Health and Technology (2022) 12:1117–1132
https://doi.org/10.1007/s12553-022-00704-4

ORIGINAL PAPER

New patch-based strategy for COVID-19 automatic identification using chest x-ray images

Jorge A Portal-Diaz[1] · Orlando Lovelle-Enríquez[2] · Marlen Perez-Diaz[3] · José D Lopez-Cabrera[4] · Osmany Reyes-Cardoso[5] · Ruben Orozco-Morales[3]

Received: 22 June 2022 / Accepted: 9 October 2022 / Published online: 10 November 2022

Abstract

Purpose The development of a robust model for automatic identification of COVID-19 based on chest x-rays has been a widely addressed topic over the last couple of years; however, the scarcity of good quality images sets, and their limited size, have proven to be an important obstacle to obtain reliable models. In fact, models proposed so far have suffered from over-fitting erroneous features instead of learning lung features, a phenomenon known as shortcut learning. In this research, a new image classification methodology is proposed that attempts to mitigate this problem.

Methods To this end, annotation by expert radiologists of a set of images was performed. The lung region was then segmented and a new classification strategy based on a patch partitioning that improves the resolution of the convolution neural network is proposed. In addition, a set of native images, used as an external evaluation set, is released.

Results The best results were obtained for the 6-patch splitting variant with 0.887 accuracy, 0.85 recall and 0.848 F1score on the external validation set.

Conclusion The results show that the proposed new strategy maintains similar values between internal and external validation, which gives our model generalization power, making it available for use in hospital settings.

Keywords Chest X-Rays · COVID-19 · Automatic classification

1 Introduction

1.1 Brief introduction about COVID-19 framework

In December 2019, in Wuhan, China, a new disease was identified, namely COVID-19. This new disease, caused by the SARS-CoV2 virus, was announced by WHO as a pandemic in March 2020 and so far, more than 413 million positive cases are reported, with more than 5.8 million deaths worldwide. SARS-CoV2 causes complications such as acute respiratory disorders, cardiac problems and secondary infections in a fairly high proportion of patients. Early identification of infected individuals is of vital importance, not only to apply medical care, but also to cut off the chain of disease transmission [1].

So far, the gold standard for disease diagnosis is by identification of viral RNA from reverse transcription-polymerase chain reaction (RT-PCR). The ability to perform this type of testing in the early months of the pandemic was limited in most countries [2]. Also, one of the disadvantages of PCR is the time it takes to provide results, in addition to not being effective when the viral load is still low in the first three or four days after infection [3]. In addition, sampling introduces a very high variability, depending on the site, the level of expertise of the personnel and the viral load of the person at the time [4]. Therefore, investigating other approaches to the diagnosis and evolution of damage to the respiratory system is an important task [5].

✉ Marlen Perez-Diaz
marlen20012001@gmail.com

José D Lopez-Cabrera
josedaniellopezcabrera@gmail.com

[1] Informatic Office, Universidad Central Marta Abreu de Las Villas, Santa Clara, Cuba

[2] Radiology Department, Hospital Comandante Manuel Fajardo, Santa Clara, Cuba

[3] Automatic Department, Universidad Central Marta Abreu de Las Villas, Santa Clara, Cuba

[4] Universidad Central Marta Abreu de Las Villas, Informatic Research Center, Santa Clara, Cuba

[5] Universidad de Ciencias Médicas de Villa Clara, Roberto Fleites Policlinics, Santa Clara, Cuba

In the case of COVID-19, lung imaging contributes to monitor the evolution of the disease, assisting human specialists in decision making, as one of the most commonly damaged organs is the lungs. In fact, several studies confirm visible lung abnormalities in patients with this disease [6]. However, the use of these techniques as a diagnostic method has shown low sensitivity and specificity in current radiological practice [7]. For example, in the work of [8] when using chest X-ray (CXR) imaging to detect SARS-CoV-2 pneumonia a sensitivity of 57% was achieved while in the work of [9] the reported sensitivity was 65%. These values demonstrate how difficult is for radiologists to diagnose COVID-19 using CXR directly.

So as with PCR, the study of lung radiographs could be very beneficial, to save time in the proper management of positive patients. Specifically, CXR imaging can be used to obtain relatively immediate diagnostic information. Moreover, such equipment is available in almost all medical settings, being portable and allowing the acquisition of these images quickly and at relatively low cost.

1.2 CXR imaging as a diagnostic method

Multiple studies have been published claiming the possibility of diagnosing COVID-19 from chest radiographs using machine learning techniques, with very high accuracy [10] or quantification of severity [11, 12]. Even, sensitivity values of 100% in automatic classification are reported [13, 14]. However, other studies demonstrate the lack of generalization of the models, by notably lowering the performance index when trying to classify images that do not come from the same distribution (ood) with which they were trained [15–21]. In other words, these proposed models suffer from the inherent drawbacks of low generalization capability, derived from the sparse labeled COVID-19 data [22]. Therefore, most of the proposed models are affected by shortcut learning. Thus, instead of focusing on radiographic COVID-19 findings, the models are based on features that do not relate to the pathologies they are trying to classify. A review on the effect of shortcut learning applied to COVID-19 identification appears in [23]. In that study, it was evidenced that, studies that applied explainable artificial intelligence (XAI) methods to determine the most important regions on which deep learning (DL) models are based were outside the lung region.

One of the strategies to address the problem of shortcut learning and lack of generalization of models is to use a larger number of images in training. However, although a multitude of COVID-19 CXRs are performed daily worldwide, the available datasets remain limited due to the lack of expert labels and difficulties in sharing patient data outside the hospital due to privacy concerns. The situation is further exacerbated in the current pandemic situation, making collaboration between different hospitals in different countries difficult. As an alternative, works agree on applying lung segmentation to eliminate biases that may be contained in the images, related to textual labels [24]. In fact, in the study of [19] it was demonstrated that when using the whole image, the textual labels turned out to be among the regions that most influenced decision making.

On the other hand, lung segmentation does not guarantee that the models actually focus on the right regions. Even after applying histogram equalization to homogenize gray level intensities in CXR images, there is a high risk that the algorithms will rely on spurious features that are not related to the disease [15].

1.3 Validation of models from external assessment sets

Studies agree that one of the strategies to eliminate the possible biases contained in the CXR image sets is to use images containing only the segmented lung region [25]. In this way, the textual labels that have been so determinant to exercise classification, are eliminated since, in reality, they are not significant to assign class membership. Likewise, applying image processing techniques to homogenize the images as a stage prior to classification can alleviate the bias that occurs in the images related to their acquisition.

The validity of the models obtained so far has been challenged by recent work in which validation is performed with ood images [15–18, 20, 21]. In these cases, the performance indices when using the ood set have decreased considerably when compared to the results obtained using sets coming from the same distribution (iid). These studies, although few in number, have begun to make researchers aware of the need to use external validation sets to determine the validity of their models. In fact, one of the indispensable requirements for establishing some kind of COVID-19 identification system in any clinical setting is that it possesses stable generalization capability over unseen images.

However, special attention must be paid to conform the set of ood images, since due to the abundant naming ambiguity and overlap between current sets, a set that is not really ood could be end up being used to evaluate the models. This occurred in the work of [26], where there was an overlap in the images used in training and the external validation set. This statement can be arrived at by carefully analyzing the composition of the sets, taking as a reference the works [27], where the provenance of many of the current sets used in the scientific literature to identify COVID-19 through CXR is described. In addition, the pneumonia/normal class dataset was based on a pediatric dataset (patient age 1–5 years) as is the case in other works such as [28, 29]. In contrast, the mean age of the COVID-19 class was older than 40 years. Looking at the pneumonia image, it is evident that the sizes of the rib cage and thoracic structures in the pneumonia dataset are different from the COVID-19 cases due to the age difference. It is likely that these studies used age-related features to differentiate pneumonia/normal cases and COVID-19 cases as an indicator of age, rather than pathologic diagnosis.

The aim of this work is to evaluate the effectiveness of a new model, to identify COVID-19 using CXR, following the recommendations collected in the scientific literature. For this purpose, the possible biases existing in the image sets are eliminated by segmenting the lung region. In addition, a new patch training methodology for the models, never reported before, is proposed. In this way, the problem of using the same source in the training and test sets for model evaluation is analyzed. Also, a new set of balanced CXR images for different classes, of common origin for research purposes, belonging to native patients is released.

1.4 Contributions of the paper

- A new methodology is proposed to address the problem of automatic COVID-19 classification from CXR images. In doing so, many of the claims of other research have been taken into account. For example, segmentation of the lung region is applied to eliminate any bias of the image set related to the textual labels of the radiographs. Also, a lung field partitioning structure is proposed to provide the network with the ability to work with higher resolution images.

- A rigorous evaluation of the proposed models from external image sets is performed to determine the generalization power of the proposed model.

- A new set of images containing radiographs from COVID-19 positive patients is released. Thus, this set is made available to the international scientific community for research purposes. Thus, it contributes to the creation of a set of images of diverse origin that makes the proposed algorithms more robust, in addition to improving the validation process of the existing ones.

