or if the dataset has less than 200 images, it means that the information may not be enough for generalizations, so we classify it as an unbalanced dataset bias.

The following example shows bias associated with mixing datasets. We used the two most popular datasets, Cohen in COVID-19 datasets and Chest X-ray Images (pneumonia), which are also the paramount combination, because eight of the nine papers that use Chest X-ray Images (pneumonia) also use Cohen[8,9,28,34,37,38,45,58]. First, it is essential to see how in Fig. 7 the Cohen dataset only enters in one bias that is the Device used, and also Fig. 3 shows that the Study date has some problems in terms of uniformity; still, it can be handled with some extra work. We notice new biases appear when combining Cohen and Chest X-ray Images (pneumonia) datasets. Deep down on each label sex on Chest X-ray Images (pneumonia) is not found and has more images than Cohen, so the Other label has a higher number of images, which also implies that the dataset is unbalanced because there are 654 images in Cohen. In contrast, 5860 is almost a 9:1 proportion. Also, in terms of sex, we see that the mean Age in Cohen is 54 ± 17 years old, and in Chest X-ray Images (pneumonia), the study uses children, so the mean could be around 11 years old, meaning the age proportion does not match. Also, Cohen contains mainly European images, while Chest X-ray Images (pneumonia) is a Chinese dataset. Finally, even if we subset the Chest X-ray Images (pneumonia) to avoid unbalanced data, age and demographic characteristics bias can not be avoided, and sex can be similar only by luck, so mixing datasets enhances the chances of having bias if the data is not uniform. For a graphical way to show the new bias, the reader can see Fig. 5 of supplementary material.

Experiment with an ethical tool

As mentioned before, the most used database is the Cohen database. It contains not only COVID-19 cases, so we used this dataset to develop the experiment and its validation with an ethical tool. The ethical tools need metadata associated, so we choose Cohen, which has most of the relevant the less missing data you could validate this in Table 1 of supplementary material. We did not mix it with Chest X-ray (pneumonia) dataset because this last one does not provide any metadata, resulting in the ethical algorithm will fail not because of a bias but instead of a lack of information.

Experiment. We used only the Cohen dataset and made a few considerations to avoid some bias and focus on the more visible ones according to the metadata's ethical tool. That includes several labels such as age, sex, location, study date, and if the patient went to ICU, which is vital because of the change in the device used. That leaves the Cohen dataset with two possible biases, the association with mixing the same patient in train and test sets and the unbalanced amount of non-COVID-19 images. To solve this, we executed the following:

- For the division of the images, instead of separating all the images, we split all the patients in the metadata using sklearn.model_selection.train_test_split over the patient amount leaving the training set with 348 patients and 704 images and the test set with 87 patients and 162 images.
- We used sklearn.utils.class_weight to ponder the model weights in the training process for the unbalanced problem.
- Both considerations ensure the model and the dataset can be as much as possible to avoid this type of bias.

Then as a model, we used a pre-trained VGG19 with the Imagenet weights, and as a result, we got a training accuracy of 100

Ethical tool. For the ethical tool, we used Aequitas[19]. Aequitas is a toolkit to analyze a dataset of AI projects and is available as a web page or desktop program. It also has a Python library[87] that was also used for this study with a CSV document with a particular structure in which there is a **score** column that corresponds to the binary classification prediction, **label_value** that is the actual class of the classification. A series of attribute columns correspond to categorical strings representing a specific attribute. There are a few recommendations for this format.

- All attribute values have to be a string.
- If the attribute corresponds to a continuous space such as age, we recommend grouping it in intervals.
- The theoretical system works with a high amount of classes on each attribute, but in these case, we found that the optimal amount for a full report and that the graph support all classes is five classes maximum.
- NaN values sometimes are a problem; we suggest group NaN, Blank spaces, and similar in a unique class such as "Not Found" or "Other."

This tool works with a reference class on each attribute. In our case, we used the class of the higher amount of images, but on the web application, there is an option to select the class automatically by the higher number or, the less bias. The metrics that the tool evaluates are mainly six. These correspond to six types of parity, equality parity, proportionality parity, false-positive rate parity (FPR), false-negative rate parity (FNR), false discovery

parity (FDR), and false omission parity (FOR). These metrics can be found in more detail in the Github repository ((https://github.com/BioAITeam/Bias-Covid).

Results of aequitas. We used the Python library to find the FPR and FDR on each one of the attributes in Fig. 8; we see results for the test. The red dots are the classes that fail the test, the blue ones are the classes that pass, and the gray one is the reference group; the disparity fairness threshold was set 1.25 times. This graph is dynamic, but in this case, Table 3 contains the information for each class. A deep-down on the Location attribute is in Fig. 9 which it shows how far each attribute is to a parity state.

In Table 3, all the comparisons are between the respective FPR or FDR of each class with the reference group, which is a division, so if FPR or FDR on a class different from the reference is zero, we will have a zero division so because this relationship trend goes to infinity for the sake of visualization we replace values of zero with 0.0002 because if we use this value tendency, it is 5000 times smaller and 5000 is the maximum value plots, as shown in Fig. 9, could support.

Using the web report generator, we changed some aspects of our attributes. First, we skipped the attribute Date because it contains many metrics that tend to NaN, and we deleted the class Africa from the Location attribute and unite it with the Not Found class. Even though the characteristics are fewer and the parameters on each class. You can find the results of metric failures in the GitHub repository ((https://github.com/BioAITeam/Bias-Covid).

