Health and Technology (2021) 11:1331–1345
https://doi.org/10.1007/s12553-021-00609-8

REVIEW PAPER

Current limitations to identify covid-19 using artificial intelligence with chest x-ray imaging (part ii). The shortcut learning problem

José Daniel López-Cabrera[1] · Rubén Orozco-Morales[2] · Jorge Armando Portal-Díaz[2] · Orlando Lovelle-Enríquez[3] · Marlén Pérez-Díaz[2]

Received: 28 July 2021 / Accepted: 5 October 2021 / Published online: 10 October 2021

Abstract

Since the outbreak of the COVID-19 pandemic, computer vision researchers have been working on automatic identification of this disease using radiological images. The results achieved by automatic classification methods far exceed those of human specialists, with sensitivity as high as 100% being reported. However, prestigious radiology societies have stated that the use of this type of imaging alone is not recommended as a diagnostic method. According to some experts the patterns presented in these images are unspecific and subtle, overlapping with other viral pneumonias. This report seeks to evaluate the analysis the robustness and generalizability of different approaches using artificial intelligence, deep learning and computer vision to identify COVID-19 using chest X-rays images. We also seek to alert researchers and reviewers to the issue of "shortcut learning". Recommendations are presented to identify whether COVID-19 automatic classification models are being affected by shortcut learning. Firstly, papers using explainable artificial intelligence methods are reviewed. The results of applying external validation sets are evaluated to determine the generalizability of these methods. Finally, studies that apply traditional computer vision methods to perform the same task are considered. It is evident that using the whole chest X-Ray image or the bounding box of the lungs, the image regions that contribute most to the classification appear outside of the lung region, something that is not likely possible. In addition, although the investigations that evaluated their models on data sets external to the training set, the effectiveness of these models decreased significantly, it may provide a more realistic representation as how the model will perform in the clinic. The results indicate that, so far, the existing models often involve shortcut learning, which makes their use less appropriate in the clinical setting.

Keywords COVID-19 · Chest X-Rays · Artificial Intelligence · Deep Learning

1 Introduction

In December 2019, a new viral infection caused by Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV2), a member of the β-coronavirus single-stranded RNA [1], was discovered in China. In March 2020, the World Health Organization (WHO) proclaimed coronavirus disease 2019 (COVID-19) a pandemic. This disease has the characteristic of being highly contagious, which has led to its arrival in almost all corners of the planet. To date, more than 180 million people are reported to be infected and more than 3 million have died[1]. This situation has caused the total or partial

This article is part of the COVID-19 Health Technology: Design, Regulation, Management, Assessment

✉ José Daniel López-Cabrera
josedaniellc@uclv.cu

Rubén Orozco-Morales
rorozco@uclv.edu.cu

Jorge Armando Portal-Díaz
jportal@uclv.cu

Orlando Lovelle-Enríquez
lovelle@infomed.sld.cu

Marlén Pérez-Díaz
mperez@uclv.edu.cu

[1] Centro de Investigaciones de La Informática, Facultad de Matemática, Física y Computación, Universidad Central "Marta Abreu" de Las Villas, Villa Clara, Santa Clara, Cuba

[2] Departamento de Control Automático, Facultad de Ingeniería Eléctrica, Universidad Central "Marta Abreu" de Las Villas, Villa Clara, Santa Clara, Cuba

[3] Departamento de Imagenología, Hospital Comandante Manuel Fajardo Rivero, Villa Clara, Santa Clara, Cuba

[1] https://www.worldometers.info/coronavirus/

lockdown of many regions leading to wide-spread, adverse public health, economic and social outcomes.

Isolation of positive patients is key to cutting off the chain of infection. The gold standard for diagnosing COVID-19 is from the identification of viral RNA by Reverse Transcription-Polymerase Chain Reaction (RT-PCR). However, this method has some limitations such as its modest diagnostic performance and the delay in obtaining results. For example, the method may take between 6 and 9 h to confirm infection [2]. In addition, sampling can be quite variable, depending on the site, personnel and viral load of the individual at the time [3]. Furthermore, this test decreases its sensitivity if not applied within a specific period of time [4, 5].

The rapid spread of the coronavirus and the serious effects it causes in humans make early diagnosis of the disease imperative [6]. The fact that COVID-19 often presents with pulmonary pathology has led a large number of studies on the utility of chest radiography to determine the presence of disease [4]. However, prestigious radiological societies have questioned the role of chest imaging alone as a diagnostic method [7, 8].

A large number of studies to determine the presence of disease using chest radiography (CXR) using computer vision, often based on deep learning (DL), have been reported [9]. These papers have reported performance rates much higher than those of human expert observers. One of the potential shortcomings of these techniques is the introduction of the bias referred to as “shortcut learning” [10]. That is, the models may rely on features that are not related to the pathology they are trying to classify. This bias can lead to models with very high-performance rates when evaluated on sets coming from the same distribution as the training set (called independent and identically distributed or iid). However, the same may not hold true when to a data set not from the same distribution (called out-of-distribution or ood). In such a case, the generalizability of the model may be severely limited.

In this research we will focus on the review of articles based on CXR as this imaging modality is widely used in the diagnosis and follow-up of patients and has some advantage compared to CT modality especially in COVID-19 positive patients as will be explained in this report. This is a continuation of previous work [11], providing further evidence of errors and biases made by researchers in the automatic identification of COVID-19 using CXR. This report seeks to reveal the weaknesses of models proposed so far to diagnose COVID-19 from CXR autonomously, using artificial intelligence (AI). In particular, we seek to alert researchers and reviewers to the concerns of shortcut learning which has been ignored in almost all the papers reviewed in the context of COVID-19 [12] as well as in other field of image classification [10]. In this research, two analyzes are proposed to verify whether the methods are being affected by this issue. Specifically, studies that use explainable artificial intelligence to determine the regions that contribute the most to the classification are reviewed. Similarly, works that make use of an external validation set are analyzed to determine the generalizability of methods based on DL. In addition, studies based on traditional computer vision approaches are examined.

This paper is organized as follows. Firstly, we will discuss the “**IMPORTANCE OF CXR AND CT IMAGING IN THE TIMELY MANAGEMENT OF COVID-19**”. After that, we will state the criteria of important radiological society about the “**USE OF CXR AND CT AS A DIAGNOSTIC METHOD FOR COVID-19**”. Afterwards, in section entitled “**THE USE OF AI IN RADIOLOGICAL IMAGING**” we will show the performance index achieved by the radiologist and the artificial intelligence method for classifying COVID-19, which reflect contradictions. Thereafter, section “**DEEP LEARNING TECHNIQUES AND SHORTCUT LEARNING**” will introduce the Deep Learning techniques and how Shortcut Learning is affecting these methods. The next section, “**EVIDENCE OF SHORTCUTS LEARNING IN CXR CLASSIFICATION**” will discuss studies that show the effect of this phenomenon, specifically on CXR images. One of the ways to determine the presence of Shortcut Learning is by using Explanatory Artificial Intelligence, these methods will be reviewed in the section “**EXPLANATORY AI METHODS IN THE IDENTIFICATION OF COVID-19 USING CXR**”. On the other hand, another way to determine Shortcut Learning is by using an external dataset to validate the models, many few studies are reported in the scientific literature which are discussed in the section “**EXTERNAL VALIDATION SET TO DETERMINE GENERALIZATION CAPABILITY OF THE MODELS**”. Because Deep Learning algorithms compute their own features and therefore may exacerbate the shortcut learning phenomenon, in the section “**BEHAVIOR OF TRADITIONAL COMPUTER VISION METHODS**” the results achieved using traditional computer vision methods will be discussed. Subsequently, in section “**DISCUSSION AND FUTURE WORK**” the main limitation of the analyzed research will be explained and some issues will be suggested to avoid the encountered limitations. Finally, the “**CONCLUSIONS**” reached in the research will be provided.

2 Importance of CXR and CT imaging in the timely management of COVID-19

Undoubtedly, medical imaging of the lungs is an important tool to assist specialists, both in the management of patients with acute respiratory infections (ARIs) as well as other diseases. In the case of COVID-19, studies confirm visible abnormalities in the lung region for some patients, thus serving as a decision-making tool for human specialists [13]. It is important to take into account that, there are patients with positive PCR who do not develop signs or symptoms, so it is likely impossible to make the diagnosis using CXR alone.

