Unveiling COVID-19 from Chest X-ray with deep learning: a hurdles race with small data

Enzo Tartaglione, , Carlo Alberto Barbano, Claudio Berzovini, Marco Calandri, and Marco Grangetto,

***Abstract*—The possibility to use widespread and simple chest X-ray (CXR) imaging for early screening of COVID-19 patients is attracting much interest from both the clinical and the AI community. In this study we provide insights and also raise warnings on what is reasonable to expect by applying deep learning to COVID classification of CXR images. We provide a methodological guide and critical reading of an extensive set of statistical results that can be obtained using currently available datasets. In particular, we take the challenge posed by current small size COVID data and show how significant can be the bias introduced by transfer-learning using larger public non-COVID CXR datasets. We also contribute by providing results on a medium size COVID CXR dataset, just collected by one of the major emergency hospitals in Northern Italy during the peak of the COVID pandemic. These novel data allow us to contribute to validate the generalization capacity of preliminary results circulating in the scientific community. Our conclusions shed some light into the possibility to effectively discriminate COVID using CXR.**

***Index Terms*—Chest X-ray, Deep Learning, classification, COVID-19**

I. Introduction

COVID-19 virus has rapidly spread in mainland China and into multiple countries worldwide [1]. As of April 7th 2020 in Italy, one of the most severely affected countries, 135,586 Patients with COVID19 were recorded, and 17,127 of them died; at the time of writing Piedmont is the 3rd most affected region in Italy, with 13,343 recorded cases [2].

Early 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 Center for Disease Control (CDC) in the US 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 [3]. 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 instead easily executable and affortable 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% [4].

Current position papers from radiological societies (Fleischner Society, SIRM, RSNA) [3], [5], [6] 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) and computed tomography (CT) scans can show pathological findings. It should be noted that they are actually non specific, and overlap with other viral infections (such as influenza, H1N1, SARS and MERS): most authors report peripheral bilateral ill-defined and ground-glass opacities, mainly involving the lower lobes, progressively increasing in extension as disease becomes more severe and leading to diffuse parenchymal consolidation [7], [8]. CT is a sensitive tool for early detection of peripheral ground glass opacities; however routine role of CT imaging in these Patients is logistically challenging in terms of safety for health professionals and other patients, and can overwhelm available resources [9].

Chest X-ray can be a useful tool, especially in emergency settings: it can help exclude other possible lung "noxa", allow a first rough evaluation of the extent of lung involvement and most importantly can be obtained at patient's bed using portable devices, limiting possible exposure in health care workers and other patients. Furthermore, CXR can be repeated over time to monitor the evolution of lung disease [5].

Wong *et al.* in a study recently published on Radiology reported that x-ray has a sensitivity of 69% and that the severity of CXR findings peaked at 10-12 days from the date of symptom onset [8].

Because of their mostly peripheral distribution, subtle early findings on CXRs may be a diagnostic challenge even for an experienced thoracic radiologist: in fact, there are many factors that should be taken into account in image interpretation and that could alter diagnostic performance (such as patient body type, compliance in breath-holding and positioning, type of projection that can be executed i.e. antero-posterior in more critical patients examined at bedside, postero-anterior if the patient can be moved to radiology unit and is more collaborating, presence of other medical devices on the thorax, especially in x rays performed in intensive care units, etc.). In the challenging and never-before seen scenario that rose to attention in the last months, radiologists may look at Artificial Intelligence and deep learning applications as a possible aid for daily activity, in particular for identification of the more subtle findings that could "escape" the human eye (i.e. reduce false-negative x-rays) or, on the other side, could prompt swab repetition of further diagnostic examinations when first virus testing is negative (considering its sub-optimal sensitivity).

Given the intrinsic limits of CXR but at the same time its potential relevant role in the fight against COVID 19, in this work we set up a state of the art deep learning pipeline to investigate if computer vision can unveil some COVID fingerprints. It is evident that the answer will be given only when publicly available large image datasets will empower scientists to train complex neural models, to provide reproducible and statistically solid results and to contribute to the clinical discussion. Unfortunately, up to date, we are stuck with few labelled images. Thanks to the collaboration with the radiology unit of Città della Salute e della Scienza di Torino (CDSS) hospital in Turin in the last days of March (at the peak of epidemic in Italy), we managed to collect the Covid Radiographic images Data-set for AI (CORDA), currently comprising images from 386 Patients that underwent COVID screening. The data are still limited but, using them to train and test Convolutional Neural Network (CNN) architectures such as *resnet* [10], we contribute to shed some light into the problem. In this work we do not mean to answer whether and how CXR can be used in the early diagnosis of COVID, but to provide a methodological guide and critical reading of the statistical results that can be obtained using currently available datasets and learning mechanisms. Our main contribution is an extensive experimental evaluation of different combinations of usage of existing datasets for pre-training and transfer learning of standard CNN models. Such analysis allows us to raise some warnings on how to build datasets, pre-process data and train deep models for COVID classification of X-ray images. We show that, given the fact that datasets are still small and geographically local, subtle biases in the pre-trained models used for transfer learning can emerge, dramatically impacting on the significance of the performance one achieves.

II. Related works

It is evident that currently there is not yet a significant amount of work devoted to automatic detection of covid from medical imaging. Nonetheless, one can refer to previous epidemics caused by novel strain of coronavirus such as severe acute respiratory syndrome (SARS), first recognized in Canada in March 2003, characterised by similar lung condition, i.e. interstitial pneumonia [11]. Most results leverage on the use of high resolution CT scans. As an example, in [12] CNN are investigated for classification of interstitial lung disease (ILD). Also [13], [14] show that deep learning can be used to detect and classify ILD tissue. The authors of [14] focus on a design a CNN tailored to match the ILD CT texture features, e.g. small filters and no pooling to guarantee spatial locality.

Fewer contributions focus on classification of X-ray chest images to help SARS diagnosis: in [15] lung segmentation, followed by feature extraction and three classification algorithms, namely decision tree, shallow neural network and classification and regression tree are compared, the latter yielding the higher accuracy on the SARS detection task. However, on the pneumonia classification task, NN-based approaches show encouraging results. In [16] texture features for SARS identification in radiographic images are proposed and designed using signal processing tools.

