A Two-Step Radiologist-Like Approach for Covid-19 Computer-Aided Diagnosis from Chest X-Ray Images

A two-step radiologist-like approach for Covid-19 computer-aided diagnosis from chest x-ray images

Carlo Alberto Barbano[1 [0000−0001−9512−0440]], Enzo Tartaglione[1,2 [0000−0003−4274−8298]], Claudio Berzovini[3], Marco Calandri[1], and Marco Grangetto[1 [0000−0002−2709−7864]⋆]

[1] University of Turin, Torino, Italy
{name.surname}@unito.it
[2] LTCI, Télécom Paris, Institut Polytechnique de Paris, France
[3] Azienda Ospedaliera Citta della Salute e della Scienza, Torino, Italy
cberzovini@cittadellasalute.to.it

Abstract. Thanks to the rapid increase in computational capability during the latest years, traditional and more explainable methods have been gradually replaced by more complex deep-learning-based approaches, which have in fact reached new state-of-the-art results for a variety of tasks. However, for certain kinds of applications performance alone is not enough. A prime example is represented by the medical field, in which building trust between the physicians and the AI models is fundamental. Providing an explainable or trustful model, however, is not a trivial task, considering the black-box nature of deep-learning based methods. While some existing methods, such as gradient or saliency maps, try to provide insights about the functioning of deep neural networks, they often provide limited information with regards to clinical needs.

We propose a two-step diagnostic approach for the detection of Covid-19 infection from Chest X-Ray images. Our approach is designed to mimic the diagnosis process of human radiologists: it detects objective radiological findings in the lungs, which are then employed for making a final Covid-19 diagnosis. We believe that this kind of *structural* explainability can be preferable in this context. The proposed approach achieves promising performance in Covid-19 detection, compatible with expert human radiologists. Moreover, despite this work being focused Covid-19, we believe that this approach could be employed for many different CXR-based diagnosis.

Keywords: Deep Learning · Chest X-Ray · Radiological findings · Covid-19.

⋆ This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111, DeepHealth Project

1 Introduction

Early Covid-19 diagnosis is a key element for proper treatment of the patients and prevention of the spread of the disease. Given the high tropism of Covid-19 for respiratory airways and lung epythelium, identification of lung involvement in infected patients can be relevant for treatment and monitoring of the disease. Virus testing is currently considered the only specific method of diagnosis. The US Center for Disease Control (CDC) recommends collecting and testing specimens from the upper respiratory tract (nasopharyngeal and oropharyngeal swabs) or from the lower respiratory tract when available (bronchoalveolar lavage, BAL) for viral testing with reverse transcription polymerase chain reaction (RT-PCR) assay ([1]). Testing on BAL samples provides higher accuracy, however this test is unconfortable for the patient, possibly dangerous for the operator due to aerosol emission during the procedure and cannot be performed routinely. Nasopharingeal swabs are easily executable and affordable and current standard in diagnostic setting; their accuracy in literature is influenced by the severity of the disease and the time from symptoms onset and is reported up to 73.3% [25]. Current position papers from radiological societies (Fleischner Society, SIRM, RSNA) [1,10,16] do not recommend routine use of imaging for Covid-19 diagnosis; however, it has been widely demonstrated that, even at early stages of the disease, chest x-rays (CXR) can show pathological findings.
In the last year, many works attempted to tackle this problem, proposing deep

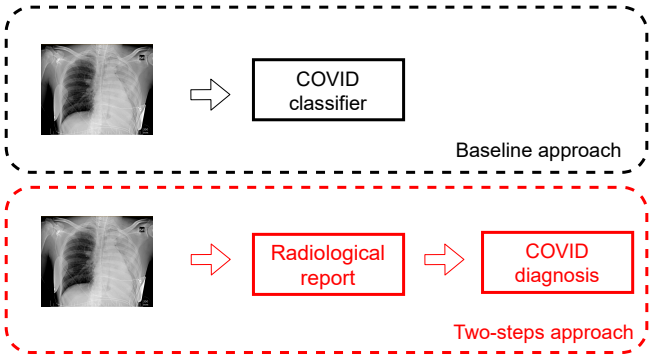

Fig. 1: **Comparison** between standard approaches for Covid-19 diagnosis and our proposed two-step method.

learning-based strategies [20,18,3,12,21]. All of the proposed approaches include some elements in common:

– the images collected during the pandemic need to be augmented with non-Covid cases from publicly available datasets;

- some standard pre-processing is applied to the images, like lung segmentation using U-Net [15] or similar models [20] or converting the pixels of the CXR scan in Hounsfield units;
- the deep learning model is trained to predict the final diagnosis using state-of-the-art approaches for deep neural networks.

