DO NOT REPEAT THESE MISTAKES - A CRITICAL APPRAISAL OF APPLICATIONS OF EXPLAINABLE ARTIFICIAL INTELLIGENCE FOR IMAGE BASED COVID-19 DETECTION

A PREPRINT

Weronika Hryniewska
Faculty of Mathematics and Information Science
Warsaw University of Technology
w.hryniewska@mini.pw.edu.pl

Przemysław Bombiński
Department of Pediatric Radiology
Medical University of Warsaw
przemyslaw.bombinski@uckwum.pl

Patryk Szatkowski
Department of Pediatric Radiology
Medical University of Warsaw
szpatryk@poczta.onet.pl

Paulina Tomaszewska
Faculty of Mathematics and Information Science
Warsaw University of Technology
paulina.tomaszewska3.stud@pw.edu.pl

Artur Przelaskowski
Faculty of Mathematics and Information Science
Warsaw University of Technology
a.przelaskowski@mini.pw.edu.pl

Przemysław Biecek
Faculty of Mathematics, Informatics and Mechanics
University of Warsaw
Faculty of Mathematics and Information Science
Warsaw University of Technology
przemyslaw.biecek@pw.edu.pl

December 17, 2020

ABSTRACT

The sudden outbreak and uncontrolled spread of COVID-19 disease is one of the most important global problems today. In a short period of time, it has led to the development of many deep neural network models for COVID-19 detection with modules for explainability. In this work, we carry out a systematic analysis of various aspects of proposed models. Our analysis revealed numerous mistakes made at different stages of data acquisition, model development, and explanation construction. In this work, we overview the approaches proposed in the surveyed ML articles and indicate typical errors emerging from the lack of deep understanding of the radiography domain. We present the perspective of both: experts in the field - radiologists and deep learning engineers dealing with model explanations. The final result is a proposed a checklist with the minimum conditions to be met by a reliable COVID-19 diagnostic model.

1 Introduction

COVID-19 is a fast spreading disease of highly contagious nature [1, 2, 3, 4, 5, 6] caused by the SARS-CoV-2 virus from the coronavirus group. At the end of January 2020, the World Health Organization (WHO) declared a global health emergency [7] and one and a half months later, a pandemic [8, 3]. By September 25, 2020, 32,110,656 confirmed cases and 980 031 deaths [9] had been documented. From a public health perspective, due to the lack of a vaccination or proper medicines, early detection of COVID-19 and patient isolation are crucial [8, 5]. Hospitals are crowded with the exponentially growing number of patients as available resources are limited [8].

Currently, reverse transcription polymerase chain reaction (RT-PCR) is the gold standard used to diagnose COVID-19 infection [4, 10, 5, 6]. However, the results of RT-PR can be affected by sampling errors and low viral load [11].

As a result, these tests suffer from high rates of false negatives [3, 12, 6, 13, 10, 14, 15] (with sensitivity of 71% [4] or 69% [5]) and may need to be conducted two or more times before finally being confirmed [16, 17, 18].

In many articles, chest imaging is considered a suitable tool for early COVID-19 screening [1, 19, 20, 21]. The point is that sensitivity of CT scan tests can reach 98%, which is much higher than RT-PCR tests [22]. Moreover, due to the fact that on CT images ground-glass opacities are visible earlier than pulmonary consolidation [23, 24], radiologists can assess the stage of COVID-19. Some studies [12, 25], indicate that distinguishing between COVID-19 and viral pneumonia is a challenging task. However, it is worth noticing that chest imaging is much faster than RT-PCR since the former takes approximately 15 seconds [26] (in terms of X-ray) or 15 [27] to 21.5 [28] minutes on average (for CT) to complete. On the other hand, RT-PCR takes from several hours up to several days [2, 29, 30, 31]. This means that the patient has to be isolated until receiving the test result [32]. In addition, X-rays are cheaper, more available worldwide, and less harmful than CTs [33] because the radiation dose is smaller [19]. Due to the existence of portable devices, X-ray imaging can be performed in isolated rooms, so the risk of contraction is significantly decreased [34, 35]. Nonetheless, especially on X-ray images, it is particularly difficult to assess the severity of the pathology [19], and, thus, only experts in radiology should interpret chest images [34, 36, 3]. In general, this process is faster than waiting for RT-PCR test results [37, 36, 6].

Recent applications of machine learning (ML) have gained popularity in the medical domain [38, 39, 40, 41, 42]. The performance achieved by neural networks is becoming similar to that reached by medical experts [37]. Deep learning techniques for medical images are present in classification (skin cancer [43, 44], brain disease[45]), detection (arrhythmia[46, 47], breast cancer[48, 49], pneumonia[50], ADHD[51]), segmentation (lung[52, 53], diabetic retinopathy[54]) and imaging reconstruction (magnetic resonance [55], Single Photon Emission Computerized Tomography (SPECT) [56]).

Considering the need for a highly accurate and fast diagnosis process, artificial intelligence (AI) can play a significant role in automating the detection of COVID-19 cases [12, 8, 3].

AI solutions are frequently based on complex, so-called black-box models [57]. For this reason, it is difficult to tell what factors lead to a particular model prediction. Such a lack of interpretability may be dangerous, as it may lead to biased results and incorrect decisions in real diagnostic procedures [3]. Recent development in the area of Explainable Artificial Intelligence (XAI) shows the importance of model explanations, which help to avoid erroneous predictions [58, 8, 6]. Nevertheless, surprisingly, in the area of COVID-19 image analysis, there are still only few results concerning the use of XAI for lung image analysis [3].

In this paper, we will summarize the most recent publications about lung imaging analysis (section II-III), and show how explainable AI techniques are used in these solutions (section IV). We will confront these approaches with the domain knowledge of radiologists and show how many of the assumptions about data, models or explanations made in many of the analyzed studies are incorrect. Finally, we will construct a checklist to help model developers assess whether they avoided the most common errors. We believe that this criticism, together with the proposed checklist, will contribute to building better models not only for the diagnosis of COVID-19 disease, but also for other applications based on lung images.

2 Methods

2.1 Literature search

This research is based on a systematic literature review. The data was collected between the 1st and 14th of August 2020. The search was performed according to Preferred Reporting Items for Systematic reviews and the Meta-Analyses (PRISMA) statement [59] presented in Figure 1 in the following academic digital databases: ArXiv, IEEE, Google Scholar, PubMed, Science Direct, Scopus, Web of Science. All studies written in English, regardless of the publication status (preprint, peer-reviewed, or published articles), were included in this review. Studies were identified by the combination of keywords: ("XAI COVID-19" OR "explainable artificial intelligence COVID-19" OR "explainable COVID-19" OR "explanations COVID-19" OR "interpretable COVID-19" OR "interpretations COVID-19" OR "transparent COVID-19") and ("X-ray" OR "radiography" OR "CT" OR "computed tomography"). Then, each study was screened for content relevance.

From 31 collected works, during eligibility checking, 4 studies were dropped due to irrelevant scope or lack of XAI parts. There were 27 studies included for qualitative synthesis. The number of studies considered in the review is vast enough to create a representative set/collection for further investigation. Some of the studies were published as preprints, not as camera-ready articles. We included them to show a variety of considered XAI approaches.
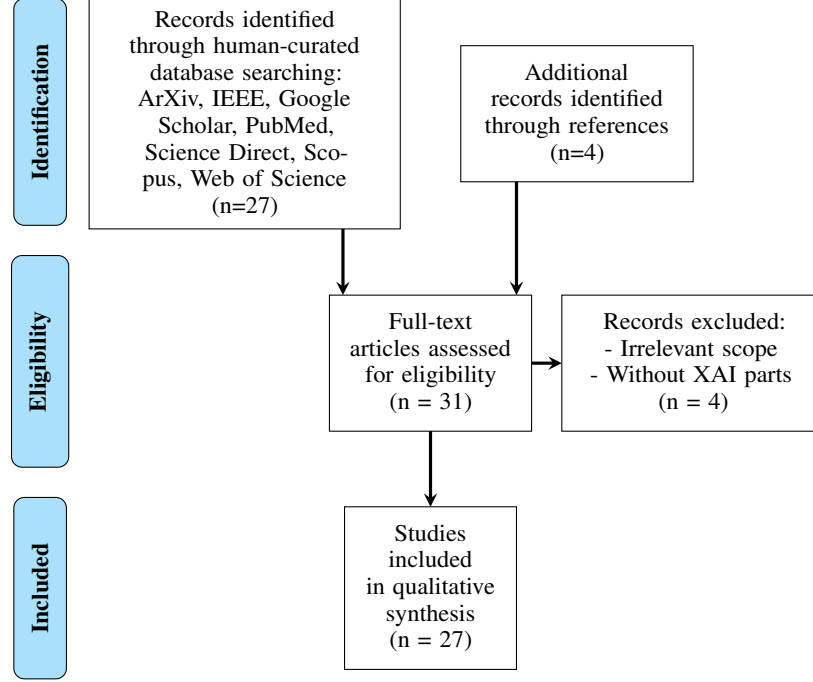
Figure 1: PRISMA Flow Diagram shows the flow of information through the different phases of a systematic review including inclusions and exclusions.

2.2 Types of COVID-19 prediction tasks

The 27 studies identified according to our methodology concern various types of ML tasks. In Figure 2, we introduce a taxonomy for these studies. In the following sections, when discussing specific solutions, it is important to remember what kind of problems they are designed for.

The first breakdown concerns the types of tasks, such as: classification (COVID-19 vs. other classes), severity assessment or segmentation. Among classification tasks, the first group is related to the detection of COVID-19 cases. The goal of the second is to assess how severe the changes caused by COVID-19 are.

In the reviewed works, segmentation can be regarded as an image preparation technique for the further classification process. Lung segmentation is present in works [4, 5, 6, 19, 27] and lesion segmentation in [13, 6, 60]. The lungs are segmented to remove the unnecessary background because, based on medical experience, the lesions caused by COVID-19 are not located outside the lungs. Lesion segmentation, also called infection mask, helps to train the model to recognize infected regions and can be beneficial for further model assessment. Only radiologists can prepare trustworthy and high-quality lung lesion masks, not an automatic tool similar to those in the reviewed works. However, manual mask preparation takes a lot of time and money and requires a high level of consistency among annotators, but, surely, it is the most valuable method for qualitative and quantitative XAI evaluation.