In the last days a number of preprints targeting covid classification with CNN on radiographic images have begun to circulate thanks to open access archives. Many approaches have been taken to tackle the problem of classifying chest X-ray scans to discriminate COVID-positive cases. For example, Sethy *et al.* compare classification performances obtained between some of the most famous convolutional architectures [17]. In particular, they use a transfer learning-based approach: they take pre-trained deep networks and they use these models to extract features from images. Then, they train a SVM on these “deep features” to the COVID classification task. A similar approach is also used by Apostopolous *et al.*: they pre-train a neural network on a similar task, and then they use the trained convolutional filters to extract features, on top of which a classifier attempts to select COVID features [18]. Narin *et al.* make use of resnet-based architectures and the recent Inception v3 and then they use a 5-fold cross validation strategy [19]. Finally, Wang *et al.* propose a new neural network architecture to be trained on the COVID classification task [20].

All of these approaches use a very small dataset, *COVID-ChestXRay* [21], consisting of approximately 100 COVID cases considering CXR only. Furthermore, in order to build COVID negative cases, typically data are sampled from other datasets (mostly, from *ChestXRay*). However, this introduces a potential issue: if any bias is present in the dataset (a label in the corners, a medical device, or other contingent factors like similar age, same sex etc.) the deep model could learn to recognize these dataset biases, instead of focusing on COVID-related features.

These works present some potential issues to be investigated:

- Transfer learning: in the literature it is widely recognized that transfer learning-based approaches prove to be effective, also for medical imaging [22]. However, it is very important to be careful on the particular task the feature extractor is trained on: if such task is very specific, or contains biases, then the transfer learning approach should be carefully carried on.
- Hidden biases in the dataset: most of the current works rely on very small datasets, due to the limited availability of public data on COVID positive cases. These few data, then, contain little or even no metadata on age, gender, different pathologies also present in these subjects, and other necessary information necessary to spot on this kind of biases. Besides these, there are other biases we can try to correct. For example, every CXR has its own image windowing parameters or other acquisition settings that a deep model could potentially learn to discriminate. For example, one model may cluster images according to the scan tool used for the exam; if some scan settings correspond to all COVID examples, these will generate a spurious correlation that the model can exploit to yield apparently optimal classification accuracy. Another

example is given by textual labeling in images: if all the negative examples are sampled from the same dataset, the deep model could learn to recognize such feature instead of focusing on the lung content etc.
- Very small test sets: as a further consequence of having very little data, test set sizes are extremely small and they do not provide any statistical certainty on learning.

III. Methodology

In this section we are going to describe the proposed deep-learning approach based on quite standard pipeline, namely chest image pre-processing and lung segmentation followed by classification model obtained with transfer learning. As we will see in this section, data pre-processing is fundamental to remove any bias present in the data. In particular, we will show that it is easy for a deep model to recognize these biases which drive the learning process. Given the small size of COVID datasets, a key role is played by the larger datasets used for pre-training. Therefore, we first discuss which datasets can be used for our goals.

A. Datasets

For the experiments we are going to show, three different datasets will be used. Each of these contain CXR images, but their purpose is different:

- *COVID-ChestXRay*: this dataset was developed by gathering CXR and CT images from various website and publications. At the time of writing, it comprises 287 images with different type of pneumonias (COVID-19, SARS, MERS, Streptococcus spp., Pneumocystis spp., ARDS) [21]. Currently, a subset of 137 CXRs (PA) containing 108 COVID positive images and 29 COVID negatives is available.[1]
- *CORDA*: this dataset was created for this study by retrospectively selecting chest x-rays performed at a dedicated Radiology Unit in a reference Hospital in Piedmont (CDSS) in all patients with fever or respiratory symptoms (cough, shortness of breath, dyspnea) that underwent nasopharingeal swab to rule out COVID-19 infection. Patients were collected over a 15-day period between the 16th and 30th March, 2020. It contains 447 CXRs from 386 patients, with 150 images coming from COVID-negative patients and 297 from positive ones. Patients' average age is 61 years (range 17-97 years old). 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.
- *Montgomery County X-ray Set*: the X-ray images in this dataset have been acquired under a tuberculosis control program of the Department of Health and Human Services of the Montgomery County, MD, USA. Such a dataset contains 138 samples: 80 are normal patients and 58 are abnormal. In these images lungs have been manually segmented. The dataset is open-sorce.[2]

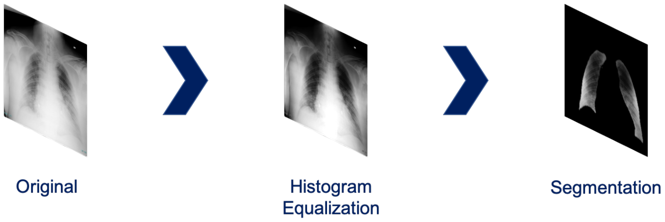


Fig. 1: CXR pre-processing steps proposed.

- *Shenzhen Hospital X-ray Set*: the X-ray images in this dataset have been collected by Shenzhen No.3 Hospital in Shenzhen, Guangdong providence, China. This dataset contains a total of 662 images: 326 images are from healthy patients while 336 show abnormalities. Such a dataset is also open-source.[3] Ground truths for this dataset have been provided by Stirenko *et al.* [23].
- *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 [24].[4]
- *RSNA*: developed by the joint effort of the *Radiological Society of North America*,*US National Institues of Health*, *The Society of Thoracic Radiology* and *MD.ai* for the RSNA Pneumonia Detection Challenge, this dataset contains pneumonia cases found in the NIH Chest X-ray dataset [25]. It comprises 20672 normal CXR scans and 6012 pneumonia cases, for a total of 26684 images.[5]

B. Pre-processing

For our simulations we propose a pre-processing strategy aiming at removing bias in the data. This step is very important in a setting in which we train to discriminate different classes belonging to different datasets: a neural network-based model might learn the distinction between the different dataset biases and from them "learn" the classification task. The proposed pre-processing chain is summarized in Fig. 1 and is based on the following steps:

- Histogram equalization: when acquiring a CXR, the so-called radiographic contrast depends on a large variety of factors, typically depending on subject contrast, receptor contrast or other factors like scatter radiations [26]. Hence, the raw acquisition has to be filtered through Value Of Interest transformation. However, due to different calibrations, different range dynamics can be covered,