CT images present greater sensitivity as a diagnostic and follow-up method compared to CXR. For example, there are reported cases of COVID-19 with visible lesions on CT but not visible on CXR [14]. In fact, one of the main CT findings in patients with COVID-19 are ground-glass opacities in the peripheral regions of the lower lobes, which may not be seen on CXR (Fig. 1). However, CT imaging capability may not be available in many medical centers where COVID-19 is diagnosed around the world. In addition, where CT equipment does exist, it is not possible to dedicate it exclusively to COVID-19 diagnosis, given the high virology of the disease and the pressure of care. On the other hand, CXR has the advantage of being available in most healthcare facilities. Its cost is much lower compared to CT imaging, and the portability of CXR can help to prevent the patient from moving about the medial center, and, thus, minimizing the possibility of spreading the virus. In many instances, this makes CXR preferable, even though it may be less sensitive for diagnosis and patient follow-up.

The most frequent findings on CXR for COVID-19 are bilateral consolidated, absence of plural effusion, bilateral ground-glass pattern, peripheral and in basal lobes, which appear as the clinical disease progresses from ten to twelve days after the onset of symptoms [16]. However, the use of this technique as a diagnostic method has shown low sensitivity and specificity in current radiological practice in asymptomatic patients with mild to medium grade disease [17]. For example, according to Ref [18] the sensitivity using CXR to detect SARS-CoV-2 pneumonia is 57%. In older patients the sensitivity was slightly higher compared to younger patients, but, in both cases, it was low. On the other hand, in [19], higher sensitivity values were recorded by radiologists, 65%. These values demonstrate the difficulty on the part of radiologists in making a diagnosis of COVID-19 using CXR alone.

3 Use of cxr and ct as a diagnostic method for covid-19

Due to the increase of COVID-19 positive cases, since March 2020 prestigious radiology organizations (Fleischner Society [7], American College of Radiology (ACR), Canadian Association of Radiologists (CAR) [8], Canadian Society of Thoracic Radiology (CSTR) and British Society of Thoracic Imaging (BSTI) [20]) issued recommendations on the use of CT and CXR as a method of screening, diagnosis and patient management for COVID-19. These organizations agree that, the use of these chest images alone should not be used to diagnose COVID-19, nor should they be used routinely in all patients with suspected COVID-19. These imaging techniques should also not be used to inform the decision to test a patient for COVID-19, as normal chest imaging findings do not exclude the possibility of COVID-

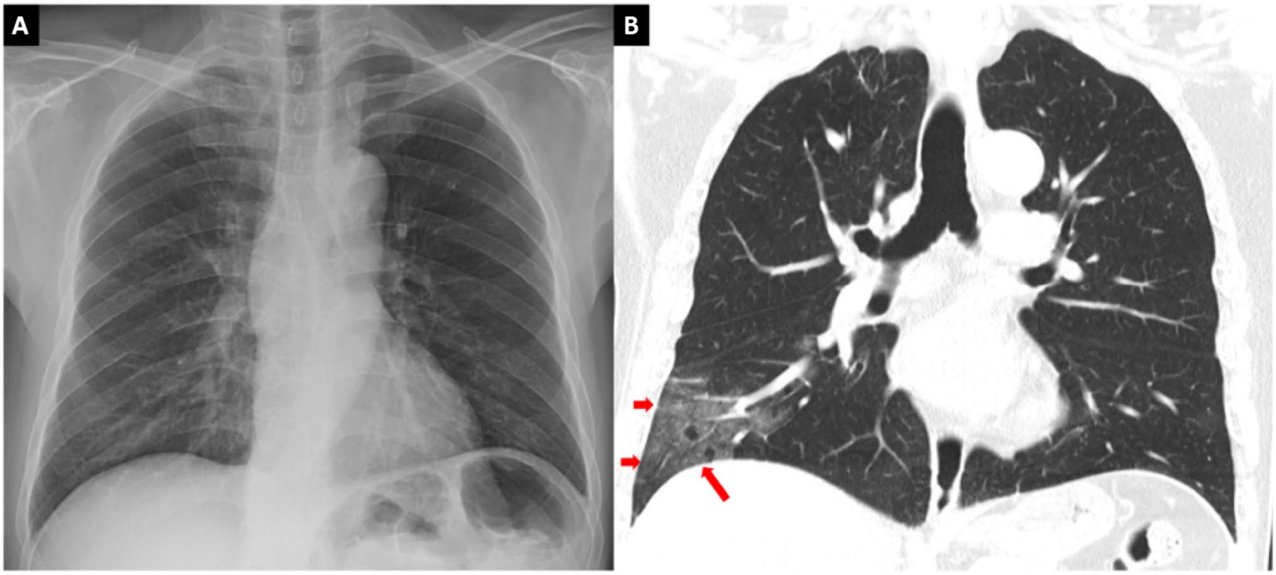

Fig. 1 Example of CXR image (**A**) and CT image (**B**) for a COVID-19 positive patient. Red arrows show a lesion visible on CT, but not detectable using CXR, extracted from [15]

19 infection. In addition, abnormal chest imaging findings are not specific for diagnosis of COVID-19. In general, chest imaging findings in COVID-19 are nonspecific and can overlap with other infections, such as influenza, H1N1, SARS and MERS. There are patients who present with positive PCR, but do not develop signs and symptoms of disease and thus may have a normal CXR or CT. Therefore, they cannot be diagnosed as positive using lung imaging alone [21]. Consequently, CXR or CT may be used to assess the status of patients at risk of disease progression and with worsening respiratory status, but should not be used as the primary diagnostic screening tool.

4 The use of ai in radiological imaging

The non-specific, subtle and difficult manifestations on both CT and CXR makes it difficult to achieve a high success rate in the diagnosis of COVID-19. Despite this, investigators from the fields of AI, DL and computer vision, have published a number of reports in this arena [9]. Articles in peer-reviewed journals have reported on automatic disease identification using CT and CXR. Such investigations should be considered with caution so as not to create false expectations, since, in many cases, the results reported far exceed those achieved by expert observers such as radiologists [22]. In general, radiologists may consider AI as a useful diagnostic support tool, but are concerned that the diagnostic accuracy using these techniques alone is overstated. [23].

Advances in automatic COVID-19 identification using CXR and CT imaging are reviewed in many studies [9, 24–33] the average accuracy of AI methods of approximately 90%. On the other hand, the values for CXR compared to CT are even higher at 96%. In some studies, the reported sensitivity is 100% [34, 35]. These reported values contradict, firstly, the fact that CT, in general, has higher sensitivity than CXR. Secondly, the lack of specificity of CXR would lead many radiology expert opinions, these images should not be used as a diagnostic tool for these patients. Note that radiologists only achieve a sensitivity between 57 and 65% [19] in the review of similar cases. None of the aforementioned review papers, question or analyze these high results achieved. In fact, in one report [36] it is stated that their method was able to identify with 100% effectiveness patients presenting lesions visible using CT, while not detectable on CXR.

According to one report [37] the advances achieved in the automatic classification of COVID-19 from CXR have little or no utility in clinical practice. Despite reporting encouraging results, the use of these models as specialist decision support systems must undergo more rigorous investigations and meet regional regulatory and quality control requirements. In particular, their performance must be validated and their efficacy demonstrated in the clinical workflow. On the other hand, in many investigations, the image sets used were small and poorly balanced. One review on the current limitations of studies using CXR to perform diagnosis [11] indicated that the use of datasets from different sources leads to models that learn features not related to the disease they are trying to identify, i.e. they demonstrate the phenomenon known as shortcut learning.

5 Deep learning techniques and shortcut learning

Most of the techniques used in the COVID-19 automatic identification task based on CXR are based on DL that specifically rely on convolutional neural networks (CNN). These approached have achieved substantial, recent success in biomedical applications [38]. CNNs specialize in classifying images autonomously, without the need to consider previously defined features to perform the classification, as with traditional CV methods. As a result, the process of feature extraction and classification can be performed in a single stage. In short, CNNs consist of the serial connection of a feature extraction network and a classification network. Through the training process, the weights of both networks are determined. The feature extraction stage contains the filters for convolution, clustering, normalization, evaluation of an activation function, and so on. Meanwhile, the fully connected layers in the last stage act similarly to a conventional Multi-Layer Perceptron (MLP)[39]. That is, a CNN in its training phase learns the coefficients that minimize the classification error, having to adjust millions of parameters. This explosion in the use of CNNs, even for complicated applications such as the analysis of medical images, has been made possible by the increase in computational power and capability [40] [41].