There are not any strict guidelines on how many classes the classification should be conducted on. In classification problems, the number of classification classes varies between studies. This distinction is particularly important when comparing model performance. Binary classification task detecting COVID-19 and non-COVID-19 is the most popular [1, 8, 2, 4, 10, 25, 12, 61, 58, 62, 63, 60]. Another frequently used method is three-class classification [10, 36, 64, 13, 61, 58, 63, 32, 34, 65]: no infection/no pneumonia, pneumonia (bacterial or none-COVID-19 viral infection), and COVID-19 (COVID-19 viral infection). In the four-class classification problem, there are two different approaches to splitting images. The first one divides images into: normal, bacterial pneumonia, non-COVID-19 viral pneumonia, and COVID-19 [37, 63], while the second one into: normal, bacterial pneumonia, tuberculosis (TB), and viral pneumonia. In the second approach, study [5] refers to [66] and claims that differentiation between viral pneumonia and COVID-19 is challenging, because similar radiological features exist between them. In classification tasks with the biggest number of classes (5), the authors distinguish five different classes: normal, pneumonia, virus, bacteria, and COVID-19 [67]. In [3] the authors applied multi-label classification. For example, for an image with the lungs of a patient suffering from COVID-19, the correct classification should predict all three labels: pneumonia, viral pneumonia, and COVID-19.
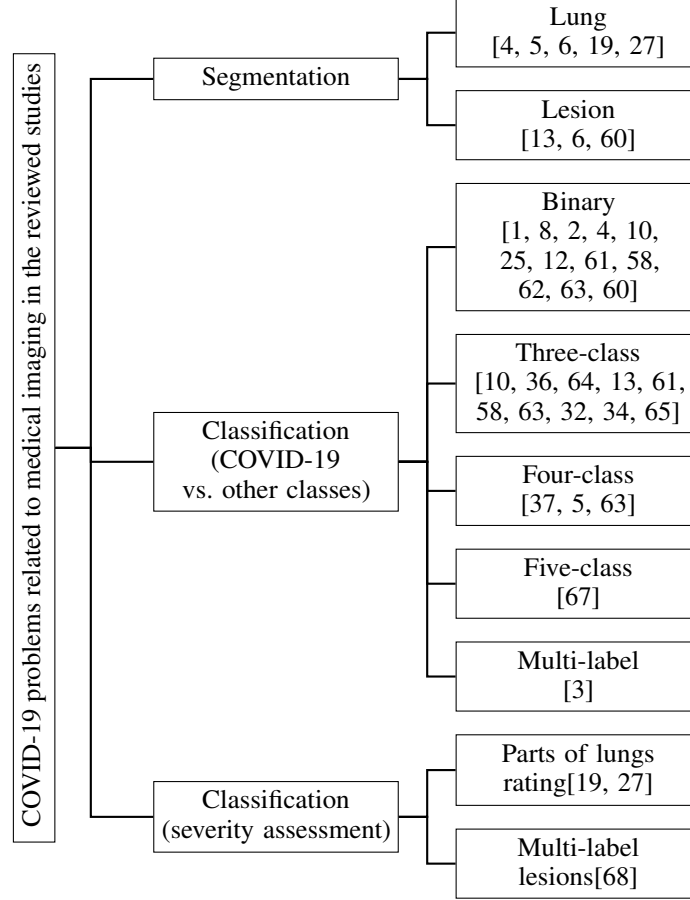
Figure 2: Taxonomy of AI applications in 27 reviewed studies

In the segmentation for severity assessment, there are two different approaches. One rates how severe the lesions are in parts of the lungs. Another one assigns labels to lesions' names that point out which changes in the lungs are present in the image.

In two studies [27, 19] images or parts of images were divided into classes that correspond to the severity of COVID-19 effects on the lungs. The authors [27] propose dividing each lung horizontally into 4 parts, and giving them a grade of 1 if it contained any lesions, such as consolidation or ground-glass opacities, or 0 otherwise. Then, the grades from all parts were summed up. Based on grades, the following scale was prepared: Normal-PCR+ 0, Mild 1-2, Moderate 3-5, and Severe 6-8. A similar solution was introduced by [19]. In this study, each lung was divided horizontally into 3 parts and a 4 grade scale was used for each part. The division of each lung into 3 parts (upper, middle, and lower lung field) resembles natural lung structure and common radiological practice.

An example of multi-label classification is described in [68] where eighteen outputs of neural network (such as atelectasis, consolidation, infiltration, pneumothorax, edema, emphysema, fibrosis, fibrosis, effusion, pneumonia, pleural thickening, cardiomegaly, nodule, mass, hernia, lung lesion, fracture, lung opacity, and enlarged cardiomediastinum) give information about lesions in lungs. Some of them are characteristic for specific illnesses. This is a very promising approach because with XAI visualizations it should be clearly visible for radiologists if a model learned to recognize proper lesions.

3 Chest image data

3.1 Data resources

The COVID-19 virus is a relatively new disease, and, in several articles, the lack of high-quality medical imaging databases is indicated [13, 27]. To check this, we gathered the data sources used in the reviewed studies in Table 1.

The last column specifies the format of images presented in the selected database. For medical imaging, the standard format for storing data is DICOM (Digital Imaging and Communications in Medicine). The most important feature of this format is a 12 to 16-bit grayscale color palette and the transferability between different medical devices. This means that image conversion from DICOM to a format such as JPG and PNG drastically reduces its quality, for example, from $2^{12} = 4096$ or $2^{16} = 65536$ to $2^8 = 256$ pixel intensity for JPG and PNG.

No publicly available COVID-19 data sources with images in DICOM format. Surprisingly, the only four repositories with the DICOM extension presented in Table 1 contain only non-COVID-19 cases. COVID-19 images shared in databases are only in JPG or PNG formats. Thus, the applied image compression may cause difficulties in further analysis. As expected, some images in databases have very poor quality, which may be the result of format conversion. The other reason may be the poor technical quality of X-ray image itself. Such images should be rejected as non-diagnostic.

Too few images with low and moderate severity cases. Most studies are based on data sources publicly available on the Internet on a popular sharing platform, such as GitHub and Kaggle. The most commonly used data source with COVID-19 cases was created at the beginning of the epidemic. The first publicly available repository was published on January 15, 2020. In [27] it is stressed that available data sources contain too few images with low and moderate severity cases. Most of the data sources have only class labels without fine-grained pixel-level annotations [60].

Class imbalance. Image format is one problem, while the class imbalance in these databases is another problem. Table 2 shows the number of cases in particular classes. The last row with aspect ratio shows the proportion of the COVID cases to non-COVID cases.

Relatively low number of COVID-19 images. The median number of COVID-19 images in the considered data resources is 250. With so little data, it is difficult to train a deep neural network (DNN). The use of imbalanced datasets requires more attention during the model training. Either proper data resampling (oversampling, undersampling) should be applied, or an appropriate loss function should be chosen, unless acquiring a greater amount of less common data is possible. However, most ML algorithms do not work very well with imbalanced datasets [78].

The data sources lack descriptions. At a minimum, the description of the dataset should include the following factors. First of all, the total number of images and the number of images in each class should be given. Additionally, the balance in terms of age and sex is another important factor because of the differences in anatomy. Information about smokers or previous lung diseases is also relevant. For analyzing model responses, the information about concurrent diseases, the severity of COVID-19, and the number of days between the exposure and the acquisition of the image of the chest are also useful.

Mix of CT and X-ray images. The problem that we found in these datasets is the data purity. If we look closer at the images presented in study [25], it appears that CT and X-ray images are mixed in the X-ray dataset. These two techniques are so different that networks for CT and X-ray images should be trained separately.

Inappropriate CT windows. For COVID-related lung analysis, it is essential to have Hounsfield Units equivalent for "lung" window (width: 1500, level: -600). Otherwise, the lung structures are obscured or not visible at all, such as some examples in studies [1, 8, 10]. This is a basic, but key, issue because we do not want to assess soft tissues or bones. Photos taken in other windows do not have any real diagnostic value.

"Children are not small adults". The next problem is related to the mixture of images of patients of different ages. There are crucial differences between the X-rays of children's and adults' chests: technical (hands are often located above the head), anatomical (different shape of the heart and mediastinum, as well as bone structures), and pathological (different pathologies). This will also include a different course of infectious diseases, with the most vivid example of round pneumonia mimicking tumors [79, 80, 81]. When we go back to the databases, it appears that, in some cases, the X-rays of children and adults are mixed [82].

Incorrect categorization of pathologies. We have noticed that many images are incorrectly categorized - into normal or pathologic, and also within the class of pathology. An additional problem is that, from a medical point of view, some images should be multi-categorized. This means that there is more than one pathology in one image. For instance, pneumonia (main class) can manifest itself as lung consolidations, which, however, can also appear with pleural effusion or atelectasis (two additional classes). On the other hand, atelectasis itself, with a mediastinal shift, can be a sign of a different pathology, such as a lung tumor. Thus, databases should be verified by experienced radiologists for proper categorization and maybe a rejection of multi-class images. This, however, would be time-consuming and - what is more important - very difficult or impossible with non-DICOM images.

Lack of information about chest projection for X-ray imaging. There are two main chest projections, see Figure 3, Posterior-Anterior (PA) and Anterior-Posterior (AP). The first one is acquired while the patient is standing. The X-ray beam comes through the patient's chest from its back (posterior) to front (anterior) - i.e., PA projection. The second one

Table 1: This table presents the data sources used in studies reviewed in this survey. For each data source, we list articles that use it. In the case of COVID-Net, please note that it is not a data source, but a study collating 5 datasets. Some other studies refer to it instead of referring to the original source.

