BMC Medical Research Methodology

RESEARCH Open Access

Does imbalance in chest X-ray datasets produce biased deep learning approaches for COVID-19 screening?

Lorena Álvarez-Rodríguez[1,2], Joaquim de Moura[1,2*], Jorge Novo[1,2] and Marcos Ortega[1,2]

Abstract

Background: The health crisis resulting from the global COVID-19 pandemic highlighted more than ever the need for rapid, reliable and safe methods of diagnosis and monitoring of respiratory diseases. To study pulmonary involvement in detail, one of the most common resources is the use of different lung imaging modalities (like chest radiography) to explore the possible affected areas.

Methods: The study of patient characteristics like sex and age in pathologies of this type is crucial for gaining knowledge of the disease and for avoiding biases due to the clear scarcity of data when developing representative systems. In this work, we performed an analysis of these factors in chest X-ray images to identify biases. Specifically, 11 imbalance scenarios were defined with female and male COVID-19 patients present in different proportions for the sex analysis, and 6 scenarios where only one specific age range was used for training for the age factor. In each study, 3 different approaches for automatic COVID-19 screening were used: Normal vs COVID-19, Pneumonia vs COVID-19 and Non-COVID-19 vs COVID-19. The study was validated using two public chest X-ray datasets, allowing a reliable analysis to support the clinical decision-making process.

Results: The results for the sex-related analysis indicate this factor slightly affects the system in the Normal VS COVID-19 and Pneumonia VS COVID-19 approaches, although the identified differences are not relevant enough to worsen considerably the system. Regarding the age-related analysis, this factor was observed to be influencing the system in a more consistent way than the sex factor, as it was present in all considered scenarios. However, this worsening does not represent a major factor, as it is not of great magnitude.

Conclusions: Multiple studies have been conducted in other fields in order to determine if certain patient characteristics such as sex or age influenced these deep learning systems. However, to the best of our knowledge, this study has not been done for COVID-19 despite the urgency and lack of COVID-19 chest x-ray images. The presented results evidenced that the proposed methodology and tested approaches allow a robust and reliable analysis to support the clinical decision-making process in this pandemic scenario.

Keywords: CAD system, Chest X-ray, COVID-19 screening, Data analysis, Deep learning

*Correspondence: joaquim.demoura@udc.es
[1] Centro de Investigación CITIC, Universidade da Coruña, Campus de Elviña, 15071 A Coruña, Spain
Full list of author information is available at the end of the article

Background

In March 2020, the World Health Organization (WHO) declared the COVID-19 outbreak a pandemic. This highly contagious disease caused by the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) overwhelmed the healthcare system of many countries, forcing them to take drastic measurements to control the incessant flow of infected patients such as lockdown and curfew, among others health measures. This health crisis resulting from the global COVID-19 pandemic caused more than 346 million confirmed cases and more than 5.5 million deaths worldwide cite[1], highlighting more than ever the necessity of rapid, reliable and safe methods of diagnosing and monitoring respiratory diseases. COVID-19 is a demonstration of the impact that these diseases can have on society, with direct repercussions on public health and the global economy. Due to its particularities, these diseases present a very high transmission rate, as they can be easily transmitted by air. In this context, early detection and assessment of the evolution of patients with these diseases is vital, since many of them in their most severe phases, can lead to symptoms including acute respiratory failure, requiring the use of assisted breathing systems or admission to an intensive care unit (ICU).

Efforts in the deep learning domain have been devoted to improving COVID diagnostics in several fronts, like by combining RT-PCR and pseudo-convolutional machines to characterize virus sequences cite[2]. In order to study lung involvement in detail, one of the most common resources is to use different lung imaging modalities (such as chest X-ray) to explore the possible affected areas. This requires a detailed analysis to identify and characterize the different pathological structures on the chest X-ray image, which should be performed by a professional with many years of experience. In this sense, the need to have a set of computational methodologies that allow detailed analysis of a chest X-ray image for diagnostic purposes is critical, especially in the current pandemic scenario. As reference, Fig. 1 shows 3 representative examples of chest X-ray images for 3 different scenarios: normal (patient without pulmonary conditions), patient with pneumonia (others than COVID-19) and patient with COVID-19.

Given the great relevance of this topic, different authors have developed methodologies to support the diagnosis of COVID-19 using X-ray imaging [3, 4]. As reference, Wang et al. [5] developed an open access customized convolutional neural network (CNN) that detects COVID-19 signs in chest X-ray images. Along with this system, they also provided a public dataset named COVIDx that combines images from the main COVID-19 public datasets. In the work of Hammoudi et al. [6], the author proposed a deep learning system that distinguished bacterial pneumonia from viral pneumonia which could be caused by COVID-19. COVIDX-Net is a framework presented by Hemdan et al. [7] whose purpose is organizing seven different chest X-ray classifiers in order to diagnose COVID-19. In the work of Zhang et al. [8], the authors used Confidence Aware Anomaly Detection (CAAD) models to differentiate viral pneumonia from non-viral pneumonia and non-infected patients. Ozturk et al. [9] designed the DarkCovidNet, a deep learning architecture based on DarkNet, and their work was validated by a radiologist who reviewed heatmaps that showed where their system was identifying anomalies related to COVID-19. Gomes et al. cite[10] created IKONOS, a tool to support diagnosis of COVID-19 by texture analysis of X-ray images. Ismael et al. cite[11] used multiresolution approaches, like Wavelet, Shearlet and Contourlet transforms, for feature extraction for chest X-ray image based COVID-19 detection to prove these traditional methods are still effective. Shelke et al. [12] proposed a methodology that classified chest X-ray into normal, pneumonia, tuberculosis and COVID-19 classes, being able to rate severity. Yoo et al. [13] proposed a methodology based on classification trees that categorized X-ray images between normal and anomalies, and COVID-19 and non-COVID-19, respectively. Ismael et al. cite[14] considered deep feature extraction from pretrained deep networks, fine-tuning of a pretrained CNN model, and end-to-end training of a CNN model to classify chest X-rays into NORMAL and COVID-19 classes. In the work of Li et al. [15], the authors made predictions about a COVID-19 infected patient outcome by using a Siamese convolutional neural network [16] to estimate the disease severity. They used chest X-ray images to prognosticate patient's intubation or death, which is a useful resource for hospital resources management. De Moura et al. [17] presented 3 complementary approaches based on Dense Convolutional Network architectures specifically designed for the classification of chest X-ray images into normal, pathological and COVID-19. Waheed et al. [18] addressed the lack of COVID-19 chest X-ray and they tried to solve this by developing CovidGAN, a model based on Auxiliary Classifier Generative Adversarial Network that generates synthetic COVID-19 images. In the work of Morís et al. [19], the authors proposed a strategy to improve the performance of COVID-19 screening [20] by using 3 CycleGAN architectures to generate synthetic images from portable chest X-ray devices.

Nowadays, there is no doubt that deep learning methods are useful resources in the field of medical image analysis. However, these methods require a large amount of data for the developed systems to be used in a real scenario. This problem is known as data scarcity and exists even for more researched and common diseases, such as cancer or pneumonia, whose public datasets are scarce and, some of them, unbalanced, containing only certain types of patients. For instance, the Kaggle Pneumonia dataset [21] that was widely used in the development of different

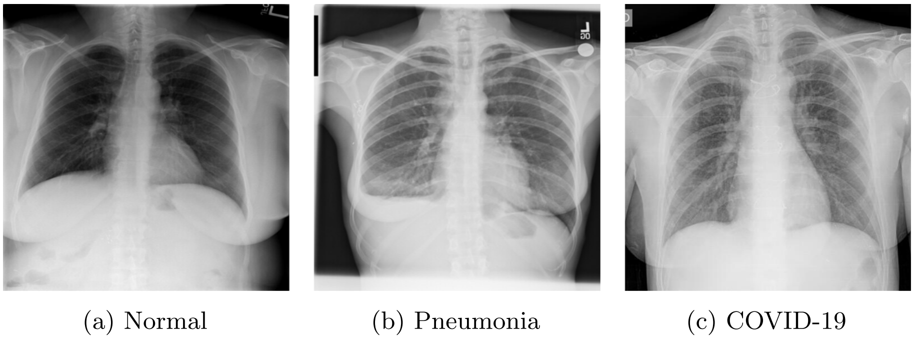


Fig. 1 Representative examples of chest X-ray images of normal (patient without pulmonary conditions), patient with pneumonia (others than COVID-19) and patient with COVID-19

systems for automatic COVID-19 screening only contains pediatric chest X-ray images. This problem was commented by Cirillo et al. [22] in their work, as they describe how biased systems produce discriminatory results in the medical field. They focus on the sex and gender factors, as they consider these aspects to affect diseases, risks, treatments, symptoms, etc. In the work of Larrazabal et al. [23], the authors analysed how imbalance related to gender slightly biases deep learning systems when diagnosing some lung pathologies and abnormalities through chest X-ray images, even though observed worsening was not large. In the work of Vidal et al. [24], the authors proposed a methodology that attempts to alleviate this data scarcity problem in the COVID-19 domain by a two-step knowledge transfer to obtain a robust system able to segment lung regions from portable X-ray devices despite the scarcity of samples and lesser quality. However, to date, to the best of our knowledge, no such study, specifically for sex and age, has been performed for COVID-19 despite all the advances, number of articles and studies, the urgency and lack of COVID-19 chest-x ray images.

Therefore, in this work, we performed a comprehensive analysis of sex and age factors in the COVID-19 datasets. As mentioned above, these characteristics might influence the diagnosis of a disease of this type, where there is a clear problem of data scarcity, which may take us away from the goal of having systems that are as representative as possible and gaining more knowledge about the pathology itself. By thoroughly studying these patient characteristics, we made sure to answer the question of whether these factors produce bias in COVID-19 deep learning-based systems. For this purpose, we analyzed 3 different computational approaches for COVID-19 screening using chest X-ray images: (I) Normal vs COVID-19, (II) Pneumonia vs COVID-19 and (III) Non-COVID-19 vs COVID-19. The proposed study was validated using two state-of-the-art datasets publicly available to the scientific community.