However, these DL methods are beginning to be evaluated critically and some limitations have been reported. One of the difficulties studied is the bias known as shortcut learning where decision rules may that perform well on standard benchmark sets, but do not transfer to more difficult test conditions, such as real-world scenarios. For example, models achieve superhuman performance in object recognition, but even small changes invisible to humans [42] or modifications to the image background [43, 44]. Furthermore, models can correctly classify an image, but worryingly, they may do so without taking into account what actually confers that classification [45]. That is, the models use features that are capable of correctly separating the class, but are not directly related to the task at hand. For instance, these models rely on differences in the background rather than the object to be classified. For example, the approach may be able recognize faces accurately, but show high error rates for faces from marginalized groups that were not adequately represented in the training set [46]. These types of observations are beginning to cause concern in the scientific community.

Shortcut learning can present a major obstacle to achieving more reliable models. Overcoming this issue entirety may be exceedingly difficult if not impossible, but any progress in mitigating it will lead to more reliable solutions. The hope is that the models can behave in a similar way even in situations outside their learning. In other words, we want the model to have high generalizability outside of its training set. Currently, research on shortcut learning and its mitigation remains fragmented. Many studies do not address these limitations and do not take into account this important issue as evident in their research. However, there are others that attempt to foster discussions and raise awareness of these issues among researchers, trying to make the rule of what has so far been only the exception. For example, it is recommended [10] that the results be considered carefully, using explainable AI techniques. It is also proposed as an essential rule to determine the generalization power of the model, using a set that does not come from any of the sets used in the training stage. These recommendations can apply to the COVID-19 identification task using CXR as will be discussed.

6 Evidence of shortcuts learning in cxr classification

The use of DL methods has been extensively studied in the field of CXR imaging [30]. As mentioned above, the evaluation of the generalizability of the proposed method from an ood set has been limited. However, some have reported evidence about existence of shortcut learning [47–49]. In one case [47], irregularities are reported when training a model on an image set and evaluating on an ood set. Specifically, from four sets A, B, C and D, it was observed that when training and evaluating on set A, the results are superior compared to training using sets B, C and D, and evaluating with set A. Another study demonstrated the presence of shortcut learning [50] when the model was able to identify the originating hospital with more than 95% accuracy. According to the network activation map, it was observed that, to achieve this result, the model relied on the text labels on the CXR images, instead of the lung region. This demonstrates that the performance of CNNs in disease diagnosis using radiographs may reflect not only their ability to identify specific disease findings in the image, but also their ability to exploit confounding information such as text labels.

Current DL models for identifying COVID-19 using CXR do not escape of shortcut learning. For example, one study [51] performed a classification with more than 90% accuracy without using the lung region demonstrating that the models have a lot of information to exploit that is not related to the disease manifestations in the lung region particularly when the entire image is utilized. In the same study, the absence of an ood set to assess generalizability is strongly criticized. On the other hand, another report [52] recognized that most of published work has not performed any analysis to demonstrate the reliability of network predictions. In the context of medical tasks, this is particularly relevant. That is, most of the state-of-the-art studies have validated their results with datasets containing tens or a few hundred COVID-19 samples, which may limit the general impact of the proposed solution. As proposed in this study, one of the ways to obtain greater reliability of the methods is to use techniques that visualize the regions on which the findings of the models are centered.

Most of the research published on the application of AI and DL in the context of COVID-19 is based on images from different sources. After the publication of the GitHub-Cohen [53] image dataset, in which a set of COVID-19 positive images was made freely available to the international scientific community, there have been numerous reports applying AI techniques for automatic disease classification. That is, to date, this has been the most widely used source of COVID-19 positive images by the scientific community. The formula used by most research to increase the number of negative (non-COVID-19) images has been to add images from sets available from other sources. A detailed explanation of the current sets, as well as their limitations, has been reported [54, 55]. In fact, in one study [54] it was determined that only five of the 256 datasets identified met the criteria for an adequate assessment of the risk of bias. In that study, it was observed that most of the data sets used in 78 published articles are not among these five data sets, resulting in models with a high risk of shortcut learning and other forms of bias.

7 Explanatory ai methods in the identification of covid-19 using cxr

Automatic diagnostic methods rely on interpretations from expert human observers on which they base their decisions. One of the current lines of research is the development of explainable artificial intelligence (XAI) methods [56]. Specifically, in the field of image-based medical applications, an adequate explanation of the decision obtained is essential. That is, a decision support system should be able to suggest the diagnosis and show, to best of its ability, what image content contributed to the decision reached by the algorithm. Such methods allow for the assessment of the veracity of the models. Therefore, through these techniques it is possible to verify if the decisions determined by the models are centered on regions that should be used for diagnosis. For example, is the determination of presence of pulmonary complications from COVID-19 based on an analysis of CXR findings in the lungs?

XAI techniques have also been applied in the environment of automatic detection of COVID-19 from CXR. Table 1 lists some of the papers published to date that make use of these XAI tools. As can be seen, several techniques are reported, among the most used are LIME [57], Grad-Cam [58] and Grad-Cam + + [59]. The table also records the presence of segmentation methods to determine the lung region. This is of vital importance, since as will be shown below, when not only the lung region is used, the models tend to focus on regions where its association with disease in question is unclear. Figure 2 (extracted from [52]) shows an example of how when using the whole image, CNNs may use as most important regions for classification areas that are not within the lungs. This means that there are regions that provide enough information to adequately separate the classes with features not related to the disease that they are trying to classify. This is likely a case where the model is using shortcut learning.

An example of lung segmentation necessity is reported in one study [61], where the outcome of attention maps was evaluated by two radiologists. Reportedly, the model half focused on regions outside the lungs to perform the classification in half of the cases. The recommendation put forward by the authors was to train on a much larger data set, so that the model would show a more robust performance in that aspect. In response to this recommendation, another study [12] image sets of a greater number of were used. The objective was to determine which regions were most used by the models to assign an image to a class. It was evident that, at times, saliency maps marked lung fields as important, suggesting that the models took into account genuine pathology of COVID-19. However, the saliency maps in some cases, highlighted regions outside the lung fields that may represent confounds. For example, the saliency maps frequently highlighted laterality markers as differing between the COVID-19 negative and COVID-19 positive datasets, and similarly highlighted arrows and other annotations that appear exclusively in the GitHub-Cohen dataset. Also, by applying the CycleGAN technique, images were generated that showed textual markings as important patterns for determining class. It is worth noting that this study made use of an external validation set. In this case, it was evident that the performance of the models decreased drastically when evaluated in the external (ood) set.

On the other hand, applying lung segmentation prior to classification and thus eliminating the use of regions outside the lungs leads to models that perform the classification based on features unrelated to the disease. Therefore, studies that use the complete image to perform the classification and achieve spectacular results (note that they are more than 30 percentage points in relation to specialists in radiology) but, in truth, are not valid.

For example, in one report [52] the imaging regions that contributed most to the identification of COVID-19 were evaluated using three imaging variants. In the first experiment, the full image was used and again, it was observed that the model took regions outside the lungs to perform classification. In the second experiment, the bounding box region of the lungs was used, where the same problem as in the previous experiment also appeared. Finally, in the third experiment an image of the segmented lungs was used, which forced the method to find the features within these regions. This time the results obtained indicated lower performance than the previous variants demonstrating that when using the previous variants, the models use features that are not related with the pathology in question.

Another attempt to visually assess the regions that a model uses to determine class was reported [63] showing that, when using the whole image as input, the CNNs pointed out as important regions those that did not belong to the lungs. In this case, the models focused their attention on the text labels present on the images. As a result, the models were able to identify with high accuracy the site where the images were acquired and thus whether they were likely to be cases of COVID-19. This occurred even after applying lung segmentation. Therefore, there are hidden features in the images that can be exploited by the models to perform classification and need to be handled coutiously to achieve reliable models.