Despite some very optimistic results [3,12,21,13,19], the proposed approaches exhibit significant limitations that deserve further analysis. For example, a recent work [20] showed that augmenting Covid-19 datasets with negative cases from publicly-available datasets can inject dangerous biases, where the trained model learn to discriminate different data sources rather than actual radiological features related to the disease. These unwanted effects are difficult to spot when using a *black box* model like deep neural networks, without having control on the decision process.
In this work, we propose an explainable approach, mimicking the radiologists' decision process. We break-down the Covid-19 diagnosis problem into two separate sub-problems. First, we train a model to detect anomalies in the lungs. These anomalies are widely known and, following [6], they comprise 14 objective radiological observations which can be found in lungs. Then, the employ this information to train a decision tree model, such that the Covid-19 diagnosis is fully explainable and grounded to objective radiological findings (Fig. 1). Mimicking the radiologist's decision is more robust to biases and aims at building trust for the physicians and patients towards the AI tool, which can be useful for a more precise COVID diagnosis. Thanks to the collaboration with the radiology units of Città della Salute e della Scienza di Torino (CDSS) and San Luigi Hospital (SLG) in Turin, we collected the COvid Radiographic images DAta-set for AI (CORDA), comprising both positive and negative COVID cases as well as a ground truth on the human radiological reporting, and it currently comprises almost 1000 CXRs.
The rest of the paper is structured as follows. Section 2 introduces and describes the datasets used in this work; Section 3 discusses the possible radiological reports, objective findings in the radiographies; Section 4 discusses the approach thanks to, from the radiological reports generated, the COVID diagnosis is being elaborated; Section 5 discusses the results achieved and finally Section 6 draws the conclusions.

2 Datasets

In this section we introduce the datasets that we employed for building our proposed method. As shown in recent works such as [20], augmenting COVID datasets with negative cases from publicly-available datasets (such as ChestXRay or RSNA) can drive the models towards spourious correlations, such as discriminating only between healthy and unhealthy lung, or also be influenced by information related to the image acquisition sites. This is why we focus on the recognition of objective pathologies and radiological findings (CheXpert), which we then use for training a model to elaborate the final Covid-19 diagnosis (CORDA).

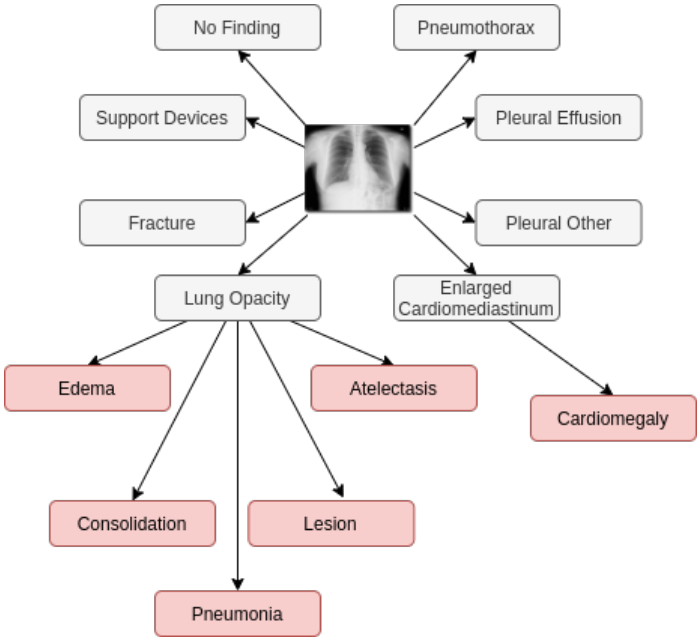

Fig. 2: **Radiological findings provided by CheXpert**, conforming to the medical known categorization [6].