Institution	Link to dataset	Used in article	Format
University of Waterloo	`github.com/lindawangg/COVID-Net` [34]	[13, 58, 19, 34]	
University of Waterloo	`github.com/agchung/Figure1-COVID-chestxray-dataset`	[64, 58]	JPG & PNG
University of Waterloo	`github.com/agchung/Actualmed-COVID-chestxray-dataset`		PNG
Qatar & Bangladesh Universities	`kaggle.com/tawsifurrahman/covid19-radiography-database` [69]	[64]	PNG
University of Montreal	`github.com/ieee8023/covid-chestxray-dataset` [70]	[67, 2, 3, 68, 37, 25, 64, 12, 5, 61, 58, 19, 63, 32, 65]	JPG & PNG
National Institutes of Health	`kaggle.com/c/rsna-pneumonia-detection-challenge` [71]	[68, 64, 58, 19]	DICOM
National Institutes of Health	`nihcc.app.box.com/v/ChestXray-NIHCC` [71]	[68, 61, 58]	PNG
National Institutes of Health	`kaggle.com/nih-chest-xrays/sample` [71]	[2]	PNG
University of Montreal	`kaggle.com/praveengovi/coronahack-chest-xraydataset` [70]	[67]	JPG & PNG
University of California San Diego	`kaggle.com/paultimothymooney/chest-xray-pneumonia`	[3, 37, 5, 63, 65]	JPG
University of California San Diego	`github.com/UCSD-AI4H/COVID-CT` [72]	[1, 8, 25, 12]	JPG & PNG
University of California San Diego	`data.mendeley.com/datasets/rscbjbr9sj/2` [69]	[32]	JPG
Elazig in Turkey	`github.com/muhammedtalo/COVID-19` [61]	[2, 64]	JPG & PNG
National Library of Medicine	`openi.nlm.nih.gov`	[68, 5]	DICOM & PNG
Stanford University School of Medicine	`stanfordmlgroup.github.io/competitions/chexpert` [73]	[68, 19]	JPG
Hospital San Juan de Alicante - University of Alicante	`bimcv.cipf.es/bimcv-projects/padchest` [74]	[68]	PNG
Hospital Universitario San Cecilio	`github.com/ari-dasci/OD-covidgr` [27]	[27]	JPG
Beth Israel Deaconess Medical Center in Boston	`physionet.org/content/mimic-cxr/2.0.0` [75]	[68]	DICOM
Società Italiana di Radiologia Medica e Interventistica	`sirm.org/category/senza-categoria/covid-19`	[62]	JPG & PNG
National Cancer Institute	`wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI` [76]	[4]	DICOM
open-edit radiology resource	`radiopaedia.org`	[62]	JPG
generated using data augmentation	`kaggle.com/nabeelsajid917/covid-19-x-ray-10000-images`	[64]	JPG & PNG
offline database or from hospital		[4, 10, 6, 60]	
lack of information		[36, 77]	

is the opposite - the beam enters through the front (anterior) of the chest and exits out of the back (posterior) - i.e., AP projection. This type of examination is mostly conducted in more severe cases, with lying patients, with comorbidities, often in Intensive Care Units. As the X-ray beam is cone-shaped, both projections have one very important difference, which is the size of the heart. In PA projection, the heart is close to the detector, resulting in a similar heart size on the X-ray as in reality. In AP projection, the heart is away from the detector, resulting in a larger heart size on the X-ray, which can be confused with cardiomegaly. In databases, AP and PA images are often mixed, which can cause bias because AP projections are performed on severely ill patients [27]. From a medical point of view, it is

Table 2: Class balance in the reviewed studies. The class balance is crucial developing an accurate model. In the following rows there are presented: the number of COVID-19 images in the study, the total number of images, the number of classes into which images were divided, and aspect ratio. The aspect ratio is calculated by dividing the number of COVID-19 images by the total number of images and then multiplying it by the number of classes. The biggest collected COVID-19 dataset and the largest total number of images are marked in bold. The smaller the aspect ratio, the less COVID-19 cases participate in the whole study's dataset, and vice versa. For this reason, the best-balanced dataset (nearest 1) is marked in bold. Studies which do not include full information about the number of cases were excluded.

Study	[1]	[67]	[2]	[3]	[37]	[4]	[10]	[36]	[64]	[13]	[12]	[5]	[6]	[61]	[62]	[27]	[63]	[32]	[34]	[60]	[65]
Number of COVID-19 images	400	58	250	234	68	829^{*2}	230	200	**855**	99	345	502	**3 389**	127	120	377	112/137	76	358	400	269
Total number of images	800	2 800	6 523	1 234	5 941	$1\ 865^{*2}$	460	4 200	**15 959**	18 529	720	1 004	2 186	1 125	239	754	366	5 949	**13 975**	750	5 801
Number of classes	2	4	2/3	12^{*1}	4	2	3	3	3	3	2	4	2	3	2	$\frac{5}{4}^{*3}$	3/2	3	3	2	3
Aspect ratio	**1.00**	0.08	0.08/0.11	2.28	0.05	0.89	1.50	0.14	0.16	0.02	**0.96**	2.00	1.36	0.34	**1.00**	0.63	0.92/0.75	0.04	0.08	1.07	0.14

^{*1} multilabel classification, ^{*2} training slides (106 COVID-19 images), ^{*3} four COVID-19 classes (NormalPCR+: 76 cases, Mild cases: 80, Moderate: 145 cases, Severe: 76 cases) and 1 Negative (377 cases)

impossible to perform chest X-rays in only one projection as this depends on patients condition. However, projection should be specified for every X-ray in dataset, and possible bias in model classification should be evaluated.
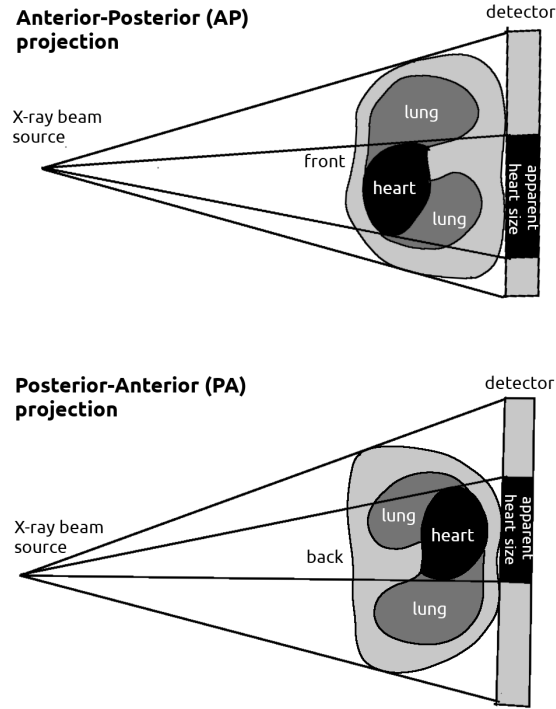

Figure 3: Differences between AP and PA chest projections

To sum up, this section shows that data sources have several weaknesses. First of all, images available for COVID-19 in public databases are in a low-quality format because DICOM images are still not publicly available for this disease. Secondly, in the data sources, there is missing data (i.e., chest projection) or poor quality data (i.e., poor image quality, inappropriate CT window, mixed CT and X-ray images or incorrect pathology categorization). Lastly, during data preparation, it should be taken into account that DNNs work better when the class balance is maintained.

3.2 Image preprocessing methods

The aim of preprocessing is to make the images from different data resources look homogeneous and coherent. This process reduces the possibility of bias via eliminating some artifacts from images, such as captions, annotations, which may lead to deceiving the model. The model should learn how to differentiate labels by focusing on image features, not by recognizing from which database the image comes from. During preprocessing, irrelevant image features that are easier to learn are removed. This is because in some databases there are no cases of people suffering from COVID-19, while in others there are, for example, only serious cases. These differences, which are insignificant from a human perspective, must be eliminated. For machines, even the information that images from one data resource are relatively darker might be relevant.

However, due to a large amount of data, automation of preprocessing is necessary. Preprocessing cannot introduce any changes in an image which will add or remove some relevant information. Its purpose is to make it impossible to identify the machine or characteristic machine's calibration parameters, e.g., the dose of exposure.

Table 3 lists preprocessing techniques used in the reviewed studies. The most common was resizing an image to the same size. It is the most basic operation needed to train DNN when images have different sizes. Other techniques applied frequently to images were: normalizing pixel intensity, changing color space, eliminating noise, equalizing histogram, and performing image enhancement. Unfortunately, in 11 out of 27 studies there was no information about preprocessing steps provided. [83] also stressed that many studies do not contain sufficient information about preprocessing, such as cropping of images.

Table 3: Image preprocessing techniques in the reviewed studies

Preprocessing technique	Reference
Resize to the same size	[67, 8, 3, 68, 37, 25, 64, 12, 5, 6, 62, 63, 32, 65]
Normalize pixel intensity	[3, 68, 37, 64, 62, 19, 63]
Change color space	[25, 64, 62, 32]
Eliminate noise	[64, 12, 19, 27]
Use Perona-Malik filter	[64]
Limit image intensity	[3, 19]
Equalize histogram	[67, 64, 5, 19]
Perform image enhancement	[64, 12, 5, 6]
Crop image	[67, 3, 27]
Cast data type	[5, 19]
Zoom image / augmentation	[62, 8]
Add pixels	[27]
Feature encoding	[65]
Rotate image	[8]
Use 2D wavelet transform	[12]
Feature extraction	[13, 65]
Lack of preprocessing or description	[1, 2, 4, 10, 36, 13, 61, 77, 58, 34, 60]

Cropping, changing color space, proportionally resizing, or zooming can be helpful to adjust images for training on specific network architecture, or the easiest way to remove some descriptions from the edges of images. If not required, resizing ought to be omitted. Normalizing pixel intensity or equalizing histograms are required to eliminate strong correlations with specific machine settings. From a technical perspective, spot changes, such as noise removal, are not desirable. Images in JPEG or PNG formats are highly compressed, so the noise should not be present. For DICOM images, these techniques are not forbidden, but they should be used carefully in order not to remove important features, such as lesions or parts of them.

To sum up, preprocessing is an important step preceding model training. It should reduce the possibility of bias and guarantee more homogeneous images without the elimination of any medically significant features.

3.3 Data augmentation

Data augmentation for ML is a technique that artificially multiplies the number of images through cropping and transforming existing images or creating new synthetic images thanks to generative adversarial networks. This procedure may help to reduce model overfitting and the problem of class imbalance [84]. It helps in achieving a larger training dataset and more robust models.

In Table 4, we summarized data augmentation techniques from the reviewed studies. The most popular augmentation techniques in the reviewed studies are affine transformations, such as rotation, scaling or zooming, flip, and shifting or translation. On the contrary, splitting a radiological image into overlapping patches, or generating new content via a type of Generative Adversarial Network are rarely used.

Table 4: Data augmentation techniques used in studies. Some techniques are parametrizable, so the table indicates the techniques and parameters used. An indentation is used to show the subtypes of the method.

Data augmentation technique	Values and studies
Affine transformations	[10]:
Rotation	[4, 58, 63, 34], 5°[67], 15°[2, 64, 62], 20°[37], 25°[19], 5-35°[36]
Scaling / Zooming	[37, 58, 34], 10% [67, 25, 19], 20% [63]
Horizontal, vertical	20% [36]
Flip	[37, 25, 13]
Horizontal	[67, 4, 58, 34, 60]
Vertical	
Shifting / Translation	[37, 58, 34], height 5% [67], 10% [19]
Shearing	[37, 63, 32]
Crop	[4, 60]
Color jitter (brightness, contrast, saturation, and hue)	[10]
Brightness change	[58, 34] +/-30 [32], 10% [67]
Gaussian noise	[32]
ZCA whitening transformation	[37]
Elastic transformation	α=60, σ=12 [19]
Grid distortion	steps=5, limit=0.3 [19]
Optical distortion	distort=0.2, shift=0.05 [19]
Warping	10% [25]
Multiple patches from each image	[5]
Class-inherent transformations Network*	[27]
Lack	[1, 3, 68, 12, 6, 61, 77, 65]
Not specified	[8]

* inspired by Generative Adversarial Networks

However, not all of them are appropriate from a medical point of view. Before an augmentation, it is recommended to consider the 'safety' for the chosen domain. For example, the rotation should be done carefully, because some parts of the lungs, such as costophrenic recesses, may be placed outside the image. Also, change of brightness or contrast should be performed only in a limited manner, as greater manipulation may obscure lung structure. Moreover, an image can be cropped or proportionally scaled/zoomed to an extent that it displays only the lungs without a background or other parts of the body.