This paper is organized as follows: "Methods" section describes the resources and deep learning approaches employed for the analysis of sex and age factors in the COVID-19 datasets; "Results" section presents the obtained results; and finally, "Discussion" and "Conclusions" sections conclude the manuscript, discussing the results and their impact in relation to the state of the art.

Methods

Datasets

In this section, we describe the 2 public chest X-ray datasets used for this research: (I) HM Hospitals COVID-19 dataset "Covid data saves lives" and (II) RSNA Pneumonia Challenge dataset. Both are described in detail below.

HM hospitals COVID-19 dataset

HM Hospitals made available to the scientific community an anonymous dataset with all clinical information of patients treated in their hospitals by the COVID-19 virus [25]. This dataset is available upon request and must be approved by the HM Hospitals Research Ethics Committee. It consists of 2,310 patients with a diagnosis of "COVID-19 positive" or "COVID-19 pending" admitted to HM Hospitals. Chest X-rays are available for some of the patients, and these were taken during the time they were hospitalized. In this sense, we used 5,493 posteroanterior chest X-ray images from 1,832 different patients whose age and sex are distributed as indicated in Fig. 2 for our COVID-19 class.

RSNA Normal/Pneumonia dataset

The RSNA Pneumonia Challenge dataset [26] is a subset of the ChestX-ray8 dataset [27] created for the Kaggle challenge on the MD.ai platform in collaboration with the Radiological Society of North America (RSNA). This dataset consists of 16,248 X-ray images, considering only the posteroanterior chest view, resulting in 9,452 images for normal cases and 6,796 images for patients diagnosed with pneumonia. In this dataset, we also have information about the age and sex of the patients. These characteristics are distributed in our subset as indicated in Fig. 3 for normal and pneumonia cases.

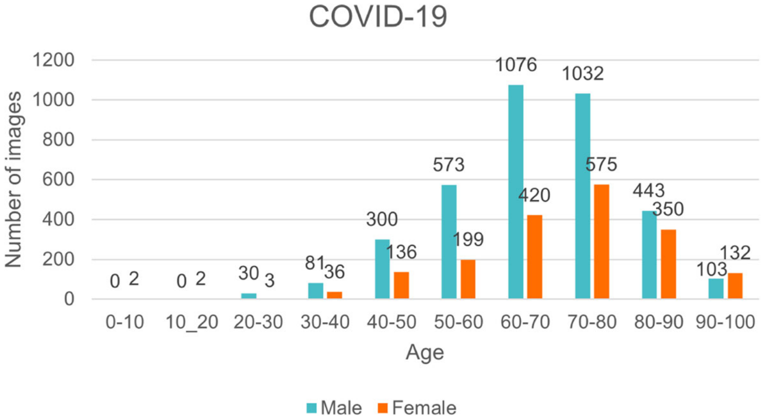
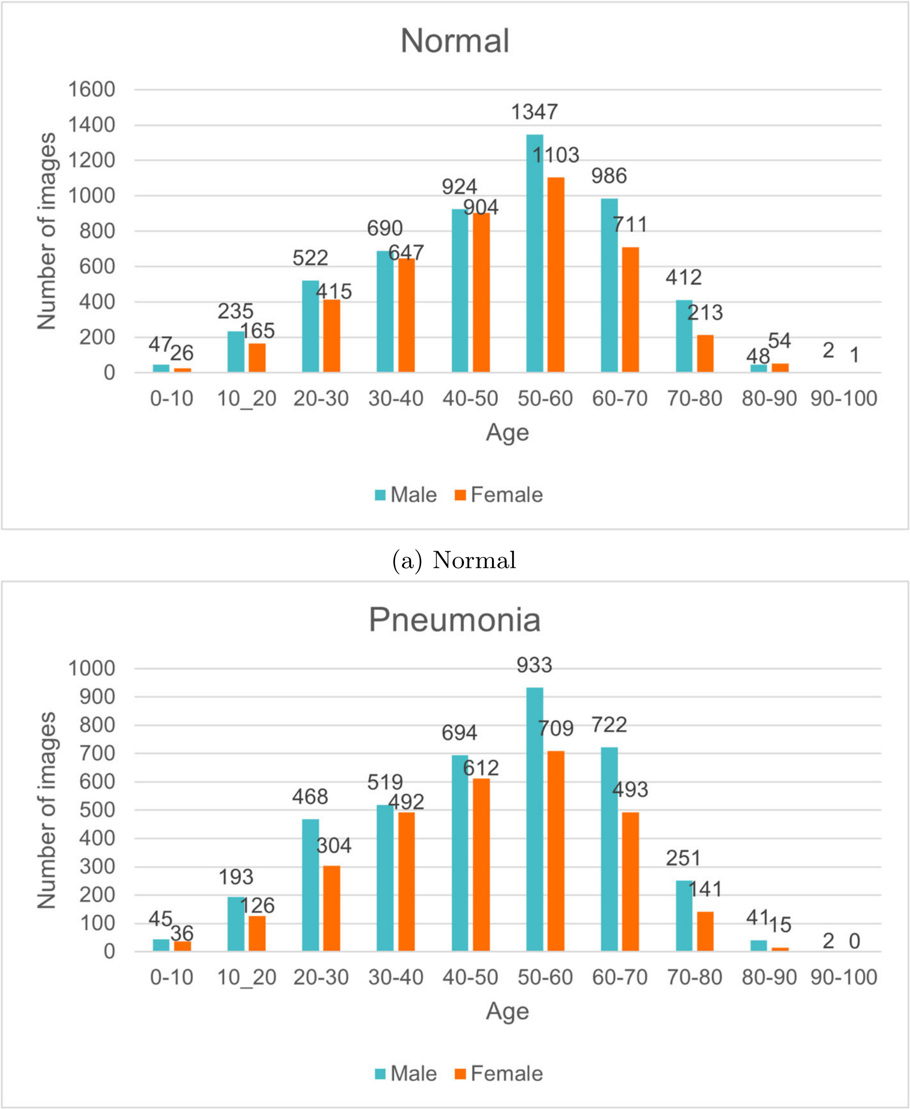


Fig. 2 Age and sex distribution for chest X-ray images of the HM Hospitals COVID-19 dataset



(a) Normal

(b) Pneumonia

Fig. 3 Age and sex distribution for chest X-ray images of the RSNA dataset

Software and hardware resources

In this work, we used Python (version 3.6.6) for the implementation of the conducted studies and machine learning libraries PyTorch (version 0.4.1) and Scikit-learn (version 0.24.2) were used to train, validation and test the obtained models, as well as to get the metrics of their performances.

In addition, in order to facilitate the replication of our studies, we present in Table 1 the main specifications of the hardware used to perform the experiments.

Architecture

In this work, we exploited the potential of the DenseNet-161 architecture [28]. This architecture is composed of dense blocks linked by transition layers, which in turn are formed by convolution and pooling layers. These dense blocks have layers with their own feature maps which consists of a batch normalisation operation, a ReLu operation and a 3 x 3 convolution with k filters, where k is the growth rate. Each of them receives the feature maps of all the previous layers, so that the collective knowledge of all the predecessor layers is preserved. In our case, this growth ratio k is 48, and the depth of the architecture L is 161. However, we modified its original structure to support the binary output defined in our computational approaches, as depicted in Table 2. This architecture provided satisfactory results in similar works aimed at classifying chest X-rays of patients with COVID-19 [17, 19, 20], which led us to choose it for this work.

Computational approaches for screening tasks

As illustrated in Fig. 4, we present 3 different approaches which classify X-ray images into 2 categories to differentiate COVID-19 patients from certain types of patients, as normal and pneumonia ones. Each of these approaches will be explained in more detail below, but in general these 3 different approaches cover a wide range of scenarios in which we can study in depth how gender and age factors affect the diagnosis of COVID-19 in deep learning systems. In this way, we will be able to draw more solid and contrasted conclusions, as most of the cases where a COVID-19 screening task is performed are taken into account and a bias could be more clearly detected.

Table 1 Specifications of the equipment used throughout the project to carry out the experiments

Name	Description
OS	DEBIAN GNU/Linux 10
Kernel	Linux 4.18.0-2-amd64
Architecture	x86-64
CPU	Intel(R) Xeon(R) CPU E5-2650 v4 @ 2.20GHz
Motherboard	Lenovo NeXtScale nx360 M5
RAM	16 GB de RAM GDDR5
HDD	IBM ServeRAID M5210 930 GB
GPU	NVIDIA Tesla P100
Driver Version	396.44
CUDA Version	9.2

Table 2 DenseNet-161 adapted structure

Layers	Output size	DenseNet-161
Convolution	112 x 112	Conv. 7 x 7, stride 2
Pooling	56 x 56	Max pool 3 x 3, stride 2
Dense block (1)	56 x 56	[1 × 1 *conv*. 3 × 3 *conv*.] x 6
Transition layer (1)	56 x 56	Conv. 1 x 1
	28 x 28	2 x 2 average pool, stride 2
Dense block (2)	28 x 28	[1 × 1 *conv*. 3 × 3 *conv*.] x 12
Transition layer (2)	28 x 28	Conv. 1 x 1
	14 x 14	2 x 2 average pool, stride 2
Dense block (3)	14 x 14	[1 × 1 *conv*. 3 × 3 *conv*.] x 36
Transition layer (3)	14 x 14	Conv. 1 x 1
	7 x 7	2 x 2 average pool, stride 2
Dense block (4)	7 x 7	[1 × 1 *conv*. 3 × 3 *conv*.] x 24
Classification layer	1 x 1	7 x 7 global average pool
		2D fully-connected, softmax

1^{st} *approach: Normal vs. COVID-19*

In this first scenario, we trained a model to obtain a consolidated approach to distinguish between normal cases (control patients without lung conditions but who may have other systemic pathologies) and COVID-19. We consider this scenario to be very useful as it is realistic and complex, as it is more difficult than distinguishing only between healthy patients and COVID-19. Moreover, this approach is present in the literature [29]. Both the fact that it is a situation that can occur in a clinical context and that it is a case that can be widely found in the state of the art make the casuistry present in this approach interesting when studying the influence of our target factors.