We now provide a short description of the main datasets, additional information can be found in the supplementary material.

CheXpert. This is a large dataset comprising about 224k CXRs. This dataset consists of 14 different observations on the radiographic image: differently from many other datasets which are focused on disease classification based on clinical diagnosis, the main focus here is "chest radiograph interpretation", where anomalies are detected [9]. The learnable radiological findings are summarized in Fig. 2.

ChestXRay. This dataset contains 5857 X-ray images collected at the Guangzhou Women and Children's Medical Center, Guangzhou, China. In this dataset, three different labels are provided: normal patients (1583), patients affected by bacterial pneumonia (2780) and affected by viral pneumonia (1493). This dataset is granted under CC by 4.0 and is part of a work on Optical Coherence Tomography [11].[4]

RSNA. Developed for the RSNA Pneumonia Detection Challenge, this dataset contains pneumonia cases found in the NIH Chest X-ray dataset [22]. It comprises 20672 normal CXR scans and 6012 pneumonia cases, for a total of 26684

[4] `https://data.mendeley.com/datasets/rscbjbr9sj/2/files/f12eaf6d-6023-432f-acc9-80c9d7393433`

images.[5]

CORDA. This dataset was created for this study by retrospectively selecting chest x-rays performed at a dedicated Radiology Unit in CDSS and at SLG in all patients with fever or respiratory symptoms (cough, shortness of breath, dyspnea) that underwent nasopharingeal swab to rule out COVID-19 infection. Patients' average age is 61 years (range 17-97 years old). It contains a total of 898 CXRs and can be split by different collecting institution into two similarly sized subgroups: CORDA-CDSS, which contains a total of 447 CXRs from 386 patients, with 150 images coming from COVID-negative patients and 297 from positive ones, and CORDA-SLG, which contains the remaining 451 CXRs, with 129 COVID-positive and 322 COVID-negative images. Including data from different hospitals at test time is crucial to doublecheck the generalization capability of our model.

The data collection is still in progress, with other 5 hospitals in Italy willing to contribute at time of writing. We plan to make CORDA available for research purposes according to EU regulations as soon as possible.

3 Radiological report

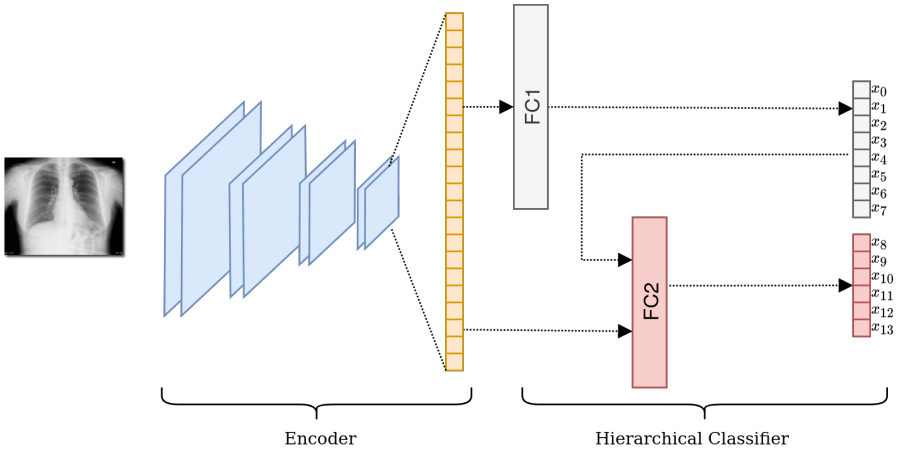

Fig. 3: **Radiological report framework.** After the convolutional encoder extracts features from the CXR, the Hierarchical Classifier provides outcome for the different lung pathologies.