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

[2] http://openi.nlm.nih.gov/imgs/collections/NLM-MontgomeryCXRSet.zip

[3] http://openi.nlm.nih.gov/imgs/collections/ChinaSet_AllFiles.zip

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

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

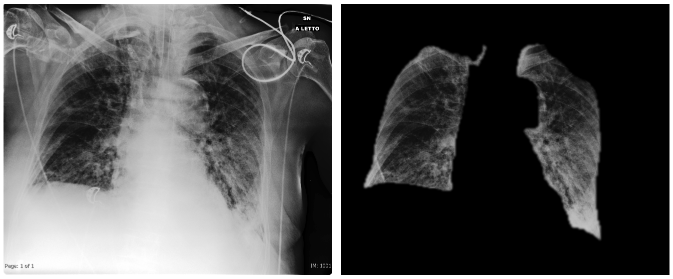


Fig. 2: Original image (a) and extracted lung segmented image (b). Many possible bias sources like all the writings and medical equipment is naturally removed.

and this potentially is a bias. Histogram equalization is a simple mean to guarantee quite uniform image dynamic in the data.

- Lung segmentation: the lung segmentation problem has been already faced and successfully tackled [27]–[29]. Being able to segment the lungs only, discarding all the rest of the CXRs, potentially prunes away possible bias sources, like for example the presence of medical devices (typically correlated to sick patients), various text which might be embed in the scan etc. In order to address this task, we train a U-Net [30] on *Montgomery County X-ray Set* and *Shenzhen Hospital X-ray Set*. The lung masks obtained are then blurred to avoid sharp edges using a 3 pixel radius. An example of the segmentation outcome is shown in Fig. 2.
- Image intensity normalization in the range $[0, 1]$.

C. Training

After data have been pre-processed, a deep model will be trained. Towards this end, the following choices have been taken:

- Pre-training on the feature extractor, i.e. convolutional layers of the CNN, will be performed. In particular, the pre-training will be performed on a related task, like pneumonia classification for CXRs. It has been shown that such an approach can be effective for medical imaging [12], in particular when the amount of available data is limited as in our classification task. Clearly, pre-training the feature extractor on a larger dataset containing related features may allow us to exploit deeper models, potentially exploiting richer image feature.
- The feature extractor will be fine-tuned on COVID data. Freezing it will certainly prevent over-fitting the small COVID data; however, we have no warranty that COVID related features can be extracted at the output of a feature extractor trained on a similar task. Of course, its initialization on a similar task helps in the training process, but in any case a fine-tuning is still necessary [31].
- Proper sizing of the encoder to-be-used is an issue to be addressed. Despite many recent works use deeper architectures to extract features on the COVID classification task, larger models are prone to over-fit data. Considering the minimal amount of data available, the choice of the appropriate deep network complexity significantly affects the performance.
- Balancing the training data is yet another extremely important issue to be considered. Unbalanced data favor biases in the learning process. Such balancing issue can be addressed in a number of ways: the most common and simple way to solve this issue is adding or removing data from the training-set. Removing data from a tiny dataset is not a viable approach; considering that the COVID datasets are built mainly of positive cases, one solution is to augment them with negative cases from publicly-available datasets. However, this is a very delicate operation and needs to be done very carefully: if all the negative cases are made of non-pathological patients, the deep model will not necessarily learn COVID features. It may simply discriminate between healthy and unhealthy lung. Providing a good variety of conditions in the negative data is not an easy task. The choice of the images may turn to be critical and, just like in the pre-training phase, one can include unwanted biases: again the model can end up classifying new images (that are positive to covid) exploiting discriminative biases present in different datasets.
- Testing with different data than those used at training time is also fundamental. Excluding from the test-set exams taken from patients already present in the training-set is important to correctly evaluate the performance and to exclude the deep model has not learned a “patient’s lung shape” feature.
- Of course many other issues have to be taken into account at training time, like the use of a validation-set to tune the hyper-parameters, using a good regularization policy etc. but these very general issues have been exhaustively discussed in many other works [32]–[34].

IV. Discussion

In this section we present and comment the experimental results obtained on a combination of different datasets (introduced in Sec. III-A). All the simulations have been run on a Tesla T4 GPU using PyTorch 1.4.[6] The performance obtained on a comprehensive number of experiments is presented in Tab. II and Tab. III. In these tables, in particular, three factors will be evaluated:

- Pre-training of the feature extractor: the feature extractor can be pre-trained on large generic CXR datasets, or can not be pre-trained.
- Composition of the training-set: the CORDA dataset is un-balanced (in fact, there is a prevalence of positive COVID cases) and some data balancing is possible, borrowing samples from publicly available non-COVID datasets. A summary of the dataset composition is desplayed in Table I. For all the datasets we used 70% of

[6] The source code is available at https://github.com/EIDOSlab/unveiling-covid19-from-cxr

TABLE I: Datasets composition. The datasets used at training and test time are in the rows, and the total data are in the last two columns.

COMPOSED DATASET		**BORROWED DATA FROM ORIGINAL DATASETS**								**TOTAL**	
		CORDA		**RSNA**		**ChestXRay**		**COVID-ChestXRay**			
		COVID+	COVID-	COVID+	COVID-	COVID+	COVID-	COVID+	COVID-	COVID+	COVID-
CORDA	train	126	105	-	-	-	-	-	-	126	105
	test	90	45	-	-	-	-	-	-	90	45
CORDA&ChestXRay	train	207	105	-	-	-	102	-	-	207	207
	test	90	45	-	-	-	45	-	-	90	90
CORDA&RSNA	train	207	105	-	102	-	-	-	-	207	207
	test	90	45	-	45	-	-	-	-	90	90
CORDA&COVID-ChestXRay	train	116	105	-	-	-	-	49	24	165	129
	test	90	45	-	-	-	-	10	5	100	50
COVID-ChestXRay	train	-	-	-	-	-	-	98	24	98	24
	test	-	-	-	-	-	-	10	5	10	5

data at training time and 30% as test-set. Training data are then further divided in training-set (80%) and validation-set (20%). Training-set data are finally balanced between COVID+ and COVID-: where possible, we increased the COVID- cases (CORDA&ChestXRay, CORDA&RSNA), where not possible we subsampled the more populated class. This percentages were not used for the COVID-ChestXRay dataset: in this case only 15 samples are used for testing in order to compare with other works [17]–[19] that use the same partitioning.