2^{nd} *approach: pneumonia vs. COVID-19*

Given the similarities between COVID-19 and both viral and bacterial pneumonia, this second approach aims to differentiate between patients with COVID-19 and patients with pneumonia not caused by COVID-19. Thus, 2 different categories are predicted: pneumonia and COVID-19. Similar approaches have been studied in related works [12, 30]. Again, this is a complex situation that could be found in a real screening task and it is broadly studied in the state of art as well, so we find here a number of interesting cases where to explore the impact that sex and age could have.

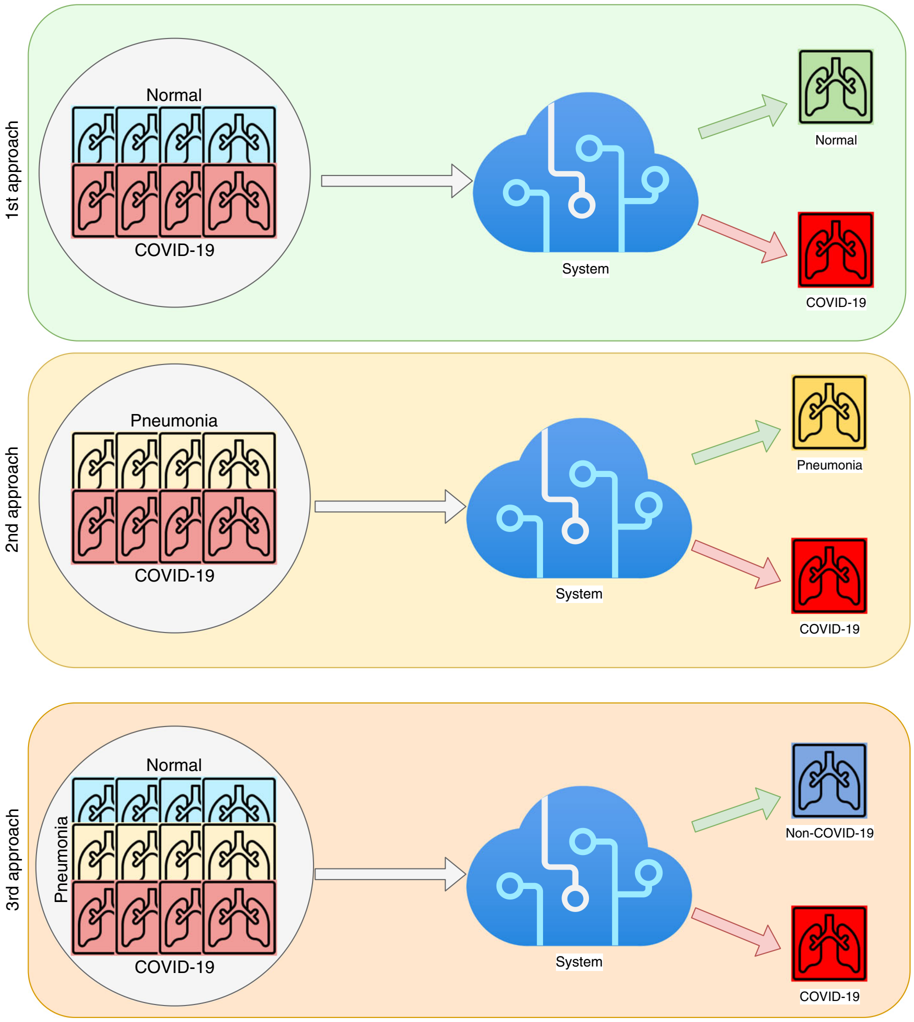


Fig. 4 Schematic representation of computational approaches for COVID-19 screening using X-ray images

3^{rd} *approach: non-COVID-19 vs. COVID-19*

In this third approach, two categories are considered: one that has normal and pneumonia patients, named Non-COVID-19, and another one that has only COVID-19 patients. In this way, we can analyse the degree of separability between COVID-19 patients from all other cases. This kind of approach is common in related works [5, 7, 31]. Thus, this approach allows us, again, to investigate how our target factors could affect a wide number of real and complex cases taken into account here.

Training details

The final dataset for each experiment where we will study the sex and age factors was divided into mutually exclusive subsets, being (60%, 20%, and 20%) for training, validation, and testing, respectively. Regarding the training, we started from the DenseNet-161 model pre-trained with the ImageNet [32] dataset, making use of the transfer learning strategy, but modifying the output layer to adapt it to our specific classification problem. In this way, the training process will be more efficient due to the faster convergence of the training and validation curves. It also reduces the number of labeled images necessary for the process to be adequate [24]. On the other hand, a cross-entropy loss function is performed on the output class and the ground truth for the target X-ray image. The optimization during the training is carried out by Stochastic Gradient Descent (SGD) [33] with a learning rate constant of 0.01, a mini-batch size of 4, and a first-order momentum of 0.9, all of them obtained by exhaustive experimentation. This optimiser has proven to be very efficient, despite its simplicity, for the discriminative learning of classifiers under convex loss functions, defined as follows, where Y represents the ground truth values and $\hat{Y}$ represents the estimated values for each identified category:

$$L = -Y \cdot log(\hat{Y}) \tag{1}$$

A complete training epoch includes a run through all the samples of the training set. Each training process had 200 epochs, since a larger number of epochs would not produce of epochs did not produce significant improvements neither in the loss function nor in the accuracy metrics. In addition, to ensure the generalization capability of the approaches presented, each experiment was repeated 5 times independently of each other with random sample selection, so it was necessary to calculate the means of these repetitions to evaluate the overall global performance. To compensate for the lack of available X-ray images and thus avoid problems of overfitting and to increase the generalization capacity, data augmentation was performed to obtain more robust and stable models. Thus, scaling and horizontal rotation operations were performed, which are appropriated given the symmetrical nature of the chest X-ray image, so the variability of the data used was increased. We consider this configuration to be suitable enough for our sex and age study, as it has provided satisfactory results in similar works [17, 19, 20].

Evaluation

The performance of the presented computational approaches was evaluated by comparing the predictions provided by the models with the ground truth labels annotated in the X-ray image datasets. Then, the values of True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN) were considered to calculate different metrics that are commonly used in the literature [17, 19, 20] to assess the stability of computational methods for medical imaging problems. Following the reference of these similar works, we also decided to use these metrics for our analysis of the sex and age factors. Thus, Precision, Recall, F1-score, and Accuracy were calculated for each approach as follows.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{4}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

Results

In this section, we present the experimental results of the proposed computational approaches for the classification of COVID-19 in chest X-ray images, covering a wide range of cases that will allow us to draw more contrasted and solid conclusions regarding the studied factors of sex and age. In particular, we perform two different and complementary studies on the COVID-19 dataset. The first one analyses the influence of the sex factor for each of the 3 approaches: (I) Normal VS COVID-19, (II) Pneumonia VS COVID-19 and (III) Non-COVID-19 VS COVID-19. The second one performs a similar analysis, but in this case considering patients by age ranges. Both studies are described below.

Sex-related imbalance analysis

One of the main characteristics of a patient that can influence a diagnostic system is sex [22, 23]. Especially in chest x-rays, we might think that differences in size, in addition to other typical sex characteristics such as the presence of breasts, could imply taking the images in different postures or certain abnormalities in the samples that could be mistaken for signs of a pathology, in this case this being COVID-19 [34]. In Fig. 5, we exemplify these differences with 2 patients of different sexes who have COVID-19. Considering how important is to identify a bias related to the sex of the patient, we designed the following study in order to test whether this characteristic influences diagnosing COVID-19.

In this first analysis, we explored intermediate imbalance scenarios in which female and male patients diagnosed with COVID-19 were analysed in different proportions with 10% intervals, ranging from 0% male patients and 100% female patients to 100% male patients and 0% female patients. Thus, we conducted a comprehensive

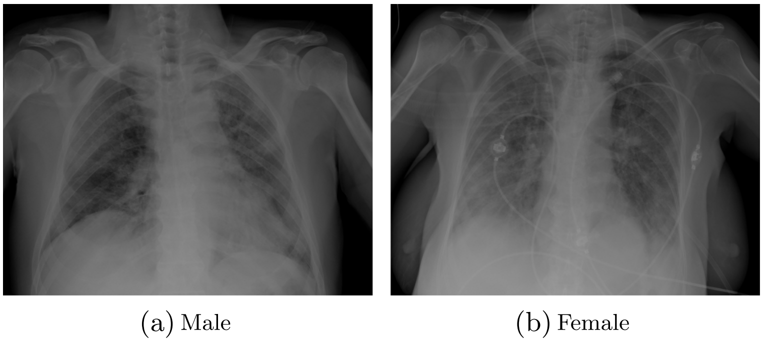

(a) Male (b) Female

Fig. 5 Example of two representative chest X-ray images of male and female patients diagnosed with COVID-19

Table 3 Distribution of randomly selected X-ray images for each computational approach in the sex-related imbalance analysis

Approach	Normal	Pneumonia
Normal VS COVID-19	350 M + 350 F	0
Pneumonia VS COVID-19	0	350 M + 350 F
Non-COVID-19 VS COVID-19	175 M + 175 F	175 M + 175 F

analysis with 11 different configurations for each computational approach. For each imbalance case, we get a model that is then tested using the remaining images not used during training. Afterwards, we compare the results obtained for each scenario with our baseline (50% female and 50% male). Regarding the amount of images considered for each approach, we used 700 COVID-19 images from 700 different patients. Although the dataset considered in this study consists of 5,493 COVID-19 images, it includes several COVID-19 images obtained from the same patient over time. Furthermore, in terms of gender, the dataset is composed of 1,132 male and 730 female patients. Finally, we have discarded 30 female patients because they did not have a chest X-ray image but another type of medical image, such as a lung CT scan. Therefore, in order to perform a more honest and unbiased analysis, we only have 700 patients in sex-related imbalance analysis. To maintain balance between this COVID-19 class and the other classes, 700 X-ray images were randomly selected and divided according to the sex of the patient, as indicated in Table 3. Therefore, each of the 11 experiments was performed using 1400 chest X-ray images.

