mathematics

MDPI

Article

Reliable Learning with PDE-Based CNNs and DenseNets for Detecting COVID-19, Pneumonia, and Tuberculosis from Chest X-ray Images

Anca Nicoleta Marginean [1,*], Delia Doris Muntean [2,3], George Adrian Muntean [2,4], Adelina Priscu [5], Adrian Groza [1], Radu Razvan Slavescu [1], Calin Lucian Timbus [6], Gabriel Zeno Munteanu [2], Cezar Octavian Morosanu [7], Maria Margareta Cosnarovici [8] and Camelia-M. Pintea [9,*]

1 Computer Science Department, Technical University of Cluj-Napoca, 400114 Cluj-Napoca, Romania; adrian.groza@cs.utcluj.ro (A.G.); radu.razvan.slavescu@cs.utcluj.ro (R.R.S.)
2 County Clinical Emergency Hospital of Cluj-Napoca, 400006 Cluj-Napoca, Romania; muntean.delia.doris@elearn.umfcluj.ro (D.D.M.); georgemuntean99@elearn.umfcluj.ro (G.A.M.); ; munteanu.gabriel.zeno@elearn.umfcluj.ro (G.Z.M.)
3 Department of Ophthalmology, Iuliu Haţieganu University of Medicine and Pharmacy, 400006 Cluj-Napoca, Romania
4 Department of Ophthalmology, Iuliu Hatieganu University of Medicine and Pharmacy, 400006 Cluj-Napoca, Romania
5 Department of Internal Medicine, Indiana University Health Ball Memorial Hospital, Muncie, IN 47303, USA; apriscu@iuhealth.org
6 Department of Mathematics, Technical University of Cluj-Napoca, 400114 Cluj-Napoca, Romania; calin.timbus@math.utcluj.ro
7 University Hospital Bristol, Bristol BS2 8HW, UK; cezar.morosanu@uhbw.nhs.uk
8 The Oncology Institute Prof. Dr. Ion Chiricuta, 400015 Cluj-Napoca, Romania; maria.cosnarovici@umfcluj.ro
9 Department of Mathematics and Informatics, Technical University of Cluj-Napoca, 400114 Cluj-Napoca, Romania
* Correspondence: anca.marginean@cs.utcluj.ro (A.N.M.); camelia.pintea@mi.utcluj.ro or dr.camelia.pintea@ieee.org (C.-M.P.)

check for updates

Citation: Marginean, A.N.; Muntean, D.D.; Muntean, G.A.; Priscu, A.; Groza, A; Slavescu, R.R.; Timbus, C.L.; Munteanu, G.Z.; Morosanu, C.O.; Cosnarovici, M.M.; Pintea, C.-M. Reliable Learning with PDE-Based CNNs and DenseNets for Detecting COVID-19, Pneumonia, and Tuberculosis from Chest X-ray Images. *Mathematics* **2021**, *9*, 434. https://doi.org/10.3390/math9040434

Academic Editor: Tudor Barbu

Received: 10 January 2021
Accepted: 11 February 2021
Published: 22 February 2021

Abstract: It has recently been shown that the interpretation by partial differential equations (PDEs) of a class of convolutional neural networks (CNNs) supports definition of architectures such as parabolic and hyperbolic networks. These networks have provable properties regarding the stability against the perturbations of the input features. Aiming for robustness, we tackle the problem of detecting changes in chest X-ray images that may be suggestive of COVID-19 with parabolic and hyperbolic CNNs and with domain-specific transfer learning. To this end, we compile public data on patients diagnosed with COVID-19, pneumonia, and tuberculosis, along with normal chest X-ray images. The negative impact of the small number of COVID-19 images is reduced by applying transfer learning in several ways. For the parabolic and hyperbolic networks, we pretrain the networks on normal and pneumonia images and further use the obtained weights as the initializers for the networks to discriminate between COVID-19, pneumonia, tuberculosis, and normal aspects. For DenseNets, we apply transfer learning twice. First, the ImageNet pretrained weights are used to train on the CheXpert dataset, which includes 14 common radiological observations (e.g., lung opacity, cardiomegaly, fracture, support devices). Then, the weights are used to initialize the network which detects COVID-19 and the three other classes. The resulting networks are compared in terms of how well they adapt to the small number of COVID-19 images. According to our quantitative and qualitative analysis, the resulting networks are more reliable compared to those obtained by direct training on the targeted dataset.