- Testing on different datasets: in order to observe the possible presence of hidden biases, testing on different, qualitatively-similar datasets is a necessary step.

For all of these trained models, a number of metrics [35] will be evaluated:

- Accuracy.
- AUC (area under the ROC curve), provides an aggregate measure of performance across all possible classification thresholds. For every other metric, the classification threshold is set to 0.5.
- Sensitivity.
- Specificity.
- F-score.
- BA (balanced accuracy), since the test-set might be unbalanced.
- DOR (diagnostic odds ratio).

Results are shown in Table II, Table III and Table IV.

A. To pre-train or not to pre-train?

One very important issue to pay attention to is whether to pre-train the feature extractor or not. Given the large availability of public data for pneumonia classification (for example, in this scope we used ChestXRay and RSNA), it could be a good move to pre-train the encoder, and effectively this is what we observe looking at Table II. For example, if we focus on the results obtained training on the CORDA dataset, without a pre-trained encoder, BA and DOR are lower than pre-training with ChestXRay or RSNA. Despite the sensitivity remains very similar, pre-training the encoder helps in improving the specificity: on the test-set extracted from CORDA, using a pre-trained encoder on RSNA, the specificity is 0.80, while it is only 0.58 with no pre-trained feature extractor. Similar improvements in the specificity can be observed also on test-sets extracted from all the other datasets, except for ChestXRay. In general, a similar behavior can be observed when comparing results for differently pre-trained encoders trained on the same dataset.

Pre-training is important; however, we can not just “freeze” the encoder on the pre-trained values. Since the encoder is pre-trained on a similar, but different task, there is no warranty the desired output features are optimal for the given classification task, and a fine-tuning step is typically required [36].

B. Pre-training on different datasets

Focusing on pre-trained encoders, we show results for encoders pre-trained on two different datasets: ChestXRay and RSNA. While RSNA is a more generic pneumonia-segmentation dataset, ChestXRay contains information also about the type of pneumonia (bacterial or viral); so, at a first glance it looks a better fit for the pre-training. However, if we look at training on the CORDA dataset, we see that for the same sensitivity value, we get typically higher specificity scores for RSNA pre-training. This is not the same we observe when we compare results on the publicly-available COVID-ChestXRay: in this case, sensitivity and specificity are higher when we pre-train on ChestXRay. Looking at the same pre-trained encoder, let us say ChestXRay, we can compare results training on CORDA and on COVID-ChestXRay, which are the two COVID datasets: CORDA shows a lower sensitivity, but in general a higher specificity, except for the ChestXRay dataset. Having very little data at training time, pre-training introduces some priors in the choice of the features to be used, and depending on the final classification task, performance changes, yielding very good metric in some cases. Pre-training on more general datasets, like RSNA, in general looks a slightly better choice than using a more specific dataset like ChestXRay.

C. Augmenting COVID- data with different datasets

For each and every simulation, performance on different test-sets is evaluated. This gives us hints on possible biases introduced by different datasets used at training time.

A general trend can be observed for many COVID- augmented

TABLE II: Results obtained training ResNet-18.