It is also worth noting, that in the case of CT and X-ray images, the augmentation based on rotation does not generate photos that we could naturally expect in the dataset, because the process of taking the photo itself is standardized.

In general, all augmentation methods should be consulted with radiologists, as domain knowledge is crucial. In every project, it is important to know the field of research to avoid a situation in which instead of solving the problem, bias is accidentally introduced.

3.4 Model architecture

In the studies different approaches of modeling were applied. Some benefited from machine learning methods, whereas the rest used deep learning. In the first case, simple classifiers or their ensembles were applied: AdaBoost [8], Deep tree [36], KNN [8], Naive Bayes [8], SVM [8].

In the reviewed studies, lung-specific model architectures (own models) were relatively often used for classification, whereas the existing architectures were frequently fine-tuned. The following model architectures or their fine-tuned, modified versions were investigated: ResNet [64] (ResNet18 [3, 13, 5, 19], ResNet34 [3, 25, 6], ResNet50 [1, 67, 4, 36, 25, 12]), DenseNet [68, 64, 58, 19] (DenseNet121 [67, 62], DenseNet-161 [3], DenseNet-201 [1, 25, 62]), VGG [64, 19] (VGG-16 [1, 2, 62, 65],VGG-19 [1, 25, 62]), Inception [19], InceptionV3 [3, 36, 63], InceptionResNetV2 [1, 62], MobileNetV2 [1, 62], NasNetMobile [1, 62], EfficientNet-B0 [25], Efficient TBCNN [67], MobileNet [62],

NasNetLarge [62], Res2Net [60], Residual Attention Net [62], ResNet15V2 [1], ResNet50V2 [37], ResNeXt [3], WideResNet [3], Xception [62], own model [8, 10, 25, 61, 19, 27, 32, 34]. It is clearly visible that there are numerous types of neural networks. Different neural networks can catch different dependencies in the data. For solving a problem, many types of model architectures are tested to find the best one for a specific task. Recommendations on how the explanations should look do not depend on the neural network architecture.

For segmentation, the following architectures were used: U-net [4, 19, 27], AutoEncoder [13], VGG-16 backbone + enhanced feature module [60], (FC)-DenseNet-103 [5], Nested version of Unet (Unet++) [19], VB-Net [6]. During the segmentation process, it is important that the lungs are accurately segmented. Otherwise, this can be an indication of pathology. In study [5], the authors were aware that their segmentation cut pathological changes in lungs. In study [19] segmentation for non-domain experts appears accurate. However, radiologists noticed that also other structures (i.e., bowel loops) were interpreted as lungs in that segmentation.

There are multiple purposes for creating new model architectures. The most common is adjusting existing architectures for better explainability or scalability for training on medical COVID-19 imaging [8, 10]. For example, in studies [25, 34], the authors conducted tests and chose the advantages of many architectures while creating their own. The proposed architectures are usually smaller and require a lower number of trainable parameters than in well-known DNN architectures [27, 32].

Five studies published their code on GitHub: [67, 68, 58, 63, 34]. Other studies did not include any reference to their code or model.

Often the prediction from multiple models is combined to improve the overall performance. However, surprisingly, in the reviewed studies, there were not many ensemble models.

3.5 Transfer learning

Transfer learning is an ML technique about reusing gained knowledge from one problem to a new one. In the reviewed studies, it is commonly used when the neural network has a large number of parameters or the number of collected samples is too small for a specific task. In such a case, fewer training epochs are needed to adjust the model to a particular task. There are several popular image databases: ImageNet and, NoisyStudent for which various architectures of pre-trained neural networks are available. Transfer learning on ImageNet database was utilized in the following studies: [1, 67, 8, 2, 25, 5, 62, 27, 63, 32, 34, 65]. Twelve out of 27 studies decided to use a neural network pre-trained on ImageNet for transfer learning. Therefore, it can be said that this is a very common procedure.

However, as [85] shows, it is not clear whether using ImageNet for transfer learning in medical imaging is the best strategy. ImageNet consists of natural images. Meanwhile, medicine is an entirely different field and is completely unrelated. [86] also stressed the fact that the features which are extracted by models pre-trained on ImageNet can introduce bias.

Only in three reviewed studies was transfer learning conducted on lung images. The chosen datasets included 112,120 in [58], 88,079 in [68], and 951 in [19] non-COVID-19 lung images. The study [64] did not perform any transfer learning because chest lung images lack colorful patterns, specific geometrical forms, or similar shapes. In another study [67], the authors discovered that the model has better performance when pre-trained on ImageNet than without it. However, the authors found out that their models pre-trained on ImageNet were using irrelevant markers on lung images while making a prediction.

Especially when the model is trained on a small amount of data, the usage of completely irrelevant features from another pre-trained model may increase model accuracy/result. For this reason, it is crucial to find a large database with images similar in domain and appearance to limit the possibility of irrelevant markers that take part in a prediction. It is recommended to train a neural network on this database and then use transfer learning to adjust it to the target task.

For transfer learning, it is recommended to take into consideration the following X-ray data sources with DICOM images (consider the fact that, in some of them, children and adults lungs are mixed): U.S. National Library of Medicine[1] (7 470 images), Radiological Society of North America[2] (29 684 images), Society for Imaging Informatics in Medicine[3] (3 209 images), Medical Information Mart for Intensive Care[4] (377 110 images). For transfer learning on CT, the following data sources are available: The Reference Image Database to Evaluate Therapy Response[5] (15 419 images), A Large-

---

[1]pubmed.ncbi.nlm.nih.gov/25525580

[2]kaggle.com/c/rsna-pneumonia-detection-challenge

[3]kaggle.com/c/siim-acr-pneumothorax-segmentation

[4]physionet.org/content/mimic-cxr/2.0.0

[5]wiki.cancerimagingarchive.net/display/Public/RIDER+Lung+CT

Scale CT and PET/CT Dataset for Lung Cancer Diagnosis[6] (260 826 images), The National Lung Screening Trial[7] (21 082 502 images).

3.6 Training parameters

The selection of hyperparameters has a large impact on model results. Nevertheless, the process of tuning parameters is empirical and depends on the model architecture. For this reason, it is difficult to present a set of parameters adequate for every model architecture. However, there are several tips which can be used for most models.

Often the learning rate is decreased during the training process. Sometimes callback functions are used to halt training, when the result of a model is optimum, and during the training process, to save and store the best model and its parameters. The most typically used optimizer is Adam[87]. The batch size of images during model training is between 2 and 81 with the most common value 8.

The whole image dataset is typically divided into 3 or 2 sets, most commonly into: training set 80%, validation set 10% and testing set 10% [58, 27, 32, 65]. Proportion 80% to 20% was the most frequently used among divisions into training and testing set respectively [1, 37].

In study [1], the recommendation to conduct external validation is indicated, meaning an evaluation on an independent database. Another public dataset will be the best choice for cross-database validation [5]. However, in the reviewed studies cross-validation is the most frequently used. It is a common choice for training on a small amount of data resources. The problem which may occur during cross-validation is overfitting to the data. For this reason, validation on an external resource is the most trustworthy method.

3.7 Model performance

Evaluation metrics are commonly used to compare different models. For DNN image classification, there are many metrics frequently used for model quality assessment. In the reviewed studies, we discovered a large variability in the number of reported metrics. It is a common situation due to the fact that there are no detailed recommendations as to which performance metric should be used. We recommend the instructions presented in [88], but, unfortunately, in almost all the reviewed studies, at least one metric out of these recommendations was missing.

There is another important factor which indicates why more than one evaluation metric should be used. It provides the opportunity to compare model architectures and then choose the best one for a given problem. Nevertheless, the models were not trained on the same images. Some databases contained only severe cases which were easier to classify [27]. Even if studies refer to the same data resources, it is possible that the amount of data has increased over time. For this reason, it is rather difficult to make a reliable comparison. The most trustworthy way to compare different model architectures is to look at studies which tested many of them, i.e. [62, 1, 8, 25, 19, 27].

4 Explainable artificial intelligence

4.1 The need of model explanations

Recently, in many applications, we observed an increasing number of complex AI models. They work below expectations, discriminate, or do not work at all on the new dataset. Very often it turns out that the reason for these problems is insufficient validation of models or relying on simple quality metrics for black-box models, which creates an illusion of effectiveness.

The answer to these problems is an in-depth diagnosis of complex models, which is offered by solutions from the area of Explainable Artificial Intelligence. It enables the verification of models by field specialists who can check if the model works according to their expectations [89, 90, 91], provide more insight for model decisions [6] and check model robustness [61]. In this section, we will see how these solutions work in case of image analysis for COVID-19.