2 Materials and methods

2.1 Image set description

Three sets of images were used in this work. The first one, which will be called Data 1, contains a total of 1613 images. The distribution by classes of the three data bases is shown in Table 1. In the case of COVID-19 positive images, 276 images were obtained from COVIDx [28], the rest of the positive images belong to autochthonous hospitals (national repository). These last images are private for the moment. In the case of the images with pneumonia and normal cases, they belong to COVIDx.

Table 1 Description of the image sets used

	COVID-19	Pneumonia	Normal	Total
Data 1	733	633	247	1613
Data 2	548	578	914	2040
Data 3	113	113	113	339

The second set of images (Data 2, part of which is released in https://github.com/jportal93/De-COVID-Rx) consists of 2040 images distributed in the three classes. These come from the same analog X-ray machine, belongs to Hospital A (autochthonous). A detailed description of Data 2 appears below. The set of images (Data 3, also released) contains images from a portable digital X-ray equipment, divided into COVID-19 113 (Hospital A), Pneumonias 113 (Hospital B, also autochthonous) and Normal 113 (Hospital A and B).

Data 2 was obtained at both anteroposterior (AP) and posteroanterior (PA) configurations, using conventional (analog) CXR equipment, to obtain images on acetate plate. Digitization was performed using a Nikon-D80 camera with 18—135 mm zoom lens, set at f/3.5–5.6G with focal length between 20 and 38 mm, AF-S focus mode. The shutter speed varied according to the characteristics of the images from 1/40 s to 1/1.3 s. In all cases sensitivity ISO 100, color space black-and-white was used. The images were saved in raw format (NEF type) with 12 bpp and size 3872 × 2592 pixels. No flash illumination was used, only the light provided by the negatoscope.

The first pre-processing was to convert the images from NEF to TIF format with 8 bpp and cropping to a size ranging from 1524 × 1450 to 1095 × 1320 pixels, using the computational tool Nikon ViewNX-i, ver. 1.4.2 64bit. The diversity of sizes is due to the fact that the acetate plates were of various dimensions and that all the part of the image that is not the thorax is discarded. No adjustment (restoration, or enhancement) was performed at this stage.

The composition of Data 2 and Data 3 is shown in Table 2.

2.2 Annotation and patching of the image set

One of the shortcomings of DL-based methods to identify COVID-19 from CXR is the low resolution of the images used to train the networks. In most of the works when using pre-trained networks, 229 × 229 pixels images are used. This image resizing process results in the loss of important features in the images. An alternative to overcome this drawback is the patch partitioning what is proposed in this work. This alternative allows the network to be trained with higher resolution images, so that the network can learn patterns that are better related to the disease it is trying to diagnose.

A total of 1613 images with diverse origin were annotated. These images belong to the Data 1 set. A total of 733

Table 2 Characteristics of subjects included in Data 2 and Data 3

Data 2 composition								
Variable	**COVID-19**		**Pneumonia**		**Normal**		**Total**	
Sex	**Cases**	**%**	**Cases**	**%**	**Cases**	**%**	**Cases**	**%**
Male	276	50.4	285	49.3	452	49.5	1013	49.7
Female	272	49.6	293	50.7	462	50.5	1027	50.3
Total	548	100	578	100	914	100	2040	100
Age								
Less than 20	89	16.2	21	3.6	186	20.4	296	14.5
20–39	170	31	109	18.9	302	33	581	28.4
40–59	174	31.8	222	38.4	359	39.3	755	37
60+	115	21	226	39.1	67	7.3	408	20
Total	548	100	578	100	914	100	2040	100
Data 3 composition								
Variable	**COVID-19**		**Pneumonia**		**Normal**		**Total**	
Sex	**Cases**	**%**	**Cases**	**%**	**Sex**	**Cases**	**%**	**Cases**
Male	71	62.8	63	55.8	49	43.4	183	54
Female	42	37.2	50	44.2	64	56.6	156	46
Total	113	100	113	100	113	100	339	100
Age								
Less than 20	12	10.7	2	1.8	18	15.9	32	9.5
20–39	34	30	26	23	32	28.3	92	27.1
40–59	41	36.3	35	31	46	40.7	122	36
60+	26	23	50	44.2	17	15.1	93	27.4
Total	113	100	113	100	113	100	339	100

COVID-19 positive images, 633 images with pneumonias due to other causes and 247 normal images were annotated. This task was performed by 3 expert radiologists independently, with more than 10 years in the specialty. In the case of doubtful images, consensus was sought among all of them. The criteria for annotation of the image data base (DB) are shown in Table 3 and are taken from the based on common international criteria. To annotate the DB, the specialists observe the image based on a division of the lower, middle and upper lung fields for the left and right lung. In this way, the CXR image is divided into six regions as shown in Fig. 1. For each of the regions, the presence or absence of some of the criteria reported in Table 3 is determined. These annotations constituted the starting point for a division into patches of this set of images as explained below.

The first step performed to divide the CXR image into regions was to segment the lung region. A pre-trained convolution network based on the U-Net architecture [30] was used. This network was trained with CXR images and masks segmented by hand by human specialists, belonging to the Montgomery [31] and JSRT [32] ensembles. This model reports a Dice similarity coefficient for the aforementioned

Table 3 Criteria for database annotation by radiologists

Typical findings of COVID-19	Peripherally distributed leaflet opacities Multifocal veil opacities Veil opacities associated with areas of segmental consolidation Interstitial pattern Centrally distributed opacities
Unlikely lesion of viral pneumonia	Pleural effusion
Typical findings of non-viral pneumonia	Lobar alveolar consolidation with bronchogram Segmental alveolar consolidation with bronchogram Lobar alveolar consolidation without bronchogram

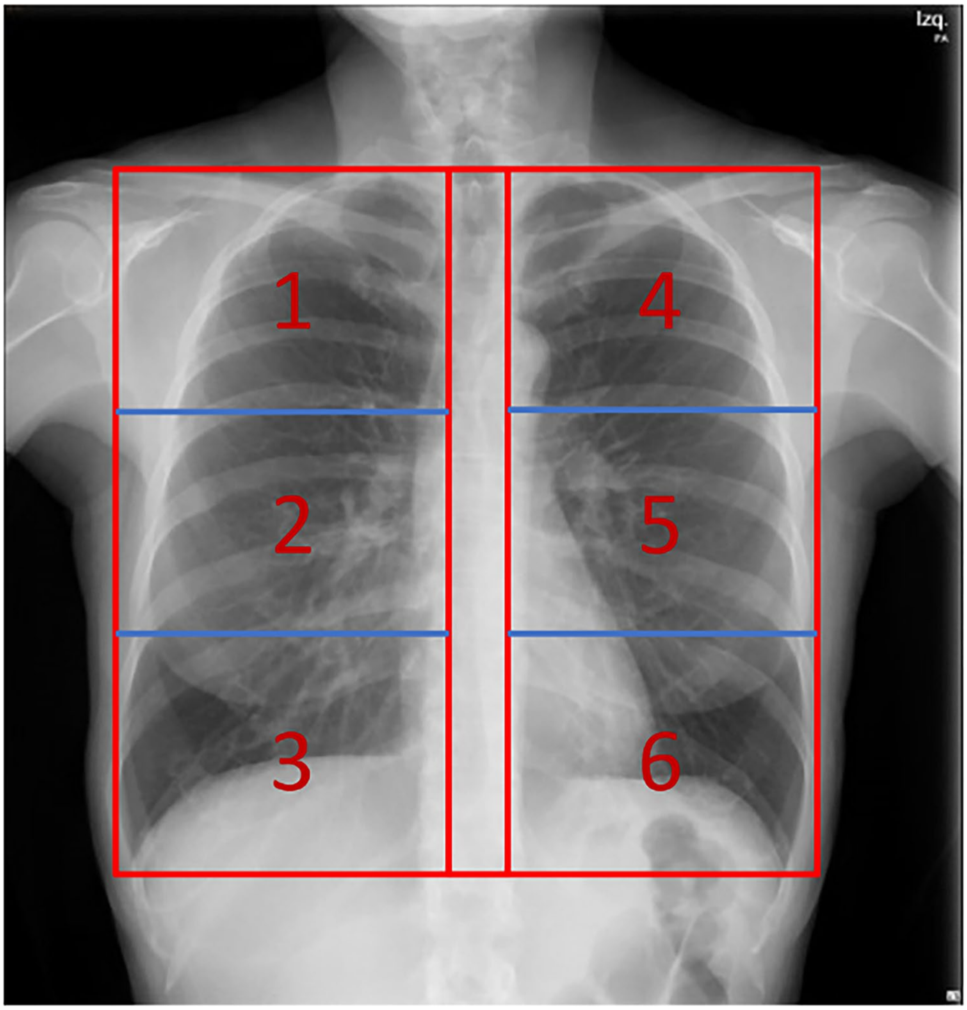


Fig. 1 Division into six regions of a CXR based on lung fields, which is the starting point for manual annotation by radiologists

image bases of 0.985 and 0.972 respectively. In order to improve the segmentation results obtained by the network, has been used as a post-processing criterion, leaving only the two largest morphological structures detected (Lungs). Then, the Convex Hull operation is performed on each detected lung to eliminate the irregularities in the edges obtained from the segmentation mask (Fig. 2). In this way, both physiological structures other than lungs and textual labels are removed from the image. In addition, the holes that may remain in the lung region due to failures in the segmentation algorithm are filled from morphological operations. This is intended to decrease the probability of learning from shortcuts or features outside the lung region [25].