Keywords: partial differential equations (PDEs); COVID-19; convolutional neural network (CNN); imbalanced dataset

1. Introduction

Can one explain the prediction of a neural network? Why are some architectures for neural networks more suitable for certain tasks? How can one decide on a good design without an extensive trial and error process? Can neural architecture search (NAS) be guided by theoretical results? Does one need very deep neural networks, or is it better to run an ablation study? These are only some of the open questions in the world of deep neural networks. Despite having enormous success, the lack of a well-established theory of neural networks, with provable properties, is a drawback.

One line of research to bringing theoretical results into CNNs is based on partial differential equations (PDEs). In [1], starting from the skip-connections from the ResNet block [2], a space-time continuous interpretation was given. Based on parabolic and hyperbolic PDEs, additional constraints were imposed on the network's weights, and parabolic and hyperbolic CNNs [1] were defined. It was asserted in that paper that well-known properties of PDEs are transferred to the newly-introduced CNNs, such as the property of stability. The size of the networks and their theoretical properties make them recommended to solve risk-sensitive classification tasks where only small datasets are available. Consequently, we consider parabolic and hyperbolic networks as good candidates to address the current necessity to rapidly identify COVID-19 from chest X-rays.

With the outbreak of COVID-19, deep learning has been applied in various tasks: predicting the number of cases and detecting COVID-19 from the sounds of coughs or from images, either X-rays [3] or CT scans. Sciacca et al. showed that the most frequent findings in the radiological data from people with COVID-19 are airspace opacities, with a predominant bilateral, peripheral, and lower zone distribution. Sciacca et al. also found that in contrast to parenchymal abnormalities, pleural effusion is rare [4].

According to a Fleischner Society consensus statement published on 7 April 2020, imaging is recommended for (i) patients with COVID-19 and worsening respiratory status and (ii) patients with features of moderate to severe COVID-19, regardless of the COVID-19 test result, but not for patients with suspected COVID-19 and mild clinical features, except when the risk of disease progression is present. CT is preferred to X-rays, but in a resource-constrained environment, the medical triage of people with suspected COVID-19 could include radiography. Additionally, it has been stated that findings suggestive of COVID-19 on a CT scan warrant COVID-19 testing [4].

In this context, the detection of COVID-19 from chest X-rays remains a challenging task. During 2020, an extremely large number of researchers tackled the problem, either considering X-rays or CT scans, with different results. As is already known from deep learning, good numerical evaluations do not guarantee a good model due to the limitations of the training and validation data. Since the problem is novel and the datasets are subject to continuous development, we argue for a cautious approach when assessing a classification model.

We follow two research hypotheses: First, the theoretical properties of PDE-inspired CNNs, like stability and reversibility, could facilitate the correctness of the learning. Second, domain-specific transfer learning can address the issue of very small datasets. Therefore, we exploit the available large datasets for the abnormalities existing in X-rays before COVID-19. Instead of transferring the ImageNet pre-trained weights, our research hypothesis is that transferring the weights trained on these large datasets will positively impact the generality of the solutions. We verify these two hypotheses on a dataset compiled by us from publicly available datasets, including COVID-19 datasets.

The rest of the article is organized as follows: In Section 2, we introduce the theoretical description of the parabolic and hyperbolic networks as formalized by Ruthotto and Haber [1]. Then, we describe the steps to build our dataset and how we apply the PDE-inspired networks to our dataset. Section 3 describes and assesses the experiments with parabolic and hyperbolic networks and the experiments with DenseNets [5] in a two-phase learning based on the large chest X-ray dataset CheXpert. Section 4 compares the PDE-inspired models and CheXpert-based model, while Section 5 concludes the paper.

2. Materials and Methods

2.1. Theoretical Aspects for PDE Inspired Neural Networks

While designing and training CNNs is a trial-end-error based process, partial differential equations (PDE) provide solid mathematical properties and results. Data interpreted as discretization of multivariate functions support PDE-based data processing. Moving forward, interpreting CNNs as a discretization of a space-time differential equation could bring the theoretical results of PDE into the world of CNNs. CNNs can benefit from the PDE theory in two ways: first, by transferring provable properties, like stability, from PDE to networks; and second, new CNNs architecture can be designed based on the PDE formalism.