Discussion

Understanding why these datasets could present each bias is vital to deep-down physiological changes and correlations on patients with particular image characteristics that differ among each dataset. First, if we focus on changes in Chest X-ray images among the age, we find that in the elderly, a reduction in the thickness of the parietal muscles is typical. This generates an increase in pulmonary transparency[88]. This characteristic specifically changes some color range on images. Other characteristics that do not change image transparency but change forms and features of the patients are "barrel chest," produced by a pronounced dorsal kynopsis and more convexity on the sternum and is common phenotypic in chest elderly, this is typically, but not exclusively, it is also expected on pulmonary emphysema and bronchiectasis, there is also an increase on left ventricular muscles on

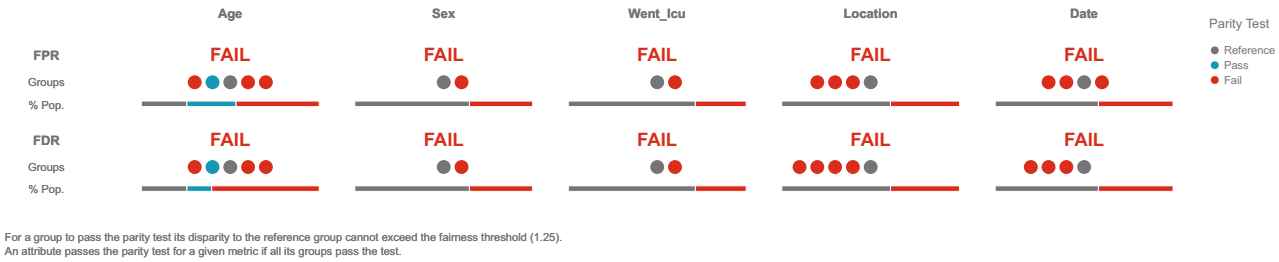


Figure 8. Aequitas results on FPR and FDR for each attribute.

Attribute	Class	FPR/FDR	Disparity	Parity test
Age	Not found	0.016/0.019	1.60 times smaller in FPR, 1.50 times smaller in FDR	Fail, Fail
Age	56–76	0.023/0.060	1.08 times smaller in FPR, 2.04 times larger in FDP	Pass, Fail
Age	36–56	0.025/0.029	Reference Group	NA
Age	17–36	0.041/0.027	1.65 times larger in FPR, 1.07 times smaller in FDP	Fail, Pass
Age	76–95	0.052/0.133	2.09 times larger in FPR, 4.49 times larger in FDR	Fail, Fail
Sex	M	0.022/0.033	Reference Group	NA
Sex	F	0.039/0.046	1.71 times larger in FPR, 1.41 times larger in FDR	Fail, Fail
Went ICU	Y	0.011/0.019	Reference Group	NA
Went ICU	N	0.051/0.105	4.49 times larger in FPR, 5.53 times larger in FDR	Fail, Fail
Location	Africa	NaN/0.000	5000 times smaller in FDR	Fail
Location	America	0.000/0.000	5000 times smaller in FPR, 5000 times smaller in FDR	Fail, Fail
Location	Asia	0.008/0.033	4.62 times smaller in FPR, 1.70 times smaller in FDR	Fail, Fail
Location	Europe	0.036/0.056	Reference Group	NA
Location	Oceania	0.000/0.000	5000 times smaller in FPR, 5000 times smaller in FDR	Fail, Fail
Date	2000-2019	0.000/0.000	5000 times smaller in FPR, 5000 times smaller in FDR	Fail, Fail
Date	2019-2020	0.000/0.000	5000 times smaller in FPR, 5000 times smaller in FDR	Fail, Fail
Date	2020	0.024/0.095	Reference Group	NA
Date	Not found	0.032/0.020	1.32 times larger in FPR, 4.70 times smaller in FDR	Fail, Fail

Table 3. FPR and FDR on each attribute detailed to complement Fig. 8.