Images of difficult segmentation, where only a single region was obtained, were eliminated as they were considered a bias for training. Next, a dilation of the segmentation mask is performed, using a disk-shaped structuring element of dimensions proportional to the image. In our case, 2% of the image was used as the radius of the structuring element. This operation aims to increase the obtained segmentation mask and to provide the network with more information about the lungs, since in the case of COVID-19 lesions it is known that they can be lodged in the border regions of both lungs.

Then, the image is divided into a set of patches, where 1 patch represents the two lungs, 2 patches equals each lung separately, 4 patches represent each lung divided into two, and 6 patches equals each lung divided into three. Starting from a binary image of 1 patch that includes the two silhouettes of the lungs, the bounding box of each one is determined independently. In this way, the 2-patch representation is obtained. The 4-patch images are obtained by dividing in half each of the connected components found in the previous step. The same is done to obtain the 3-patch images, this time dividing into 3 regions instead of 2. Figure 2 represents this process.

2.3 Description of phase a of the proposed system

In the work of [33] a patch-based classification is proposed. In contrast to this research, a model based on a random selection of patches belonging to the lung region is trained

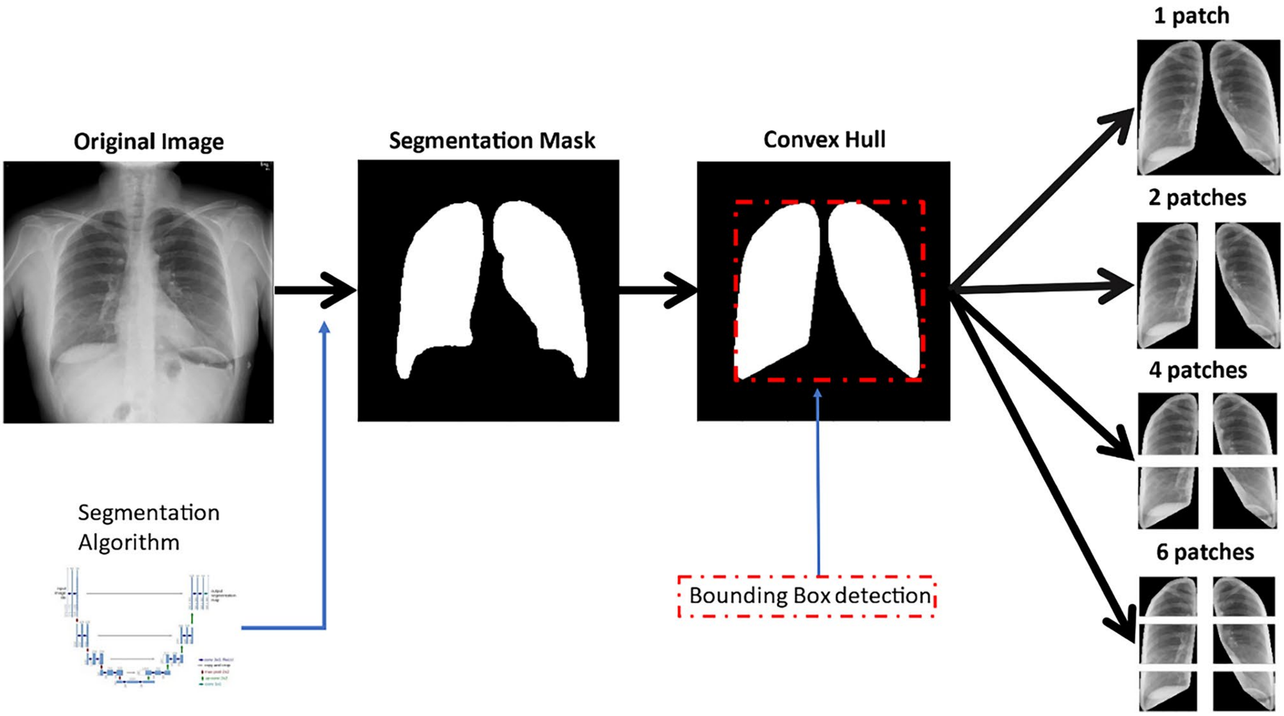


Fig. 2 Segmentation process and division into patches

in that study. Such methodology assumes that each selected patch presents manifestations of COVID-19. Assuming that in each patch selected from a COVID-19 positive image or from another type of pneumonia there is some manifestation of the disease does not hold true in all images, since the lesions appear localized in specific regions and are not always distributed throughout the image. Therefore, it is possible to select patches that do not belong to the class which is trying to classify because they do not contain the radiographic manifestation which are trying to classify. To alleviate this difficulty in this study, manual annotation was performed by expert radiologists in a fixed division by patch regions as explained in the previous section.

The first task performed was the organization in folders of the patches according to the classification received by the radiologists. For this purpose, each image is assigned to a class (normal, pneumonia or COVID-19). Thus, for example, an image with COVID-19 will be placed in the COVID-19 folder of a patch, but in the divisions of 2, 4 and 6 patches, there is the possibility of dividing the images depending on the region they occupy in the lung, according to the number of patches into which they were divided. In other words, if Fig. 1 is taken as a reference, in this new representation 6 divisions will be generated, corresponding to each of the numbers shown in the figure.

In phase A of the system, the training of the models is carried out exclusively from the radiologists' annotations. This will be called Initial Models (IM). Therefore, as many models are obtained as patches are generated per division. Figure 3 shows the process described above.

2.4 Description of Phase B of the proposed system

The manual annotation work performed by radiologists is a tedious and time consuming process. In order to increase the number of images for model training, Phase B is proposed. In this new stage the Initial Models are used to automatically classify a new set of images (Data 2). With this procedure, a division into patches with the same hierarchical structure as that obtained in Phase A is obtained. That is, the manifestations of COVID-19 and pneumonia are not seen in the whole image region. Therefore, the following heuristic was used to perform the division. The images divided into 2, 4 and 6 patches where there was no match between the model and the class to which the whole image belonged, the label prevailed. That is, in case the model classified one of the patches as pneumonia when the image was actually COVID-19, this patch would be assigned to the COVID-19 folder (the label that the image had) instead of the model prediction. This is illustrated in Fig. 4.

2.5 Description of Phase C of the proposed system

In Phase C, the models are re-trained, using as training set the one obtained in Phase B (Data 1 + Data 2), using transfer

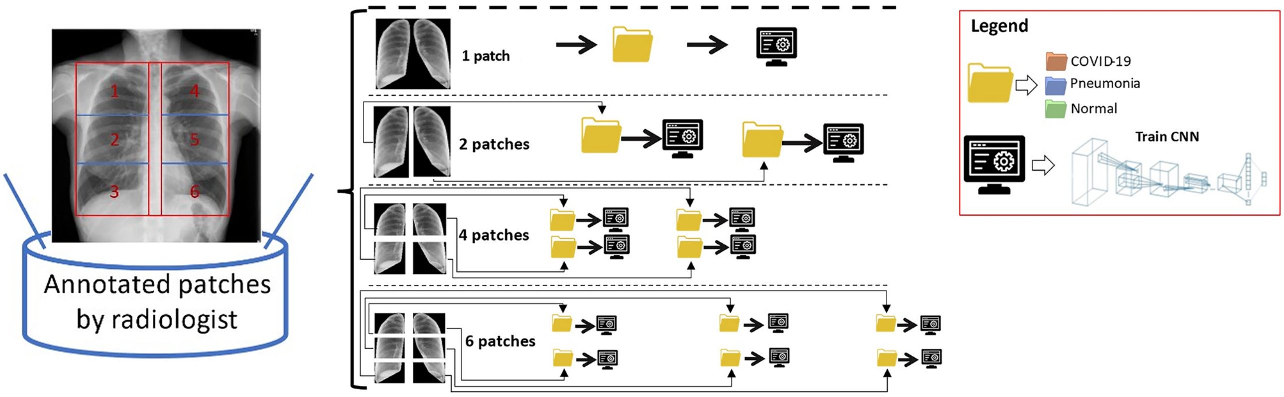


Fig. 3 Phase A of the system, division into patches of the images annotated by the radiologists and obtaining Initial Models from the new partitions generated

learning from Phase A. In this stage, what was learned in the previous stage is used with the objective of improving the performance of the models for the four classification strategies. The 90% percent of the data was used to train the system and 10% to validate it.

2.6 Description of the architecture and parameters

Pytorch was used as a framework for training, testing and evaluation, starting from the pre-trained network Resnet34 and then using transfer learning. In the first stage, a training with 400 epochs is done in order to adjust the weights of the layers of the pre-trained network to the new problem. This value was set because the training tests showed that the parameters remained stable and maintained convergence from 300 epochs.