In [2], the authors conclude that XAI is necessary to establish where the network is "looking" in the input image, which series of neurons are activated in the forward-pass during inference/prediction, and how the network arrived at its final output. It is also important to verify that the model is not making a decision based on inappropriate regions of the medical image [58] and to identify the critical regions on the patients' chest [64] to aid clinicians in making faster yet accurate assessments [34]. It may seem bold that a critical assessment of explanations can potentially help

---

[6]wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=70224216
[7]wiki.cancerimagingarchive.net/display/NLST

clinicians discover new insights into the key visual indicators associated with COVID-19 infection, which then can improve screening accuracy [34].
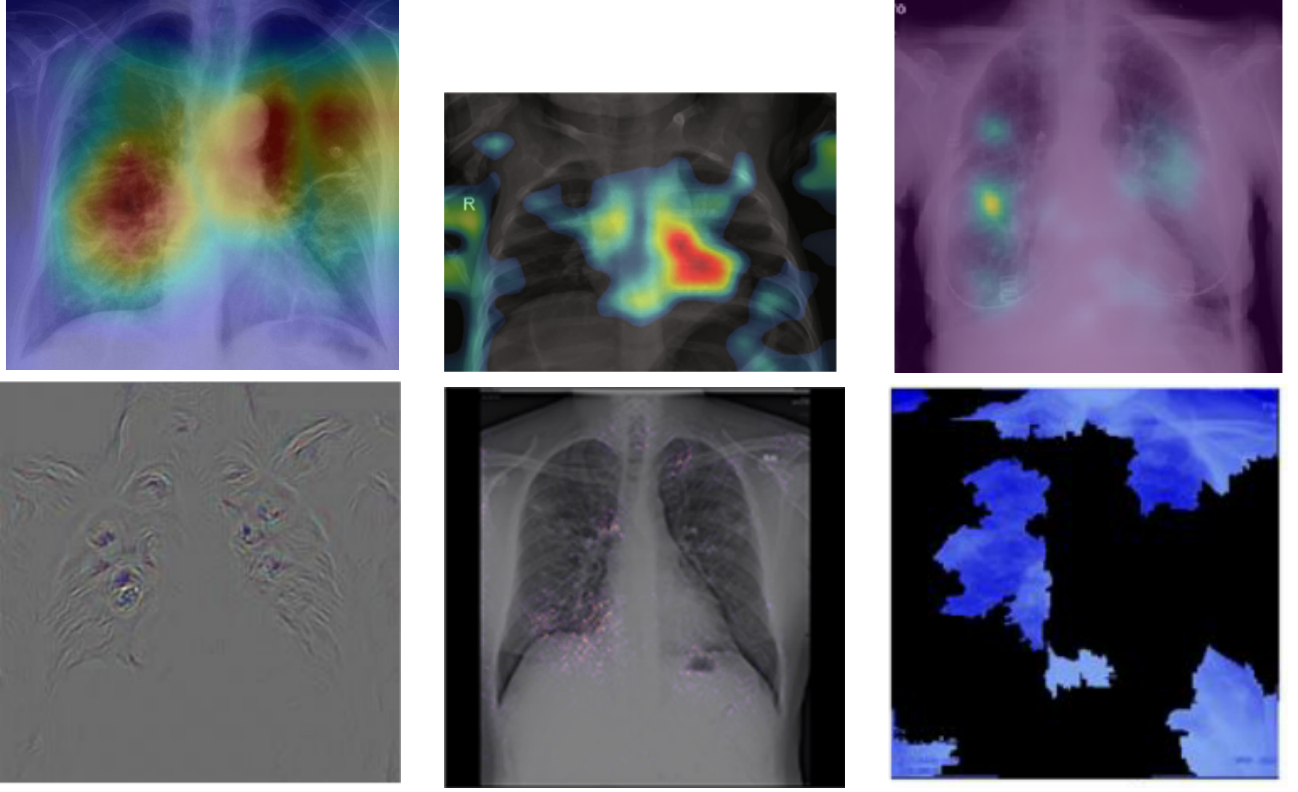

Figure 4: Examples of explanations for COVID-related models from studies: [27, 32, 68, 37, 3, 1]. The following explanations are used (in a reading direction): Grad-CAM, CAM, saliency, guided backpropagation, integrated gradients, LIME. Such explanations can be divided into 4 types: heat maps (image from 1 to 3), contour lines (image 4), points (image 5), and image pieces (image 6).

4.2 XAI methods used in the reviewed studies

The area of model explanations and the number of methods that can be used for this purpose are increasing rapidly [90, 91]. Such methods differ in properties; they work either for a single image (so-called instance level methods) or globally for the whole dataset. Some of them are based on gradients, others on interpretable features, some are intrinsic or post-hoc, model-specific (class-discriminative, high-resolution, multi-layer) or model-agnostic.

Table 5 shows which approach to model explanation was used in which study. The most popular in the reviewed studies was Grad-CAM (Gradient-weighted Class Activation Mapping)[92]. Grad-CAM is a gradient-based, post-hoc, model-specific, and class-discriminative explanation method. It uses the gradient information which flows to the first convolutional layer of the CNN network. This method shows important regions on a coarse heat map. Its popularity may be related to the fact that colorful heat maps are easy to implement and seem to be readable. An example of an implementation for Grad-CAM is available online, and its use on melanoma images shows great results.

Another very popular method is LIME (Local Interpretable Model-agnostic Explanations)[93]. This method finds a set of superpixels that are most closely associated with the predictive label. However, as clearly visible in studies [1, 25], some large superpixels include different structures (i.e., lung tissue and chest wall). However, However, this method is not accurate enough for an interpretable representation of space due to the lack of semantic meaning.

Some threads related to the application of XAI in the analyzed publications are questionable. Contrary to what [64] states, explanations of ensemble models are possible. A single best model does not have to be selected for the visualization of the prediction. There are many model-agnostic interpretation methods which can be easily used for explanations. In most XAI methods, it is possible to adjust them to suit ensemble models. Moreover, ensemble models usually outperform a single model in terms of accuracy.

Table 5: Applied XAI techniques

Grad-CAM (gradient-weighted class activation mapping)	[67, 2, 10, 36, 64, 12, 5, 6, 61, 58, 19, 27, 63]
LIME (local interpretable model-agnostic explanations)	[1, 36, 25, 19]
CAM (class activation mapping)	[1, 37, 6, 32]
Saliency (saliency map)	[67, 3, 68, 37]
Guided Backpropagation	[3, 37]
LRP (layer-wise relevance propagation)	[64, 27]
Occlusion (occlusion sensitivity)	[3, 12]
AM (activation mapping)	[60]
Attribution maps	[13]
DeepLIFT	[3]
Feature maps	[62]
Grad-CAM++	[64]
Guided Grad-CAM	[37]
GSInquire	[34]
Input X Gradient	[3]
Integrated Gradients	[3]

According to [61], the model makes incorrect decisions in poor quality X-rays. This is because the low quality or very low-resolution images do not show enough details even for the models. Such images should be removed while checking the database contents.

The study [12] noticed that the region of the lesion is marked correctly, but that model prediction is wrong. Unless we perform a quantitative and qualitative evaluation of XAI results, we will not have the opportunity to assess the trustworthiness of our model. The model may take into greater consideration other image features than it should. To explore this kind of a model mistake, other XAI methods ought to be used to obtain a better comparison possibility.

4.3 Expert based evaluation of XAI methods

In most of the reviewed studies, the application of XAI comes down to the series of colorful images without any assessment about how valid these explanations are. Colored explanations obscure the original image, which makes it even more difficult to assess their correctness. In images with XAI heat maps, it is often hard or impossible to see pathologies and guess if the model works well. Raw lung images shall be put next to explanations. Also the explanations should be interpreted or validated by radiologists. Otherwise, they are redundant and contribute nothing to the trustworthiness of the model.

Together with the radiologists, we analyzed the explanations from the discussed works. In the following paragraphs, we discuss the most common mistakes or inappropriate explanations.

In the first example, in Figure 5a), the model focuses on clavicles, scapulas, and soft tissues, which are outside the lungs. Very likely, the model predicts illness based on an improper part of the image. Location of the areas marked by explanation should be inside the chest on the lung tissue because COVID-19 lesions are not located on, e.g., lymph nodes. Moreover, there are some elements that cannot be considered as decision factors like imaging artifacts (cables, breathing tubes, image compression) or embedded markup symbols [34]. To prevent the model from focusing on irrelevant features, in some studies, the lungs were segmented, and their background was removed [4, 5, 6, 19, 27]. However, it may not help when some imaging artifacts are present in the area of the lungs.

The second example, in Figure 5b) shows that the model does not take the lesions into account. The model states that parts of the lungs other than the ones marked by the radiologist are relevant for model prediction. Explanations that "roughly indicate the infection location" [6] are not acceptable for the robust model. The model should do this with the accuracy of the pixel marked by radiologists as relevant.

The third example, in Figure 5c), visualization is not clear. The study describes a different XAI method than the one present in the image. Moreover, this visualization highlights the whole image, and it is not possible to guess which features took part in the prediction. It is important to point out that some explanation methods can give clearer results for a specific type of DNN and for a specific domain.

The last example, in Figure 5d) is blurred. The image of the lungs is improperly taken, and the process should be repeated. The current image is useless for the accurate diagnosis process. Such images should be removed during data resource verification before model training.
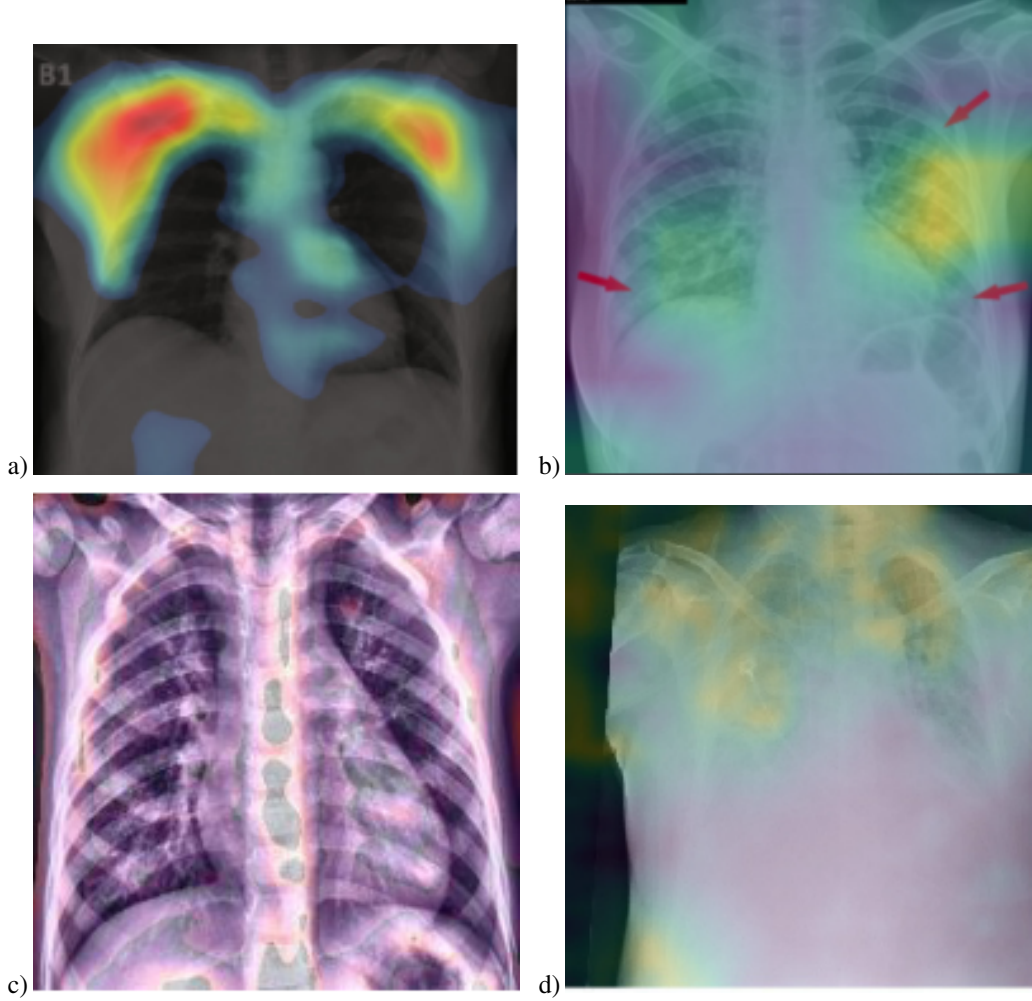
Figure 5: Examples of biased model explanations [32, 2, 67, 13] Red arrows in the image b) are marked by a radiologist to help locate the lesions. They were not present in the training set.