Analysis of the 1st approach: Normal vs. COVID-19

In Table 4 we present a comparative analysis of the performance at the test stage using precision, recall, and F1-score measures, where we highlight our baseline as we are going to use it to compare our metrics. As for the mean accuracy obtained at each scenario, our values ranged from 0.9757± 0.0105 at the 40%M 60%F case, to 0.9835± 0.0105 at the 90%M 10%F case. The standard deviation of these metrics was always below 2.1%, being the highest at the 60%M 40%F case, and the lowest at 30%M 70%F with 0.58%. In general, it can be observed

Table 4 Mean ± standard deviation of the results obtained in the test stage for the classification of chest X-ray images between Normal VS COVID-19 after 5 independent repetitions. The baseline is highlighted in grey

Experiment	Class	Precision	Recall	F1-Score
0%M 100%F	Normal	0.9872± 0.0076	0.9831± 0.0124	0.9851± 0.0090
	COVID-19	0.9827± 0.0132	0.9869± 0.0080	0.9848± 0.0095
10%M 90%F	Normal	0.9827± 0.0185	0.9882± 0.0082	0.9854± 0.0092
	COVID-19	0.9888± 0.0079	0.9833± 0.0181	0.9859± 0.0090
20%M 80%F	Normal	0.9957± 0.0063	0.9821± 0.0139	0.9888± 0.0078
	COVID-19	0.9810± 0.0154	0.9956± 0.0065	0.9882± 0.0085
30%M 70%F	Normal	0.9830± 0.0113	0.9800± 0.0112	0.9814± 0.0053
	COVID-19	0.9800± 0.0117	0.9827± 0.0121	0.9813± 0.0063
40%M 60%F	Normal	0.9839± 0.0120	0.9670± 0.0121	0.9754± 0.0108
	COVID-19	0.9677± 0.0115	0.9843± 0.0115	0.9759± 0.0102
50%M 50%F	Normal	0.9902± 0.0104	0.9817± 0.0095	0.9859± 0.0077
	COVID-19	0.9813± 0.0093	0.9896± 0.0112	0.9854± 0.0082
60%M 40%F	Normal	0.9818± 0.0198	0.9874± 0.0240	0.9846± 0.0208
	COVID-19	0.9868± 0.0257	0.9811± 0.0213	0.9839± 0.0224
70%M 30%F	Normal	0.9928± 0.0100	0.9803± 0.0175	0.9865± 0.0130
	COVID-19	0.9800± 0.0188	0.9927± 0.0103	0.9863± 0.0139
80%M 20%F	Normal	0.9843± 0.0135	0.9843± 0.0128	0.9842± 0.0083
	COVID-19	0.9843± 0.0123	0.9843± 0.0135	0.9842± 0.0081
90%M 10%F	Normal	0.9843± 0.0135	0.9957± 0.0063	0.9899± 0.0090
	COVID-19	0.9955± 0.0065	0.9845± 0.0136	0.9899± 0.0094
100%M 0%F	Normal	0.9826± 0.0139	0.9840± 0.0091	0.9833± 0.0104
	COVID-19	0.9844± 0.0095	0.9831± 0.0138	0.9837± 0.0107

that the differences between the metrics are small when compared to our baseline and their values are maintained regardless of the studied scenario.

Analysis of the 2nd approach: pneumonia vs. COVID-19

The second group of experiments deals with the analysis of sex-related imbalance in the second approach. In this line, Table 5 show a comparative analysis of the performance at the test stage using precision, recall, and F1-score measures. Here, we highlight our baseline as we are going to use it to compare our metrics. As we can see, the results show a similar tendency to the previous set of experiments of the first approach, with values for the mean accuracy ranged from 0.9721± 0.0187 at the 0%M 100%F case, to 0.9892± 0.0091 at the 100%M 0%F case. The standard deviation of these metrics was always below 1.8%, being the highest at the 0%M 100%F case, and the lowest at 10%M 90%F with 0.86%.

Analysis of the 3rd approach: non-COVID-19 vs. COVID-19

In this third set of experiments, we analyzed the behavior of the sex factor imbalance in the data on separability between the Non-COVID-19 vs. COVID-19 classes. Table 6 shows the results of the test stage in terms of precision, recall and F1-Score for each class, after performing the proposed experiments, and we highlighted our baseline as we are going to use it to compare our metrics. As we can see, these results reflect that all models are able to accurately separate samples from both classes. As for the mean accuracy obtained at each scenario, our values ranged from 0.9700± 0.0117 at the 40%M 60%F case, to 0.9857± 0.0035 at the 100%M 0%F case. The standard deviation of these metrics was always below 1.3%, being the highest at the 60%M 40%F case, and the lowest at 100%M 0%F with 0.35%.

Age-related imbalance analysis

Age-related deterioration of both the skeleton and the musculature of the body is visible on chest X-rays, which may affect the diagnosis obtained from them [22, 35]. In addition, older COVID-19 patients often require more medical equipment that appears on chest X-ray images, such as intravenous lines, ventilators, pacemakers, and so on, which may again affect the diagnosis obtained from

Table 5 Mean ± standard deviation of the results obtained in the test stage for the classification of chest X-ray images between Pneumonia VS COVID-19 after 5 independent repetitions. The baseline is highlighted in grey

Experiment	Class	Precision	Recall	F1-Score
0%M 100%F	Pneumonia	0.9769± 0.0174	0.9675± 0.0224	0.9721± 0.0191
	COVID-19	0.9671± 0.0220	0.9771± 0.0167	0.9720± 0.0184
10%M 90%F	Pneumonia	0.9838± 0.0166	0.9756± 0.0116	0.9796± 0.0090
	COVID-19	0.9761± 0.0118	0.9848± 0.0155	0.9803± 0.0082
20%M 80%F	Pneumonia	0.9830± 0.0216	0.9898± 0.0110	0.9864± 0.0157
	COVID-19	0.9899± 0.0109	0.9829± 0.0214	0.9864± 0.0155
30%M 70%F	Pneumonia	0.9815± 0.0089	0.9815± 0.0125	0.9815± 0.0096
	COVID-19	0.9813± 0.0133	0.9812± 0.0102	0.9812± 0.0108
40%M 60%F	Pneumonia	0.9897± 0.0065	0.9854± 0.0116	0.9875± 0.0076
	COVID-19	0.9860± 0.0108	0.9902± 0.0062	0.9881± 0.0071
50%M 50%F	Pneumonia	0.9830± 0.0107	0.9766± 0.0206	0.9797± 0.0135
	COVID-19	0.9751± 0.0223	0.9825± 0.0109	0.9787± 0.0143
60%M 40%F	Pneumonia	0.9856± 0.0088	0.9779± 0.0191	0.9817± 0.0119
	COVID-19	0.9766± 0.0212	0.9855± 0.0086	0.9810± 0.0131
70%M 30%F	Pneumonia	0.9868± 0.0121	0.9854± 0.0145	0.9861± 0.0130
	COVID-19	0.9859± 0.0140	0.9874± 0.0114	0.9866± 0.0124
80%M 20%F	Pneumonia	0.9816± 0.0094	0.9778± 0.0151	0.9797± 0.0094
	COVID-19	0.9766± 0.0170	0.9811± 0.0095	0.9788± 0.0104
90%M 10%F	Pneumonia	0.9769± 0.0174	0.9675± 0.0224	0.9721± 0.0191
	COVID-19	0.9671± 0.0220	0.9771± 0.0167	0.9720± 0.0184
100%M 0%F	Pneumonia	0.9769± 0.0174	0.9675± 0.0224	0.9721± 0.0191
	COVID-19	0.9671± 0.0220	0.9771± 0.0167	0.9720± 0.0184

Table 6 Mean ± standard deviation of the results obtained in the test stage for the classification of chest X-ray images between Non-COVID-19 VS COVID-19 after 5 independent repetitions. The baseline is highlighted in grey

Experiment	Class	Precision	Recall	F1-Score
0%M 100%F	Non-COVID-19	0.9813± 0.0103	0.9826± 0.0167	0.9819± 0.0090
	COVID-19	0.9832± 0.0156	0.9815± 0.0108	0.9823± 0.0084
10%M 90%F	Non-COVID-19	0.9767± 0.0093	0.9795± 0.0196	0.9781± 0.0125
	COVID-19	0.9805± 0.0182	0.9775± 0.0092	0.9790± 0.0117
20%M 80%F	Non-COVID-19	0.9814± 0.0228	0.9900± 0.0121	0.9855± 0.0124
	COVID-19	0.9899± 0.0117	0.9819± 0.0217	0.9858± 0.0117
30%M 70%F	Non-COVID-19	0.9854± 0.0106	0.9838± 0.0150	0.9846± 0.0121
	COVID-19	0.9847± 0.0142	0.9859± 0.0096	0.9853± 0.0112
40%M 60%F	Non-COVID-19	0.9680± 0.0093	0.9707± 0.0213	0.9692± 0.0122
	COVID-19	0.9722± 0.0196	0.9692± 0.0095	0.9706± 0.0112
50%M 50%F	Non-COVID-19	0.9902± 0.0093	0.9885± 0.0121	0.9893± 0.0051
	COVID-19	0.9777± 0.0222	0.9793± 0.0197	0.9782± 0.0099
60%M 40%F	Non-COVID-19	0.9813± 0.0183	0.9785± 0.0108	0.9798± 0.0126
	COVID-19	0.9786± 0.0116	0.9816± 0.0188	0.9800± 0.0134
70%M 30%F	Non-COVID-19	0.9782± 0.0150	0.9896± 0.0084	0.9838± 0.0062
	COVID-19	0.9903± 0.0077	0.9792± 0.0146	0.9846± 0.0057
80%M 20%F	Non-COVID-19	0.9818± 0.0218	0.9707± 0.0148	0.9760± 0.0106
	COVID-19	0.9698± 0.0153	0.9812± 0.0230	0.9753± 0.0110
90%M 10%F	Non-COVID-19	0.9813± 0.0103	0.9826± 0.0167	0.9819± 0.0090
	COVID-19	0.9832± 0.0156	0.9815± 0.0108	0.9823± 0.0084
100%M 0%F	Non-COVID-19	0.9813± 0.0103	0.9826± 0.0167	0.9819± 0.0090
	COVID-19	0.9832± 0.0156	0.9815± 0.0108	0.9823± 0.0084