Pre-trained encoder	Training dataset	Test dataset	Sensitivity	Specificity	F-Score	Accuracy	BA	AUC	DOR
none	CORDA&RSNA	CORDA	0.56	0.42	0.60	0.51	0.49	0.52	0.91
		CORDA&ChestXRay	0.56	0.22	0.15	0.26	0.39	0.33	0.36
		CORDA&RSNA	0.56	0.96	0.49	0.95	0.76	0.95	34.23
		CORDA&COVID-ChestXRay	0.52	0.48	0.58	0.51	0.50	0.53	1.00
	CORDA	CORDA	0.56	0.58	0.63	0.56	0.57	0.59	1.71
		CORDA&ChestXRay	0.56	0.37	0.18	0.39	0.46	0.43	0.74
		CORDA&RSNA	0.56	0.38	0.08	0.39	0.47	0.46	0.76
		CORDA&COVID-ChestXRay	0.56	0.58	0.63	0.57	0.57	0.59	1.76
	CORDA&COVID-ChestXRay	CORDA	0.58	0.64	0.66	0.60	0.61	0.63	2.48
		CORDA&ChestXRay	0.58	0.63	0.27	0.63	0.61	0.63	2.37
		CORDA&RSNA	0.58	0.54	0.11	0.54	0.56	0.58	1.62
		CORDA&COVID-ChestXRay	0.57	0.66	0.66	0.60	0.61	0.64	2.57
	COVID-ChestXRay	CORDA	0.91	0.11	0.77	0.64	0.51	0.54	1.28
		CORDA&ChestXRay	0.91	0.66	0.41	0.69	0.78	0.87	19.56
		CORDA&RSNA	0.91	0.11	0.09	0.14	0.51	0.45	1.22
		CORDA&COVID-ChestXRay	0.91	0.18	0.78	0.67	0.55	0.58	2.22
	CORDA&ChestXRay	CORDA	0.88	0.18	0.77	0.64	0.53	0.58	1.55
		CORDA&ChestXRay	0.88	0.94	0.76	0.93	0.91	0.97	112.93
		CORDA&RSNA	0.88	0.14	0.09	0.17	0.51	0.42	1.14
		CORDA&COVID-ChestXRay	0.87	0.20	0.77	0.65	0.54	0.60	1.67
RSNA	CORDA&RSNA	CORDA	0.68	0.44	0.69	0.60	0.56	0.61	1.68
		CORDA&ChestXRay	0.68	0.22	0.18	0.27	0.45	0.49	0.59
		CORDA&RSNA	0.68	0.90	0.37	0.89	0.79	0.90	19.82
		CORDA&COVID-ChestXRay	0.67	0.50	0.70	0.61	0.58	0.63	2.03
	CORDA	CORDA	0.54	0.80	0.66	0.63	0.67	0.72	4.78
		CORDA&ChestXRay	0.54	0.31	0.16	0.34	0.43	0.48	0.54
		CORDA&RSNA	0.54	0.55	0.10	0.55	0.55	0.61	1.48
		CORDA&COVID-ChestXRay	0.57	0.76	0.67	0.63	0.67	0.72	4.20
	CORDA&COVID-ChestXRay	CORDA	0.70	0.49	0.72	0.63	0.59	0.67	2.23
		CORDA&ChestXRay	0.70	0.30	0.20	0.34	0.50	0.59	0.98
		CORDA&RSNA	0.70	0.37	0.10	0.39	0.53	0.61	1.37
		CORDA&COVID-ChestXRay	0.71	0.52	0.73	0.65	0.61	0.70	2.65
	COVID-ChestXRay	CORDA	0.94	0.09	0.79	0.66	0.52	0.57	1.66
		CORDA&ChestXRay	0.94	0.61	0.39	0.65	0.78	0.92	26.24
		CORDA&RSNA	0.94	0.08	0.09	0.12	0.51	0.58	1.50
		CORDA&COVID-ChestXRay	0.95	0.14	0.80	0.68	0.54	0.62	3.09
	CORDA&ChestXRay	CORDA	0.82	0.38	0.77	0.67	0.60	0.63	2.81
		CORDA&ChestXRay	0.82	0.95	0.75	0.94	0.89	0.97	89.14
		CORDA&RSNA	0.82	0.30	0.10	0.32	0.56	0.59	1.98
		CORDA&COVID-ChestXRay	0.83	0.38	0.78	0.68	0.60	0.64	2.99
ChestXRay	CORDA&RSNA	CORDA	0.86	0.31	0.78	0.67	0.58	0.60	2.67
		CORDA&ChestXRay	0.86	0.29	0.24	0.36	0.58	0.48	2.47
		CORDA&RSNA	0.86	0.95	0.61	0.95	0.90	0.97	122.64
		CORDA&COVID-ChestXRay	0.82	0.38	0.77	0.67	0.60	0.61	2.79
	CORDA	CORDA	0.54	0.58	0.62	0.56	0.56	0.67	1.64
		CORDA&ChestXRay	0.54	0.37	0.17	0.39	0.46	0.49	0.70
		CORDA&RSNA	0.54	0.73	0.15	0.72	0.64	0.72	3.21
		CORDA&COVID-ChestXRay	0.56	0.62	0.64	0.58	0.59	0.70	2.08
	CORDA&COVID-ChestXRay	CORDA	0.71	0.49	0.72	0.64	0.60	0.67	2.35
		CORDA&ChestXRay	0.71	0.25	0.20	0.31	0.48	0.51	0.83
		CORDA&RSNA	0.71	0.47	0.11	0.48	0.59	0.64	2.16
		CORDA&COVID-ChestXRay	0.73	0.52	0.74	0.66	0.62	0.70	2.93
	COVID-ChestXRay	CORDA	0.91	0.20	0.79	0.67	0.56	0.61	2.56
		CORDA&ChestXRay	0.91	0.70	0.44	0.73	0.81	0.89	24.38
		CORDA&RSNA	0.91	0.15	0.09	0.19	0.53	0.55	1.83
		CORDA&COVID-ChestXRay	0.92	0.28	0.81	0.71	0.60	0.66	4.47
	CORDA&ChestXRay	CORDA	0.88	0.24	0.78	0.67	0.56	0.66	2.32
		CORDA&ChestXRay	0.88	0.94	0.77	0.94	0.91	0.97	122.67
		CORDA&RSNA	0.88	0.24	0.10	0.27	0.56	0.67	2.26
		CORDA&COVID-ChestXRay	0.88	0.26	0.78	0.67	0.57	0.68	2.58

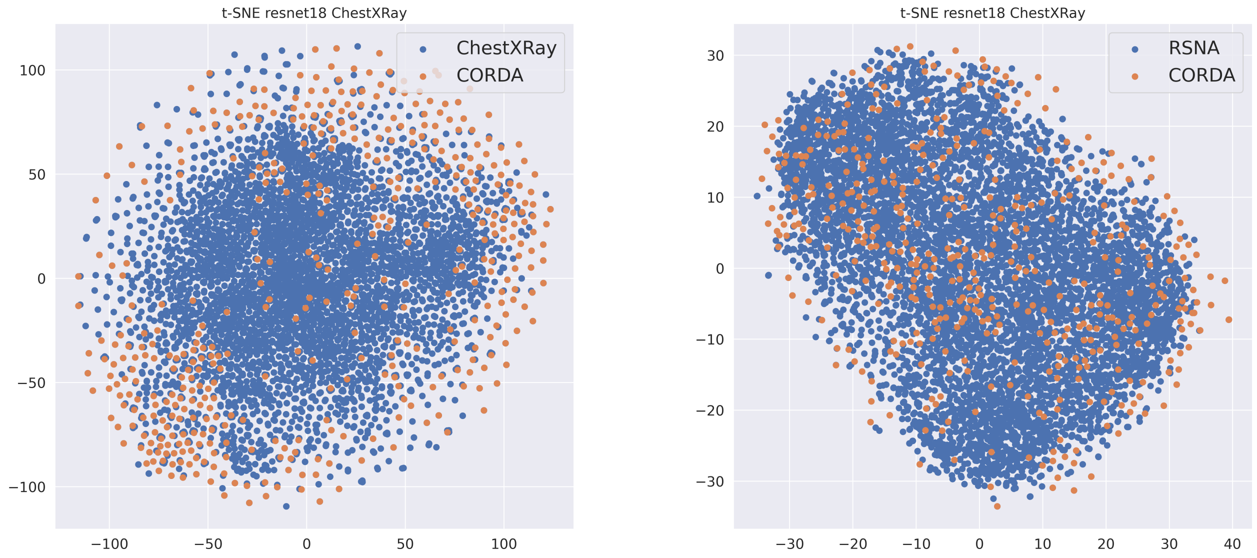
TABLE III: Results obtained training a ResNet-50 model.