Ruthotto et al. have introduced the interpretation of residual CNNs as a discretization of a space-time differential equation [1]. Starting from the heat equation or wave equation, parabolic and hyperbolic CNNs have been formalized. Properties of PDEs like stability and reversibility have been proven for parabolic and hyperbolic CNNs. To describe the PDE inspired networks, we rely here on the work of Ruthotto et al. [1].

For a machine learning researcher, ResNets (Residual networks) [2] use skip-connections through which the signal jumps over one or several layers. A ResNet block [2] is described by Equation (1):

$$a = W_2 g(W_1 X) + X \tag{1}$$

W_1 and W_2 stand for the weight vectors, g is the activation function, X is the input vector for the block, a is the activation of the block, and $+$ is the element-wise addition. DenseNet [5] extends this type of connections and alleviates the vanishing-gradient problem. Even though the number of parameters is substantially reduced, feature propagation is strengthened.

The same residual block is defined in [1] through the linear operators K_1 and K_2, while the parameter vector Θ has three parts: $\Theta^{(1)}$ and $\Theta^{(3)}$ for weights W_1 and W_2 from Equation (1), respectively $\Theta^{(2)}$ for parameters of the normalization layer $\mathcal{N}$. The residual block output is defined by Equation (2), where σ is the activation function, and Y is the input vector. The transformation of a network with several layers with ResNet blocks, input Y_0 and N layers is described in (3).

$$F(\Theta, Y) = K_2(\Theta^{(3)})\sigma(\mathcal{N}(K_1(\Theta^{(1)})Y, \Theta^{(2)})) \tag{2}$$

$$Y_{i+1} = Y_i + F(\Theta^{(i)}, Y_i), \;\; i = 0, 1.., N-1 \tag{3}$$

A major advantage of describing the filtering made by ResNets in this form is the fact that the continuous time interpretation (Equation (4)) of the networks becomes more obvious: the ResNets filtering can be seen as a forward Euler discretization of the initial value problem [1].

$$\begin{aligned} \partial_t Y(\Theta, t) &= F(\Theta(t), Y(t)), \text{ for } t \in (0, T] \\ Y(\Theta, 0) &= Y_0 \end{aligned} \tag{4}$$

While considering K_1 and K_2 as convolutional operators, the continuous space-time interpretation of a convolutional neural network (CNN) is possible. In order to have networks for which the stability condition is present, either forward and backward in time, K_1 and K_2 are constrained to $K_2 = -K_1^T$ in Equation (2), resulting in Equation (5).

$$F_{sym}(\Theta, Y) = -K(\Theta)^T \sigma(\mathcal{N}(K(\Theta)Y, \Theta)) \tag{5}$$

Based on the these observations, the following CNNs architectures were defined in [1]:

– Parabolic CNNs. The forward propagation of parabolic CNNs (Equation (6) is equivalent to the heat equation when the action function is the identity function, i.e. $\mathcal{N}(Y) = Y$, and $K(t) = \nabla$. Parabolic PDEs are robust to perturbations of initial conditions, therefore parabolic-like CNNs are also stable when the activation function σ is monotonically nonde-

creasing. This makes them more appropriate for smoothing type learning tasks, since they are not strongly affected by the noise.

$$\partial_t Y(\Theta, t) = F_{sym}(\Theta(t), Y(t)), \ for\ t \in (0, T] \tag{6}$$

– Hyperbolic CNNs. Hyperbolic PDEs have the property of reversibility. In order to improve memory efficiency, the hyperbolic CNNs: Hamiltonian CNNs and second-order CNNs, were introduced. The dynamics changes the Equations (7) and (8), where Y is the input vector, and Z is an auxiliary variable.