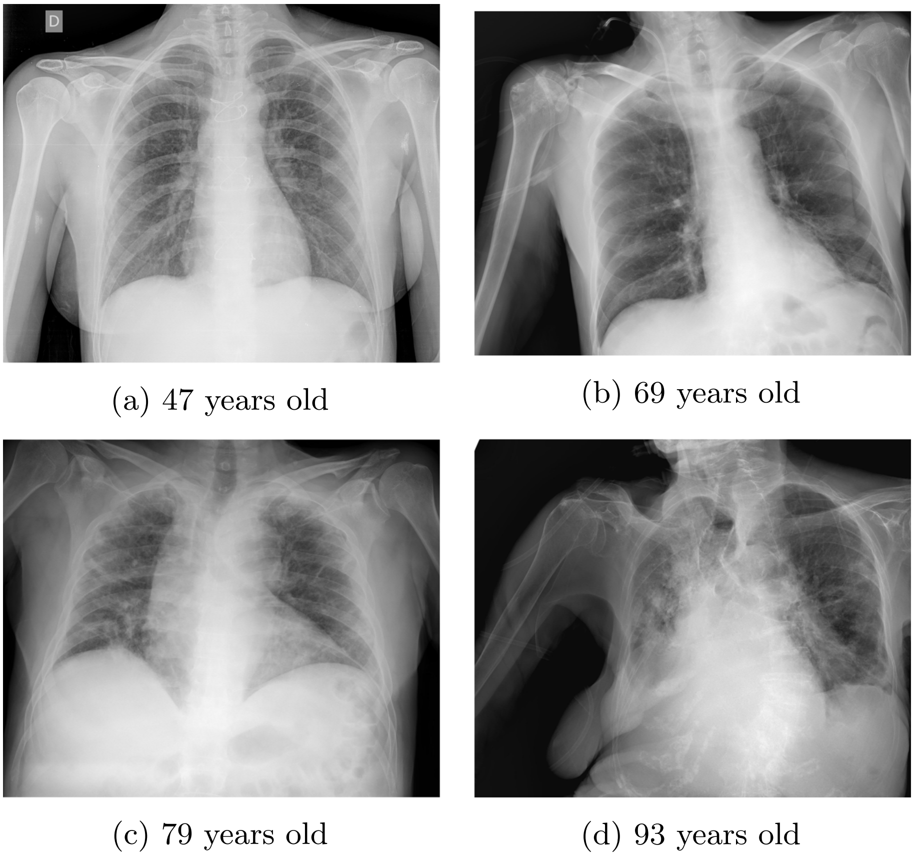
the X-rays [34]. To illustrate these characteristics associated with different ages, Fig. 6 shows representative examples of different COVID-19 patients ranging in age from 47 to 93 years old. These differences raise the need for a detailed study of how the patient age affects the diagnosis of COVID-19. Therefore, we describe below the analysis we have carried out for this purpose.

For the age-related imbalance study, we defined 6 different age ranges: 0-40, 40-50, 50-60, 60-70, 70-80, ≥ 80. For each range, we used only images from patients in that age spectrum for training and then tested it with the remaining images. We analysed the differences between the age group used for training, which acts as our baseline, and all other ages. Regarding the exact number of samples used for each class in our 3 computational approaches, we present our distribution in Table 7. Using this amount of images of each class, we sought to emphasise the older age groups, who suffer more from the disease and have to go through a more critical diagnostic process, but also adapting to the number of samples we had available from the studied Normal, Pneumonia and COVID-19 classes of interest.

In the following sections, we will show the results of our six baselines (one per age range) for each approach. However, the details of how these baselines responded to the different age groups will be discussed in the Discussion section in order to simplify this section and facilitate understanding.

Analysis of the 1st approach: Normal vs. COVID-19

For this first approach, we present in Table 8 precision, recall and F1-score means and their standard deviation obtained at test for each experiment training with only one age group. These results for our six baselines were satisfactory and mainly stable, as the metrics were over 90% in most cases and standard deviation was under 8%. Regarding the mean accuracy obtained for each one of these baselines, we obtained the following values: 0.9587± 0.0298, 0.9748± 0.0012, 0.9877± 0.0001, 0.9876± 0.0001, 0.9808± 0.0004 and 0.9429± 0.0086,


(a) 47 years old

(b) 69 years old

(c) 79 years old

(d) 93 years old

Fig. 6 Example of four representative chest X-ray images of patients of different ages diagnosed with COVID-19

ordering them from the youngest to the oldest age group. In general, this indicates that our baselines are acceptable and stable, since the accuracy was above 94% and the standard deviation kept under 8.6%.

Analysis of the 2nd approach: pneumonia vs. COVID-19

For our second set of experiments, we summarized in Table 9 the metrics and their standard deviation obtained for our baseline models at the test stage for each experiment training with only one age group. Again, these models had acceptable results, as they were above 90% in nearly all cases and its standard deviations were below 10%. As for the mean accuracy obtained for each one of these baselines, we obtained these values for every baseline ordered by age: 0.9396± 0.0027, 0.9760± 0.0004, 0.9800± 0.0005, 0.9919± 0.0001, 0.9772± 0.0004 and 0.9083± 0.043. Overall, these metrics are satisfactory and

Table 7 Number of samples of each class considered per approach

Age	Normal VS COVID-19	Pneumonia VS COVID-19	Non-COVID-19 VS COVID-19
<40	154 vs 154	154 vs 154	(77+77) vs 154
40-50	436 vs 436	436 vs 436	(218+218) vs 436
50-60	772 vs 772	772 vs 772	(386 + 386) vs 772
60-70	1,496 vs 1,496	1,215 vs 1,215	(748+748) vs 1,496
70-80	625 vs 625	392 vs 392	(392+392) vs 784
≥ 80	105 vs 105	58 vs 58	(58+58) vs 116

Table 8 Mean ± standard deviation of the results obtained in the test stage for the classification of chest X-ray images between Normal VS COVID-19 after 5 independent repetitions

Exp.	Class	Precision	Recall	F1-Score
<40	Normal	0.9453± 0.0514	0.9749± 0.0349	0.9588± 0.0277
	COVID-19	0.9742± 0.0355	0.9438± 0.0588	0.9573± 0.0276
40-50	Normal	0.9673± 0.0288	0.9814± 0.0141	0.9742± 0.0193
	COVID-19	0.9816± 0.0125	0.9693± 0.0250	0.9753± 0.0156
50-60	Normal	0.9911± 0.0057	0.9846± 0.0080	0.9878± 0.0053
	COVID-19	0.9844± 0.0066	0.9907± 0.0058	0.9875± 0.0043
60-70	Normal	0.9925± 0.0055	0.9826± 0.0062	0.9875± 0.0053
	COVID-19	0.9828± 0.0064	0.9926± 0.0055	0.9877± 0.0054
70-80	Normal	0.9780± 0.0171	0.9846± 0.0086	0.9812± 0.0102
	COVID-19	0.9832± 0.0111	0.9774± 0.0173	0.9802± 0.0115
≥ 80	Normal	0.9373± 0.0691	0.8912± 0.0849	0.9125± 0.0690
	COVID-19	0.8859± 0.0805	0.9255± 0.0800	0.9043± 0.0734

Table 9 Mean ± standard deviation of the results obtained in the test stage for the classification of chest X-ray images between Pneumonia VS COVID-19 after 5 independent repetitions

Exp.	Class	Precision	Recall	F1-Score
<40	Pneumonia	0.9603± 0.0440	0.9292± 0.0241	0.9438± 0.0188
	COVID-19	0.9199± 0.0303	0.9498± 0.0625	0.9336± 0.0356
40-50	Pneumonia	0.9811± 0.0177	0.9717± 0.0161	0.9762± 0.0105
	COVID-19	0.9689± 0.0219	0.9821± 0.0172	0.9752± 0.0122
50-60	Pneumonia	0.9802± 0.0118	0.9815± 0.0116	0.9808± 0.0112
	COVID-19	0.9796± 0.0126	0.9784± 0.0129	0.9790± 0.0122
60-70	Pneumonia	0.9942± 0.0046	0.9895± 0.0036	0.9918± 0.0036
	COVID-19	0.9893± 0.0040	0.9944± 0.0045	0.9919± 0.0035
70-80	Pneumonia	0.9869± 0.0132	0.9678± 0.0101	0.9772± 0.0097
	COVID-19	0.9668± 0.0123	0.9874± 0.0122	0.9770± 0.0097
≥ 80	Pneumonia	0.8850± 0.1117	0.9000± 0.1732	0.8893± 0.1380
	COVID-19	0.9346± 0.1084	0.9095± 0.0831	0.9195± 0.0840

steady, being above 90% and with a standard deviation under 4.3%.