In this section we describe our proposed method for detecting and classifying radiological findings from CXRs. For this task, we leverage the large scale

[5] `https://www.kaggle.com/c/rsna-pneumonia-detection-challenge`

dataset *CheXpert*, which contains annotation for different kinds of common radiological findings that can be observed in CXR images (like opacity, pleural effusion, cardiomegaly, etc.). Given the high heterogeneity and the large size of CheXpert, its use is perfect for our purposes: in fact, once the model is trained on this dataset, there is no need to fine-tune it for the Covid-19 diagnosis, since it will already extract objective radiological findings.
CheXpert provides 14 different types of observations for each image in the dataset. For each class, the labels have been generated from radiology reports associated with the studies with NLP techniques, conforming to the Fleischner Society's recommended glossary [6], and marked as: negative (N), positive (P), uncertain (U) or blank (N/A). Following the relationship among labels illustrated in Fig. 2, as proposed by [9], we can identify 8 top-level pathologies and 6 child ones.

3.1 Architecture

As the backbone of our model, we use the widely known ResNet [7] and DenseNet [4] convolutional architecture. The encoder is followed by a fully connected classifier. The classifier architecture that we design reflects the hierarchy of the different lung pathologies presented in Fig. 2. As shown in Fig. 3, the classifier is constructed by stacking two fully-connected layers, and makes use of connectivity patterns similar to "dense connections" as proposed by [8]. The first fully-connected layer (*FC1*) is used to classify the 8 top-level classes from the extracted features. Output logits from FC1 are then concatenated with the extracted image features, and the second fully-connected layer (*FC2*) is used to predict the remaining 6 children pathologies. Finally, the logits from FC1 and FC2 are concatenated, and a final sigmoid layer is used to obtain the probability for each class. We call this architecture *Hierarchical Classifier* (HC).

3.2 Dealing with uncertain labels

In order to extract the radiological findings from CXRs, a deep learning model is trained on the 14 observations. For this purpose, given the possibility of having multiple findings in the same CXR, we employ the weighted binary cross entropy loss for training the model. Typically, weights are used to compensate class unbalancing, giving higher importance to less represented classes. Within *CheXpert*, however, we also need to tackle another issue: how to treat the samples with the U label. Towards this issue, multiple approaches have been suggested by [9]. The most popular is to just ignore all the uncertain samples, excluding them from the training process and considering them as N/A.
We propose to include the U samples in the learning process, mapping them to the maximum uncertainty (probability 0.5 to be P or N). Then, we assign a weight to the P and N outcomes:

$$w_n = \begin{cases} 1 + S_n^+ / S_n^- & \text{if } y_n = 0 \\ 1 + S_n^- / S_n^+ & \text{if } y_n = 1 \\ 1 & \text{if } y_n = 0.5 \end{cases} \tag{1}$$

where S_n^- and S_n^+ respectively represent the cardinality of negative and positive samples for the n-th class. These weights are then plugged in the weighted loss entropy loss to be minimized:

$$\mathcal{L}_n = -w_n \cdot [y_n \log(x_n) + (1 - y_n) \log(1 - x_n)] \quad (2)$$

Hence, uncertain samples will have a lower influence during the training process, while being pushed either towards 0 or 1 by the higher weight certain samples in the same class. All of the remaining blank labels are ignored when computing the BCE loss, considering them as missing labels.
Table 1 shows a performance comparison between the standard approach as proposed by [9] and our proposal (U-label use), for 5 salient radiological findings, using the same setting as in [9]. We observe an overall improvement in the performance, which is expected by the inclusion of the U-labeled examples. For all our experiments, we will use models trained using the U labeled samples.

Table 1: Performance (AUC) for a DenseNet-121 trained on CheXpert.

Method	**Atelectasis**	**Cardiomegaly**	**Consolidation**	**Edema**	**Pleural Eff.**
Baseline [9]	0.79	**0.81**	0.90	0.91	0.92
U-label	**0.83**	0.79	**0.93**	**0.93**	**0.93**

4 COVID diagnosis

The second step of the proposed approach is building the model which can actually provide a clinical diagnosis for COVID. We freeze the model obtained from Sec. 3 and use its output as input features to train a new binary classifier on the CORDA dataset. We test two different types of classifiers: a decision tree (Tree) and a two-layers fully-connected classifier (FC). The decision tree is trained on the probabilities output of the radiological reports, using the standard CART Algorithm implementation provided by the Python scikit-learn [14] package. The fully-connected classifier, comprising one hidden layer of size 512 and the output layer, is instead trained on the encoder latent space. The reason is that training it on the output probabilities would result in a loss of explainability compared to the decision tree, hence it makes more sense to maximize the achievable performance by training on the full latent space as discussed in Sec.5.