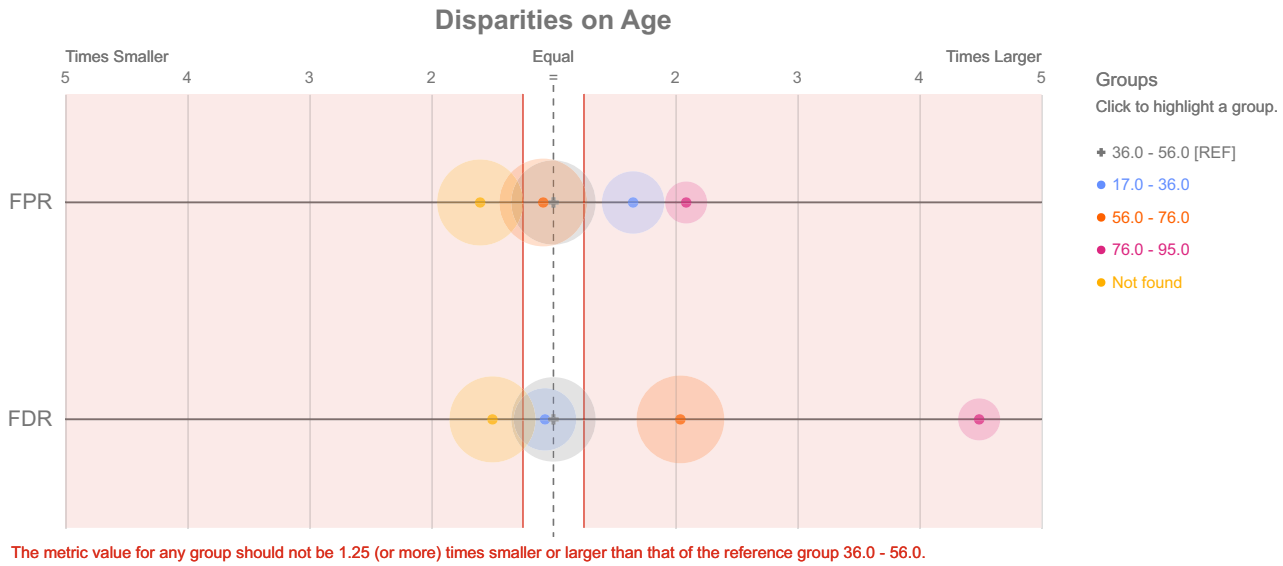


Figure 9. Aequitas results on FPR and FDR the attribute Age.

elderly[88]. But there are not only changes in the elderly if we take into account the development of the respiratory system, we find that the maximum number of alveoli are presented around 10–12 years old, and the maturation of the respiratory system usually ends at 20 years old on females and on 25 years old on males meaning we find less complex pulmonary structures on early stages of life[88], other common changes are also a decrease on chest wall compliance with the age and dilatation of alveolar duct, so air apace is enlargement with an irregular distribution of air[88] and as we know the air spaces are fundamental on radio lucid images.

If we focus on pediatric signs and COVID-19 is shown that most symptomatic children with COVID-19 show abnormalities in chest X-rays, but these findings are typically non-specific, so the use of chest x-ray images could not lead to a first diagnostic test for the identification of COVID-19[89]. It is difficult to quality assurance on children's Chest X-rays. The main factors that affect the quality, especially on young children and babies, are the patient rotation that is inevitable on some babies, the images taken under inspiration because it is challenging to coordinate with the patient respiration timing because you can guide a baby to inspire or expire when you indicate, and finally because of the movement is expected to get a bat scapula position[90].

Other authors as Albrandt-Salmeron et al.[91] agree that there is a correlation between age and some symptoms and image findings, but they find out that upon the Mexican-mestizo community, there are no significant different in terms of patient sex[91], meanwhile Borghesi and Maroldi, 2020 refutes this idea in a study over Italy that finds significantly higher pulmonary insolvents in males than in females[92], this information can be interpreted in different ways first Mexican study use patients of only one hospital, but they have 1000 chest x-ray images, and Italian study use information of 100 hospitals without specifying the number of images, and both are using the CXR scoring system for COVID-19 pneumonia, proposed by Borghesi and Maroldi of the Italian study[92] that is also used On BrixIA dataset for the severity label, the information shows two main possibilities as an overview, first using 100 hospitals shows more generalization on the population than only using one, meaning there is a tendency on pulmonary involvement's grater in males rather than females, but also we can argue that both studies are equally valid, but the differences between the results depends on the population that the study focus on meaning that on Italian or even European population the COVID-19 findings are more usual in males while in Mexico there is not a marked difference.

Some pathologies, especially ground glass densities, are typical in COVID-19 cases but extremely difficult to detect on portable CXR images but easy to catch on CT[93]. Complementing the visualizations in Fig. 6 we saw that the image format affects, and some datasets contained only images in PNG or JPEG format as Cohen that is contaminated with additional elements such as arrows, numbers, or letters different from the ones provided by the device used, that by the way it is also different there are devices that show the letter R for mark the patient right. Still, it can also be a letter D or A, and there are devices that show a P for portable. There are also cases of DICOM format images that depend on the preprocess treatment. Finally, we found formats that try to leave more information available, like the 16 bits PNG images from the BIMCV dataset shown in Fig. 6 and if we see there are visibly different from usual PNG images.

One important thing to highlight is that this study is focused on the bias on COVID-19 classification or diagnosis, so even if most of the databases have a bias on this task, this does not mean they are useless, foremost the main general problem of the datasets presented is that the datasets are created focused on gathering COVID-19 confirmed patients images. Still, a COVID-19 dataset alone can have some powerful uses as triage or prognosis; if there is a certainty that a patient has COVID-19, it is possible that the patient could have a low chance of getting worst, or it is possible that the pneumonia is severe. Identifying these characteristics in less time, a radiologist could identify helps with the patient's treatment and if the task is to find the probability of a patient going to the ICU or even passing away helps to make strategies to avoid this result. The task mentioned before has less probability to acquire bias because it is not necessary to mix the dataset with others and have mismatches in

terms of age distribution and sex, among others, instead even if the dataset is from only one hospital, these could lead to the generation of a helpful software inside this institution. Also, COVID-QU has the highest number of images in the COVID-19 datasets, which could have much bias for classification but is one of the most extensive Lung segmentation datasets available online. The fact that images come from many sources makes that, in these cases, generalization could be better for many formats in chest X-rays for proper lung segmentation and, this way, enhance algorithms in other tasks. BrixIA, for example, has its own severity classification, which groups specific pathologies in certain image zones. The Cohen dataset author recently published a paper that uses this dataset for severity classification[59].