Analysis of the 3rd approach: non-COVID-19 vs. COVID-19

Finally for this third approach, we show in Table 10 precision, recall and F1-score means and their standard deviation obtained at test for each experiment training with only one age group. Following the trend that we have already seen in the two previous approaches, our baseline models had adequate metrics, as they were above 90% in all scenarios and the corresponding standard deviation was below 6%. The results obtained for the mean accuracy from the youngest to the oldest baseline were the following: 0.9683± 0.0020, 0.9760± 0.0002, 0.9819± 0.0002, 0.9913±0.8 × 10^{-5}, 0.9898±0.8 × 10^{-4} and 0.9234± 0.0077. As we can see, all baselines remained above 96% and their standard deviation was under 7.7%, which make these metrics satisfactory and mainly stable.

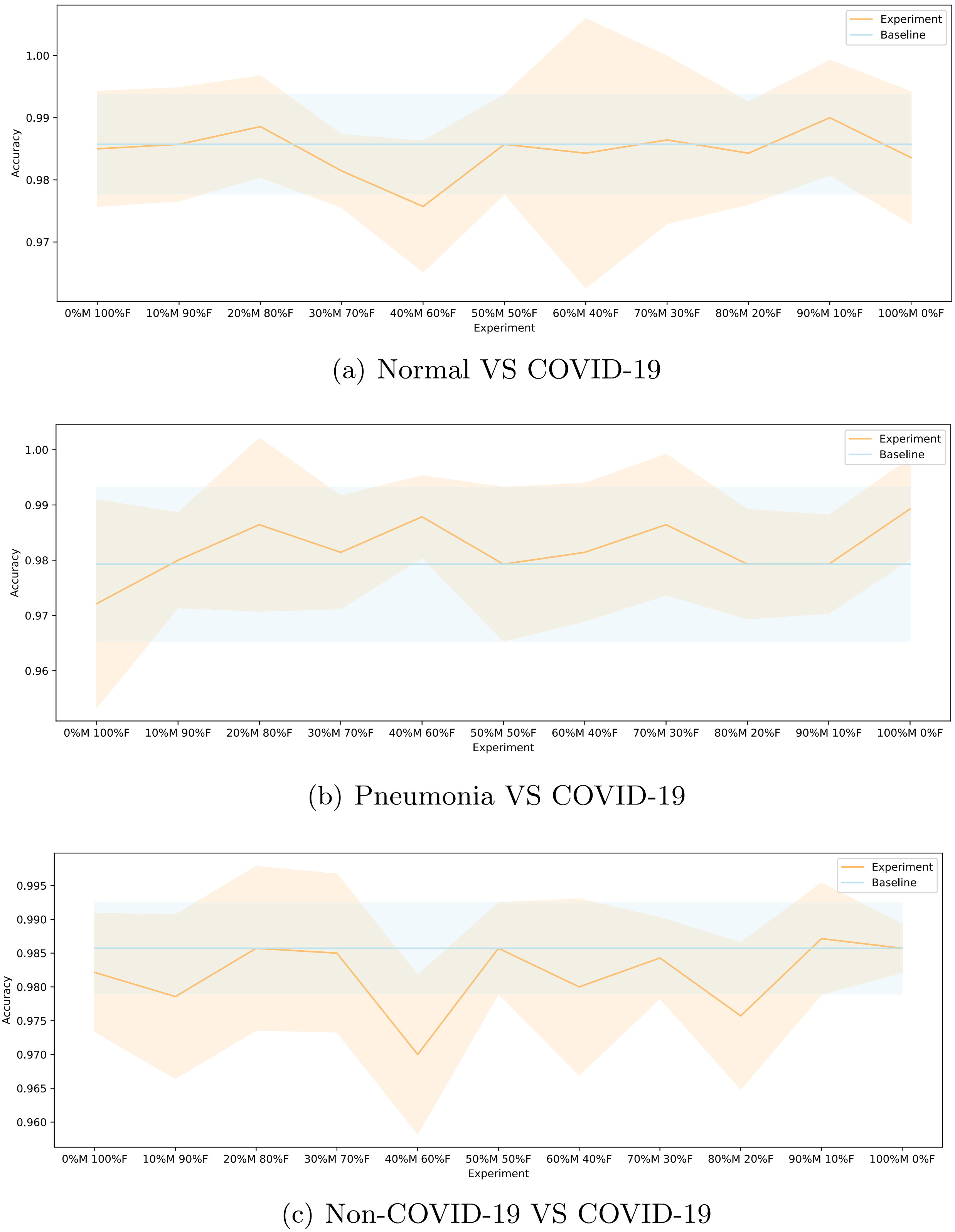
Table 10 Mean ± standard deviation of the results obtained in the test stage for the classification of chest X-ray images between Non-COVID-19 VS COVID-19 after 5 independent repetitions

Exp.	Class	Precision	Recall	F1-Score
<40	Non-COVID-19	0.9754± 0.0141	0.9625± 0.0344	0.9688± 0.0231
	COVID-19	0.9617± 0.0352	0.9725± 0.0157	0.9669± 0.0226
40-50	Non-COVID-19	0.9707± 0.0155	0.9821± 0.0154	0.9762± 0.0059
	COVID-19	0.9812± 0.0189	0.9701± 0.0184	0.9754± 0.0093
50-60	Non-COVID-19	0.9849± 0.0096	0.9802± 0.0105	0.9825± 0.0078
	COVID-19	0.9785± 0.0123	0.9840± 0.0096	0.9812± 0.0084
60-70	Non-COVID-19	0.9925± 0.0043	0.9898± 0.0041	0.9911± 0.0016
	COVID-19	0.9901± 0.0039	0.9927± 0.0043	0.9914± 0.0011
70-80	Non-COVID-19	0.9950± 0.0052	0.9850± 0.0069	0.9900± 0.0046
	COVID-19	0.9846± 0.0072	0.9947± 0.0053	0.9896± 0.0047
≥ 80	Non-COVID-19	0.9429± 0.0547	0.9107± 0.0633	0.9250± 0.0426
	COVID-19	0.9068± 0.0573	0.9380± 0.0695	0.9206± 0.0480

Discussion

Regarding the sex-related imbalance analysis, the precision, recall and F1-score measures shown in Results section were in every experiment in all the approaches above 96%, which is a satisfactory result. As for accuracy, we summarized the obtained measures for every experiment for each approach in Fig. 7. We can see here how there are no extreme peaks in either the accuracy or its standard deviation in none of the approaches, and differences between experiments and approaches are around 5%. Although the Normal VS COVID-19 approach has a bigger standard deviation peak at the 60% male and 40% female experiment, all values remain closer and similar to our baseline. The same occurs for the Pneumonia VS COVID-19 approach, as accuracy continues to be stable and alike our baseline. In the Non-COVID-19 VS COVID-19 approach we have a slightly different scenario, since most of the obtained values are under our baseline, especially in experiments 40% male and 60% female, and 80% male and 20% female. Despite these differences, we can observe how accuracy remains stable and similar to other approaches. All these satisfactory results, together with the stability observed in all the scenarios considered in each of our approaches, indicate that this factor has not clearly affected the diagnosis offered by our system. If it had, we would have seen graphs with more evident differences between each of the different sex ratios with which we experimented. Thereby, no influence caused by the sex factor was observed. Although male and female patients may have differentiating features that allow us to identify their sex on chest x-rays, such as breasts, differences in shape and size, etc., these typically sex-associated features do not influence their COVID-19 diagnosis and do not favour one sex over the other, as they do not interfere with the lung assessment. For example, differences in shape and size do not difficult the finding of suspicious densities in the lung itself, and those densities related to the mammary glands are easily discarded, as they are present in most female patients and do not usually obscure COVID-19 related findings.

Regarding the age-related imbalance analysis, the precision, recall and F1-score measures shown in Results section were in every experiment in all approaches above 96%, which is a satisfactory result. As for accuracy, we summarized the obtained results for each approach in Fig. 8, taking as a reference the baselines metrics shown in the Results section. In this accuracy comparative across all six age ranges it is presented how its standard devia-