Pre-trained encoder	Training dataset	Test dataset	Sensitivity	Specificity	F-Score	Accuracy	BA	AUC	DOR
RSNA	CORDA&RSNA	CORDA	0.74	0.49	0.74	0.66	0.62	0.65	2.79
		CORDA&ChestXRay	0.74	0.40	0.24	0.44	0.57	0.64	1.92
		CORDA&RSNA	0.74	0.92	0.43	0.91	0.83	0.93	31.76
		CORDA&COVID-ChestXRay	0.70	0.54	0.73	0.65	0.62	0.66	2.74
	CORDA	CORDA	0.61	0.71	0.70	0.64	0.66	0.67	3.87
		CORDA&ChestXRay	0.61	0.40	0.20	0.43	0.51	0.53	1.06
		CORDA&RSNA	0.61	0.58	0.12	0.58	0.60	0.63	2.20
		CORDA&COVID-ChestXRay	0.62	0.74	0.71	0.66	0.68	0.69	4.64
	CORDA&COVID-ChestXRay	CORDA	0.53	0.64	0.62	0.57	0.59	0.64	2.07
		CORDA&ChestXRay	0.53	0.56	0.22	0.56	0.55	0.58	1.47
		CORDA&RSNA	0.53	0.57	0.10	0.57	0.55	0.58	1.53
		CORDA&COVID-ChestXRay	0.55	0.68	0.64	0.59	0.61	0.66	2.60
	COVID-ChestXRay	CORDA	0.97	0.04	0.79	0.66	0.51	0.57	1.35
		CORDA&ChestXRay	0.97	0.45	0.32	0.51	0.71	0.89	23.29
		CORDA&RSNA	0.97	0.09	0.09	0.13	0.53	0.56	2.91
		CORDA&COVID-ChestXRay	0.97	0.10	0.80	0.68	0.54	0.62	3.59
	CORDA&ChestXRay	CORDA	0.76	0.33	0.72	0.61	0.54	0.65	1.55
		CORDA&ChestXRay	0.76	0.95	0.72	0.93	0.85	0.97	63.61
		CORDA&RSNA	0.76	0.36	0.10	0.38	0.56	0.63	1.75
		CORDA&COVID-ChestXRay	0.76	0.32	0.72	0.61	0.54	0.64	1.49
ChestXRay	CORDA&RSNA	CORDA	0.73	0.40	0.72	0.62	0.57	0.58	1.83
		CORDA&ChestXRay	0.73	0.25	0.20	0.31	0.49	0.44	0.92
		CORDA&RSNA	0.73	0.96	0.58	0.95	0.85	0.97	68.71
		CORDA&COVID-ChestXRay	0.70	0.46	0.71	0.62	0.58	0.60	1.99
	CORDA	CORDA	0.64	0.56	0.69	0.61	0.60	0.65	2.27
		CORDA&ChestXRay	0.64	0.49	0.24	0.51	0.57	0.61	1.72
		CORDA&RSNA	0.64	0.63	0.14	0.63	0.64	0.69	3.06
		CORDA&COVID-ChestXRay	0.67	0.60	0.72	0.65	0.64	0.69	3.05
	CORDA&COVID-ChestXRay	CORDA	0.63	0.38	0.65	0.55	0.51	0.63	1.05
		CORDA&ChestXRay	0.63	0.46	0.22	0.48	0.55	0.61	1.46
		CORDA&RSNA	0.63	0.62	0.14	0.62	0.63	0.70	2.86
		CORDA&COVID-ChestXRay	0.65	0.44	0.67	0.58	0.55	0.66	1.46
	COVID-ChestXRay	CORDA	0.98	0.13	0.81	0.70	0.56	0.61	6.77
		CORDA&ChestXRay	0.98	0.72	0.48	0.75	0.85	0.90	112.57
		CORDA&RSNA	0.98	0.11	0.10	0.15	0.55	0.61	5.59
		CORDA&COVID-ChestXRay	0.98	0.20	0.82	0.72	0.59	0.65	12.25
	CORDA&ChestXRay	CORDA	0.81	0.29	0.75	0.64	0.55	0.64	1.74
		CORDA&ChestXRay	0.81	0.94	0.73	0.93	0.88	0.97	73.35
		CORDA&RSNA	0.81	0.25	0.09	0.28	0.53	0.57	1.43
		CORDA&COVID-ChestXRay	0.80	0.30	0.74	0.63	0.55	0.64	1.71

TABLE IV: Results of COVID-Net and ResNet-18 training on COVID-ChestXRay.

Architecture	Test dataset	Sensitivity	Specificity	F-Score	Accuracy	BA	AUC	DOR
COVID-Net	CORDA	0.12	0.98	0.22	0.41	0.55	0.55	6.68
COVID-Net	COVID-ChestXRay	0.90	0.80	0.90	0.85	0.85	0.85	36.00
ResNet-18	CORDA	0.91	0.20	0.79	0.67	0.56	0.61	2.56
ResNet-18	COVID-ChestXRay	1.00	1.00	1.00	1.00	1.00	1.00	∞

training-sets: the BA and DOR scores measured on the test-set built from the same dataset used at training time are typically very high. Let us focus on the ChestXRay pre-trained encoder. When we train on CORDA&ChestXRay, the BA score measured on the test-set from the same dataset is 0.9 and the DOR is 122.67. However, its generalization capability for a different composition of the test-set, let us say, CORDA&RSNA, is way lower: the BA is 0.56 and the DOR 2.26 only. The same scenario can be observed when we train on CORDA&RSNA: on its test-set the BA is 0.90 and DOR 122.64, while on the test-set of CORDA&ChestXRay the BA is 0.59 and DOR 2.47. The key to understand these results lies again in the specificity score: this score is extremely high for the test-set of the same dataset the training is performed on (for example, for CORDA&RSNA is 0.95 and for CORDA&ChestXRay is 0.94) while for the others is extremely low. Such a behavior is due to the presence of some common features in all the data belonging to the same augmenting dataset. This can be observed, for example, in Fig. 3a, where the extracted features from an encoder pre-trained on ChestXRay and trained on CORDA&ChestXRay are clustered using t-SNE [37] (blue and orange dots represent ChestXray and CORDA data samples respectively, regardless of the COVID label). It can be noted that CORDA samples, regardless the COVID+ or COVID- label, are clearly separable from ChestXRay data. Of course, all ChestXRay images have COVID- label, so someone could argue that the COVID feature has been captured. Unfortunately we have a counterexample: in Fig. 3b we compare CORDA vs. RSNA samples, using the same ChestXRay pre-trained encoder and now RSNA and



(a)

(b)

Fig. 3: t-SNE on ChestXRay trained encoder. (a) shows ChestXRay data vs CORDA data, (b) instead shows RSNA vs CORDA.