As a review of the datasets that are best for COVID-19 diagnosis, it is essential to point out that the main datasets used are the ones entirely free that do not even need a request for their usage. Usually, those datasets are the ones with more problems and missing information that could lead to some bias; then let's avoid the use of control datasets because it is difficult to get a database that matches and generate the least possible amount of conflict in terms of the image characteristics that could lead into a possible bias, the more presented problem is the device used. It is not something we could solve by filtering because there is no information available, and the cable or tubes can be mitigated by removing the ICU patients of both data frames, but this might enhance the unbalance on the dataset, so it is necessary will need clinical validation before the final deployment on the Health institute or organization.

The experiment performed using the Cohen dataset shows different things. First, we see a high accuracy on the dataset in general, meaning the system could differ images from both labels, been zero COVID-19 and one NON-COVID-19, but the Aequitas analysis did not show good results. Let's suppose that the unbalanced dataset and the patient bias were well avoided; it would be possible to say that Aequitas metrics that depend on balanced datasets could be avoided in attributes such as date because before 2019, there was no COVID-19 so is an inevitable condition if I work with datasets developed before this date. However, still Figs. 8 and 9, and Table 3 shows that in general terms, the dataset fails in all the metrics, even the ones that do not depend on balanced data. There is a particular case that is Age intervals 17–36 and 56–76; on both, there is one metric (FDR and FPR, respectively) that passes the ethical test. Yet, it is one of the two evaluated, meaning it still has issues even though it was the only one to pass. Also, the 5000 times is in real infinitesimal smaller than the reference group meaning these classes are not generating information to generalize, or there are too small that the bias can not be compared, same with Africa on the location that has a NaN value, so is not possible to get a reasonable interpretation of the result. But if we avoid these two cases, those attributes still fail this test. As mentioned before the case that, in general, the metric failed, the detail on Fig. 8 shows that not all classes in all metrics fail. Still, each attribute has more failure metrics, so it generally fails. For more details, see the GitHub repository (https://github.com/BioAITeam/Bias-Covid).

Conclusion and recommendations

After reviewing the different databases of chest X-rays that are being used to study COVID-19, it is observed that due to the newness of this virus and the eagerness to obtain results in the rapid and early detection of the disease has motivated the release of numerous databases that may have biases such as those associated with: patient information, capture conditions, imbalance, database mixtures, among others. The above can generate high percentages of accuracy in classifying this pathology. Still, when performing an exhaustive analysis of the information, it is found that the AI algorithms may only be calibrated to identify the characteristic features of the disease if the different biases are managed. In addition, the lack of information in the metadata often does not allow a correct choice of the dataset or the identification of different types of biases, especially when mixing highly heterogeneous databases without such information is desired. Therefore, it is recommended to make a deep analysis of the data and its metadata, such as performing a statistical analysis of all the information to be clear about the quality of the database and additionally to perform a visual examination of the information in the case that the type of data used is images, all this to observe possible early biases and try to mitigate them. It is recommended to perform analysis using ethical tools such as Aequitas to ensure that the database does not have biases of age, gender, or race, among others, and thus obtain results with ethical and responsible standards. For the construction and release of new databases in COVID-19 or any other type of problem, it is recommended to consider the experiments and analyses performed in this work to deliver information as homogeneous as possible, where the only difference is, for example, the detection of pathology. Mixing existing databases to increase the volume of information may not be recommended in COVID-19 because it may introduce biases such as those mentioned in this work. Many exposed databases may work well for other problems unrelated to COVID-19. For future work, we propose to perform lung segmentation on existing databases to focus on the disease area of interest to help AI algorithms identify disease-specific features and mitigate potential biases. The identification of COVID-19 using chest X-rays is an area that is still under construction, and there is a long way to go before AI systems can reliably classify the disease. However, building and releasing high-quality databases with as few biases as possible is necessary to achieve that goal.

In this paper, we provide a series of arguments that show some terrible aspects of using and creating Chest X-ray datasets for COVID-19 classification purposes. In addition, we use an AI ethical tool to make a more profound validation of some characteristics in terms of the bias using a simple model and the most used datasets, hoping this could be an example of how further we can validate a model's performance.

As recommendations for further studies and database creation, it is vital to create homogeneous data. For a hospital, it is almost impossible to acquire the same amount of images from some age groups or sex. Still, at least the use of the device can be homogeneous for positive and control cases, also do not mix positive with control datasets, also validate the results of a group with an expert radiologist, and make detailed metadata of the images for preventing, patient mix in the different sets, and as an optional parameter, please avoid ICU images, the

portable devices quality and the severe state of the patients is evident. These algorithms should be guided more in making a fast test to have an initial presumptuous diagnosis and be able to take rapid actions and avoid a rapid patient health deterioration. Making an ICU image is not so helpful because, in this step, the options for treatment are limited. Finally, we recommend using AI ethical tools or frameworks to find possible bias in the model.

Future work

As a result of this investigation, we are developing a structured dataset taking into account age, sex, and device distribution of positive and negatives cases of COVID-19 for further experiments in terms of classification and the development of software for an actual COVID-19 classification application that has an ethical validation. In this line, some preliminary results can be found here[15].

Additional information

The authors affirm no one has a competing financial interest or personal issues that could influence the work developed in this paper. Code and information available on (https://github.com/BioAITeam/Bias-Covid).

Data availability

The datasets analyzed during the current study are available on multiple repositories that can be accessed throw the links on Table 1 and 2 or Data Availability on supplementary material that contains a summary of all articles and links for downloading the public dataset and this paper repository.