In other cases, the demographic characteristics of the population can be a strong confounding factor. Several papers have used image sets where one of the classes belongs to children [55]. On the other hand, patients with COVID-19 often showed artifacts such as electrodes and their wires while other patients are intubated. Also, the position of the patients can have an effect, since in healthy patients, the X-ray view wass usually AP while in the COVID-19 patients, patients were more often supine, and the view is PA.

XAI methods have been used to determine the regions of the image that contribute most to the classification and thus build more reliable models. Furthermore, it became evident that, when using the whole image, the regions marked as important may not be related with the classification label, invalidating the results achieved. On the other hand, the segmentation of the lungs does not guarantee that the models really focus on appropriate regions and may contain underlying features that may still mask the good performance of the models. Nevertheless, as can be seen in Table 1, there are studies that, reporting the regions on which their models

Table 1 Main studies using XAI techniques to identify COVID-19 using CXR

Ref	Lung Segmentation	XAI method used	Performance index	Evaluation using ood set

[64]	No	Grad-Cam, Grad-Cam + +, LRP	Precision = 92%
			Recall = 92%
			Fscore = 0.92
				No

[65]	No	Grad-Cam	Acc = 95.57%	No

[66]	No	Grad-Cam, Grad-Cam + +	Precision = 96.58%
			Recall = 96.59%
			Fscore = 0.96
				No

[67]	No	Grad-Cam	Precision = 96.44%
			Recall = 96.33%
			Fscore = 0.96
			Acc = 96.33%
				No

[61]	No	Grad-Cam	Two class
			Acc = 100%
			Sensitivity = 99%
			Specificity = 100%
			AUC = 1
			Three class
			Acc = 98%
			Sensitivity = 96%
			Specificity = 99%
			AUC = 0.99
				No

[68]	No	Occlusion, Saliency, Input X Gradient, Integrated Gradients, Guided Backpropagation, DeepLIFT	Micro-F1 = 0.89	No

[69]	No	RISE	Sensitivity = 100%
			Acc = 90.5%
				No

[70]	No	LIME, Saliency Map, Grad-Cam	Two class
			Acc = 98.02%
			Three class
			Acc = 97.12%
				No

[71]	No	Grad-Cam + +	Acc = 91.26%	No

[12]	No	CycleGAN, Expected Gradients	Internal Partition (iid)
			AUC = 0.99
			External Dataset (ood)
			AUC = 0.76
				Yes

[72]	No	Grad-Cam	Acc = 96.3%	No

[73]	-	Grad-Cam	Positivity Predicted Value = 95%
			Sensitivity = 94%
			Fscore = 0.95
				No

[52]	Yes	Grad-Cam	Acc = 91.67%
			Fscore = 0.94
				No

[62]	Yes (bounding box of the lungs)	Grad-Cam, LIME	Fscore = 0.92	No

[74]	Yes	Saliency Map, Guided Backpropagation, Grad-Cam	Acc = 97.94%
			AUC = 0.984
				No

[75]	Yes	Grad-Cam,	Acc = 98.67%
			Fscore = 0.98
				No

[63]	Yes	LIME, Grad-Cam	Fscore = 0.88	No

[76]	Yes	Grad-Cam	Acc = 88.9%
			Fscore = 0.84
			Specificity = 96.4%
				No

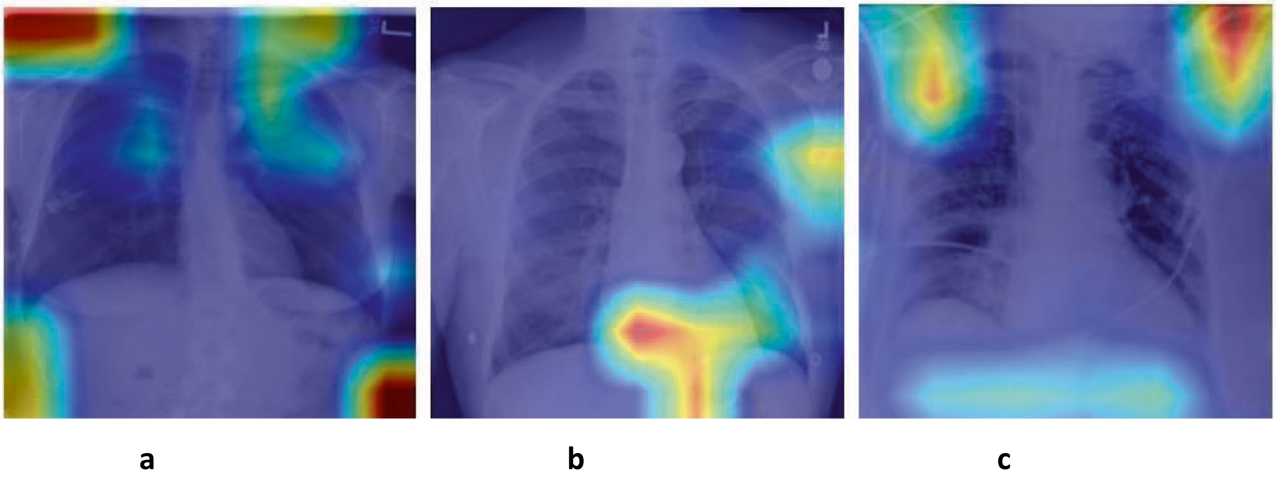

Fig. 2 Activation map for a modification of the CNN COVID-Net [60], obtained from the Grad-Cam method, by using the whole image to perform the classification. Image "a" belongs to the normal class, "b" belongs to the pneumonia class and "c" to COVID-19 class. In all cases, the regions on which the network is basing its decision are outside the lungs

are based and knowing that they do not correspond to the disease they are trying to identify, report high rates of effectiveness in the classification. Again, this is evidence of the presence of shortcuts learning, as well as the omission of this issue by scientific community. Hence, an external evaluation set is needed as a complement to demonstrate that the models maintain their behavior. Despite this, an evaluation methodology is not reported in any of the studies using the XAI techniques.

8 External validation set to determine generalization capability of the models

One way to eliminate biases in CXR image sets is use image processing techniques to pre-process the image before applying the AI and DL methods. One approach is to automatically limit the portion of the image to be analyzed to a bounding box region enclosing the lungs. A second approach is to segment the lung region automatically. With these techniques, spurious labeling marks that could artificially assist the model with classification are removed. However, the removal of these marks does not guarantee improvement in the model's generalizability. One way to test the validity and generalizability of the model is to evaluate it with an external, ood data set. To date, few studies report the use of an ood validation.

A discussion regarding validation on a set of external, ood image set has been reported. Table 2 presents an update of the published studies that when evaluated on external validation sets showed evidence of a lack of generalizability. These studies demonstrate that the algorithms learn features related to the source dataset, rather than the disease they are trying to classify, that is, studies being affected by shortcut learning. Note that the results of studies using an internal validation set report extremely good performance. However, when using the external evaluation set, these resulting performance decreases considerably. In fact, the reported performance measures have values close to that of a random classifier in most cases. Table 2 also presents the link to the image sets used by these investigations, which can constitute a starting point to carry out a more rigorous evaluation of the proposed models.

The creation of an appropriate evaluation strategy to address such biases is imperative. In other words, making a correct assessment reveals the existence of an issue that may otherwise remain hidden. Understanding the existence of the problem is the first step towards a solution. This issue needs to be taken seriously, especially since these systems are intended for use in clinical settings for the identification of COVID-19.

9 Behavior of traditional computer vision methods

According to the review studies analyzed, the majority of the investigations (27 articles) used CNN to identify COVID-19, most commonly ResNet, using different amounts of layers. DL techniques may tend to overfitting the classification models by generating their own features in the training process. Therefore, the use of traditional computer vision (CV) methods could lead to models with greater generalizability, especially, when using data sets that present marked differences [55].