$$\begin{aligned} \partial_t Y(t) &= -F_{sym}(\Theta^1(t), Z(t)), & Y(0) &= Y_0 \\ \partial_t Z(t) &= F_{sym}(\Theta^2(t), Y(t)), & Z(0) &= Z_0. \end{aligned} \tag{7}$$

$$\begin{aligned} \partial_t^2 Y(t) &= F_{sym}(\Theta(t), Y(t)) \\ Y(0) &= Y + 0, \ \ \partial_t Y(0) = 0 \end{aligned} \tag{8}$$

– Stable CNNs. Parabolic and hyperbolic CNNs are CNNs with restricted weights. The restrictions are introduced in such a way that training the resulting networks is similar to solving known classes of PDEs. The spatial-time dependent PDEs selected for the parabolic and hyperbolic networks are stable with respect to perturbations of the initial conditions [1]. In the definition of the stability from Equation (9), Y and Y_1 are solutions of the Equation (4), which gives the continuous interpretation of the computation in ResNets. This means that if we have a solution associated with a perturbed input image, then the norm of the perturbation at any moment T is bound by the norm of the initial perturbation. The stability property of a CNN would prevent the effect of adversarial attacks. Moreover, since the classification of medical images are meant to reduce the risks of misdiagnosis, a provable property of stability could not only reduce the risks, but also increase the trust in the learned models.

$$||Y(\Theta, T) - Y_1(\Theta, T)||_F \leq M||Y(0) - Y_1(0)||_F \tag{9}$$

2.2. Compiling the Dataset

With the 2020 outbreak of the coronavirus SARS-CoV-2, there were different initiatives gathering chest X-rays or CT scans in order to apply artificial intelligence (AI) for rapid and precise diagnoses [6]. From the public datasets, we mention Cohen dataset [7] and Hannover dataset [8]. We conducted an analysis of the existing COVID-19 X-rays datasets and we decided to compile our own dataset with images for COVID-19, pneumonia, and tuberculosis, together with X-rays with normal findings. We used several public datasets, as follows.

2.2.1. Collecting COVID-19 X-rays

For COVID-19 positive X-rays, we used the Cohen dataset. This gathers X-ray and CT images from Radiopaedia (https://radiopaedia.org/, accessed on 9 January 2021), SIRM (https://www.sirm.org/category/senza-categoria/COVID-19/, accessed on 9 January 2021), and other research articles. It also includes scripts for the integration of the Hannover dataset. For the current experiments, we extracted from the Cohen dataset only the X-ray AP (anteroposterior) and PA (posteroanterior), which were gathered from Radiopaedia and SIRM. Out of all the 319 COVID-19 samples from the dataset (both X-ray and CT), only 121 met our filtering conditions. The reason for this decision was the quality of the images.

We underline two aspects which are present in this set of COVID-19 X-rays and which could affect the learning process:

1. Patients in evolution. Especially for the cases extracted from Radiopaedia, there are several X-rays done for the same patient in different stages of the disease: from the moment when they were admitted into the hospital to the moment when the

symptoms improved. On one hand, this is quite useful in order to see whether the learned model could also predict the disease severity. On the other hand, all the X-rays are annotated as COVID-19, including the ones from the end of the monitoring.

2. Train/test split. Having several X-rays for the same patient requires extra care at splitting into train/test set. X-rays from the same patient should not be included both in the train and test set.

2.2.2. Collecting Pneumonia X-rays

In our first experiments for COVID-19 detection, we used the COVID-19 Radiography database [9] (https://www.kaggle.com/tawsifurrahman/COVID19-radiography-database, accessed on 9 January 2021), which includes X-rays for COVID-19, pneumonia and normal findings, with Kermany dataset [10] as the source for the normal findings and viral/bacterial pneumonia X-rays. Our initial experiments show that the dataset is not adequate for detection of COVID-19, since the Kermany dataset is meant for diagnosing pediatric pneumonia; consequently, it includes X-rays from children. By using this dataset, we obtained very good results for predicting COVID-19/non-COVID pneumonia, but the reason was that our network learned age-related patterns.

Based on these preliminary experiments, we decided to use RSNA pneumonia challenge (https://www.kaggle.com/c/rsna-pneumonia-detection-challenge, accessed on 9 January 2021) as the source for the pneumonia and normal X-ray images. This challenge was organized by the Radiological Society of North America. It includes data from [11], and it aims for the detection and the localization of the pneumonia in chest radiographs (as it was perceived before the COVID-19 outbreak). From this dataset, we extracted 8851 normal X-rays and 6012 X-rays of patients suffering from pneumonia.