Received: 24 June 2022; Accepted: 17 February 2023
Published online: 01 March 2023

References

1. Wang, D. *et al.* Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus-infected pneumonia in Wuhan, China. *JAMA* **323**, 1061–1069. https://doi.org/10.1001/jama.2020.1585 (2020).
2. Ducharme, J. World health organization declares COVID-19 a 'pandemic.' here's what that means. https://time.com/5791661/who-coronavirus-pandemic-declaration/ (2020).
3. Tahamtan, A. & Ardebili, A. Real-time rt-pcr in COVID-19 detection: Issues affecting the results. *Exp. Rev. Mol. Diagn.* **20**, 453–454. https://doi.org/10.1080/14737159.2020.1757437 (2020).
4. Long, C. *et al.* Diagnosis of the coronavirus disease (COVID-19): rrt-pcr or ct?. *Eur. J. Radiol.* **126**, 108961. https://doi.org/10.1016/j.ejrad.2020.108961 (2020).
5. Albahri, O. S. *et al.* Systematic review of artificial intelligence techniques in the detection and classification of COVID-19 medical images in terms of evaluation and benchmarking: Taxonomy analysis, challenges, future solutions and methodological aspects. *J. Infect. Public Health* **13**, 1381–1396. https://doi.org/10.1016/J.JIPH.2020.06.028 (2020).
6. Ai, T. *et al.* Correlation of chest ct and rt-pcr testing for coronavirus disease 2019 (COVID-19) in china: A report of 1014 cases. *Radiology* **296**, E32–E40. https://doi.org/10.1148/radiol.2020200642 (2020).
7. Balaha, H. M., El-Gendy, E. M. & Saafan, M. M. A complete framework for accurate recognition and prognosis of COVID-19 patients based on deep transfer learning and feature classification approach. *Artif. Intell. Rev.* **55**, 5063–5108. https://doi.org/10.1007/s10462-021-10127-8 (2022).
8. Tartaglione, E., Barbano, C. A., Berzovini, C., Calandri, M. & Grangetto, M. Unveiling COVID-19 from chest X-ray with deep learning: A hurdles race with small data. *Int. J. Environ. Res. Public Health*https://doi.org/10.3390/ijerph17186933 *(2020).*
9. Ghoshal, B. & Tucker, A. Estimating uncertainty and interpretability in deep learning for coronavirus (Covid-19) detection. https://doi.org/10.48550/arxiv.2003.10769 (2020).
10. Malhotra, A. *et al.* Multi-task driven explainable diagnosis of COVID-19 using chest XX-ray images. *Pat. Recogn.*https://doi.org/10.48550/arxiv.2008.03205 *(2020).*
11. Cruz, B. G. S., 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. https://doi.org/10.1016/j.media.2021.102225 (2021).
12. Roberts, M. *et al.* Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and ct scans. *Nat. Mach. Intell.* **3**, 199–217. https://doi.org/10.1038/s42256-021-00307-0 (2021).
13. Gao, J. *et al.* Medml: Fusing medical knowledge and machine learning models for early pediatric COVID-19 hospitalization and severity prediction. *iScience* **25**, 104970. https://doi.org/10.1016/j.isci.2022.104970 (2022).
14. Arias-Garzón, D. *et al.* COVID-19 detection in X-ray images using convolutional neural networks. *Mach. Learn. Appl.* **6**, 100138. https://doi.org/10.1016/j.mlwa.2021.100138 (2021).
15. Alzate-Grisales, J. A. *et al.* Cov-caldas: A new COVID-19 chest X-ray dataset from state of caldas-colombia. *Sci. Data* **9**, 757. https://doi.org/10.1038/s41597-022-01576-z (2022).
16. Hagendorff, T. The ethics of ai ethics: An evaluation of guidelines. *Minds Mach.* **30**, 99–120. https://doi.org/10.1007/s11023-020-09517-8 (2020).
17. Floridi, L. *et al.* Ai4people-an ethical framework for a good ai society: Opportunities, risks, principles, and recommendations. *Minds Mach.* **28**, 689–707. https://doi.org/10.1007/s11023-018-9482-5 (2018).
18. Tabares-Soto, R. *et al.* Analysis of ethical development for public policies in the acquisition of ai-based systems. https://doi.org/10.4018/978-1-6684-5892-1.ch010 (2022).
19. Saleiro, P. *et al.* Aequitas: A bias and fairness audit toolkit. https://doi.org/10.48550/arXiv.1811.05577 (2018).
20. Cohen, J. P., Morrison, P. & Dao, L. COVID-19 image data collection. arXiv https://doi.org/10.48550/arXiv.2003.11597 (2020).
21. Zhang, R. *et al.* Diagnosis of coronavirus disease 2019 pneumonia by using chest radiography: Value of artificial intelligence. *Radiology* **298**, E88–E97. https://doi.org/10.1148/radiol.2020202944 (2021).
22. Afifi, A., Hafsa, N. E., Ali, M. A. S., Alhumam, A. & Alsalman, S. An ensemble of global and local-attention based convolutional neural networks for COVID-19 diagnosis on chest X-ray images. *Symmetry* **13**, 113 (2021).
23. Imagawa, K. & Shiomoto, K. Performance change with the number of training data: A case study on the binary classification of COVID-19 chest X-ray by using convolutional neural networks. *Comput. Biol. Med.* **142**, 105251. https://doi.org/10.1016/J.COMPBIOMED.2022.105251 (2022).
24. Bassi, P. R. A. S. & Attux, R. A deep convolutional neural network for Covid-19 detection using chest X-rays. https://doi.org/10.1007/s42600-021-00132-9 (2020).
25. Jain, G., Mittal, D., Thakur, D. & Mittal, M. K. A deep learning approach to detect COVID-19 coronavirus with X-ray images. *Biocybern. Biomed. Eng.* **40**, 1391–1405. https://doi.org/10.1016/j.bbe.2020.08.008 (2020).