5 Experiments

In this section we compare the COVID diagnosis generalization capability through a direct deep learning-based approach (baseline) and our proposed two-step diagnosis, where first we detect the radiological findings, and then we discriminate

Table 2: Results for COVID diagnosis. Abbreviations: ResNet-18 (RN-18), DenseNet-121 (DN-121), Pretrain dataset (Pretrain), Chest X-Ray (CXR), CheXpert (ChXp), Train dataset (Train), CORDA-CDSS (CDSS), CORDA-SLG (SLG). We denote fully-explainable methods with †.

Method	**Baseline** [20]			**Two-step**			
Backbone	RN-18	RN-18	RN-18	RN-18	DN-121	DN-121	DN-121
Classifier	FC	FC	FC	FC	FC	Tree†	FC
Pretrain	-	RSNA	CXR		ChXp		ChXp
Train		CDSS			CDSS		SLG
Sensitivity	**0.56**	0.54	0.54	0.69	0.72	**0.77**	0.79
Specificity	0.58	**0.80**	0.58	0.73	**0.78**	0.60	0.82
BA	0.57	**0.67**	0.56	0.71	**0.75**	0.68	0.81
AUC	0.59	**0.72**	0.67	0.76	**0.81**	0.70	0.84

patients affected by COVID using a decision tree-based diagnosis (Tree) or a deep learning-based classifier from the radiological findings (FC). The performance is tested on a subset of *patients* not included in the training / validation set. The assessed metrics are: balanced accuracy (BA), sensitivity, specificity and area under the ROC curve (AUC). For all of the methods we adopt a 70%-30% train-test split. For the deep learning-based strategy, SGD is used with a learning rate 0.01 and a weight decay of 10^{-5}. All of the experiments were run on NVIDIA Tesla T4 GPUs using PyTorch 1.4.[6]

Table 2 compares the standard deep learning-based approach [20] to our two-step diagnosis. Baseline results are obtained pre-training the model on some of the most used publicly-available datasets. We observe that the best achievable performance is very low, consisting in a BA of 0.67. A key takeaway is that trying to directly diagnose diseases such as COVID-19 from CXRs might be currently infeasible, probably given the small dataset sizes and strong selective bias in the datasets. We can clearly see how the two-step method outperforms the direct diagnosis: using the same network architecture (ResNet-18 as backbone and a fully-connected classifier on top of it), we obtain a significant increase in all of the assessed metrics. Even better results are achieved by using a DenseNet-121 as backbone and the fully-connected classifier.

Focusing on the fully-explainable decision tree method, we found that a maximum depth of 4 gave the best results in terms of model complexity and generalization ability. Fig. 4 graphically shows the learned decision tree (whose performance is shown in Table 2): this provides a very clear interpretation for the decision process. From the clinical and radiological perspective, these data are consistent with the COVID-19 CXR semiotics that radiologists are used to deal with. The edema feature, although unspecific, is strictly related to the interstitial involvement that is typical of COVID-19 infections and it has been largely

[6] The source code is available at `https://github.com/EIDOSlab/covid-two-step`.