2.2.3. Collecting Tuberculosis X-rays

We faced two challenges when learning to detect COVID-19. First, even though the COVID-19 disease caused a pandemic, the number of publicly available X-rays remained small. Consequently, the dataset obtained by joining COVID-19 X-rays with normal findings and pneumonia X-rays is highly imbalanced. Second, some clinical signs identified in patients suffering from COVID-19 (e.g., airspace opacities, whether described as consolidation or, less commonly, ground-glass opacification) are present for other pulmonary diseases than COVID-19 or non-COVID pneumonia. For COVID-19, the distribution of these lung abnormalities is most often bilateral, peripheral, and lower zone predominant [4].

Because of these two difficulties, we decided to include into our dataset an additional class, whose detection should not be altered by the presence of COVID-19 class. Therefore, X-rays from patients suffering from tuberculosis were added into our dataset. Since the differentiation between tuberculosis vs. COVID-19, respectively tuberculosis vs. non-COVID-19 pneumonia is easier to assess, we consider that this class could help in checking the generality of the approach.

We used the dataset from Shenzhen No. 3 People's Hospital in China [12]. We extracted not only X-rays for the tuberculosis, but also normal X-rays: 335 X-rays for tuberculosis and 326 normal X-rays (see Table 1). Note that the compiled dataset is imbalanced and limited in terms of the number of images for COVID-19.

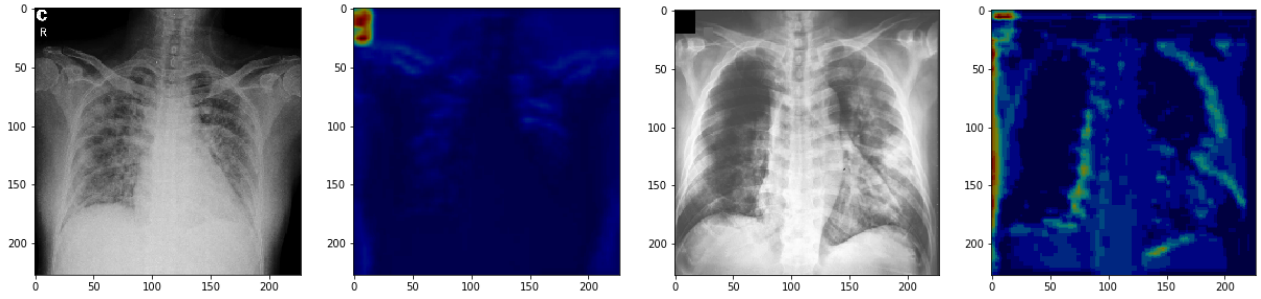
Table 1. Building the dataset of COVID-19 X-ray images.

Source	COVID-19	Pneumonia	Tuberculosis	Normal
Cohen dataset (RadiopaediaSIRM)	121	-	-	-
RSNA Pneumonia Challenge	-	6012	-	8851
Shenzhen	-	-	335	326
Total	121	6012	335	9177

2.3. Method

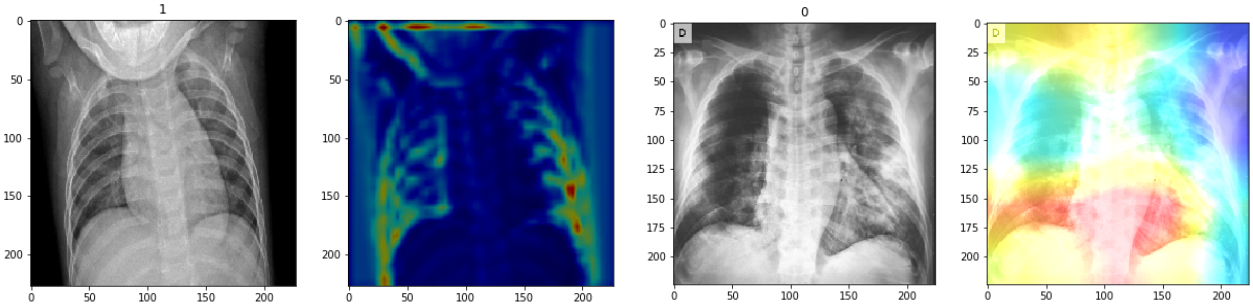
Our preliminary experiments on training from scratch on COVID-19 Radiography database [9] displayed good performance for the detection of COVID-19 (precision and recall above 0.90), but we considered that the result was not reliable. Even with transfer learning from ImageNet, the good performance was due to the properties of the dataset and not to the capability of the models to identify patterns in the COVID-19 images. As already mentioned, this is the reason for compiling our own dataset.