Table 2 Summary of research using an external image set (ood) as a method of evaluating their models

Ref	Region used in the image	Performance index on iid set	Performance index on ood set	Sets of images

[12]	Whole Image	Dataset 1
		AUC = 0.992
		Dataset 2
		AUC = 0.995
			Dataset 1
			AUC = 0.76
			Dataset 2
			AUC = 0.70
				Dataset 1
				(GitHub-COVID)[2]
				(ChestX-ray14)[3]
				Dataset 2
				(BIMCV-COVID-19 +)[4] (PadChest)[5]

[21]	Bounding box of lungs	Dataset 1
		Using COVID-Net CXR model [60]
		Sensitivity = -
		Specificity = -
		Acc = 93.33%
		Using COVID-CAPS model [77]
		Sensitivity = 90%
		Specificity = 95.8%%
		Acc = 95.3%
			Dataset 2
			Using COVID-Net CXR model [60]
			Sensitivity = 99.29%
			Specificity = 0.23%
			Acc = 49.76%
			COVID-CAPS model [77]
			Sensitivity = 69.01%
			Specificity = 26.30%
			Acc = 47.66%
				Dataset 1
				(COVIDx)[6]
				Dataset 2
				(COVIDGR)[7]

[78]	Bounding box of lungs	Dataset 1
		AUC = 1
		Dataset 2
		AUC = 0.96
			Dataset 1
			AUC = 0.38
			Dataset 2
			AUC = 0.63
				Dataset 1
				(V2-COV19-NII)[8]
				(ChestX-ray14)[3]
				Dataset2
				(COVID-19-AR)[9]
				(BIMCV-COVID-19 +)[4],
				(Chexpert)[10]
				(Padchest)[5]

[79]	Segmented Lungs	Dataset 1
		Sensitivity = 100%
		Specificity = 100%
		AUC = 1
		Acc = 100%
			Dataset 2
			Sensitivity = 56%
			Specificity = 58%
			AUC = 0.59
			Acc = 57%
				Dataset 1
				(GitHub-COVID)[2]
				Dataset 2
				CORDA (Private)

[80]	Segmented Lungs	Dataset 1
		State 1 (classify pneumonia)
		Sensitivity = 92.85%
		Specificity = 90.05%
		AUC = 0.9672
		State 2 (classify COVID-19)
		Sensitivity = 85.26%
		Specificity = 85.86%
		AUC = 0.8804
			Dataset 2
			State 1 (classify pneumonia)
			Sensitivity = 63.64%
			Specificity = 90.48%
			AUC = 0.9394
			State 2 (classify COVID-19)
			Sensitivity = 50%
			Specificity = 40%
			AUC = 0.4
				Dataset1
				(GitHub-COVID)[2]
				(Padchest)[5]
				(RSNA)[11]
				Dataset2
				Private image sets from Taiwanese hospitals

[2] https://github.com/ieee8023/covid-chestxray-dataset

[3] https://nihcc.app.box.com/v/ChestXray-NIHCC

[4] https://bimcv.cipf.es/bimcv-projects/bimcv-covid19/

[5] https://bimcv.cipf.es/bimcv-projects/padchest/

[6] https://github.com/lindawangg/COVID-Net

[7] https://dasci.es/es/transferencia/open-data/covidgr/

[8] https://data.uni-hannover.de/dataset/cov-19-img/resource/38e72a9b-30a9-422a-a481-c7491e655437

[9] https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=70226443

[10] https://stanfordmlgroup.github.io/competitions/chexpert/

[11] https://www.kaggle.com/c/rsnapneumonia-detection-challenge/data

Traditional CV algorithms involve four main stages: 1) image preprocessing is performed by applying noise filtering, enhancement, resizing techniques, etc., 2) the detection of regions of interest is performed based on different sampling strategies or using segmentation techniques, 3) feature extraction is performed by means of some generally hand-constructed descriptor, e.g., SIFT [81], Local Binary Pattern (LBP) [82] among others, 4) the features describing the image are used by automatic classification algorithms to find the boundaries separating each class. These computed features can have high dimensionality, something that deters the good performance of the methods. One of the ways to eliminate this problem has been through feature selection techniques [83]

Table 3 Summary of works using traditional Computer Vision approach to identify COVID-19 using chest X-Ray imaging

Ref	Feature Extraction Method	Feature Selection/Reduction Method	Classification Algorithm	Performance Index	Image	Use of ood
[84]	New Orthogonal Exponent Moments of Fractional Orders Derived	New feature selection method: Manta Ray; Foraging Optimization (MRFO) using Differential evolution	Knn	Acc = 93%	Whole Image	No
[72]	VGG-19 + DenceNet-121	-	SVM	Acc = 98.28%	Whole Image	No

[87]	Each pixel as Feature;
	CNN;
	LBP;
	Gray Level Co-occurrence
		-	MLP
			CNN
				AUC = 0.93	Whole Image	No

[88]	Alexnet VGG-16 VGG-19 Xception Resnet18 Resnet50 Resnet101 Inceptionv3 Inceptionresnetv2 GoogleNet Densenet201	-	SVM	Acc = 95.38% Sensitivity = 97.29% Specificity = 93.47%	Whole Image	No
[89]	ChexNet [90]	PCA	MLP SVM Knn SRC-Dalm SRC-Hom CRC-light CRC CSEN1 CSEN2 ReconNet ResNet-50 Inception-v3	Acc = 99.26% Sensitivity = 97.14% Specificity = 99.49%	Whole Image, and lateral view	No
[91]	MobileNet DesnseNet121 DenseNet201 Xception InceptionV3 InceptionResNetV2 ResNet50 ResNet152 VGG16 VGG19 NASNetLarge NASNetMobile ResNet50V2 ResNet101V2 ResNet152V2	-	Decision Tree Random Forest XGBoost AdaBoost Bagging LightGBM	Acc = 98.00 Precision = 98.00 Recall = 98.00	Whole Image	No
[92]	New CNN	-	SVM Decision Tree Knn	Acc = 98.97% Sensitivity = 89.39% Specificity = 99.75% Fscore = 0.96	Whole Image	No

[93]	New architecture of CNN;
	Texture-based;
	FFT;
	Wavelet;
	GLCM;
	GLDM
		DNE
		Relief
		LPP
		Fast-ICA
		recursive feature elimination
		variable ranking techniques
			SVM
			GLM
			Random Forest
				Precision = 95%
				Sensitivity = 94%
				Fscore = 0.94
					Whole Image	No

Table 3 (continued)

Ref	Feature Extraction Method	Feature Selection/Reduction Method	Classification Algorithm	Performance Index	Image	Use of ood
[86]	Inception-v3 LBP LPQ LDN EQP LETRIST BSIFT OBIF	-	SVM Knn MLP Decision Tree Random Forest Ensemble (Sum rule, Product rule, Voting Rule) Clus-HMC	Fscore = 0.83	Lung bounding box	No

Such approaches have also been used in the COVID-19 automatic classification task using CXR. Table 3 presents a summary of some of studies that make use of this methodology. There is a tendency in these studies to use pre-trained CNN networks as the method for feature extraction. However, other studies use traditional methods to extract features such as LBP and GLCM, among others. Likewise, in another study [84], a new descriptor based on orthogonal moments is proposed. Also, the use of algorithms for dimensionality reduction has also been studied, although it has not been a common practice. On the other hand, a great diversity of classification methods is also observed in Table 3. In fact, Support Vector Machine (SVM) and Random Forest (RF) are the most used, [85], where these classifiers are reported as the best performing ones. The performance indices achieved are comparable with those achieved by the CNNs analyzed in previous sections, also with values above what is reported by expert observers such as radiologists.

These studies have not taken into consideration the elimination of features that are not related to the disease since, in all cases, the complete image was used to extract the features. Thus, the same mistakes associated with the use of the whole image and the marks it contains can be made. Only one study addressed this issue [86], in which a manual segmentation of the images was performed such that the bounding box region enclosed the lungs and thus eliminated the labels from the analysis. That study also evaluated class imbalance distribution issues using resampling techniques. However, the authors of that study themselves in a new investigation [63] state that, although the experimental results achieved in [86] showed that it may be possible to identify COVID-19 using CXR, it was a challenge to ensure that other patterns not belonging to the lungs did not contribute to the classification.

Finally, Table 3 also shows that none of these investigations make use of an external validation set. In all cases a partition of the training set was used. It should be noted that, in all cases, the image sets used were obtained in a similar way as in the studies using CNN. That is, the image datasets present the same issues and biases discussed above.