CORDA samples no longer form clear clusters. Hence, the deep model specializes not in recognizing COVID features, but in learning the common features in the same dataset used at training time. We would like to remark that for all the data used at training or at test time, all the pre-processing presented in Sec. III-B has been used. We ran the same experiments without that pre-processing and performance on different datasets than the one used at training time gets even worse. For example, pre-training and training on CORDA&ChestXRay without pre-processing lowers the BA to 0.73 and the DOR to 8.31 on CORDA, while from Table II we have higher scores on the test set (BA of 0.91 and DOR of 122.67).

Dealing with generality of the results is a very delicate matter: what it is possible to see in Tab. II is that augmenting data with COVID- data needs to be very thoughtful since the classification performance may vary from very high accuracy down to almost useless discriminative power. Nonetheless, training using only COVID datasets yields some promising scores: for example, using ChestXRay pre-trained encoder and CORDA for training and testing, the BA we achieve is 0.56 and the DOR is 1.64. Including also COVID-ChestXRay for training (which consists in having more COVID+ and COVID- examples) improves the BA to 0.62 and the DOR to 2.93. In this case, however, the specificity is an issue, since we lack of COVID- data. However, these results show some promise that can be confirmed only by collecting large amount of data in the next months.

D. *How deep should we go?*

After reviewing results on ResNet-18, we move to similar experiments run on the deeper ResNet-50 shown in ab. III. The hope is that a deeper network could extract more representative features for the classification task. Given the discussion in Sec. IV-A, we show only the cases with pre-training of the feature extractor. Using this deeper architecture, we can observe that all the discussions made for ResNet-18 still holds. In some cases performance impairs slightly: for example, the DOR score on CORDA&ChestXRay for ResNet-18 was 122.67 while for ResNet-50 drops to 73.35. This is a sign of over-fitting: given the very small quantity of data currently available, using a small convolutional neural network is sufficient and safer. Taking an opposite approach, we tried to use a smaller artificial neural network, made of 8 convolutional layers and a final fully-connected layer, which takes inspiration from the ALL-CNN-C architecture [38]. We call this architecture Conv8. The results on this smaller architecture are similar to those observed in Table II. For example, training the model on CORDA dataset, on Conv8 we have a BA of 0.61 and DOR of 2.38 while for ResNet-18 with encoder pre-trained on RSNA we have BA of 0.67 and DOR 4.78. We can conclude that using a smaller architecture than ResNet-18 does not give relevant training advantages, while by using larger architectures we might over-fit data.

E. *Comparison between deep networks trained on COVID-ChestXRay*

All the observations on train and test data made above are also valid for the recently published results on the COVID classification from CXR [17]–[20]. One very promising approach is COVID-Net [20]. They also share the source code and the trained model.[7] In Tab. IV we compare the classification metrics obtained with COVID-Net and our ResNet-18 model: both models have been trained using COVID-ChestXRay, and tested on both CORDA and COVID-ChestXRay. In line with the discussion above we can note that both COVID-Net and ResNet-18 yields surprising results when the same dataset is used for traning and testing: The performance of COVID-Net on the COVID-ChestXRay test-set (the same dataset used at training time) is very high (BA of 0.85 and DOR of 36.0) while it drops significantly when tested on CORDA, where BA is 0.55 only and DOR is 6.68. This drop is explained looking at sensitivity and specificity: it is evident that the model classifies as COVID- almost all the data. A similar behavior can be observed also in ResNet-18 model: the observed performance apparently looks incredible (since that the BA on the test-set is 1.0), and in fact similar numbers are also claimed in the other works on ResNet-like architectures [17]–[19]. However, testing on CORDA reveals that the deep model is likely to have learned some hidden biases in COVID-ChestXRay and tends to mis-classify COVID- samples as COVID+ (given that the specificity is here 0.20).

V. Conclusions

One of the very recent challenges for both clinical and AI community is to use deep learning to learn to discriminate COVID from CXR. Some recent works highlighted the possibility of successfully tackle this problem, despite the currently small quantity of publicly available data. In this work we have highlighted possible obstacles in successfully training a deep model, ranging from the proper choice of the architecture to-be-trained to handling removable biases in medical datasets. Extensive experiments show that extracting a "COVID" feature from CXR is not an easy task. Such a problem should be addressed very carefully: it is very easy to misinterpret very good results on test-data, still showing poor generalization on new data in the same domain. We could perform such a test thanks to the possibility of using CORDA, a larger dataset comprising COVID cases. Of course, the quantity of available data is still limited but allowed us to find some promising seminal classification results. The ongoing collection and sharing of large amount of CXR data is the only way to further investigate if promising CNN results can aid in the fight to COVID pandemic.

Acknowledgment

The authors wish to thank Prof. Paolo Fonio (Radiology and Interventional Radiology department of Città della Salute e della Scienza di Torino), Dr. Giorgio Limerutti (head of Radiology 2) and Marco Grosso for letting the CORDA data collection start, Stefano Tibaldi, Simona Capello e Patrizia Sardo for their valuable help in CXR anonymization and labeling.

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

References

[1] Z. Y. Zu, M. D. Jiang, P. P. Xu, W. Chen, Q. Q. Ni, G. M. Lu, and L. J. Zhang, "Coronavirus disease 2019 (covid-19): A perspective from china," *Radiology*, p. 200490, 2020.

[2] "Covid-19: situation report by italian ministry of health," http://www.salute.gov.it.

[3] "ACR recommendations for the use of chest radiography and computed tomography (CT) for suspected COVID-19 infection," https://www.acr.org/.

[4] Y. Yang, M. Yang, C. Shen, F. Wang, J. Yuan, J. Li, M. Zhang, Z. Wang, L. Xing, J. Wei *et al.*, "Laboratory diagnosis and monitoring the viral shedding of 2019-ncov infections," *medRxiv*, 2020.

[5] Italian Radiology Society, "Utilizzo della Diagnostica per Immagini nei pazienti Covid 19," https://www.sirm.org/.