To mitigate the problem of low resolution of pre-trained networks, an image size of 512×512 was used, as well as image patching. The network output has three classes, COVID-19, Pneumonia and Normal. The Adam optimizer [34] was used, which is an extension of the down gradient to adjust the weights during training and minimize losses. Cross entropy is used as the loss function. As the losses decrease in training, the obtained learning models are saved up to the point, where it is seen that the next epoch increases the losses and therefore training continues until the stopping requirements are met. In phase A of the study, training is performed from zero. Training stops after 300 epochs or when the losses are less than 10^{-3}. In the case of Phase C, the transfer learning technique was applied to obtain the retrained models using 35 epochs. The learning rate used was 0.01 and the batch size was 32. It should be noted that initially the hyperparameters were established following the recommendations of

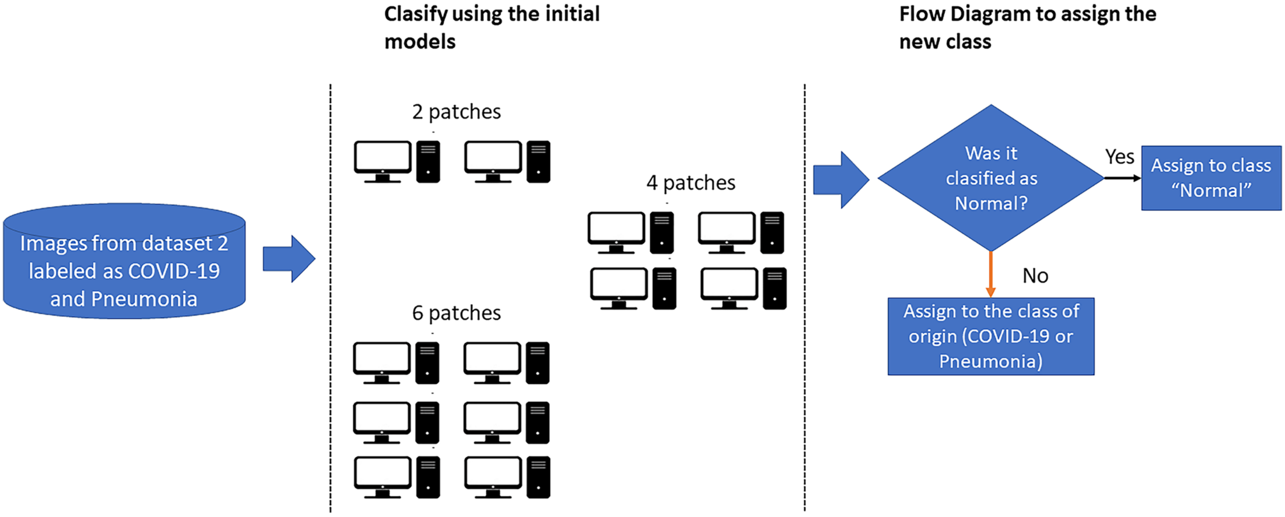

Clasify using the initial models
Flow Diagram to assign the new class
Images from dataset 2 labeled as COVID-19 and Pneumonia
2 patches
4 patches
6 patches
Was it clasified as Normal?
Yes
Assign to class "Normal"
No
Assign to the class of origin (COVID-19 or Pneumonia)

Fig. 4 Scheme followed to increase the set of annotated images

[35] and [36]. Later, they were tuned up from multiple tests with various data sets taken from data 1 and 2.

During all the trainings, the data augmentation technique was applied, starting from generating more patches for each original patch, based on left and right rotations in angles up to 10 degrees, zoom of 1.05% and image brightness variations up to 0.8%. In the case of the 2-, 4- and 6-patch variants, the product of the probabilities of each of the independent models was used as the fusion strategy to output the classification of an image.

2.7 Evaluation of the proposed models

The first evaluation of the models is performed in Phase A, this is done using 10% of Data 1 (internal validation), and Data 3 (external test). Then in the next stage the previous results are improved, so Phase B and C are introduced. As part of the Phase C evaluation, two sets were used newly. The first one contains images belonging to 10% of the union of Data 1 and Data 2 sets (internal validation). Finally, the external evaluation is performed using Data 3 set.

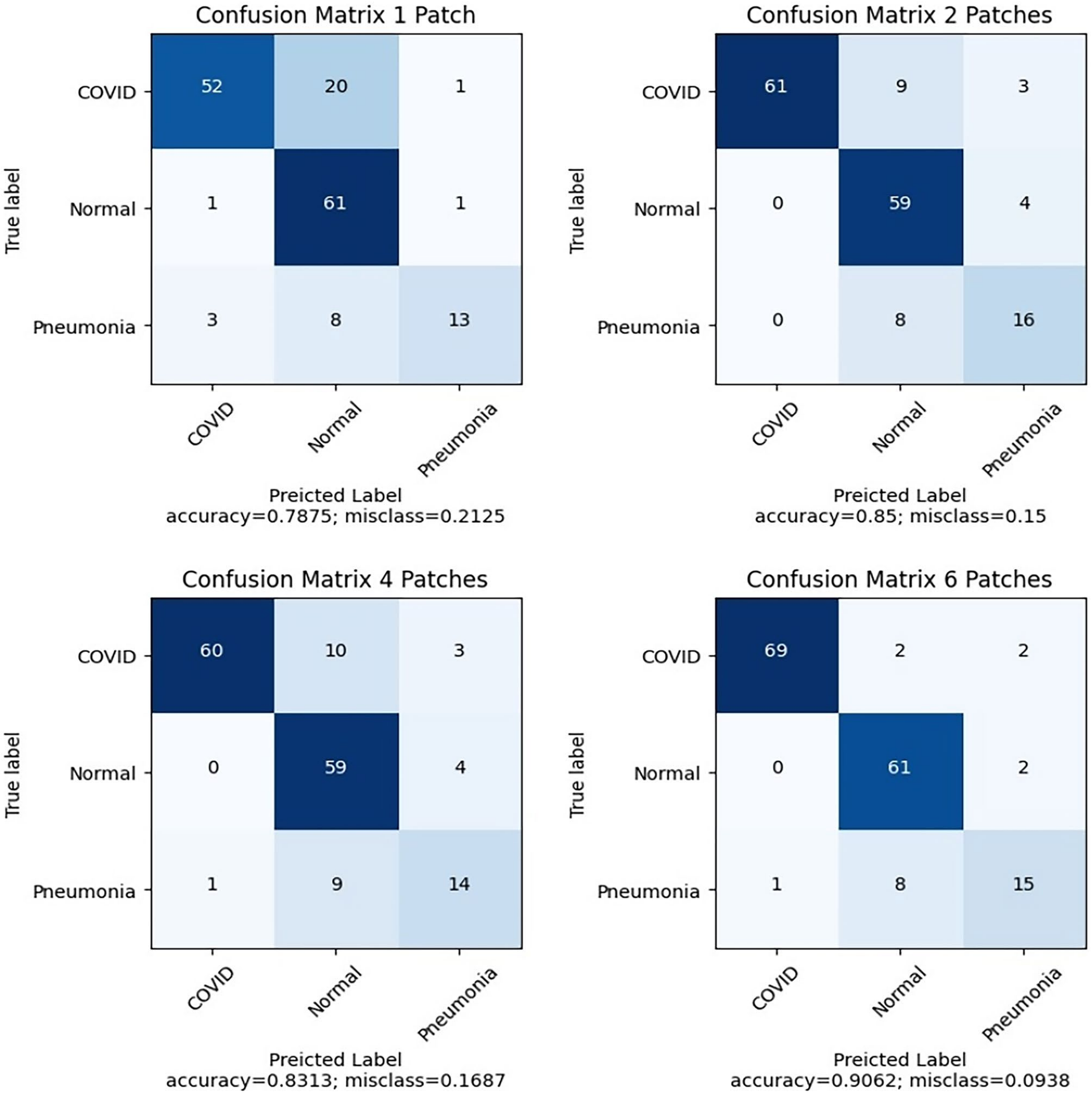


Fig. 5 Confusion matrix for the different variants of patch splitting in Phase A

Table 4 Performance indices obtained for the different patch splitting evaluation strategies in the Phase A patching evaluation

Classification strategy	Metrics	Class			Macro Metrics
		COVID-19	**Normal**	**Pneumonia**	
patch	Precision	0.929	0.685	0.867	0.827
	Recall	0.712	0.968	0.542	0.741
	F1score	0.806	0.803	0.667	0.759
2 patches	Precision	1.000	0.859	0.594	0.818
	Recall	0.877	0.873	0.792	0.847
	F1score	0.934	0.866	0.679	0.826
4 patches	Precision	0.984	0.756	0.667	0.802
	Recall	0.822	0.937	0.583	0.781
	F1score	0.896	0.837	0.622	0.785
6 patches	Precision	0.986	0.859	0.790	0.878
	Recall	0.945	0.968	0.625	0.846
	F1score	0.965	0.910	0.698	0.858

This strategy followed for the evaluation of the models proposed in each of the phases allows to determine the generalization power of the system. Likewise, by making use of an external evaluation set, one of the claims made by the scientific community on the validity of the proposed models is addressed.