26. Kana, E. B. G., Kana, M. G. Z., Kana, A. F. D. & Kenfack, R. H. A. A web-based diagnostic tool for COVID-19 using machine learning on chest radiographs (cxr). medRxiv https://doi.org/10.1101/2020.04.21.20063263 (2020).
27. Zokaeinikoo, M., Kazemian, P., Mitra, P. & Kumara, S. Aidcov: An interpretable artificial intelligence model for detection of COVID-19 from chest radiography images. medRxiv https://doi.org/10.1101/2020.05.24.20111922 (2020).
28. Tamal, M. *et al.* An integrated framework with machine learning and radiomics for accurate and rapid early diagnosis of COVID-19 from chest X-ray. *Exp. Syst. Appl.* **180**, 115152. https://doi.org/10.1016/J.ESWA.2021.115152 (2021).
29. Ezzat, D., Hassanien, A. E. & Ella, H. A. An optimized deep learning architecture for the diagnosis of COVID-19 disease based on gravitational search optimization. *Appl. Soft Comput.* **98**, 106742. https://doi.org/10.1016/J.ASOC.2020.106742 (2021).
30. Wang, Z. *et al.* Automatically discriminating and localizing COVID-19 from community-acquired pneumonia on chest X-rays. *Pat. Recogn.* **110**, 107613. https://doi.org/10.1016/J.PATCOG.2020.107613 (2021).
31. Khan, A. I., Shah, J. L. & Bhat, M. M. Coronet: A deep neural network for detection and diagnosis of COVID-19 from chest X-ray images. *Comput. Methods Program. Biomed.* **196**, 105581. https://doi.org/10.1016/j.cmpb.2020.105581 (2020).
32. Apostolopoulos, I. D. & Mpesiana, T. A. COVID-19: Automatic detection from X-ray images utilizing transfer learning with convolutional neural networks. *Phys. Eng. Sci. Med.* **43**, 635–640. https://doi.org/10.1007/s13246-020-00865-4 (2020).
33. Mangal, A. et al. Covidaid: COVID-19 detection using chest X-ray. https://doi.org/10.48550/arxiv.2004.09803 (2020).
34. Sayyed, A. Q. M. S., Saha, D. & Hossain, A. R. Covmunet: A multiple loss approach towards detection of COVID-19 from chest x-ray. https://doi.org/10.48550/arxiv.2007.14318 (2020).
35. 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.*https://doi.org/10.1016/j.media.2020.101794 *(2020).*
36. Rahaman, M. M. *et al.* Identification of COVID-19 samples from chest X-ray images using deep learning: A comparison of transfer learning approaches. *J. X-Ray Sci. Technol.* **28**, 821–839. https://doi.org/10.3233/XST-200715 (2020).
37. Tsiknakis, N. *et al.* Interpretable artificial intelligence framework for COVID-19 screening on chest X-rays. *Exp. Therap. Med.* **20**, 727–735. https://doi.org/10.3892/etm.2020.8797 (2020).
38. Elaziz, M. A. *et al.* New machine learning method for image-based diagnosis of COVID-19. *PLOS ONE* **15**, e0235187. https://doi.org/10.1371/journal.pone.0235187 (2020).
39. Yamac, M. *et al.* Convolutional sparse support estimator based COVID-19 recognition from X-ray images. *IEEE Tran. Neural Netw. Learn. Syst.*https://doi.org/10.48550/arxiv.2005.04014 *(2020).*
40. Fan, Y., Liu, J., Yao, R. & Yuan, X. COVID-19 detection from X-ray images using multi-kernel-size spatial-channel attention network. *Pat. Recogn.* **119**, 108055. https://doi.org/10.1016/J.PATCOG.2021.108055 (2021).
41. Farooq, M. & Hafeez, A. Covid-resnet: A deep learning framework for screening of COVID19 from radiographs. https://doi.org/10.48550/arxiv.2003.14395 (2020).
42. 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. https://doi.org/10.1038/s41598-020-76550-z (2020).
43. Ahmed, K. B., Goldgof, G. M., Paul, R., Goldgof, D. B. & Hall, L. O. Discovery of a generalization gap of convolutional neural networks on COVID-19 X-rays classification. *IEEE Access* **9**, 72970–72979. https://doi.org/10.1109/access.2021.3079716 (2021).
44. Gil, D., Díaz-Chito, K., Sánchez, C. & Hernández-Sabaté, A. Early screening of sars-cov-2 by intelligent analysis of X-ray images. https://doi.org/10.48550/arxiv.2005.13928 (2020).
45. Heidari, M. *et al.* Improving the performance of cnn to predict the likelihood of COVID-19 using chest X-ray images with preprocessing algorithms. *Int. J. Med. Inform.* **144**, 104284. https://doi.org/10.1016/J.IJMEDINF.2020.104284 (2020).