10 Discussion and future work

Automatic COVID-19 classification using CXR imaging is an active topic by the scientific community. Most papers report high performance (Tables 1, 2 and 3). The majority of these studies use the DL approach, although the use of traditional CV methods to address the task has also been reported. In both cases, the results are far superior to those achieved by experienced radiologists. However, most of the studies using automated approaches utilized internationally available image sets. In these data sets, the positive and negative cases may have come from different sources, and the methods may learn to recognize the source rather than the disease. This can result in a lack of generalizability of the models as seen in Table 2.

The main concern has been the absence of a correct evaluation protocol on the proposed models. In the studies analyzed, the results of using images that do not belong to any of the sources of the image sets used in the training of the models are rarely presented. In the case of making use of an ood set, a notable decrease in performance has been reported. In one review [94], it was determined that none of the articles analyzed in their research met the requirements to be considered reliable. The authors found no sufficiently documented manuscript describing a reproducible method. Also, no method was identified that follows best practices for developing a machine learning model with sufficient external validation to justify the applicability of the model. These are issues that should be taken into account to ensure the development of better quality, reproducible models that were free from biases such as shortcut learning.

An important step in this process lies in the proper selection by computer vision specialists together with radiologists and medical physicists of an adequate training set. Special care must be taken to select the training set minimizes the potential for biases in the resulting models such as those that learn by shortcuts. Otherwise, the models may yield good results in the iid sets but poor results in the ood sets, as has often been the case. In addition, one should be aware of the need to demonstrate as well as possible what the decisions reached by model are based upon. This will make the decision process of AI techniques more transparent from which human specialists can learn. So far, it seems unlikely that CXR alone can provide an accurate diagnosis of COVID-19. In fact, radiologists typically rely on other patient characteristics and information to make a diagnosis. Thus, the union of several clinical features seems to be the way forward to achieve a system that really helps human specialists.

11 Conclusions

This paper reviewed the main approaches presented in the scientific literature to address the issues of automatic COVID-19 classification using CXR. According to the reviewed papers, the performance rates reported by automatic classifiers outperform human specialists by more than 30 percentage points. However, a review of published papers using XAI that, CNNs base more of their classifications on regions outside the lung area. This suggests that these networks are performing shortcut learning. One approach to test the generalizability of these models is to base evaluation on an external, ood data set. However, this methodology has not been applied in most of the studies reviewed. In fact, the papers that have evaluated models on ood sets report performance rates close to random classification. This is evidence that the models proposed so far learn patterns that are not related to the disease they are trying to classify. That is, evaluating the performance of the models on an iid as a validation set (as most current benchmark tests do) is insufficient to distinguish the generalization power of the models. Therefore, as a fundamental step in model evaluation is to require the use an external, ood data set. Studies based on traditional computer vision methods showed the same issues as DL approaches. Hence, ood generalization tests should become the rule rather than the exception, especially in biomedical solutions where inadequate diagnoses may be applied to patients that negatively impact the choice of treatment for serious diseases such as COVID-19. When properly validated, AI and DL methods can provide the radiologist with valuable tools to assist in the diagnosis and classification of these diseases.

Acknowledgements We thank Professor Frederic Fahey for his valuable comments in the preparation of the final manuscript.

Funding Not applicable.

Data availability Not applicable.

Code availability Not applicable.

Declarations

Conflicts of interest The authors declare that they have no conflict of interest.

References

1. Lai C-C, Shih T-P, Ko W-C, Tang H-J, Hsueh P-R. 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(3): 105924. https://doi.org/10.1016/j.ijantimicag.2020.105924.
2. Narayan N, Das N, Kumar, Kaur M, Kumar V, Singh D. "Automated Deep Transfer Learning-Based Approach for Detection of COVID-19 Infection in Chest X-rays". IRBM 2020. https://doi.org/10.1016/j.irbm.2020.07.001.
3. Liu R, 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.
4. Ai T, et al. Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases. Radiology. 2020;296(2):E32–40.
5. Kucirka LM, Lauer SA, Laeyendecker O, Boon D, Lessler J. Variation in False-Negative Rate of Reverse Transcriptase Polymerase Chain Reaction-Based SARS-CoV-2 Tests by Time Since Exposure. Ann Intern Med. 2020;173(4):262–7. https://doi.org/10.7326/M20-1495.
6. Xie M, Chen Q. Insight into 2019 novel coronavirus — An updated interim review and lessons from SARS-CoV and MERS-CoV. Int J Infect Dis. 2020;94:119–24. https://doi.org/10.1016/j.ijid.2020.03.071.
7. Rubin GD, et al. The Role of Chest Imaging in Patient Management During the COVID-19 Pandemic: A Multinational Consensus Statement From the Fleischner Society. Chest. 2020;158(1):106–16. https://doi.org/10.1016/j.chest.2020.04.003.
8. Dennie C, et al. Canadian Society of Thoracic Radiology/Canadian Association of Radiologists Consensus Statement Regarding Chest Imaging in Suspected and Confirmed COVID-19. Can Assoc Radiol J. 2020. https://doi.org/10.1177/0846537120924606.
9. Islam MM, Karray F, Alhajj R, Zeng J. A Review on Deep Learning Techniques for the Diagnosis of Novel Coronavirus (COVID-19). IEEE Access. 2021;9:30551–72. https://doi.org/10.1109/ACCESS.2021.3058537.
10. Geirhos R, et al. "Shortcut learning in deep neural networks". Nat Mach Intell 2020;2(11). https://doi.org/10.1038/s42256-020-00257-z
11. López-Cabrera JD, Orozco-Morales R, Portal-Diaz JA, Lovelle-Enríquez O, Pérez-Díaz M. Current limitations to identify COVID-19 using artificial intelligence with chest X-ray imaging. Health Technol. 2021;11(2):411–24. https://doi.org/10.1007/s12553-021-00520-2.