To determine the performance of the models obtained, the Precision, Recall (sensitivity) and F1score metrics per class and their averages have been used, which are called macro metrics (MacroPrecision, MacroRecall and MarcoF1score). All the metrics used are defined below in Eqs. 1, 2, 3, 4, 5 and 6. These performance measures are obtained from the confusion matrix. Their use allows the performance of the models to be evaluated by class and globally.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F1_{score} = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{3}$$

$$MacroPrecision = \frac{1}{3}\sum\nolimits_{c=1}^{3} Precision_c \tag{4}$$

$$MacroRecall = \frac{1}{3}\sum\nolimits_{c=1}^{3} Recall_c \tag{5}$$

$$MacroF1_{score} = \frac{1}{3}\sum\nolimits_{c=1}^{3} F1_{scorec} \tag{6}$$

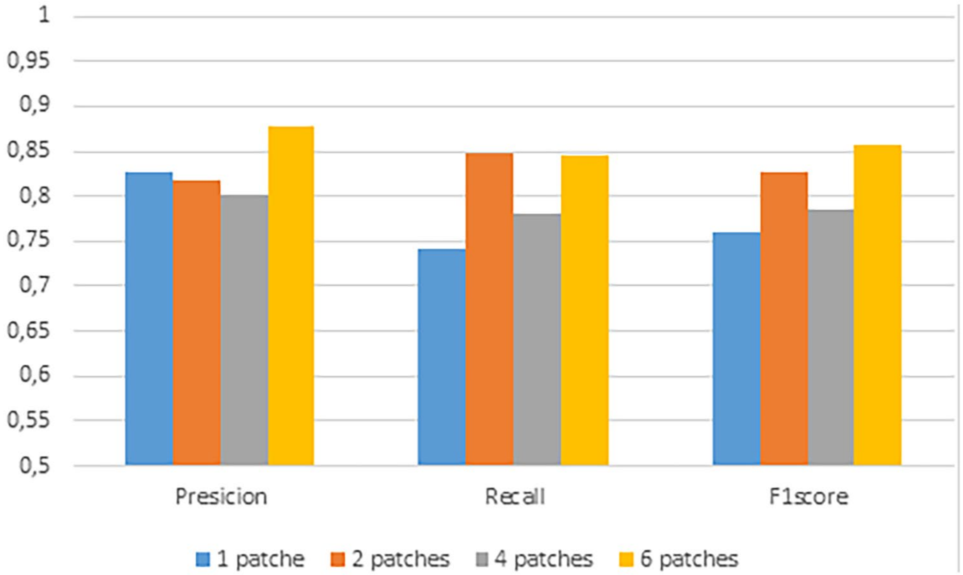
where TP are the true positives, FP are the false positive and FN are the false negatives.

3 Results

3.1 Evaluation of Phase A

The first stage of the system is performed with the images labeled by the radiologists. For this purpose, the training of the models belonging to each of the division variants is performed. Thus, four different systems are obtained for image classification (1, 2, 4 and 6 patches). In the case of variants 2, 4 and 6 patches contain as many models as patch splits are performed.

Fig. 6 Macrometrics for the different variants of the Phase A patch division



Then, in order to issue the classification of the complete image, the multiplication of the probabilities of memberships obtained by each model that composes the scheme is used as an aggregation function.

Figure 5 shows the confusion matrices obtained at this stage for the different patch division strategies. Also, Table 4 shows the performance measures obtained from the aforementioned confusion matrices. It is observed that, the best performance indices are obtained for the 6-patch variant, followed by the 2-patch variant as presented in Fig. 6.

3.2 Evaluation of Phase A using the external set

This section presents the evaluation of the MI on the external data set. Figure 7 presents the confusion matrices obtained for each of the patch partitioning strategies. As can be seen in Table 5, the models created do not present generalization capacity. That is, the models obtained are unable to maintain the results presented in the previous stage, this time in the presence of a set that does not come from any of the sources used in the training. It is observed that in the case of variants

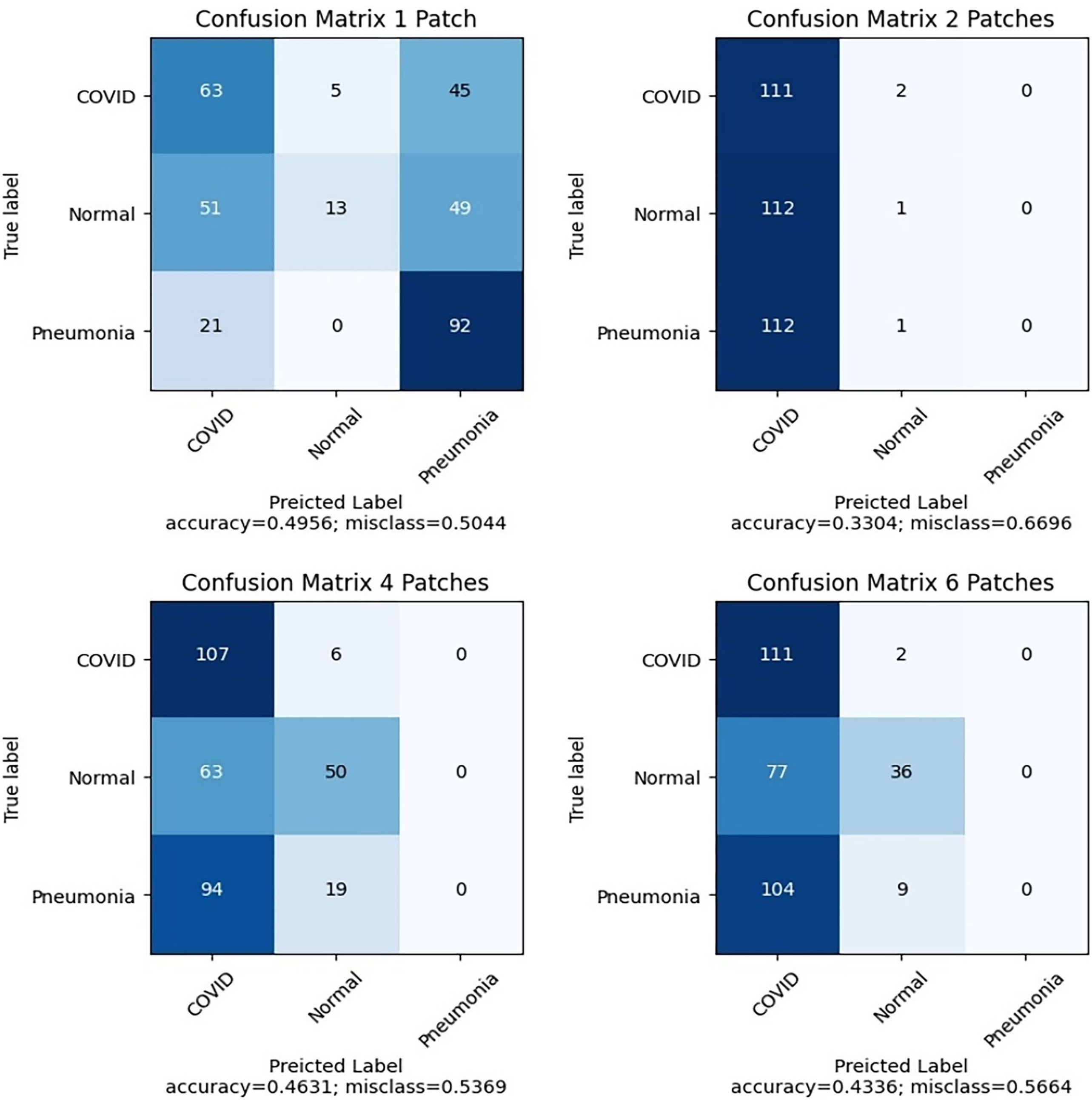


Fig. 7 Confusion matrix for the different patch splitting variants in Phase A for the external evaluation set

Table 5 Performance indices obtained for the different patch splitting evaluation strategies using the external evaluation set in Phase A

Classification strategy	Metrics	Classes			Macro Metrics
		COVID-19	**Normal**	**Pneumonia**	
1 patch	Precision	0.467	0.722	0.495	0.561
	Recall	0.558	0.115	0.814	0.496
	F1score	0.508	0.198	0.615	0.441
2 patches	Precision	0.331	0.250	<0.001	0.194
	Recall	0.982	0.009	<0.001	0.330
	F1score	0.496	0.017	<0.001	0.171
				<0.001	
4 patches	Precision	0.405	0.667	<0.001	0.357
	Recall	0.947	0.442	<0.001	0.463
	F1score	0.568	0.532	<0.001	0.367
				<0.001	
6 patches	Precision	0.380	0.766	<0.001	0.382
	Recall	0.982	0.319	<0.001	0.434
	F1score	0.548	0.450	<0.001	0.332

2, 4 and 6 patches the systems are unable to detect the class "Pneumonia" (note that the performance measures are 0). On the other hand, variant 1 patch shows the best results for determining the "Pneumonia" class, to the detriment of the performance indices for the "Normal" class. That is, the behavior of these models is close to the random classification and far from the values obtained in the previous stage, showing a clear overfitting effect by the models. These results show that the models obtained fail to perform the correct classification of this new set of images.

3.3 Evaluation of Phase C

This section shows the results obtained in the Phase C evaluation stage. The evaluation set used was 10% of the images belonging to this stage. Figure 8 presents the confusion matrices obtained when evaluating the models with the four splitting variants. Table 6 presents the metrics calculated from the confusion matrices presented above. Figure 9 presents the performance measures for the different classification variants. It can be seen that again the 6-patch variant presents slightly higher values than the rest of the strategies.