46. Qi, X., Foran, D. J., Nosher, J. L. & Hacihaliloglu, I. Multi-feature semi-supervised learning for COVID-19 diagnosis from chest X-ray images (2021).
47. Degerli, A., Kiranyaz, S., Chowdhury, M. E. H. & Gabbouj, M. Osegnet: Operational segmentation network for COVID-19 detection using chest X-ray images. Arxiv abs/2202.10185 (2022).
48. Guarrasi, V., D'Amico, N. C., Sicilia, R., Cordelli, E. & Soda, P. Pareto optimization of deep networks for COVID-19 diagnosis from chest X-rays. *Pat. Recogn.* **121**, 108242. https://doi.org/10.1016/J.PATCOG.2021.108242 (2022).
49. Luz, E. *et al.* Towards an effective and efficient deep learning model for COVID-19 patterns detection in X-ray images. *Res. Biomed. Eng.*https://doi.org/10.1007/s42600-021-00151-6 *(2020).*
50. Pereira, R. M., Bertolini, D., Teixeira, L. O., Silla, C. N. & Costa, Y. M. COVID-19 identification in chest X-ray images on flat and hierarchical classification scenarios. *Comput. Methods Program. Biomed.* **194**, 105532. https://doi.org/10.1016/J.CMPB.2020.105532 (2020).
51. Moura, J. D. *et al.* Deep convolutional approaches for the analysis of COVID-19 using chest X-ray images from portable devices. *IEEE Access* **8**, 195594–195607. https://doi.org/10.1109/ACCESS.2020.3033762 (2020).
52. Kassania, S. H., Kassanib, P. H., Wesolowskic, M. J., Schneidera, K. A. & Detersa, R. Automatic detection of coronavirus disease (COVID-19) in X-ray and ct images: A machine learning based approach. *Biocybern. Biomed. Eng.* **41**, 867–879. https://doi.org/10.1016/j.bbe.2021.05.013 (2021).
53. Teixeira, L. O. *et al.* Impact of lung segmentation on the diagnosis and explanation of COVID-19 in chest X-ray images. *Sensors* **21**, 7116. https://doi.org/10.3390/s21217116 (2021).
54. Maguolo, G. & Nanni, L. A critic evaluation of methods for COVID-19 automatic detection from X-ray images. *Inf. Fus.* **76**, 1–7. https://doi.org/10.1016/j.inffus.2021.04.008 (2021).
55. Singh, K. K. & Singh, A. Diagnosis of COVID-19 from chest X-ray images using wavelets-based depthwise convolution network. *Big Data Min. Anal.* **4**, 84–93. https://doi.org/10.26599/BDMA.2020.9020012 (2021).
56. Li, X., Li, C. & Zhu, D. Covid-mobilexpert: On-device COVID-19 patient triage and follow-up using chest X-rays. https://doi.org/10.48550/arxiv.2004.03042 (2020).
57. Signoroni, A. *et al.* Bs-net: Learning COVID-19 pneumonia severity on a large chest X-ray dataset. *Med. Image Anal.* **71**, 102046. https://doi.org/10.1016/J.MEDIA.2021.102046 (2021).
58. Bararia, A., Ghosh, A., Bose, C. & Bhar, D. Network for subclinical prognostication of COVID 19 patients from data of thoracic roentgenogram: A feasible alternative screening technology. *medRxiv*https://doi.org/10.1101/2020.09.07.20189852 *(2020).*
59. Cohen, J. P. *et al.* Predicting COVID-19 pneumonia severity on chest X-ray with deep learning. *Cureus*https://doi.org/10.48550/arxiv.2005.11856 *(2020).*
60. Irmak, E. COVID-19 disease severity assessment using cnn model. *IET Image Process.* **15**, 1814–1824. https://doi.org/10.1049/ipr2.12153 (2021).
61. Tahir, A. M. *et al.* COVID-19 infection localization and severity grading from chest X-ray images. *Comput. Biol. Med.* **139**, 105002. https://doi.org/10.1016/J.COMPBIOMED.2021.105002 (2021).
62. Park, S. *et al.* Multi-task vision transformer using low-level chest X-ray feature corpus for COVID-19 diagnosis and severity quantification. *Med. Image Anal.* **75**, 102299. https://doi.org/10.1016/J.MEDIA.2021.102299 (2022).
63. de la Iglesia Vayá, M. *et al.* Bimcv covid-19+: a large annotated dataset of rx and ct images from COVID-19 patients. 1–22 (2020).
64. Desai, S. *et al.* Chest imaging representing a COVID-19 positive rural u.s. population. *Sci. Data* **7**, 414. https://doi.org/10.1038/s41597-020-00741-6 (2020).
65. Winther, H. B. *et al.* Dataset: Covid-19 image repository, https://doi.org/10.6084/m9.figshare.12275009 (2020).
66. Chung, A. Actualmed-covid-chestxray-dataset: Actualmed COVID-19 chest X-ray dataset initiative (2020).