[6] G. D. Rubin, C. J. Ryerson, L. B. Haramati, N. Sverzellati *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.

[7] H. Shi, X. Han, N. Jiang, Y. Cao, O. Alwalid, J. Gu, Y. Fan, and C. Zheng, "Radiological findings from 81 patients with covid-19 pneumonia in wuhan, china: a descriptive study," *The Lancet Infectious Diseases*, 2020.

[8] H. Y. F. Wong, H. Y. S. Lam, A. H.-T. Fong, S. T. Leung, T. W.-Y. Chin, C. S. Y. Lo, M. M.-S. Lui, J. C. Y. Lee, K. W.-H. Chiu, T. Chung *et al.*, "Frequency and distribution of chest radiographic findings in covid-19 positive patients," *Radiology*, p. 201160, 2020.

[9] M. D. Hope, C. A. Raptis, A. Shah, M. M. Hammer, and T. S. Henry, "A role for ct in covid-19? what data really tell us so far," *The Lancet*, 2020.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[11] Z.-W. Lang, L.-J. Zhang, S.-J. Zhang, X. Meng, J.-Q. Li, C.-Z. Song, L. Sun, Y.-S. Zhou, and D. E. Dwyer, "A clinicopathological study of three cases of severe acute respiratory syndrome (sars)," *Pathology*, vol. 35, no. 6, pp. 526–531, 2003.

[12] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[13] N. Bondfale and D. S. Bhagwat, "Convolutional neural network for categorization of lung tissue patterns in interstitial lung diseases," in *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, 2018, pp. 1150–1154.

[14] M. Anthimopoulos, S. Christodoulidis, L. Ebner, A. Christe, and S. Mougiakakou, "Lung pattern classification for interstitial lung diseases using a deep convolutional neural network," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1207–1216, 2016.

[15] Xie Xuanyang, Gong Yuchang, Wan Shouhong, and Li Xi, "Computer aided detection of sars based on radiographs data mining," in *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, 2005, pp. 7459–7462.

[16] Xiaoou Tang, Dacheng Tao, and G. E. Antonio, "Texture classification of sars infected region in radiographic image," in *2004 International Conference on Image Processing, 2004. ICIP '04.*, vol. 5, 2004, pp. 2941–2944 Vol. 5.

[17] P. K. Sethy and S. K. Behera, "Detection of coronavirus disease (covid-19) based on deep features," 2020.

[18] I. D. Apostolopoulos and T. Bessiana, "Covid-19: Automatic detection from x-ray images utilizing transfer learning with convolutional neural networks," *arXiv preprint arXiv:2003.11617*, 2020.

[19] A. Narin, C. Kaya, and Z. Pamuk, "Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks," *arXiv preprint arXiv:2003.10849*, 2020.

[20] L. Wang and A. Wong, "Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest radiography images," *arXiv preprint arXiv:2003.09871*, 2020.

[21] J. P. Cohen, P. Morrison, and L. Dao, "Covid-19 image data collection," *arXiv 2003.11597*, 2020. [Online]. Available: https://github.com/ieee8023/covid-chestxray-dataset

[22] U. Lopes and J. F. Valiati, “Pre-trained convolutional neural networks as feature extractors for tuberculosis detection,” *Computers in biology and medicine*, vol. 89, pp. 135–143, 2017.

[23] S. Stirenko, Y. Kochura, O. Alienin, O. Rokovyi, Y. Gordienko, P. Gang, and W. Zeng, “Chest x-ray analysis of tuberculosis by deep learning with segmentation and augmentation,” in *2018 IEEE 38th International Conference on Electronics and Nanotechnology (ELNANO)*. IEEE, 2018, pp. 422–428.

[24] D. Kermany, K. Zhang, and M. Goldbaum, “Labeled optical coherence tomography (oct) and chest x-ray images for classification,” *Mendeley data*, vol. 2, 2018.

[25] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. Summers, “Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2017, pp. 3462–3471.

[26] J. T. Bushberg and J. M. Boone, *The essential physics of medical imaging*. Lippincott Williams & Wilkins, 2011.

[27] S. Candemir, S. Jaeger, K. Palaniappan, J. P. Musco, R. K. Singh, Z. Xue, A. Karargyris, S. Antani, G. Thoma, and C. J. McDonald, “Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration,” *IEEE transactions on medical imaging*, vol. 33, no. 2, pp. 577–590, 2013.

[28] S. Hu, E. A. Hoffman, and J. M. Reinhardt, “Automatic lung segmentation for accurate quantitation of volumetric x-ray ct images,” *IEEE transactions on medical imaging*, vol. 20, no. 6, pp. 490–498, 2001.

[29] A. Mansoor, U. Bagci, Z. Xu, B. Foster, K. N. Olivier, J. M. Elinoff, A. F. Suffredini, J. K. Udupa, and D. J. Mollura, “A generic approach to pathological lung segmentation,” *IEEE transactions on medical imaging*, vol. 33, no. 12, pp. 2293–2310, 2014.

[30] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[31] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, “Convolutional neural networks for medical image analysis: Full training or fine tuning?” *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1299–1312, 2016.

[32] L. Prechelt, “Early stopping-but when?” in *Neural Networks: Tricks of the trade*. Springer, 1998, pp. 55–69.

[33] Y. Yao, L. Rosasco, and A. Caponnetto, “On early stopping in gradient descent learning,” *Constructive Approximation*, vol. 26, no. 2, pp. 289–315, 2007.

[34] E. Tartaglione, D. Perlo, and M. Grangetto, “Post-synaptic potential regularization has potential,” in *International Conference on Artificial Neural Networks*. Springer, 2019, pp. 187–200.

[35] A.-M. Šimundić, “Measures of diagnostic accuracy: basic definitions,” *Ejifcc*, vol. 19, no. 4, p. 203, 2009.

[36] B. Chu, V. Madhavan, O. Beijbom, J. Hoffman, and T. Darrell, “Best practices for fine-tuning visual classifiers to new domains,” in *European conference on computer vision*. Springer, 2016, pp. 435–442.

[37] L. v. d. Maaten and G. Hinton, “Visualizing data using t-sne,” *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[38] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, “Striving for simplicity: The all convolutional net,” *arXiv preprint arXiv:1412.6806*, 2014.