3.4 Evaluation of Phase C using the external set

This section presents the results obtained from the models retrained on the external evaluation (Data 3). Figure 10 shows the confusion matrix obtained for each of the patch splitting strategies. It can be seen that the best results appear again for the 6-patch splitting variant. On this occasion, it is observed that the results obtained are maintained compared to the internal validation stage. This shows that the proposed models have generalization capacity. Thus, this type of models can help human specialists in the timely identification of COVID-19 cases with visible manifestations in the lung region and apply a timely management to these patients.

4 Discussion

This work has highlighted the importance of not committing the same mistakes widely diffused in publications from 2020 and 2021 on the subject of COVID 19/pneumonia classification based on CNN. It is related to the fact that if the AI systems are trained and validated only from complete images taken from the various international or autochthonous databases, they are capable of performing a classification based on spurious characteristics, since they perform learning by shortcuts [19–23, 25]. This can be evidenced in two ways: Comparing the performance of the system against external data (of a different origin than those used during training/validation) and verifying that CNN only classifies based on features contained in the lung region. Systems without power of generalization that have based their learning on biases, will show much lower performance metrics in an external evaluation compared to those achieved during validation with data of the same origin as the training data.

The system proposed in this research has shown that if only Phase A is run (which shares many points in common – particularly training based on one patch – with various other systems published in the scientific literature, see [10–14]), when evaluated with external dataset it shows a very poor performance, which proves the presence of shortcut learning.

To overcome these obstacles, phases B and C have been implemented, not only to take advantage of using only useful information contained within the lung region and a higher number of training images, but also to carry out a correct training/validation/evaluation strategy, where images with the same origin are not overlapped at all stages of system deployment. Once this is done, similar values are achieved in the metrics during the evaluation with external data with respect to what was achieved during the validation, which proves the power of generalization achieved.

Another interesting aspect is that CXR has not proven to be efficient enough to achieve correct classification rates, close to 100% in COVID 19/pneumonia classification, as suggested by the systems that present shortcut learning reported in the literature [10–14].

During the year 2022, an enormous diversity of models tested with different databases has continued to be presented,

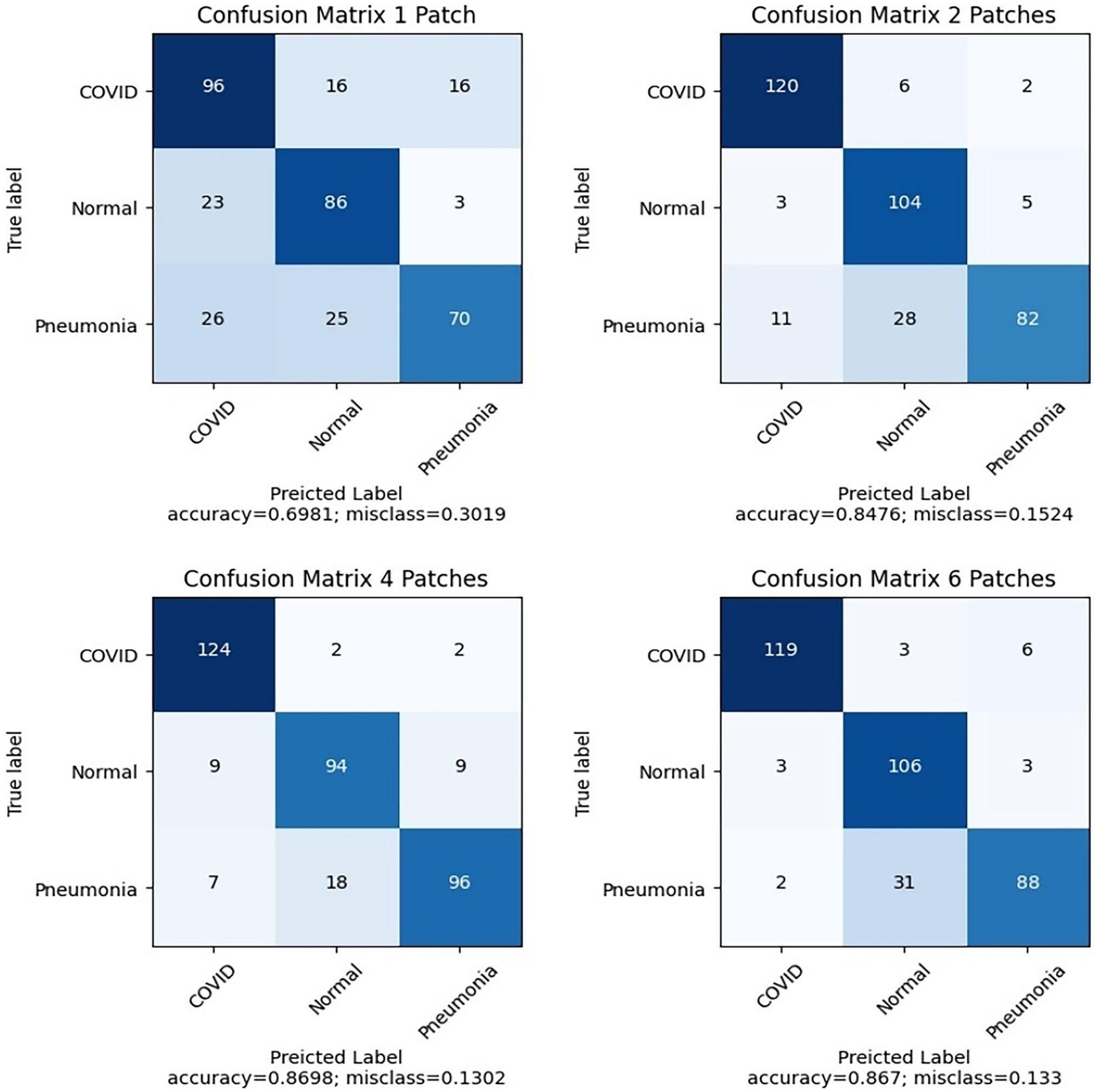


Fig. 8 Confusion matrix for the different Phase C patching variants for 10% of the internal validation data

reporting precisions between 84 and 95% for various databases, but without reporting the behavior with an external test [37]. However, more researchers have joined in recognizing the existence of biases and verifying that most of the models are not robust when tested with a source data different from the training/validation data. For example, in [38] has shown that when trained and tested on a single dataset, the performance of all CNNs is relatively poor (precision: 0.65–0.72, recall: 0.59–0.71), but remains relatively constant during external testing (precision: 0.58–0.82, recall: 0.57–0.72). In contrast, when internally trained and evaluated on combinatorial data sets, all CNNs perform well (precision: 0.94–1.00, recall: 0.77–1.00). However, when tested on the crossover data set, the results drop substantially (precision: 0.10–0.61, recall: 0.04–0.80). In this sense, the present work presents superior results to those reported in this review and using models that already have generalization power and lower level of bias.

Table 6 Performance indices obtained for the patch splitting strategies using 10% of the data for the evaluation

Classification strategy	Metric	Classes			Macro Metrics
		COVID-19	**Normal**	**Pneumonia**	
1 patch	Precision	0.662	0.677	0.787	0.709
	Recall	0.750	0.7679	0.579	0.699
	F1score	0.703	0.720	0.667	0.697
2 patches	Precision	0.896	0.754	0.921	0.857
	Recall	0.938	0.929	0.678	0.848
	F1score	0.916	0.832	0.781	0.843
4 patches	Precision	0.886	0.825	0.897	0.869
	Recall	0.969	0.839	0.793	0.867
	F1score	0.925	0.832	0.842	0.866
6 patches	Precision	0.960	0.757	0.907	0.875
	Recall	0.930	0.946	0.727	0.868
	F1score	0.944	0.841	0.807	0.864

On the other hand, it is necessary to test whether CT images are more efficient in this task by serving as input to CNNs. Normally, CT images have a higher sensitivity as a method of diagnosis and follow-up compared to chest X-ray. Cases of COVID-19 with lesions visible on CT but not visible on CXR have been described in [39]. In fact, one of the main findings on CT images of COVID-19 patients is ground-glass opacities in the peripheral regions of the lower lobes, which are sometimes not visible on CXR images. From the scientific literature it can be seen that for CT the studies also differ in terms of the metrics used, the architectures, the models, and the databases used to train/validate the networks, so comparison between them is not easy. In the work [40] for example, which focuses on the importance of CT image segmentation to achieve better results, the authors report a precision of 87% in the classification of 3 classes (Covid/Normal/Pneumonia), which is equivalent to that achieved in the present work, with a different database. In the study itself, the precision values achieved by the previous models for CT are between 73 and 85%. In the work [41] modifications are introduced on the usual architecture of ResNet. Data from COVIDx-CT (CT equivalent to COVIDx) are used for training and validation, obtaining an accuracy of 99.7%. However, they do not report external test results. In the study [42], where results are also reported with several CNNs but where it is stated that they have carried out an external test, the precision values range are between 59 and 97% with the COVIDx-CT data, depending on the model used.

Finally, it should be stated that the presented system has some limitations. In new external tests carried out looking for the causes of the present classification errors, at least two biases have been identified. One relates to the correct position of the patient when taking the radiography, which presumably comes from COVIDx training data [28], where very sick patients were studied with X-rays from portable equipment in incorrect positions. The second is related to scapular opacities that are misinterpreted by the system.