12. DeGrave AJ, Janizek JD, Lee S-I. "AI for radiographic COVID-19 detection selects shortcuts over signal". Nat Mach Intell 2021;1–10. https://doi.org/10.1038/s42256-021-00338-7
13. 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(2):E113–4. https://doi.org/10.1148/radiol.2020200527.
14. 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(3)e189–e194. https://doi.org/10.1016/j.radi.2020.04.017
15. Ng M-Y, et al. Imaging Profile of the COVID-19 Infection: Radiologic Findings and Literature Review. Radiol Cardiothorac Imaging. 2020;2(1): e200034. https://doi.org/10.1148/ryct.2020200034.
16. Wong HYF, et al. Frequency and Distribution of Chest Radiographic Findings in Patients Positive for COVID-19. Radiology. 2020;296(2):E72–8. https://doi.org/10.1148/radiol.2020201160.
17. Yoon SH, 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(4):494–500.
18. Ippolito D, 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.
19. Castiglioni I, et al. "Artificial intelligence applied on chest X-ray can aid in the diagnosis of COVID-19 infection: a first experience from Lombardy, Italy,". medRxiv 2020.
20. Nair A, et al. A British Society of Thoracic Imaging statement: considerations in designing local imaging diagnostic algorithms for the COVID-19 pandemic. Clin Radiol. 2020;75(5):329–34. https://doi.org/10.1016/j.crad.2020.03.008.
21. Tabik S, 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(12):3595–605. https://doi.org/10.1109/JBHI.2020.3037127.
22. Laghi A. Cautions about radiologic diagnosis of COVID-19 infection driven by artificial intelligence. Lancet Digit Health. 2020;2(5): e225. https://doi.org/10.1016/S2589-7500(20)30079-0.
23. Summers RM. "Artificial Intelligence of COVID-19 Imaging: A Hammer in Search of a Nail,". Radiology 2020;204226. https://doi.org/10.1148/radiol.2020204226.
24. Farhat H, Sakr GE, Kilany R. "Deep learning applications in pulmonary medical imaging: recent updates and insights on COVID-19,". Mach Vis Appl 2020;31(6). https://doi.org/10.1007/s00138-020-01101-5.
25. Ilyas M, Rehman H, Nait-ali A. "Detection of Covid-19 From Chest X-ray Images Using Artificial Intelligence: An Early Review". ArXiv Prepr. ArXiv200405436 2020.
26. Nguyen TT. "Artificial intelligence in the battle against coronavirus (COVID-19): a survey and future research directions". Prepr 10 2020.
27. Shah FM, et al. "A Comprehensive Survey of COVID-19 Detection using Medical Images" 2020. [Online]. Available: https://engrxiv.org/9fdyp/download/?format=pdf.
28. Shi F, et al. "Review of Artificial Intelligence Techniques in Imaging Data Acquisition, Segmentation and Diagnosis for COVID-19". IEEE Rev Biomed Eng 2020;1–1. https://doi.org/10.1109/RBME.2020.2987975.
29. Ulhaq A, Born J, Khan A, Gomes DPS, Chakraborty S, Paul M. COVID-19 Control by Computer Vision Approaches: A Survey. IEEE Access. 2020;8:179437–56. https://doi.org/10.1109/ACCESS.2020.3027685.
30. Albahri OS, et al. Systematic review of artificial intelligence techniques in the detection and classification of COVID-19 medical images in terms of evaluation and benchmarking: Taxonomy analysis, challenges, future solutions and methodological aspects. J Infect Public Health. 2020. https://doi.org/10.1016/j.jiph.2020.06.028.
31. Shoeibi A, et al. "Automated Detection and Forecasting of COVID-19 using Deep Learning Techniques: A Review". ArXiv200710785 Cs Eess 2020. Accessed: Aug. 14, 2020. [Online]. Available: http://arxiv.org/abs/2007.10785
32. Chen Y, et al. A Survey on Artificial Intelligence in Chest Imaging of COVID-19. BIO Integr. 2020;1(3):137–46. https://doi.org/10.15212/bioi-2020-0015.
33. Soomro TA, Zheng L, Afifi AJ, Ali A, Yin M, Gao J. Artificial intelligence (AI) for medical imaging to combat coronavirus disease (COVID-19): a detailed review with direction for future research. Artif Intell Rev. 2021. https://doi.org/10.1007/s10462-021-09985-z.
34. 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: 102365. https://doi.org/10.1016/j.bspc.2020.102365.
35. 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.
36. Keles A, Keles MB, Keles A. COV19-CNNet and COV19-ResNet: Diagnostic Inference Engines for Early Detection of COVID-19. Cogn Comput. 2021. https://doi.org/10.1007/s12559-020-09795-5.
37. Bullock J, Luccioni A, Pham KH, Lam CSN, Luengo-Oroz M. Mapping the landscape of Artificial Intelligence applications against COVID-19. J Artif Intell Res. 2020;69:807–45. https://doi.org/10.1613/jair.1.12162.
38. Meijering E. A bird's-eye view of deep learning in bioimage analysis. Comput Struct Biotechnol J. 2020;18:2312–25. https://doi.org/10.1016/j.csbj.2020.08.003.
39. Kim P. "Convolutional Neural Network", in *MATLAB Deep Learning*. Berkeley, CA: Apress; 2017. p. 121–47.
40. Goodfellow I, Bengio Y, Courville A, Bengio Y. *Deep learning*, vol. 1. MIT press Cambridge 2016.
41. Sengupta S, et al. "A review of deep learning with special emphasis on architectures, applications and recent trends". Knowl-Based Syst 2020;194:105596. https://doi.org/10.1016/j.knosys.2020.105596.
42. Szegedy C, et al. "Intriguing properties of neural networks," presented at the 2nd International Conference on Learning Representations, ICLR 2014. Accessed: Feb. 23, 2021. [Online]. Available: https://nyuscholars.nyu.edu/en/publications/intriguing-properties-of-neural-networks.
43. Beery S, Van Horn G, Perona P. "Recognition in Terra Incognita". 2018;8456–473. Accessed: Feb. 23, 2021. [Online]. Available: https://openaccess.thecvf.com/content_ECCV_2018/html/Beery_Recognition_in_Terra_ECCV_2018_paper.html.
44. Rosenfeld A, Zemel R, Tsotsos JK. "The Elephant in the Room". ArXiv180803305 Cs 2018. Accessed: Feb. 23, 2021. [Online]. Available: http://arxiv.org/abs/1808.03305.
45. Heuer H, Monz C, Smeulders AWM. "Generating captions without looking beyond objects". ArXiv161003708 Cs 2016. Accessed: Feb. 23, 2021. [Online]. Available: http://arxiv.org/abs/1610.03708.
46. Buolamwini J, Gebru T. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification". In Conference on Fairness, Accountability and Transparency 2018;77–91. Accessed: Feb. 23, 2021. [Online]. Available: http://proceedings.mlr.press/v81/buolamwini18a.html.