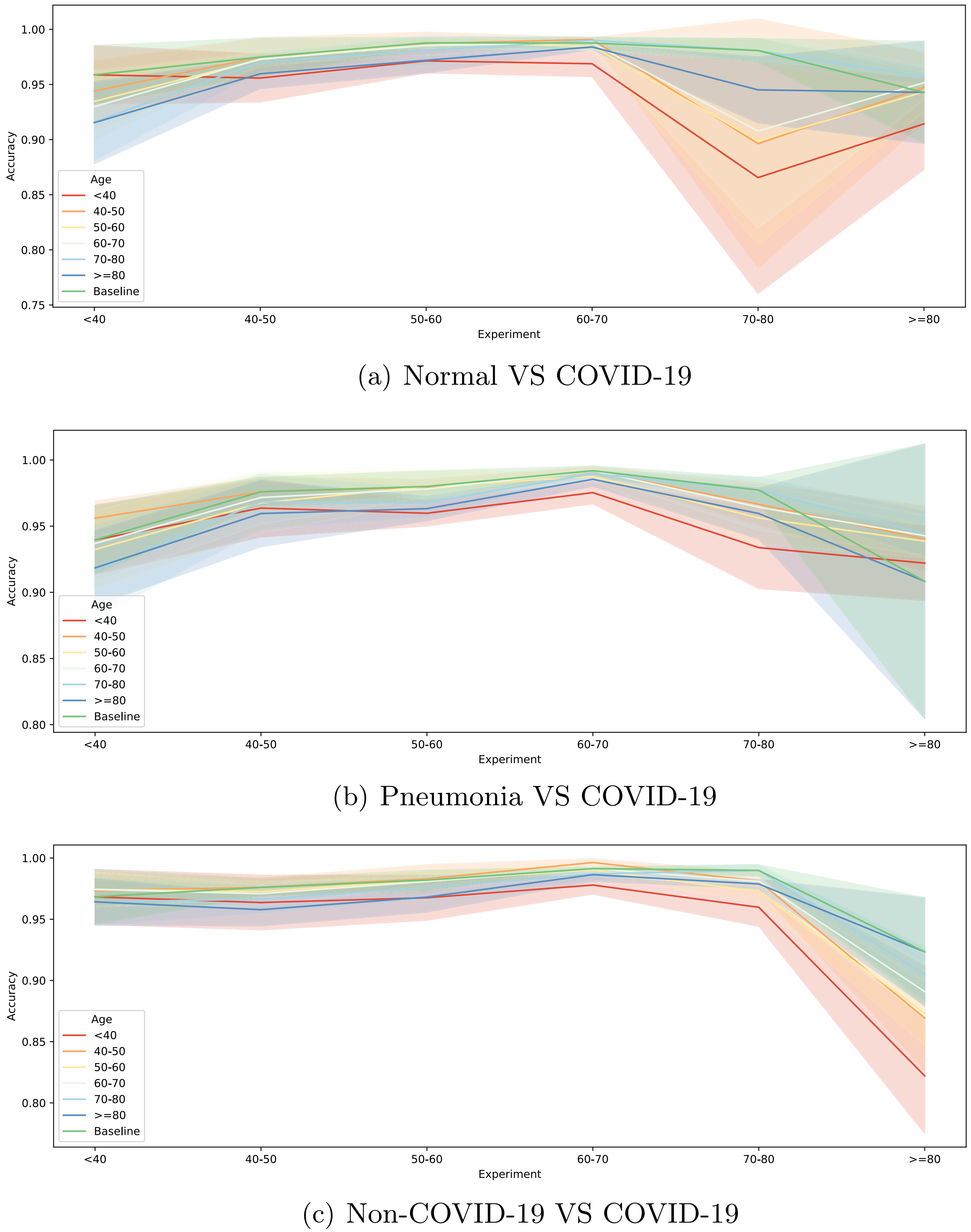


(a) Normal VS COVID-19

(b) Pneumonia VS COVID-19

(c) Non-COVID-19 VS COVID-19

Fig. 7 Mean ± standard deviation test accuracy obtained for every studied scenario in every approach



(a) Normal VS COVID-19

(b) Pneumonia VS COVID-19

(c) Non-COVID-19 VS COVID-19

Fig. 8 Mean ± standard deviation test accuracy obtained for every studied age range in every approach

tion increases as baseline patients get older than 70. The worst instability peaks are in the 70-80 range in the Normal VS COVID-19 approach and the ≥ 80 range in the Pneumonia VS COVID-19 approach, but these increases only represent a worsening of 10%. This behaviour is not as clearly observed for the Non-COVID-19 VS COVID-19 approach, since its standard deviations rises at the ≥ 80 range, but not as noticeably as in other approaches. In relation to the accuracy metric itself, it is observed how the closer to the baseline age the tested age range gets, the better accuracies are obtained. However, these differences are not of great magnitude. In general, the third approach seems like the best and most stable of the three ones considered, since its accuracy is consistently good enough at every age range, and its standard deviation has a smaller peak at the older ages. Nevertheless, both the worsening in the obtained accuracy and the its instability are not of great magnitude in any approach. Thus, we can clearly observe in these graphs the clear tendency of the diagnosis offered to be influenced by age, regardless of the age group studied or the used computational approach. Moreover, it is noteworthy that this worsening is more or less present in all the cases studied, but is more pronounced in the older age groups, which is consistent given that the most critical cases of COVID-19 are more frequent in this group, resulting in a greater variability of pathological affectations in the lungs. For example, older patients are usually easily recognized by the wide range of different damaged ribcages they might present, being these caused by diseases or by the passing of time. In this situation, these patients are typically weaker in the face of such an aggressive disease as COVID-19, so different types of medical equipment, such as pathways or thoracostomy tubes, among other cardiac and pulmonary devices, are more present in these X-rays. All of these elements can appear on these images, obscuring lung densities typical of COVID-19 or leading our systems to recognise these patients more by the irregularity of their X-rays than by the signs of disease they may manifest, both affecting their COVID-19 diagnosis. However, these characteristics do not appear as frequently in the chest X-rays of younger patients, who typically have images where abnormalities are more easily observed and their association to COVID-19 is more straightforward, because they do not have other pathologies that may cause the presence of irregularities in their images. Hence, these reasons could justify the presence of this bias. In this work, we have performed a comprehensive analysis of sex and age factors in the chest Xray images. Accordingly, we have generated 615 ROC curves from the experiments (see supplementary material available at https://doi.org/10.1186/s12874-022-01578-w).

Conclusions

In this work, we have proposed the first study to analyze whether imbalance in chest X-ray datasets produces biased deep learning approaches for COVID-19 screening with respect to the studied sex and age factors. For this purpose, 3 computational approaches using deep learning strategies that allowed us to carry out these studies of these factors in a detailed and comprehensive manner are presented and evaluated. To demonstrate the capabilities of our proposal, we perform several experiments on different public image datasets, including Normal, Pneumonia and COVID-19 cases. The presented results evidenced that the proposed methodology and tested approaches allow a robust and reliable analysis to support the clinical decision-making process in this pandemic scenario. Given the effort made to consider as many cases as possible and to make these studies as comprehensive as possible, we believe that the conclusions presented below are robust and reliable.

Regarding the sex-related imbalance analysis, we observed that this characteristic did not significantly affect the performance of our system. Whatever the sex ratio, the system performed well and provided satisfactory and stable results in all analyzed approaches. Since we performed a thorough study where we examined many different scenarios and explored different sex proportions, we can conclude that our system was not biased by this characteristic. Therefore, any difference observed between male and female patients from our dataset was not big enough to influence the system. On the other hand, regarding the age-related imbalance analysis, we observed that this characteristic did affect the performance of our system. It was clearly seen in every approach how the age used for training biased the system making it perform better for those with closer ages to the training phase one. Although obtained accuracy was good enough in every scenario as it was above 90% for most of the cases, age bias was consistent across all approaches. Again, since this analysis was conducted in a comprehensive manner, we can reliably conclude that the system was affected by the age of the patient. This could be caused by many reasons. For example, older patients have more irregular chest X-rays than younger people, since they can manifest different bone or cardiac pathologies. These differences might explain separability between the age ranges studied and their different results. Despite the fact a clear cause for this behaviour was not found, it is not necessary to emphasize how much it is needed to review the datasets being used for COVID-19 screening and identify possible bias related to the patient's age in them, since it was checked by our experiments that this factor's imbalance might affect the performance of the developed system.

As future work, it would be interesting to extend our study with patients diagnosed with other pulmonary disorders, such as emphysema, bronchitis and tuberculosis, among others. On the one hand, common pathologies affecting the lungs could represent a more challenging scenario of interest. On the other hand, expanding the dataset is of great interest to validate more completely the proposed methodologies. Other interesting future work would be to extend this analysis to other types of medical imaging modalities and correlate the results in a multimodal context to identify more precisely the influence of sex and age factors in COVID-19 screening systems. From a more technical point of view, in this work, we choose the input image size that is commonly used in the state of the art in similar problems. However, analyzing the relevance of this factor would ensure that important details are not being overlooked by reducing the image so much. In this sense, a more complete study could be done, testing with different input sizes. In addition, to facilitate the detection of biases of this type in related works, it would be interesting to implement a graphical user interface in order to make it easier for other users to test our methodology with different datasets.

Abbreviations

WHO: World Health Organization; SARS-CoV-2: Severe Acute Respiratory Syndrome Coronavirus 2; CNN: Convolutional Neural Network; CAAD: Confidence Aware Anomaly Detection; TP: True Positive; TN: True Negative; FP: False Positive; FN: False Negative

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-022-01578-w.

Additional file 1: Supplementary material.

Acknowledgements

We acknowledge all study participants, who made this research possible. We thank the HM Hospitales group and the Radiological Society of North America (RSNA) for releasing the data set used to perform this research to the scientific community.

Authors' contributions

Lorena Álvarez-Rodríguez: Conceptualization, Methodology, Software, Writing - Original Draft, Writing - Review & Editing, Visualization. **Joaquim de Moura:** Conceptualization, Methodology, Software, Validation, Investigation, Writing - Review & Editing, Supervision. **Jorge Novo:** Conceptualization, Methodology, Validation, Investigation, Writing - Review & Editing, Supervision, Project administration. **Marcos Ortega:** Conceptualization, Methodology, Validation, Investigation, Writing - Review & Editing, Supervision, Project administration, Funding Acquisition. The authors read and approved the final manuscript.

Funding

This research was funded by Instituto de Salud Carlos III, Government of Spain, DTS18/00136 research project; Ministerio de Ciencia e Innovación y Universidades, Government of Spain, RTI2018-095894-B-I00 research project; Ministerio de Ciencia e Innovación, Government of Spain through the research project with reference PID2019-108435RB-I00; Consellería de Cultura, Educación e Universidade, Xunta de Galicia, Grupos de Referencia Competitiva, grant ref. ED431C 2020/24; postdoctoral grant ref. ED481B 2021/059; Axencia Galega de Innovación (GAIN), Xunta de Galicia, grant ref. IN845D 2020/38; CITIC, Centro de Investigación de Galicia ref. ED431G 2019/01, receives financial support from Consellería de Educación, Universidade e Formación Profesional, Xunta de Galicia, through the ERDF (80%) and Secretaría Xeral de Universidades (20%).

Availability of data and materials

The data that support the findings of this study are available from Radiological Society of North America (RSNA) and the HM Hospitales group. The RSNA dataset is publicly available in the Kaggle platform (https://www.kaggle.com/c/rsna-pneumonia-detection-challenge). In the case of the HM Hospitales dataset, although publicly available, restrictions apply to the availability of these data, which were used under licence for the present study. The source code developed during this study is available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

The study was approved by the Ethics Review Board and Data Management Technical Commission of Galician Health Ministry for High Impact studies with protocol code 2020-007, following the Declaration of Helsinki Ethical Principles for Medical Research Involving Human Subjects. In this study, all our data were obtained from publicly available datasets and we did not directly recruit or have direct contact with study participants.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Centro de Investigación CITIC, Universidade da Coruña, Campus de Elviña, 15071 A Coruña, Spain. [2]Grupo VARPA, Instituto de Investigación Biomédica de A Coruña (INIBIC), Universidade da Coruña, 15006 A Coruña, Spain.