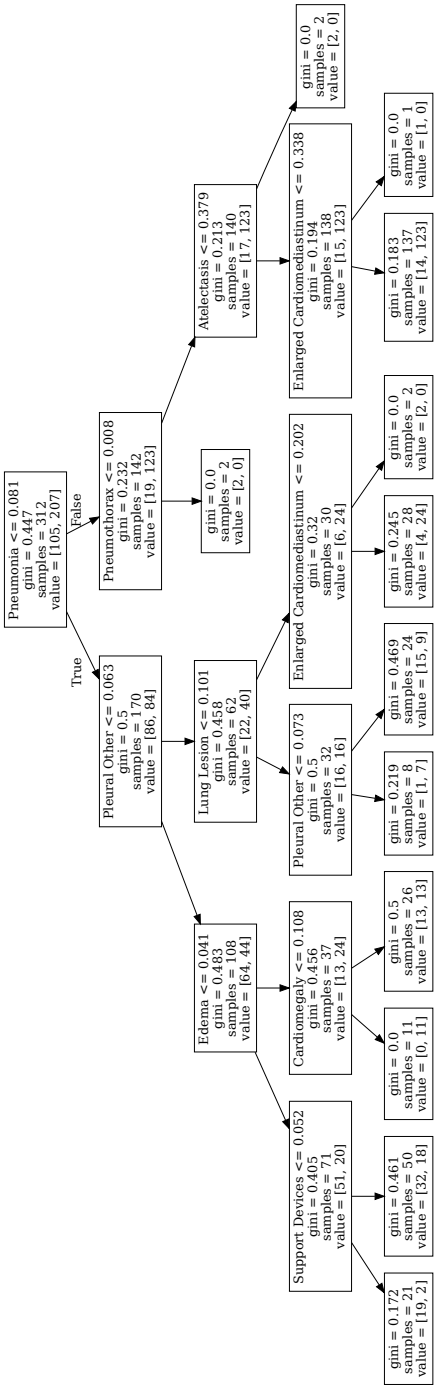

Fig. 4: Decision Tree obtained for COVID-19 classification based on the probabilities for the 14 classes of findings.

reported in the recent literature [5]. Indeed, in recent COVID-19 radiological papers, interstitial involvement has been reported as ground glass opacity appearance [24]. However this definition is more pertinent to the CT imaging setting rather than CXR; the "edema" feature can be compatible, from the radiological perspective, to the interstitial opacity of COVID-19 patients. Furthermore, the not irrelevant role of cardiomegaly (or more in general enlarged cardiomediastinum) in the decision tree can be interesting from the clinical perspective. In fact, this can be read as an additional proof that established cardiovascular disease can be a relevant risk factor to develop COVID-19 [2]. Moreover, it may be consistent with the hypotheses of a larger role of the primary cardiovascular damage observed on on preliminary data of autopsies of COVID-19 patients [23].

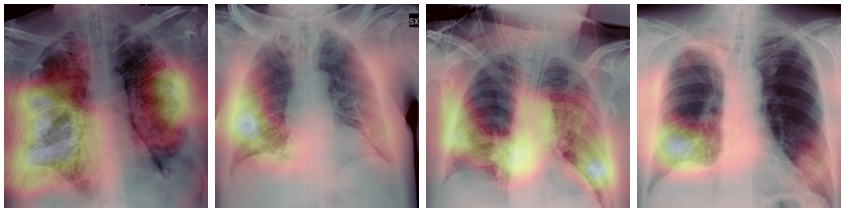

Fig. 5: GradCAM on COVID-positive images obtained from Densenet121+FC.

Although it is true that with the deep learning-based approach we observe a boost in the performance (BA of 0.75 with DN-121+FC vs 0.68 with Tree), this is the result of a trade-off between interpretability and discriminative power. Using Grad-CAM [17] we have hints on the area the model focused on to take the final diagnostic decision. From Fig. 5 we observe that on COVID-positive images, the model seems to mostly focus on the expected lung areas. However this kind of interpretability provides very limited insights when compared to the clinical-based explanation given by our decision tree.
Finally, to further test the reliability of our approach, we used our strategy also on CORDA-SLG (which are data coming from a different hospital structure), reaching comparable and encouraging results.

6 Conclusions

One of the latest challenges for both the clinical and the AI community has been applying deep learning in diagnosing COVID from CXRs. Recent works suggested the possibility of successfully tackling this problem, despite the currently small quantity of publicly available data. In this work we propose a multi-step approach, close to the physicians' diagnostic process, in which the final diagnosis is based upon detected lung pathologies. We performed our experiments on CORDA, a COVID-19 CXR dataset comprising approximately 1000 images. All of our experiments have been carried out bearing in mind that, especially

for clinical applications, explainability plays a major role for building trust in machine learning algorithms, although better interpretability can come at the cost of a lower prediction accuracy.