If the lung lesions are well described, it will be possible to prepare quantity and quality XAI assessment to score the trustworthiness of the specific model. One possible option would be to create measures for the evaluation of XAI image models based on the measures quoted in study [94]: Intersection-Over-Union and token-level, which presents measures for the evaluation of text models.

Evaluation of explanation methods is crucial for confirmation of model trustworthiness. First of all, radiologists should validate a specific model with the help of XAI. They should assess location, size, and shape of explanations. Their interpretations should contain clear references to structures and lesions in the lungs, such as posterior basal segment, ground-glass opacity, consolidation, frosted glass shadows, etc. The example of a well-prepared XAI interpretation can be found in the study [12].

4.4 The checklist for responsible analysis of lung images with deep learning models

In this work, we have shown that development of a model which analyzes lung images is a complex process. Therefore, we prepared the checklist based on the analyzed studies and the errors we found in them. In [95], it is shown that well-prepared checklists significantly improve the quality of the modeling process. They help to avoid, or quickly detect and fix, errors.

In the list below, the letter R indicates that the point should be consulted with a field expert / radiologist, and the letter D indicates that the point should be consulted with a model developer.

The points in the checklist below are grouped according to the sections' names discussed in this study. This should assist in finding a detailed description of the problem stated in the checkpoint list in the corresponding section.

- Data resources

    D Does the data format provide diagnostic quality? (DICOM is recommended)

    R Are the low quality images (i.e., blurred, too dark, or too bright) rejected?

    D Is the dataset balanced in terms of sex and age?

    R Does the dataset contain one type of images (CT or X-ray)?

    R Are the lung structures visible ("lung" window) on CT images?

    D Are children's and adults' images separate within the dataset?

    R Are images correctly categorized in relation to class of pathology?

    D Are AP/PA projections described for every X-ray image?

- Image preprocessing

    D Is the data preprocessing described?

    D Are artifacts (such as captions) removed?

- Data augmentation (if needed)

    D Are the lungs fully present after transformations?

    R Are lung structures visible after brightness or contrast transformations?

- Transfer learning (if used)

    D Is the transfer learning procedure described?

- Model performance

    D Are at least a few metrics of those proposed in [88] used?

    D Is the model validated on a different database than the one used for training?

- Domain quality of model explanations

    R Are other structures (i.e., bowel loops) misinterpreted as lungs?

    R Are areas marked as explanations inside the lungs?

    R Are artifacts (cables, breathing tubes, image compression, embedded markup symbols) misidentified as part of the explanations?

    R Are areas indicated as explanations consistent with opinions of radiologists?

    R Do explanations accurately indicate lesions?

5 Conclusion

To summarize, models without explanations create validation debt and explanations without consultations with a radiologist are just an illusion of validation. Unfortunately, only in 7 out of the 27 reviewed studies, the domain validation with the support of radiologists was used. Regarding the XAI part, the quality of output was assessed by radiologists only in 3 studies. It is important to point out that, for medical examination, the most valuable are large-resolution images, especially in DICOM format. Surprisingly, DICOM COVID-19 cases are not available in popular online datasets. However, there are 3 DICOM datasets with other lung diseases. Another difficulty is caused when raw lung images are not placed in studies. Images with visualizations (e.g., heat maps) often make the information underneath illegible.

The motivations for explaining models are commendable. Nevertheless, in many works, interpretations of explanations and summaries are missing. The XAI method is not a conclusion in itself. The fact that the model provides correct explanations for a few images does not yet show that the model works properly. It would be good to quantitatively validate XAI methods. For this purpose, the help of clinicians or proper annotations prepared by radiologists beforehand are necessary.

We believe that if the aforementioned checklist had been used in these studies, the resulting models would have had much higher quality.

References

[1] Md Manjurul Ahsan, Kishor Datta Gupta, Mohammad Maminur Islam, Sajib Sen, Md. Lutfar Rahman, and Mohammad Shakhawat Hossain. Study of Different Deep Learning Approach with Explainable AI for Screening Patients with COVID-19 Symptoms: Using CT Scan and Chest X-ray Image Dataset. 2020.

[2] Luca Brunese, Francesco Mercaldo, Alfonso Reginelli, and Antonella Santone. Explainable Deep Learning for Pulmonary Disease and Coronavirus COVID-19 Detection from X-rays. *Computer Methods and Programs in Biomedicine*, 196:105608, 2020.

[3] Soumick Chatterjee, Fatima Saad, Chompunuch Sarasaen, Suhita Ghosh, Rupali Khatun, Petia Radeva, Georg Rose, Sebastian Stober, Oliver Speck, and Andreas Nürnberger. Exploration of Interpretability Techniques for Deep COVID-19 Classification using Chest X-ray Images. 2020.

[4] Ophir Gozes, Ma Frid, Hayit Greenspan, and D Patrick. Rapid AI Development Cycle for the Coronavirus (COVID-19) Pandemic : Initial Results for Automated Detection & Patient Monitoring using Deep Learning CT Image Analysis. *arXiv:2003.05037*, 2020.

[5] Yujin Oh, Sangjoon Park, and Jong Chul Ye. Deep Learning COVID-19 Features on CXR using Limited Training Data Sets. *IEEE Transactions on Medical Imaging*, 0062(c):1–1, 2020.

[6] Xi Ouyang, Jiayu Huo, Liming Xia, Fei Shan, Jun Liu, Zhanhao Mo, Fuhua Yan, Zhongxiang Ding, Qi Yang, Bin Song, Feng Shi, Huan Yuan, Ying Wei, Xiaohuan Cao, Yaozong Gao, Dijia Wu, Qian Wang, and Dinggang Shen. Dual-Sampling Attention Network for Diagnosis of COVID-19 from Community Acquired Pneumonia. *IEEE Transactions on Medical Imaging*, 39(XX):1–1, 2020.

[7] Ahmed Al-Jabir, Ahmed Kerwan, Maria Nicola, Zaid Alsafi, Mehdi Khan, Catrin Sohrabi, Niamh O'Neill, Christos Iosifidis, Michelle Griffin, Ginimol Mathew, and Riaz Agha. Impact of the Coronavirus (COVID-19) pandemic on surgical practice, jul 2020.

[8] Plamen P Angelov and Eduardo A Soares. Explainable-By-Design Approach For Covid-19 Classification Via CT-Scan. 04 2020.

[9] World Health Organization. Coronavirus Disease (COVID-19) Situation Reports.

[10] Zhongyi Han, Benzheng Wei, Yanfei Hong, Tianyang Li, Jinyu Cong, Xue Zhu, Haifeng Wei, and Wei Zhang. Accurate Screening of COVID-19 using Attention Based Deep 3D Multiple Instance Learning. *IEEE Transactions on Medical Imaging*, XX(XX):1–1, 2020.

[11] Xingzhi Xie, Zheng Zhong, Wei Zhao, Chao Zheng, Fei Wang, and Jun Liu. Chest CT for Typical Coronavirus Disease 2019 (COVID-19) Pneumonia: Relationship to Negative RT-PCR Testing. *Radiology*, 296(2):E41–E45, 2020.

[12] Eri Matsuyama. A Deep Learning Interpretable Model for Novel Coronavirus Disease (COVID-19) Screening with Chest CT Images. *Journal of Biomedical Science and Engineering*, 13(7):140–152, 2020.

[13] Shahin Khobahi, Chirag Agarwal, and Mojtaba Soltanalian. CoroNet: A Deep Network Architecture for Semi-Supervised Task-Based Identification of COVID-19 from Chest X-ray Images. *medRxiv*, page 2020.04.14.20065722, 2020.

[14] Shuai Wang, Bo Kang, Jinlu Ma, Xianjun Zeng, Mingming Xiao, Jia Guo, Mengjiao Cai, Jingyi Yang, Yaodong Li, Xiangfei Meng, and Bo Xu. A deep learning algorithm using CT images to screen for Corona Virus Disease (COVID-19). *medRxiv*, page 2020.02.14.20023028, jan 2020.

[15] Jasper Fuk Woo Chan, Shuofeng Yuan, Kin Hang Kok, Kelvin Kai Wang To, Hin Chu, Jin Yang, Fanfan Xing, Jieling Liu, Cyril Chik Yan Yip, Rosana Wing Shan Poon, Hoi Wah Tsoi, Simon Kam Fai Lo, Kwok Hung Chan, Vincent Kwok Man Poon, Wan Mui Chan, Jonathan Daniel Ip, Jian Piao Cai, Vincent Chi Chung Cheng, Honglin Chen, Christopher Kim Ming Hui, and Kwok Yung Yuen. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *The Lancet*, 395(10223):514–523, feb 2020.

[16] Victor M. Corman, Olfert Landt, Marco Kaiser, Richard Molenkamp, Adam Meijer, Daniel K.W. Chu, Tobias Bleicker, Sebastian Brünink, Julia Schneider, Marie Luisa Schmidt, Daphne G.J.C. Mulders, Bart L. Haagmans, Bas Van Der Veer, Sharon Van Den Brink, Lisa Wijsman, Gabriel Goderski, Jean Louis Romette, Joanna Ellis, Maria Zambon, Malik Peiris, Herman Goossens, Chantal Reusken, Marion P.G. Koopmans, and Christian Drosten. Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. *Eurosurveillance*, 25(3):1–8, 2020.

[17] Daniel K.W. Chu, Yang Pan, Samuel M.S. Cheng, Kenrie P.Y. Hui, Pavithra Krishnan, Yingzhi Liu, Daisy Y.M. Ng, Carrie K.C. Wan, Peng Yang, Quanyi Wang, Malik Peiris, and Leo L.M. Poon. Molecular Diagnosis of a Novel Coronavirus (2019-nCoV) Causing an Outbreak of Pneumonia. *Clinical chemistry*, 555:549–555, 2020.