Received: 8 October 2021 Accepted: 21 March 2022
Published online: 28 April 2022

References

1. Weekly epidemiological update on COVID-19 - 25 January 2022. https://www.who.int/publications/m/item/weekly-epidemiological-update-on-covid-19---25-january-2022. Accessed Sep 2021.
2. Gomes JC, Masood AI, de S. Silva LH, da Cruz Ferreira JRB, Júnior AAF, dos Santos Rocha AL, de Oliveira LCP, da Silva NRC, Fernandes BJT, dos Santos WP. Covid-19 diagnosis by combining RT-PCR and pseudo-convolutional machines to characterize virus sequences. Sci Rep. 2021;11(1):. https://doi.org/10.1038/s41598-021-90766-7.
3. Serena Low WC, Chuah JH, Tee CAT, Anis S, Shoaib MA, Faisal A, Khalil A, Lai KW. An overview of deep learning techniques on chest x-ray and ct scan identification of covid-19. Comput Math Meth Med. 2021;2021:. https://doi.org/10.1155/2021/5528144.
4. Mohammad-Rahimi H, Nadimi M, Ghalyanchi-Langeroudi A, Taheri M, Ghafouri-Fard S. Application of machine learning in diagnosis of covid-19 through x-ray and ct images: a scoping review. Front Cardiovasc Med. 2021;8:185. https://doi.org/10.3389/fcvm.2021.638011.
5. Wang L, Lin ZQ, Wong A. Covid-net: a tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. Sci Rep. 2020;10(1):19549. https://doi.org/10.1038/s41598-020-76550-z.
6. Hammoudi K, Benhabiles H, Melkemi M, Dornaika F, Arganda-Carreras I, Collard D, Scherpereel A. Deep Learning on Chest X-ray Images to Detect and Evaluate Pneumonia Cases at the Era of COVID-19. 2020. 2004.03399. Accessed Sep 2021.
7. Hemdan EE-D, Shouman MA, Karar ME. COVIDX-Net: A Framework of Deep Learning Classifiers to Diagnose COVID-19 in X-Ray Images. 2020. 2003.11055. Accessed Sep 2021.
8. Zhang J, Xie Y, Pang G, Liao Z, Verjans J, Li W, Sun Z, He J, Li Y, Shen C, et al. Viral pneumonia screening on chest x-rays using confidence-aware anomaly detection. IEEE Trans Med Imaging. 2020;40(3):879–90.

9. Ozturk T, Talo M, Yildirim EA, Baloglu UB, Yildirim O, Rajendra Acharya U. Automated detection of covid-19 cases using deep neural networks with x-ray images. Comput Biol Med. 2020;121:103792. https://doi.org/10.1016/j.compbiomed.2020.103792.
10. Gomes JC, de F. Barbosa VA, Santana MA, Bandeira J, Valença MJS, de Souza RE, Ismael AM, dos Santos WP. IKONOS: an intelligent tool to support diagnosis of COVID-19 by texture analysis of x-ray images. Res Biomed Eng. 2020. https://doi.org/10.1007/s42600-020-00091-7.
11. Ismael AM, Şengür A. The investigation of multiresolution approaches for chest x-ray image based COVID-19 detection. Health Inf Sci Syst. 2020;8(1):. https://doi.org/10.1007/s13755-020-00116-6.
12. Shelke A, Inamdar M, Shah V, Tiwari A, Hussain A, Chafekar T, Mehendale N. Chest x-ray classification using deep learning for automated covid-19 screening. medRxiv. 2020. https://doi.org/10.1101/2020.06.21.20136598. https://www.medrxiv.org/content/early/2020/06/23/2020.06.21.20136598.full.pdf.
13. Yoo SH, Geng H, Chiu TL, Yu SK, Cho DC, Heo J, Choi MS, Choi IH, Cung Van C, Nhung NV, Min BJ, Lee H. Deep learning-based decision-tree classifier for covid-19 diagnosis from chest x-ray imaging. Front Med. 2020;7:427. https://doi.org/10.3389/fmed.2020.00427.
14. Ismael AM, Şengür A. Deep learning approaches for COVID-19 detection based on chest x-ray images. Expert Syst Appl. 2021;164:114054. https://doi.org/10.1016/j.eswa.2020.114054.
15. Li MD, Arun NT, Gidwani M, Chang K, Deng F, Little BP, Mendoza DP, Lang M, Lee SI, O'Shea A, et al. Automated assessment of covid-19 pulmonary disease severity on chest radiographs using convolutional siamese neural networks. 2020. https://doi.org/10.1101/2020.05.20.20108159.
16. Chicco D. Siamese Neural Networks: An Overview. New York: Springer; 2021, pp. 73–94.
17. de Moura J, Novo J, Ortega M. Fully automatic deep convolutional approaches for the analysis of covid-19 using chest x-ray images. medRxiv. 2020. https://doi.org/10.1101/2020.05.01.20087254.
18. Waheed A, Goyal M, Gupta D, Khanna A, Al-Turjman F, Pinheiro PR. Covidgan: Data augmentation using auxiliary classifier gan for improved covid-19 detection. IEEE Access. 2020;8:91916–23. https://doi.org/10.1109/ACCESS.2020.2994762.
19. Morís DI, de Moura Ramos JJ, Buján JN, Hortas MO. Data augmentation approaches using cycle-consistent adversarial networks for improving covid-19 screening in portable chest x-ray images. Expert Syst Appl. 2021;185:115681. https://doi.org/10.1016/j.eswa.2021.115681.
20. De Moura J, García LR, Vidal PFL, Cruz M, López LA, Lopez EC, Novo J, Ortega M. Deep convolutional approaches for the analysis of covid-19 using chest x-ray images from portable devices. IEEE Access. 2020;8:195594–607. https://doi.org/10.1109/ACCESS.2020.3033762.
21. Mooney P. Chest x-ray images (Pneumonia). https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia. Accessed Sep 2021.
22. Cirillo D, Catuara-Solarz S, Morey C, Guney E, Subirats L, Mellino S, Gigante A, Valencia A, Rementeria MJ, Chadha AS, Mavridis N. Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. NPJ Digit Med. 2020;3(1):. https://doi.org/10.1038/s41746-020-0288-5.
23. Larrazabal AJ, Nieto N, Peterson V, Milone DH, Ferrante E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proc Natl Acad Sci. 2020;117(23):12592–94. https://doi.org/10.1073/pnas.1919012117.
24. Vidal PL, de Moura J, Novo J, Ortega M. Multi-stage transfer learning for lung segmentation using portable x-ray devices for patients with covid-19. Expert Syst Appl. 2021;173:114677. https://doi.org/10.1016/j.eswa.2021.114677.
25. Covid Data Save Lives Dataset. 2021. https://www.hmhospitales.com/coronavirus/covid-data-save-lives/english-version. Accessed Sep 2021.
26. of North America RS. RSNA Pneumonia Detection Challenge. 2018. https://www.rsna.org/education/ai-resources-and-training/ai-image-challenge/rsna-pneumonia-detection-challenge-2018. Accessed Sep 2021.
27. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers R. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR); 2017. p. 3462–71. https://doi.org/arxiv.org/abs/1705.02315.
28. Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017. https://doi.org/arxiv.org/abs/1608.06993.
29. Afifi A, Hafsa NE, Ali MAS, 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. 2021;13(1):. https://doi.org/10.3390/sym13010113.
30. Wang Z, Xiao Y, Li Y, Zhang J, Lu F, Hou M, Liu X. Automatically discriminating and localizing covid-19 from community-acquired pneumonia on chest x-rays. Pattern Recog. 2021;110:107613. https://doi.org/10.1016/j.patcog.2020.107613.
31. Minaee S, Kafieh R, Sonka M, Yazdani S, Jamalipour Soufi G. Deep-covid: Predicting covid-19 from chest x-ray images using deep transfer learning. Med Image Anal. 2020;65:101794. https://doi.org/10.1016/j.media.2020.101794.
32. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE; 2009. p. 248–55. https://doi.org/10.1109/ACCESS.2021.3082638.
33. Ketkar N. Stochastic Gradient Descent. Berkeley: Springer; 2017, pp. 113–32.
34. Arias-Londoño JD, Gómez-García JA, Moro-Velázquez L, Godino-Llorente JI. Artificial intelligence applied to chest x-ray images for the automatic detection of covid-19. a thoughtful evaluation approach. IEEE Access. 2020;8:226811–27. https://doi.org/10.1109/ACCESS.2020.3044858.
35. Suri JS, Agarwal S, Gupta SK, Puvvula A, Biswas M, Saba L, Bit A, Tandel GS, Agarwal M, Patrick A, Faa G, Singh IM, Oberleitner R, Turk M, Chadha PS, Johri AM, Miguel Sanches J, Khanna NN, Viskovic K, Mavrogeni S, Laird JR, Pareek G, Miner M, Sobel DW, Balestrieri A, Sfikakis PP, Tsoulfas G, Protogerou A, Misra DP, Agarwal V, Kitas GD, Ahluwalia P, Teji J, Al-Maini M, Dhanjil SK, Sockalingam M, Saxena A, Nicolaides A, Sharma A, Rathore V, Ajuluchukwu JNA, Fatemi M, Alizad A, Viswanathan V, Krishnan PK, Naidu S. A narrative review on characterization of acute respiratory distress syndrome in covid-19-infected lungs using artificial intelligence. Comput Biol Med. 2021;130:104210. https://doi.org/10.1016/j.compbiomed.2021.104210.

Publisher's Note

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