The possible improvements to be introduced for the problems described above are based on taking the misclassified images and placing them in the folder with the correct label, performing a new training. This strategy is known as ¨Federal Learning¨. As the system is retrained with more images, its performance metrics improve.

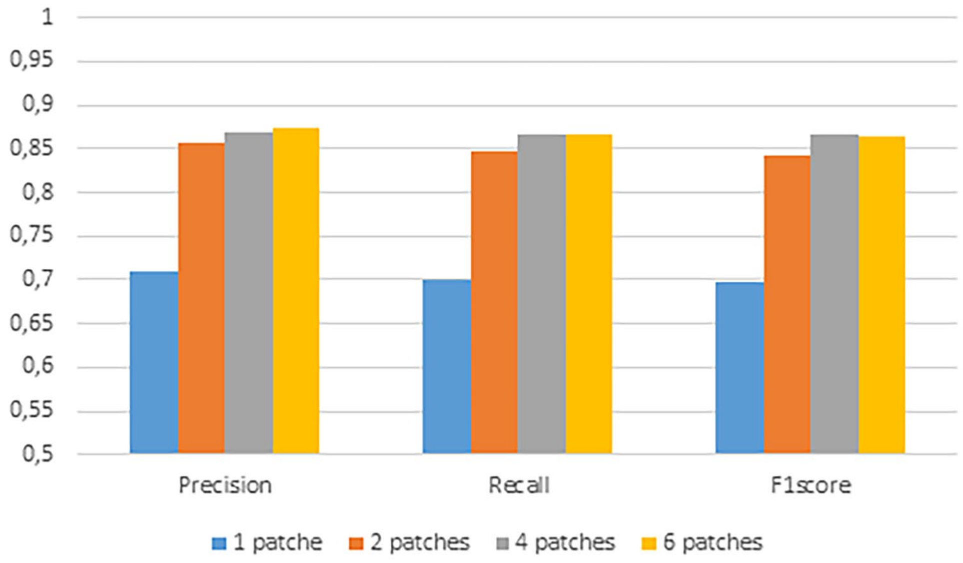


Fig. 9 Macrometrics for the different Phase C patch splitting variants using 10% of the images as evaluation

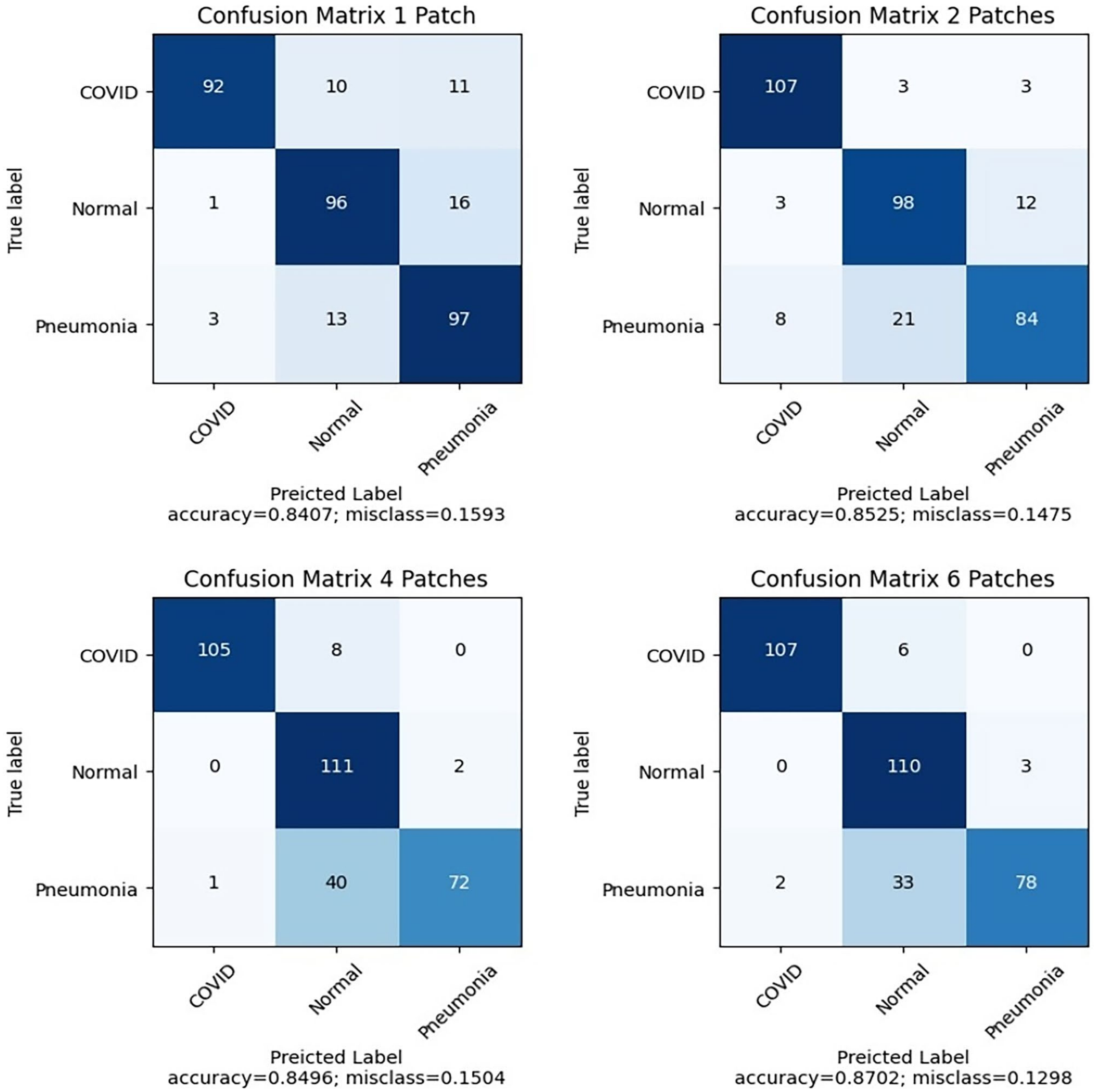


Fig. 10 Confusion matrix for the different Phase C patching variants for the external evaluation set

5 Conclusions

In this study, a model was developed to automatically classify COVID-19 from CXR images. Recommendations suggested in the literature were followed to increase the robustness of the models. For this purpose, the architecture of a resnet34 CNN with input images of size 512×512 pixels was used. In addition, the training of the networks was limited to the inner regions of the lungs, using a segmented image and a patch division. A patch partitioning step was proposed and it was shown that 6-patch partitioning yielded the best results. Specifically, this model yielded 0.887 accuracy, 0.85 recall and 0.848 F1score on the external set. The models obtained after extending the training set in stage B showed similar results in the internal and external evaluation. This endows these models with generalization power. Therefore, making them useful to be used in conjunction with human specialists in the identification of COVID-19 in clinical settings.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12553-022-00704-4.

Acknowledgements We would like to thank Professor Miguel Mendoza Reyes for his important contribution in the data preparation, as well as the Cuban Society of Imaging and the Cuban Neuroscience Center, for the contribution of criteria and data.

Funding This work has been partially funded thank to the Agence Universitaire de la Francophonie in the Caribbean and the Agency for Nuclear Energy and Advanced Technology (AENTA). All the simulations in this research were run in the HPC cluster of UCLV/Cuba, which was developed thanks to the JOINT/UGent project (CU2019JOI012A103) from VLIR-UOS.

Declarations

Ethical approval Part of the CXRs images used in this research are public and anonymous on the internet. Other images are autochthonous and they were taken from an anonymous national repository. In this case, informed consent was waived due to the retrospective character. The rest of the data is made up of anonymous in-house images, donated by two Hospitals participating in the research, once the Project was approved by the local ethics commission on May, 2020. (Both documents are attached as supplementary material).

Conflict of interest The authors declare that there is no conflict of interest.