47. Cohen JP, Hashir M, Brooks R, Bertrand H "On the limits of cross-domain generalization in automated X-ray prediction". In Medical Imaging with Deep Learning 2021;136–155. Accessed: Feb. 01, 2021. [Online]. Available: http://proceedings.mlr.press/v121/cohen20a.html.
48. Prevedello LM, et al. Challenges Related to Artificial Intelligence Research in Medical Imaging and the Importance of Image Analysis Competitions. Radiol Artif Intell. 2019;1(1): e180031. https://doi.org/10.1148/ryai.2019180031.
49. Yao L, Prosky J, Covington B, Lyman K. "A Strong Baseline for Domain Adaptation and Generalization in Medical Imaging". ArXiv190401638 Cs Eess Stat 2019. Accessed: Aug. 26, 2020. [Online]. Available: http://arxiv.org/abs/1904.01638
50. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLOS Med. 2018;15(11): e1002683. https://doi.org/10.1371/journal.pmed.1002683.
51. Maguolo G, Nanni L. A critic evaluation of methods for COVID-19 automatic detection from X-ray images. Inf Fusion. 2021;76:1–7. https://doi.org/10.1016/j.inffus.2021.04.008.
52. 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.
53. Cohen JP, Morrison P, Dao L, Roth K, Duong T, Ghassemi M. "COVID-19 Image Data Collection: Prospective Predictions are the Future". MELBA 2020;18272.
54. Garcia Santa Cruz B, Bossa MN, Sölter J, Husch A. "Public Covid-19 X-ray datasets and their impact on model bias - a systematic review of a significant problem". medRxiv 2021;02(15): 21251775. https://doi.org/10.1101/2021.02.15.21251775.
55. Garcia Santa Cruz B, Sölter J, Nicolas Bossa M, Dominik Husch A. "On the Composition and Limitations of Publicly Available COVID-19 X-Ray Imaging Datasets". ArXiv200811572 Cs Eess 2020. Accessed: Sep. 21, 2020. [Online]. Available: http://arxiv.org/abs/2008.11572.
56. Barredo Arrieta A, et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". Inf Fusion. 2020;58:82–115. https://doi.org/10.1016/j.inffus.2019.12.012.
57. Ribeiro MT, Singh S, Guestrin C. " Why should i trust you?" Explaining the predictions of any classifier. InProceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining 2016;1135-1144. https://doi.org/10.1145/2939672.2939778.
58. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. InProceedings of the IEEE international conference on computer vision 2017;618-626. https://doi.org/10.1109/ICCV.2017.74.
59. Chattopadhay A, Sarkar A, Howlader P, Balasubramanian VN. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In2018 IEEE winter conference on applications of computer vision (WACV) 2018;839-847. IEEE. https://doi.org/10.1109/WACV.2018.00097.
60. 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". Scientific Rep. 2020;10(1):1-2. https://doi.org/10.1038/s41598-020-76550-z.
61. Tsiknakis N, et al. Interpretable artificial intelligence framework for COVID-19 screening on chest X-rays. Exp Ther Med. 2020;20(2):727–35. https://doi.org/10.3892/etm.2020.8797.
62. Haghanifar A, Majdabadi MM, Choi Y, Deivalakshmi S, Ko S. "COVID-CXNet: Detecting COVID-19 in Frontal Chest X-ray Images using Deep Learning". ArXiv200613807 Cs Eess 2020. Accessed: Feb. 28, 2021. [Online]. Available: http://arxiv.org/abs/2006.13807.
63. Teixeira LO, Pereira RM, Bertolini D, Oliveira LS, Nanni L, Cavalcanti GD, Costa YM. "Impact of lung segmentation on the diagnosis and explanation of COVID-19 in chest X-ray images". ArXiv200909780 Cs Eess 2020. Accessed: Sep. 29, 2020. [Online]. Available: http://arxiv.org/abs/2009.09780
64. Karim MR, Döhmen T, Rebholz-Schuhmann D, Decker S, Cochez M, Beyan O. "DeepCOVIDExplainer: Explainable COVID-19 Diagnosis Based on Chest X-ray Images" 2020. Accessed: Jul. 10, 2020. [Online]. Available: https://arxiv.org/abs/2004.04582v3.
65. Qi X, Brown LG, Foran DJ, Nosher J, Hacihaliloglu I. Chest X-ray image phase features for improved diagnosis of COVID-19 using convolutional neural network. Int J Comput Assist Radiol Surg. 2021. https://doi.org/10.1007/s11548-020-02305-w.
66. Chowdhury NK, Rahman MdM, Kabir MA. PDCOVIDNet: a parallel-dilated convolutional neural network architecture for detecting COVID-19 from chest X-ray images. Health Inf Sci Syst. 2020;8(1):27. https://doi.org/10.1007/s13755-020-00119-3.
67. Chowdhury NK, Kabir MA, Rahman M, Rezoana N. "ECOVNet: An Ensemble of Deep Convolutional Neural Networks Based on EfficientNet to Detect COVID-19 From Chest X-rays". ArXiv200911850 Cs Eess 2020. Accessed: Mar. 13, 2021. [Online]. Available: http://arxiv.org/abs/2009.11850.
68. Chatterjee S, et al. "Exploration of Interpretability Techniques for Deep COVID-19 Classification using Chest X-ray Images". ArXiv200602570 Cs Eess 2020. Accessed: Mar. 13, 2021. [Online]. Available: http://arxiv.org/abs/2006.02570.
69. Mangal A, et al. "CovidAID: COVID-19 Detection Using Chest X-Ray". ArXiv200409803 Cs Eess 2020. Accessed: Feb. 28, 2021. [Online]. Available: http://arxiv.org/abs/2004.09803.
70. Siddhartha M, Santra A. "COVIDLite: A depth-wise separable deep neural network with white balance and CLAHE for detection of COVID-19," ArXiv200613873 Cs Eess 2020. Accessed: Feb. 28, 2021. [Online]. Available: http://arxiv.org/abs/2006.13873.
71. Liu B, Yan B, Zhou Y, Yang Y, Zhang Y. "Experiments of Federated Learning for COVID-19 Chest X-ray Images". ArXiv200705592 Cs Eess 2021. Accessed: Mar. 06, 2021. [Online]. Available: http://arxiv.org/abs/2007.05592.
72. Kedia P, Katarya R "CoVNet-19: A Deep Learning model for the detection and analysis of COVID-19 patients". Appl Soft Comput 2021;104:107184. https://doi.org/10.1016/j.asoc.2021.107184.
73. Aviles-Rivero AI, Sellars P, Schönlieb CB, Papadakis N. "GraphX-COVID: Explainable Deep Graph Diffusion Pseudo-Labelling for Identifying COVID-19 on Chest X-rays". ArXiv201000378 Cs Stat 2021. Accessed: Feb. 28, 2021. [Online]. Available: http://arxiv.org/abs/2010.00378.
74. Karthik R, Menaka R, HM. "Learning distinctive filters for COVID-19 detection from chest X-ray using shuffled residual CNN". Appl Soft Comput 2020;106744. https://doi.org/10.1016/j.asoc.2020.106744.
75. Singh RK, Pandey R, Babu RN. COVIDScreen: explainable deep learning framework for differential diagnosis of COVID-19 using chest X-rays. Neural Comput Appl. 2021. https://doi.org/10.1007/s00521-020-05636-6.
76. Oh Y, Park S, Chul Ye J. "Deep Learning COVID-19 Features on CXR using Limited Training Data Sets," IEEE Trans Med Imaging 2020. https://doi.org/10.1109/TMI.2020.2993291.
77. Afshar P, Heidarian S, Naderkhani F, Oikonomou A, Plataniotis KN, Mohammadi A. COVID-CAPS: A capsule network-based framework for identification of COVID-19 cases from X-ray images. Pattern Recognit Lett. 2020;138:638–43. https://doi.org/10.1016/j.patrec.2020.09.010.

78. 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.
79. 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 2020;17(18, Art. no. 18).
80. Yeh C-F, et al. “A Cascaded Learning Strategy for Robust COVID-19 Pneumonia Chest X-Ray Screening.” ArXiv200412786 Cs Eess 2020. Accessed: Aug. 14, 2020. [Online]. Available: http://arxiv.org/abs/2004.12786.
81. Lowe DG. Distinctive Image Features from Scale-Invariant Keypoints. Int J Comput Vis. 2004;60(2):91–110. https://doi.org/10.1023/B:VISI.0000029664.99615.94.
82. Ojala T, Pietikainen M, Maenpaa T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans Pattern Anal Mach Intell. 2002;24(7):971–87. https://doi.org/10.1109/TPAMI.2002.1017623.
83. Bolón-Canedo V, Remeseiro B. “Feature selection in image analysis: a survey”. Artif Intell Rev 2019;1–27. https://doi.org/10.1007/s10462-019-09750-3.
84. Elaziz MA, Hosny KM, Salah A, Darwish MM, Lu S, Sahlol AT. New machine learning method for image-based diagnosis of COVID-19. PLoS ONE. 2020;15(6): e0235187. https://doi.org/10.1371/journal.pone.0235187.
85. Fernández-Delgado M, Cernadas E, Barro S, Amorim D. Do we need hundreds of classifiers to solve real world classification problems? J Mach Learn Res. 2014;15(1):3133–81.
86. Pereira RM, Bertolini D, Teixeira LO, Silla Jr CN, Costa YM. “COVID-19 identification in chest X-ray images on flat and hierarchical classification scenarios”. Comput Methods Programs Biomed 2020;105532. https://doi.org/10.1016/j.cmpb.2020.105532.
87. Varela-Santos S, Melin P. A new approach for classifying coronavirus COVID-19 based on its manifestation on chest X-rays using texture features and neural networks. Inf Sci. 2021;545:403–14. https://doi.org/10.1016/j.ins.2020.09.041.
88. Pk S, Sk B. Detection of Coronavirus Disease (COVID-19) Based on Deep Features. Preprints. 2020. https://doi.org/10.20944/preprints202003.0300.v1.
89. Ahishali M, et al. “A Comparative Study on Early Detection of COVID-19 from Chest X-Ray Images”. ArXiv200605332 Cs Eess 2020. Accessed: Aug. 16, 2020. [Online]. Available: http://arxiv.org/abs/2006.05332.
90. Rajpurkar P, et al. “CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning”. ArXiv171105225 Cs Stat 2017. Accessed: Nov. 06, 2020. [Online]. Available: http://arxiv.org/abs/1711.05225.
91. Kassania SH, Kassanib PH, Wesolowskic MJ, Schneidera KA, Detersa R. “Automatic Detection of Coronavirus Disease (COVID-19) in X-ray and CT Images: A Machine Learning-Based Approach”. ArXiv200410641 Cs Eess 2020. Accessed: Aug. 08, 2020. [Online]. Available: http://arxiv.org/abs/2004.10641.
92. Nour M, Cömert MZ, Polat K. “A Novel Medical Diagnosis model for COVID-19 infection detection based on Deep Features and Bayesian Optimization”. Appl Soft Comput 2020;106580. https://doi.org/10.1016/j.asoc.2020.106580.
93. Khuzani AZ, Heidari M, Shariati SA. “COVID-Classifier: An automated machine learning model to assist in the diagnosis of COVID-19 infection in chest x-ray images”. medRxiv 2020. https://doi.org/10.1101/2020.05.09.20096560.
94. Roberts M, Driggs D, Thorpe M, Gilbey J, Yeung M, Ursprung S, Aviles-Rivero AI, Etmann C, McCague C, Beer L, Weir-McCall JR. “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(3, Art. no. 3). https://doi.org/10.1038/s42256-021-00307-0.

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