[18] Naru Zhang, Lili Wang, Xiaoqian Deng, Ruiying Liang, Meng Su, Chen He, Lanfang Hu, Yudan Su, Jing Ren, Fei Yu, Lanying Du, and Shibo Jiang. Recent advances in the detection of respiratory virus infection in humans. *J Med Virol*, page 92, 2020.

[19] Alberto Signoroni, Mattia Savardi, Sergio Benini, Nicola Adami, Riccardo Leonardi, Paolo Gibellini, Filippo Vaccher, Marco Ravanelli, Andrea Borghesi, Roberto Maroldi, and Davide Farina. End-to-end learning for semiquantitative rating of COVID-19 severity on Chest X-rays. pages 1–22, 2020.

[20] Yan Li and Liming Xia. Coronavirus disease 2019 (COVID-19): Role of chest CT in diagnosis and management. *American Journal of Roentgenology*, 214(6):1280–1286, jun 2020.

[21] Weifang Kong and Prachi P. Agarwal. Chest Imaging Appearance of COVID-19 Infection. *Radiology: Cardiothoracic Imaging*, 2(1):e200028, jan 2020.

[22] Yicheng Fang, Huangqi Zhang, Jicheng Xie, Minjie Lin, Lingjun Ying, Peipei Pang, and Wenbin Ji. Sensitivity of chest CT for COVID-19: Comparison to RT-PCR, aug 2020.

[23] Michael Chung, Adam Bernheim, Xueyan Mei, Ning Zhang, Mingqian Huang, Xianjun Zeng, Jiufa Cui, Wenjian Xu, Yang Yang, Zahi A. Fayad, Adam Jacobi, Kunwei Li, Shaolin Li, and Hong Shan. CT imaging features of 2019 novel coronavirus (2019-NCoV). *Radiology*, 295(1):202–207, feb 2020.

[24] Chaolin Huang, Yeming Wang, Xingwang Li, Lili Ren, Jianping Zhao, Yi Hu, Li Zhang, Guohui Fan, Jiuyang Xu, Xiaoying Gu, Zhenshun Cheng, Ting Yu, Jiaan Xia, Yuan Wei, Wenjuan Wu, Xuelei Xie, Wen Yin, Hui Li, Min Liu, Yan Xiao, Hong Gao, Li Guo, Jungang Xie, Guangfa Wang, Rongmeng Jiang, Zhancheng Gao, Qi Jin, Jianwei Wang, and Bin Cao. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet*, 395(10223):497–506, feb 2020.

[25] Amit Kumar Jaiswal, Prayag Tiwari, Vipin Kumar Rathi, Jia Qian, Hari Mohan Pandey, and Victor Hugo C Albuquerque. COVIDPEN: A Novel COVID-19 Detection Model using Chest X-Rays and CT Scans. *medRxiv*, page 2020.07.08.20149161, 2020.

[26] Ho Yuen Frank Wong, Hiu Yin Sonia Lam, Ambrose Ho Tung Fong, Siu Ting Leung, Thomas Wing Yan Chin, Christine Shing Yen Lo, Macy Mei Sze Lui, Jonan Chun Yin Lee, Keith Wan Hang Chiu, Tom Wai Hin Chung, Elaine Yuen Phin Lee, Eric Yuk Fai Wan, Ivan Fan Ngai Hung, Tina Poy Wing Lam, Michael D. Kuo, and Ming Yen Ng. Frequency and Distribution of Chest Radiographic Findings in Patients Positive for COVID-19. *Radiology*, 296(2):E72–E78, aug 2020.

[27] Siham Tabik, Anabel Gómez-Ríos, J. Martín-Rodríguez, I. Sevillano-García, Manuel Rey-Area, David Charte, Emilio Guirado, J. Suárez, Julián Luengo, M. Valero-González, P. García-Villanova, E. Olmedo-Sánchez, and Francisco Herrera. COVIDGR dataset and COVID-SDNet methodology for predicting COVID-19 based on Chest X-Ray images. 2020.

[28] Zixing Huang, Shuang Zhao, Zhenlin Li, Weixia Chen, Lihong Zhao, Lipeng Deng, and Bin Song. The Battle Against Coronavirus Disease 2019 (COVID-19): Emergency Management and Infection Control in a Radiology Department. *Journal of the American College of Radiology*, 17(6):710–716, jun 2020.

[29] Chunqin Long, Huaxiang Xu, Qinglin Shen, Xianghai Zhang, Bing Fan, Chuanhong Wang, Bingliang Zeng, Zicong Li, Xiaofen Li, and Honglu Li. Diagnosis of the Coronavirus disease (COVID-19): rRT-PCR or CT? *European Journal of Radiology*, 126:108961, may 2020.

[30] Subhashini A. Sellers, Kenton L. Dover, Aubrey G. Bailey, Avery Cheves, Anthony B. Eason, Elena B. Popowitch, Melissa B. Miller, David A. Wohl, Dirk P. Dittmer, and William A. Fischer. Burden of respiratory viral infection in persons with human immunodeficiency virus. *Influenza and Other Respiratory Viruses*, 14(4):465–469, jul 2020.

[31] Joungha Won, Solji Lee, Myungsun Park, Tai Young Kim, Mingu Gordon Park, Byung Yoon Choi, Dongwan Kim, Hyeshik Chang, V. Narry Kim, and C. Justin Lee. Development of a laboratory-safe and low-cost detection protocol for SARS-CoV-2 of the Coronavirus Disease 2019 (COVID-19). *Experimental Neurobiology*, 29(2):107–119, 2020.

[32] Ferhat Ucar and Deniz Korkmaz. COVIDiagnosis-Net: Deep Bayes-SqueezeNet based diagnosis of the coronavirus disease 2019 (COVID-19) from X-ray images. *Medical Hypotheses*, 140(April):109761, 2020.

[33] Geoffrey D. Rubin, Christopher J. Ryerson, Linda B. Haramati, Nicola Sverzellati, Jeffrey P. Kanne, Suhail Raoof, Neil W. Schluger, Annalisa Volpi, Jae Joon Yim, Ian B.K. Martin, Deverick J. Anderson, Christina Kong, Talissa Altes, Andrew Bush, Sujal R. Desai, Onathan Goldin, Jin Mo Goo, Marc Humbert, Yoshikazu Inoue, Hans Ulrich Kauczor, Fengming Luo, Peter J. Mazzone, Mathias Prokop, Martine Remy-Jardin, Luca Richeldi, Cornelia M. Schaefer-Prokop, Noriyuki Tomiyama, Athol U. Wells, and Ann N. Leung. The role of chest imaging in patient management during the covid-19 pandemic: A multinational consensus statement from the fleischner society. *Radiology*, 296(1):172–180, jul 2020.

[34] Linda Wang and Alexander Wong. COVID-Net: A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases from Chest X-Ray Images. pages 1–12, 2020.

[35] Adam Jacobi, Michael Chung, Adam Bernheim, and Corey Eber. Portable chest X-ray in coronavirus disease-19 (COVID-19): A pictorial review, aug 2020.

[36] M. Shamim Hossain, Ghulam Muhammad, and Nadra Guizani. Explainable AI and Mass Surveillance System-Based Healthcare Framework to Combat COVID-I9 Like Pandemics. *IEEE Network*, 34(4):126–132, 2020.

[37] Biraja Ghoshal and Allan Tucker. Estimating Uncertainty and Interpretability in Deep Learning for Coronavirus (COVID-19) Detection. pages 1–14, 2020.

[38] Justin Ker, Lipo Wang, Jai Rao, and Tchoyoson Lim. Deep Learning Applications in Medical Image Analysis. *IEEE Access*, 6:9375–9379, dec 2017.

[39] Dinggang Shen, Guorong Wu, and Heung Il Suk. Deep Learning in Medical Image Analysis. *Annual Review of Biomedical Engineering*, 19:221–248, jun 2017.

[40] Oliver Faust, Yuki Hagiwara, Tan Jen Hong, Oh Shu Lih, and U. Rajendra Acharya. Deep learning for healthcare applications based on physiological signals: A review, jul 2018.

[41] Fatma Murat, Ozal Yildirim, Muhammed Talo, Ulas Baran Baloglu, Yakup Demir, and U. Rajendra Acharya. Application of deep learning techniques for heartbeats detection using ECG signals-analysis and review, may 2020.

[42] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis, dec 2017.

[43] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, feb 2017.

[44] Noel Codella, Quoc-Bao Nguyen, Sharath Pankanti, David Gutman, Brian Helba, Allan Halpern, and John R. Smith. Deep Learning Ensembles for Melanoma Recognition in Dermoscopy Images. *IBM Journal of Research and Development*, 61(4), oct 2016.

[45] Muhammed Talo, Ozal Yildirim, Ulas Baran Baloglu, Galip Aydin, and U Rajendra Acharya. Convolutional neural networks for multi-class brain disease detection using MRI images. *Computerized Medical Imaging and Graphics*, 78:101673, dec 2019.

[46] Özal Yıldırım, Paweł Pławiak, Ru San Tan, and U. Rajendra Acharya. Arrhythmia detection using deep convolutional neural network with long duration ECG signals. *Computers in Biology and Medicine*, 102:411–420, nov 2018.

[47] Awni Y. Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H. Tison, Codie Bourn, Mintu P. Turakhia, and Andrew Y. Ng. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, 25(1):65–69, jan 2019.

[48] Douglas A. Woten, John Lusth, and Magda El-Shenawee. Interpreting artificial neural networks for microwave detection of breast cancer. *IEEE Microwave and Wireless Components Letters*, 17(12):825–827, dec 2007.

[49] Roslidar Roslidar, Aulia Rahman, Rusdha Muharar, Muhammad Rizky Syahputra, Fitri Arnia, Maimun Syukri, Biswajeet Pradhan, and Khairul Munadi. A Review on Recent Progress in Thermal Imaging and Deep Learning Approaches for Breast Cancer Detection, 2020.

[50] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. nov 2017.

[51] Shashank Jaiswal, Michel Valstar, Alinda Gillott, and David Daley. Automatic detection of adhd and asd from expressive behaviour in rgbd data. 2016.

[52] Muhammad Usman, Byoung Dai Lee, Shi Sub Byon, Sung Hyun Kim, Byung il Lee, and Yeong Gil Shin. Volumetric lung nodule segmentation using adaptive ROI with multi-view residual learning. *Scientific Reports*, 10(1):12839, dec 2020.

[53] Johnatan Carvalho Souza, João Otávio Bandeira Diniz, Jonnison Lima Ferreira, Giovanni Lucca França da Silva, Aristófanes Corrêa Silva, and Anselmo Cardoso de Paiva. An automatic method for lung segmentation and reconstruction in chest X-ray using deep neural networks. *Computer Methods and Programs in Biomedicine*, 177:285–296, aug 2019.