References

1. Lai CC, Shih TP, Ko WC, Hsueh THJ, PR,. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): The epidemic and the challenges. Int J Antimicrob Agents. 2020;55:105924. https://doi.org/10.1016/j.ijantimicag.2020.105924.
2. Peeling RW, Wedderburn CJ, Garcia PJ, Boeras D, et al. Serology testing in the COVID-19 pandemic response. Lancet Infect Dis. 2020;20:e245–9. https://doi.org/10.1016/S1473-3099(20)30517-X.
3. Weissleder R, Lee H, Ko J, Pittet MJ. COVID-19 diagnostics in context, Sci Transl Med. 2020;12:eabc1931. https://doi.org/10.1126/scitranslmed.abc1931.
4. Liu R, Han H, Liu F, Zhihua L, Kailang W, Liu Y, et al. Positive rate of RT-PCR detection of SARS-CoV-2 infection in 4880 cases from one hospital in Wuhan, China, from Jan to Feb 2020. Clin Chim Acta. 2020;505:172–5. https://doi.org/10.1016/j.cca.2020.03.009.
5. Wang W, Xu Y, Gao R, Lu R, Han K, Wu G, Tan W. Detection of SARS-CoV-2 in Different Types of Clinical Specimens. JAMA. 2020;323:1843–4. https://doi.org/10.1001/jama.2020.3786.
6. Kanne JP, Little BP, Chung JH, Elicker BM, Ketai LH. Essentials for Radiologists on COVID-19: An Update—Radiology Scientific Expert Panel. Radiology. 2020;296:E113–4. https://doi.org/10.1148/radiol.2020200527.
7. Yoon SH, Kyung HL, Jin YK, Young KL, Ko H, Ki HK, et al. Chest radiographic and CT findings of the 2019 novel coronavirus disease (COVID-19): analysis of nine patients treated in Korea, Korean. J Radiol. 2020;21:494–500. https://doi.org/10.3348/kjr.2020.0132.
8. Ippolito D, Pecorelli A, Maino C, Capodaglio C, Mariani I, Giandola T, et al. Diagnostic impact of bedside chest X-ray features of 2019 novel coronavirus in the routine admission at the emergency department: case series from Lombardy region, Eur J Radiol. 2020;129:109092. https://doi.org/10.1016/j.ejrad.2020.109092.
9. Castiglioni I, Ippolito D, Interlenghi M, Monti C B, Salvatore C , Schiaffino S, et al. Machine learning applied on chest x-ray can aid in the diagnosis of COVID-19: a first experience from Lombardy, Italy, Eur Radiol Exp. 2021;5:7. https://doi.org/10.1186/s41747-020-00203-z.
10. Dong D, Tang Z, Wang S, Lixin H, Lu Y, et al. The role of imaging in the detection and management of COVID-19: a review (2021) IEEE Rev Biomed Eng. 2021;14:16–29. https://doi.org/10.1109/RBME.2020.2990959.
11. Signoroni A, Savardi M, Benini S, Adami N, Leonardi R, Gibellini P, et al. BS-Net: Learning COVID-19 pneumonia severity on a large chest X-ray dataset. Med Image Anal. 2021;71:102046. https://doi.org/10.1016/j.media.2021.102046.
12. Cohen JP, Dao L, Roth K, Morrison P, Bengio Y, Abassi AF, et al. Predicting COVID-19 pneumonia severity on chest x-ray with deep learning. Cureus. 2020;12(7):e9448. https://doi.org/10.7759/cureus.9448.
13. Nayak SR, Nayak DR, Sinha U, Arora V, Pachori RB. Application of deep learning techniques for detection of COVID-19 cases using chest X-ray images: A comprehensive study. Biomed Signal Process Control. 2021;64.
14. Ucar F, Korkmaz D. COVIDiagnosis-Net: Deep Bayes-SqueezeNet based diagnosis of the coronavirus disease 2019 (COVID-19) from X-ray images, Med. Hypotheses. 2020;140:109761. https://doi.org/10.1016/j.mehy.2020.109761.
15. Tartaglione E, Barbano CA, 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. 2029;17:6933. https://doi.org/10.3390/ijerph17186933.
16. Tabik S, Gómez-Ríos A, Martín-Rodríguez JL, Sevillano-García I, Rey-Area M, Charte D, et al. COVIDGR Dataset and COVID-SDNet Methodology for Predicting COVID-19 Based on Chest X-Ray Images. IEEE J Biomed Health Inform. 2020;24:3595–605. https://doi.org/10.1109/JBHI.2020.3037127.
17. Ahmed KB, Goldgof GM, Paul R, Goldgof DB, Hall LO. Discovery of a Generalization Gap of Convolutional Neural Networks on COVID-19 X-Rays Classification. IEEE Access. 2021;9:72970–9. https://doi.org/10.1109/ACCESS.2021.3079716.
18. Yeh CF, Hsien TC, Wei A, Chen HM, Kuo PC, Liu KC, et al. A Cascaded Learning Strategy for Robust COVID-19 Pneumonia Chest X-Ray Screening, ArXiv200412786 Cs Eess. 2020. http://arxiv.org/abs/2004.12786. Accessed 14 Aug 2020.
19. DeGrave AJ, Janizek JD, Lee SI. AI for radiographic COVID-19 detection selects shortcuts over signal. Nat Mach Intell. 2021;3:610–9. https://doi.org/10.1038/s42256-021-00338-7.
20. Teixeira LO, Pereira RM, Bertolini D, Oliveira LS, Nanni L, Cavalcanti DC, et al. Impact of Lung Segmentation on the Diagnosis and Explanation of COVID-19 in Chest X-ray Images, Sensors. 2021;21:7116, https://doi.org/10.3390/s21217116.
21. Dhont J, Wolfs C, Verhaegen F. Automatic coronavirus disease 2019 diagnosis based on chest radiography and deep learning – Success story or dataset bias. Med Phys. 2022;49:978–987. https://doi.org/10.1002/mp.15419.
22. Roberts M, Driggs D, Thorpe M, Gilbey J, Yeung M, Ursprung S, 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. 2021;3:199–217. https://doi.org/10.1038/s42256-021-00307-0.
23. López-Cabrera JD, Orozco-Morales R, Portal-Díaz JA, Lovelle-Enríquez O, Pérez-Díaz M. Current limitations to identify covid-19 using artificial intelligence with chest x-ray imaging (part ii). The shortcut learning problem. Health Technol. 2021;11:1331–45. https://doi.org/10.1007/s12553-021-00609-8.

24. 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.
25. López-Cabrera JD, Orozco-Morales R, Portal-Díaz JA, Lovelle-Enríquez O, Pérez-Díaz M. Current limitations to identify COVID-19 using artificial intelligence with chest X-ray imaging. Health Technol (Berl). 2021;11(2):411–24. https://doi.org/10.1007/s12553-021-00520-2.
26. Ahrabi S, Scarpiniti M, Baccarelli E, y Momenzadeh A. An Accuracy vs. Complexity Comparison of Deep Learning Architectures for the Detection of COVID-19 Disease. Computation. 2021;9:3 https://doi.org/10.3390/computation9010003
27. Garcia Santa Cruz B, Bossa MN, Sölter J, Husch AD. Public Covid-19 X-ray datasets and their impact on model bias – A systematic review of a significant problem. Med Image Anal. 2021;74.
28. 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;10:19549. https://doi.org/10.1038/s41598-020-76550-z.
29. Panwar H, Gupta PK, Siddiqui MK, Morales-Menendez R, Singh V. Application of Deep Learning for Fast Detection of COVID-19 in X-Rays using nCOVnet. Chaos Solitons Fractals. 2020;138:109944. https://doi.org/10.1016/j.chaos.2020.109944.
30. Ronneberger O, Fische P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation, in Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. 2015;pp 234–241. https://doi.org/10.1007/978-3-319-24574-4_28.
31. Jaeger S, Candemir S, Antani S, Yì-Xiáng J, Lu W, Thomas G. Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. Quant Imaging Med Surg. 2014;4:475–7. https://doi.org/10.3978/j.issn.2223-4292.2014.11.20.
32. Shiraishi J, Katsuragawa S, Ikezoe J, Matsumoto T, Kobayashi T, Komatsu K, et al. Development of a Digital Image Database for Chest Radiographs With and Without a Lung Nodule. Am J Roentgenol. 2000;174:71–4. https://doi.org/10.2214/ajr.174.1.1740071.
33. Oh Y, Park S, Ye JC. Deep Learning COVID-19 Features on CXR Using Limited Training Data Sets. IEEE Trans Med Imaging. 2020;39:2688–2700. https://doi.org/10.1109/TMI.2020.2993291.
34. Kingma DP, Ba J. Adam: A method for stochastic optimization, San Diego, CA, USA. 2015.
35. Kandel I, Castelli M. The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset. ICT Express. 2020;6:312–5. https://doi.org/10.1016/j.icte.2020.04.010.
36. Nicolov M, Tsenov G, Mladenov V. COVID-19 detection with X-Ray input data COVID-19 detection with X-Ray input data. Int Conf Autom Inform 2021. 2021;Varna, Bulgaria (ICAI-21):437–442. https://doi.org/10.1109/ICAI52893.2021.9639562
37. Aslan MF, Sabanci K, Durdu A, Unlersen MF. COVID-19 diagnosis using state-of-the-art CNN architecture features and Bayesian Optimization. Comput Biol Med. 2022;42:105244. https://doi.org/10.1016/j.compbiomed.2022.105244.
38. Dhont J, Wolfs C, Verhaegen F. Automatic coronavirus disease 2019 diagnosis based on chest radiography and deep learning – Success story or dataset bias? Med Phys. 2022;49:978–87. https://doi.org/10.1002/mp.15419.
39. Cellina M, Orsi M, Toluian T, Valenti Pittino C, Oliva G. False negative chest X-Rays in patients affected by COVID-19 pneumonia and corresponding chest CT findings. Radiography. 2020;26:e189–94. https://doi.org/10.1016/j.radi.2020.04.017.
40. Karthik R, Menaka R, Hariharan M, Daehan W. Contour-enhanced attention CNN for CT-based COVID-19 segmentation. Pattern Recognit. 2022;125:108538. https://doi.org/10.1016/j.patcog.2022.108538.
41. Guangyu J, Hak-Keung L, Yujia X. Classification of COVID-19 chest X-Ray and CT images using a type of dynamic CNN modification method. Comput Biol Med. 2021;134:104425. https://doi.org/10.1016/j.compbiomed.2021.104425.
42. Xiaole F, Xiufang F, Yunyun D, Huichao H. COVID-19 CT image recognition algorithm based on transformer and CNN. Displays. 2022;72:102150. https://doi.org/10.1016/j.displa.2022.102150.

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