67. Chowdhury, M. E. H. *et al.* Can ai help in screening viral and COVID-19 pneumonia?. *IEEE Access* **8**, 132665–132676. https://doi.org/10.1109/ACCESS.2020.3010287 (2020).
68. Covid-19 database - sirm (2020).
69. Yamac, M. *et al.* Convolutional sparse support estimator-based COVID-19 recognition from X-ray images. *IEEE Trans. Neural Netw. Learn. Syst.* **32**, 1810–1820. https://doi.org/10.1109/TNNLS.2021.3070467 (2021).
70. Radiopedia.org. COVID 19 | search | radiopaedia.org (2020).
71. EuroRad. Euorad search results for COVID-19 (2020).
72. 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 Analy.* **74**, 102216 (2020).
73. Imaging, C. This is a thread of COVID-19 cxr (2020).
74. the British Society of Thoracic Imaging. COVID-19 british society of thoracic imaging database.
75. Chung, A. Figure 1 COVID-19 chest X-ray dataset initiative (2020).
76. Rahman, T. *et al.* Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images. *Comput. Biol. Med.* **132**, 104319. https://doi.org/10.1016/J.COMPBIOMED.2021.104319 (2021).
77. Summers, R. & NIH. Cxr8 | con la tecnología de box (2020).
78. Irvin, J. *et al.* Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proc. AAAI Conf. Artif. Intell.* **33**, 590–597. https://doi.org/10.1609/aaai.v33i01.3301590 (2019).
79. Bassi, P. R. & Attux, R. A deep convolutional neural network for COVID-19 detection using chest X-rays. *Res. Biomed. Eng.* **38**, 139–148. https://doi.org/10.1007/S42600-021-00132-9/FIGURES/4 (2022).
80. Bustos, A., Pertusa, A., Salinas, J. M. & de la Iglesia-Vayá, M. Padchest: A large chest X-ray image dataset with multi-label annotated reports. *Med. Image Anal.* **66**, 101797. https://doi.org/10.1016/j.media.2020.101797 (2020).
81. Kermany, D. S. *et al.* Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* **172**, 1122-1131.e9. https://doi.org/10.1016/j.cell.2018.02.010 (2018).
82. of North America, R. S. Rsna pneumonia detection challenge | kaggle (2019).
83. Shiraishi, J. *et al.* Development of a digital image database for chest radiographs with and without a lung nodule. *Am. J. Roentgenol.* **174**, 71–74. https://doi.org/10.2214/ajr.174.1.1740071 (2000).
84. Jaeger, S. *et al.* Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. *Quant. Imag. Med. Surg.* **4**, 475–477. https://doi.org/10.3978/j.issn.2223-4292.2014.11.20 (2014).
85. Kermany, D., Zhang, K. & Goldbaum, M. Large dataset of labeled optical coherence tomography (oct) and chest X-ray images. *Mendel. Data*https://doi.org/10.17632/RSCBJBR9SJ.3 *(2018).*
86. Cruz, B. G. S., 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. https://doi.org/10.1016/j.media.2021.102225 (2021).
87. The bias and fairness audit toolkit for machine learning - aequitas documentation.
88. Hochhegger, B. *et al.* O tórax e o envelhecimento: manifestações radiológicas. *J. Brasil. Pneumol.* **38**, 656–665. https://doi.org/10.1590/S1806-37132012000500016 (2012).
89. Serrano, C. O. *et al.* Pediatric chest X-ray in COVID-19 infection. *Eur. J. Radiol.* **131**, 109236. https://doi.org/10.1016/j.ejrad.2020.109236 (2020).
90. Hlabangana, L. T. *et al.* Inter-rater reliability in quality assurance (qa) of pediatric chest X-rays. *J. Med. Imag. Radiat. Sci.* **52**, 427–434. https://doi.org/10.1016/j.jmir.2021.04.002 (2021).
91. Albrandt-Salmeron, A., Espejo-Fonseca, R. & Roldan-Valadez, E. Correlation between chest X-ray severity in COVID-19 and age in mexican-mestizo patients: An observational cross-sectional study. *BioMed Res. Int.* **2021**, 5571144. https://doi.org/10.1155/2021/5571144 (2021).
92. Borghesi, A. & Maroldi, R. COVID-19 outbreak in Italy: Experimental chest X-ray scoring system for quantifying and monitoring disease progression. *La Radiol. Med.* **125**, 509–513. https://doi.org/10.1007/s11547-020-01200-3 (2020).
93. Jacobi, A., Chung, M., Bernheim, A. & Eber, C. Portable chest X-ray in coronavirus disease-19 (COVID-19): A pictorial review. *Clin. Imag.* **64**, 35–42. https://doi.org/10.1016/j.clinimag.2020.04.001 (2020).

Acknowledgements

We would like to thank Universidad Autonoma de Manizales for making this paper as part of the "Detección de COVID-19 en imágenes de rayos X usando redes neuronales convolucionales" proyecct with code 699-106, and Minciencias for fund this proyect on the call No. 874 of 2020, named "Convocatoria para el Fortalecimiento de Proyectos en ejecución de CTeI en Ciencias de la Salud con Talento Joven e Impacto Regional" also to the proyects "CH-T1246 : Oportunidades de Mercado para las Empresas de Tecnología - Compras Públicas de Algoritmos Responsables, Èticos y Transparentes" and ANID PIA/BASAL FB0002, that help with the ethical tools applications on this paper.

Author contributions

D.A.-G. developed most of the main code and wrote the manuscript, also prepared all tables and Figures except for Figures 15, 16 and 17. Ethical Code and research was developed by J.B.-S. include Figures 15, 16 and 17, R.T.-S. supervise the project development work, and G.A.R. supervise the Ethical component on the research. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-30174-1.

Correspondence and requests for materials should be addressed to G.A.R.

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.