References

1. ACR recommendations for the use of chest radiography and computed tomography (CT) for suspected COVID-19 infection. https://www.acr.org/
2. ESC Guidance for the Diagnosis and Management of CV Disease during the COVID-19 Pandemic. (2020), `https://www.escardio.org/Education/COVID-19-and-Cardiology/ESC-COVID-19-Guidance"`
3. Apostolopoulos, I.D., Bessiana, T.: Covid-19: Automatic detection from x-ray images utilizing transfer learning with convolutional neural networks. arXiv preprint arXiv:2003.11617 (2020)
4. De Fauw, J., Ledsam, J.R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O'Donoghue, B., Visentin, D., et al.: Clinically applicable deep learning for diagnosis and referral in retinal disease. Nature medicine **24**(9), 1342–1350 (2018)
5. Guan, W.j., Ni, Z.y., Hu, Y., Liang, W.h., Ou, C.q., He, J.x., Liu, L., Shan, H., Lei, C.l., Hui, D.S., et al.: Clinical characteristics of coronavirus disease 2019 in china. New England journal of medicine **382**(18), 1708–1720 (2020)
6. Hansell, D.M., Bankier, A.A., MacMahon, H., McLoud, T.C., Muller, N.L., Remy, J.: Fleischner society: glossary of terms for thoracic imaging. Radiology **246**(3), 697–722 (2008)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
8. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
9. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 590–597 (2019)
10. Italian Radiology Society: Utilizzo della Diagnostica per Immagini nei pazienti Covid 19. https://www.sirm.org/
11. Kermany, D., Zhang, K., Goldbaum, M.: Labeled optical coherence tomography (oct) and chest x-ray images for classification. Mendeley data **2** (2018)
12. Narin, A., Kaya, C., Pamuk, Z.: Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks. arXiv preprint arXiv:2003.10849 (2020)
13. Oh, Y., Park, S., Ye, J.C.: Deep learning covid-19 features on cxr using limited training data sets. IEEE transactions on medical imaging **39**(8), 2688–2700 (2020)
14. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
15. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)

16. Rubin, G.D., Ryerson, C.J., Haramati, L.B., Sverzellati, N., et al.: The role of chest imaging in patient management during the covid-19 pandemic: A multinational consensus statement from the fleischner society. RSNA Radiology (2020). https://doi.org/10.1148/radiol.2020201365
17. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
18. Sethy, P.K., Behera, S.K.: Detection of coronavirus disease (covid-19) based on deep features (2020)
19. Sitaula, C., Hossain, M.B.: Attention-based vgg-16 model for covid-19 chest x-ray image classification. Applied Intelligence **51**(5), 2850–2863 (2021)
20. Tartaglione, E., Barbano, C.A., Berzovini, C., Calandri, M., Grangetto, M.: Unveiling covid-19 from chest x-ray with deep learning: A hurdles race with small data. International Journal of Environmental Research and Public Health **17**(18), 6933 (Sep 2020). https://doi.org/10.3390/ijerph17186933, `http://dx.doi.org/10.3390/ijerph17186933`
21. Wang, L., Lin, Z.Q., Wong, A.: Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. Scientific Reports **10**(1), 1–12 (2020)
22. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). pp. 3462–3471 (2017)
23. Wichmann, D., Sperhake, J.P., Lütgehetmann, M., Steurer, S., Edler, C., Heinemann, A., Heinrich, F., Mushumba, H., Kniep, I., Schröder, A.S., et al.: Autopsy findings and venous thromboembolism in patients with covid-19: a prospective cohort study. Annals of Internal Medicine (2020)
24. Wong, H.Y.F., Lam, H.Y.S., Fong, A.H.T., Leung, S.T., Chin, T.W.Y., Lo, C.S.Y., Lui, M.M.S., Lee, J.C.Y., Chiu, K.W.H., Chung, T., et al.: Frequency and distribution of chest radiographic findings in covid-19 positive patients. Radiology p. 201160 (2020)
25. Yang, Y., Yang, M., Shen, C., Wang, F., Yuan, J., Li, J., Zhang, M., Wang, Z., Xing, L., Wei, J., et al.: Laboratory diagnosis and monitoring the viral shedding of 2019-ncov infections. medRxiv (2020)