[54] Jen Hong Tan, Hamido Fujita, Sobha Sivaprasad, Sulatha V. Bhandary, A. Krishna Rao, Kuang Chua Chua, and U. Rajendra Acharya. Automated segmentation of exudates, haemorrhages, microaneurysms using single convolutional neural network. *Information Sciences*, 420:66–76, dec 2017.

[55] Haris Jeelani, Jonathan Martin, Francis Vasquez, Michael Salerno, and Daniel S. Weller. Image quality affects deep learning reconstruction of MRI. In *Proceedings - International Symposium on Biomedical Imaging*, volume 2018-April, pages 357–360. IEEE Computer Society, may 2018.

[56] Charalambos Chrysostomou, Loizos Koutsantonis, Christos Lemesios, and Costas N. Papanicolas. SPECT Imaging Reconstruction Method Based on Deep Convolutional Neural Network. *2019 IEEE Nuclear Science Symposium and Medical Imaging Conference, NSS/MIC 2019*, oct 2020.

[57] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *NIPS 2018*, pages 1–15, nov 2018.

[58] Laboni Sarker, Mohaiminul Islam, Tanveer Hannan, Ahmed Zakaria, Zakaria Ahmed, and Ahmed Zakaria. COVID-DenseNet: A Deep Learning Architecture to Detect COVID-19 from Chest Radiology Images. *Preprints*, 2020.

[59] Alessandro Liberati, Douglas G. Altman, Jennifer Tetzlaff, Cynthia Mulrow, Peter C. Gøtzsche, John P.A. Ioannidis, Mike Clarke, P. J. Devereaux, Jos Kleijnen, and David Moher. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. *PLoS Medicine*, 6(7), 2009.

[60] Yu-Huan Wu, Shang-Hua Gao, Jie Mei, Jun Xu, Deng-Ping Fan, Chao-Wei Zhao, and Ming-Ming Cheng. JCS: An Explainable COVID-19 Diagnosis System by Joint Classification and Segmentation. pages 1–11, 2020.

[61] Tulin Ozturk, Muhammed Talo, Eylul Azra Yildirim, Ulas Baran Baloglu, Ozal Yildirim, and U. Rajendra Acharya. Automated detection of COVID-19 cases using deep neural networks with X-ray images. *Computers in Biology and Medicine*, 121(April):103792, 2020.

[62] Vishal Sharma and Curtis Dyreson. COVID-19 detection using Residual Attention Network an Artificial Intelligence approach. 2020.

[63] Nikos Tsiknakis, Eleftherios Trivizakis, Evangelia Vassalou, Georgios Papadakis, Demetrios Spandidos, Aristidis Tsatsakis, Jose Sánchez-García, Rafael López-González, Nikolaos Papanikolaou, Apostolos Karantanas, and Kostas Marias. Interpretable artificial intelligence framework for COVID-19 screening on chest X-rays. *Experimental and Therapeutic Medicine*, pages 727–735, 2020.

[64] Md. Rezaul Karim, Till Döhmen, Dietrich Rebholz-Schuhmann, Stefan Decker, Michael Cochez, and Oya Beyan. DeepCOVIDExplainer: Explainable COVID-19 Diagnosis Based on Chest X-ray Images. 2020.

[65] Maryam Zokaeinikoo, Prasenjit Mitra, Soundar Kumara, and Pooyan Kazemian. AIDCOV: An Interpretable Artificial Intelligence Model for Detection of COVID-19 from Chest Radiography Images. *medRxiv*, page 2020.05.24.20111922, 2020.

[66] Soon Ho Yoon, Kyung Hee Lee, Jin Yong Kim, Young Kyung Lee, Hongseok Ko, Ki Hwan Kim, Chang Min Park, and Yun Hyeon Kim. Chest radiographic and ct findings of the 2019 novel coronavirus disease (Covid-19): Analysis of nine patients treated in Korea. *Korean Journal of Radiology*, 21(4):498–504, apr 2020.

[67] Nikita Albert. Evaluation of Contemporary Convolutional Neural Network Architectures for Detecting COVID-19 from Chest Radiographs. *arXiv.org*, 2020.

[68] Joseph Paul Cohen, Lan Dao, Paul Morrison, Karsten Roth, Yoshua Bengio, Beiyi Shen, Almas Abbasi, Mahsa Hoshmand-Kochi, Marzyeh Ghassemi, Haifang Li, and Tim Q Duong. Predicting COVID-19 Pneumonia Severity on Chest X-ray with Deep Learning. 8(December 2019), 2020.

[69] Muhammad Chowdhury, Tawsifur Rahman, Amith Khandakar, Rashid Mazhar, Muhammad Kadir, Zaid Mahbub, Khandakar Islam, Muhammad Salman Khan, Atif Iqbal, Nasser Al-Emadi, and Mamun Bin Ibne Reaz. Can AI help in screening viral and COVID-19 pneumonia? *IEEE Access*, 8:132665–132676, 03 2020.

[70] Joseph Paul Cohen, Paul Morrison, Lan Dao, Karsten Roth, Tim Q Duong, and Marzyeh Ghassemi. COVID-19 Image Data Collection: Prospective Predictions Are the Future. *arXiv 2006.11988*, 2020.

[71] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017.

[72] Daniel S. Kermany, K. Zhang, and M. Goldbaum. Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification. 2018.

[73] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, and et al. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:590–597, Jul 2019.

[74] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. PadChest: A large chest x-ray image dataset with multi-label annotated reports. *Medical Image Analysis*, 66:101797, 2020.

[75] Alistair E.W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih Ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

[76] Armato III, Samuel G., Geoffrey McLennan, Luc Bidaut, Michael F. McNitt-Gray, Charles R. Meyer, Anthony P. Reeves, Binsheng Zhao, Denise R. Aberle, Claudia I. Henschke, Eric A. Hoffman, Ella A. Kazerooni, Heber MacMahon, Edwin J.R. Van Beek, David Yankelevitz, Alberto M. Biancardi, Peyton H. Bland, Matthew S. Brown, Roger M. Engelmann, Gary E. Laderach, Daniel Max, Richard C. Pais, David P.Y. Qing, Rachael Y. Roberts, Amanda R. Smith, Adam Starkey, Poonam Batra, Philip Caligiuri, Ali Farooqi, Gregory W. Gladish, C. Matilda Jude, Reginald F. Munden, Iva Petkovska, Leslie E. Quint, Lawrence H. Schwartz, Baskaran Sundaram, Lori E. Dodd, Charles Fenimore, David Gur, Nicholas Petrick, John Freymann, Justin Kirby, Brian Hughes, Alessi Vande Casteele, Sangeeta Gupte, Maha Sallam, Michael D. Heath, Michael H. Kuhn, Ekta Dharaiya, Richard Burns, David S. Fryd, Marcos Salganicoff, Vikram Anand, Uri Shreter, Stephen Vastagh, Barbara Y. Croft, and Laurence P. Clarke. Data from LIDC-IDRI, 2015.

[77] Md. Abdur Rahman, M. Shamim Hossain, Nabil A. Alrajeh, and Nadra Guizani. B5G and Explainable Deep Learning Assisted Healthcare Vertical at the Edge: COVID-I9 Perspective. *IEEE Network*, 34(4):98–105, 2020.

[78] Yilin Yan, Min Chen, Mei Ling Shyu, and Shu Ching Chen. Deep Learning for Imbalanced Multimedia Data Classification. In *Proceedings - 2015 IEEE International Symposium on Multimedia, ISM 2015*, pages 483–488. Institute of Electrical and Electronics Engineers Inc., mar 2016.

[79] Rosemary Arthur. Interpretation of the paediatric chest X-ray. *Paediatric Respiratory Reviews*, 1(1):41–50, 2000.

[80] Edward Y. Lee. Pediatric Interstitial (Diffuse) Lung Disease. In *Imaging in Pediatric Pulmonology*, pages 145–197. Springer International Publishing, 2020.

[81] Matthew Silver and Steven Kohler. Evolution of a round pneumonia. *Western Journal of Emergency Medicine*, 14(6):643–644, 2013.

[82] Radiological Society of North America. RSNA Pneumonia Detection Challenge | Kaggle.

[83] Laure Wynants, Ben Van Calster, Marc M.J. Bonten, Gary S. Collins, Thomas P.A. Debray, Maarten De Vos, Maria C. Haller, Georg Heinze, Karel G.M. Moons, Richard D. Riley, Ewoud Schuit, Luc J.M. Smits, Kym I.E. Snell, Ewout W. Steyerberg, Christine Wallisch, and Maarten Van Smeden. Prediction models for diagnosis and prognosis of covid-19 infection: Systematic review and critical appraisal. *The BMJ*, 369, 2020.

[84] Connor Shorten and Taghi M. Khoshgoftaar. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, 6(1):1–48, dec 2019.

[85] Veronika Cheplygina. Cats or CAT scans: transfer learning from natural or medical image source datasets? *Current Opinion in Biomedical Engineering*, 9:21–27, oct 2018.

[86] Hak Gu Kim, Yeoreum Choi, and Yong Man Ro. Modality-bridge Transfer Learning for Medical Image Classification. *Proceedings - 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics, CISP-BMEI 2017*, 2018-January:1–5, aug 2017.

[87] Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR, dec 2015.

[88] O.S. Albahri, A.A. Zaidan, A.S. Albahri, B.B. Zaidan, Karrar Hameed Abdulkareem, Z.T. Al-qaysi, A.H. Alamoodi, A.M. Aleesa, M.A. Chyad, R.M. Alesa, L.C. Kem, Muhammad Modi Lakulu, A.B. Ibrahim, and Nazre Abdul Rashid. 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. *Journal of Infection and Public Health*, 13(June), 07 2020.

[89] Przemyslaw Biecek and Tomasz Burzykowski. *Explanatory Model Analysis. Explore, Explain, and Examine Predictive Models*. Chapman and Hall/CRC, 2021.

[90] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. *CoRR*, abs/1708.08296, 2017.

[91] Andreas Holzinger, Chris Biemann, Constantinos S. Pattichis, and Douglas B. Kell. What do we need to build explainable AI systems for the medical domain? *CoRR*, abs/1712.09923, 2017.

[92] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *International Journal of Computer Vision*, 128(2):336–359, oct 2016.

[93] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, volume 13-17-August-2016, pages 1135–1144, New York, NY, USA, aug 2016. Association for Computing Machinery.

[94] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A Benchmark to Evaluate Rationalized NLP Models. *arXiv*, nov 2019.

[95] Atul Gawande. *The Checklist Manifesto: How to Get Things Right*. Profile, 2011.