To exemplify this situation with high performance but erroneously learned patterns, we bear out in Figure 1 some of the issues behind the high performance metric. These results were obtained by training DenseNets or VGG with transfer from ImageNet. Figure 1 shows activation maps for some examples from the initial dataset. The red areas are the ones that mostly influenced the CNN's decision, while the blue areas have a low impact on the CNN's decision. The learning capability of CNN was strongly affected by the small number of images with COVID-19, since the networks identified other abnormal elements present in COVID-19 samples.



(**a**) COVID-19 X-ray with an upper left corner artifact.

(**b**) The most relevant areas are outside the lungs.



(**c**) Normal X-ray of a child, the most important areas are related to the bone structure which is age-specific.

(**d**) Along the correct relevant area, the heart is erroneously considered important.

Figure 1. Pitfalls of learning to detect COVID-19 from COVID-19 Radiography database [9].

For instance, in Figure 1d, along the correct relevant area, the heart is also considered relevant, since it has abnormal size (cardiomegaly). Figure 1c shows how the CNN based its decision on age-specific structural elements. Even worse, the presence of artifacts like symbols L, R influenced the models (Figure 1a). Moreover, the difference between images in Figure 1b,d is only the upper left corner, but their activation map is completely different.

Aiming to build models able to avoid such problems, our proposed method is twofold: (i) employing PDEs inspired networks for which the stability property is provable, (ii) using a more specialized transfer learning: in a first step, we learn to interpret chest X-rays with respect to a large number of general modifications, and in the second step, we specialize the learning to a specific problem. This approach is inline with a few-shots training [13], which heavily relies on building good embeddings.

Figure 2 illustrates our proposed methodology and how it is applied on the available data. Identification of COVID-19 from X-rays is a new classification problem, therefore

(i) no stable, curated, and large dataset is available, and (ii) it raises questions not only for AI models, but also for humans. We consider that these two facts compel us to increase the attention towards reliability of the trained models by dealing with the following questions:

1. Which are the classes we should compare COVID-19 to in order to learn to identify COVID-19 from X-rays?
2. Which data source should we use for the selected classes?
3. What kind of architectures and learning should we use?

The main elements of the answer for the first two questions are captured in the left part of Figure 2. The classes are: the targeted class (COVID-19), the class which is the most similar to it (non-COVID pneumonia), the normal class, and a class, (tuberculosis), which is in some degree unrelated to the targeted class. The details about the sources for these classes were given in Section 2.2. For the third question, we propose two answers: the first one is to use architectures for which general theoretical properties are known, such as PDE-inspired CNNs (top-right of Figure 2), while the second one is to exploit the previously available knowledge about general radiological aspects and apply the transfer learning twice when going from general to specific (bottom-right of Figure 2).

We underline that any image classification task, including the diagnosis of other pulmonary diseases or, from a completely different area, health monitoring of industrial systems [14], could benefit from using PDE-inspired CNNs or going in several phases from general to specific. With the current scenario, identification of COVID-19, we advocate that completely new problems which suddenly appear in areas which include stable datasets for known problems, might be addressed with (i) restricted CNN architectures and (ii) models initially trained for known domain-specific aspects instead of models trained directly for the targeted new problem. Learning to identify COVID-19 is different compared to learning to identify non-COVID pneumonia, due to the issues related to the lack of stable and large datasets and the lack of clear medical knowledge.

The details about learning with PDE-inspired CNNs and 2-phase transfer learning from CheXpert are given in the next section: which type of data is used, which type of network, and whether or not pretrained weights are used, together with details about the performance.

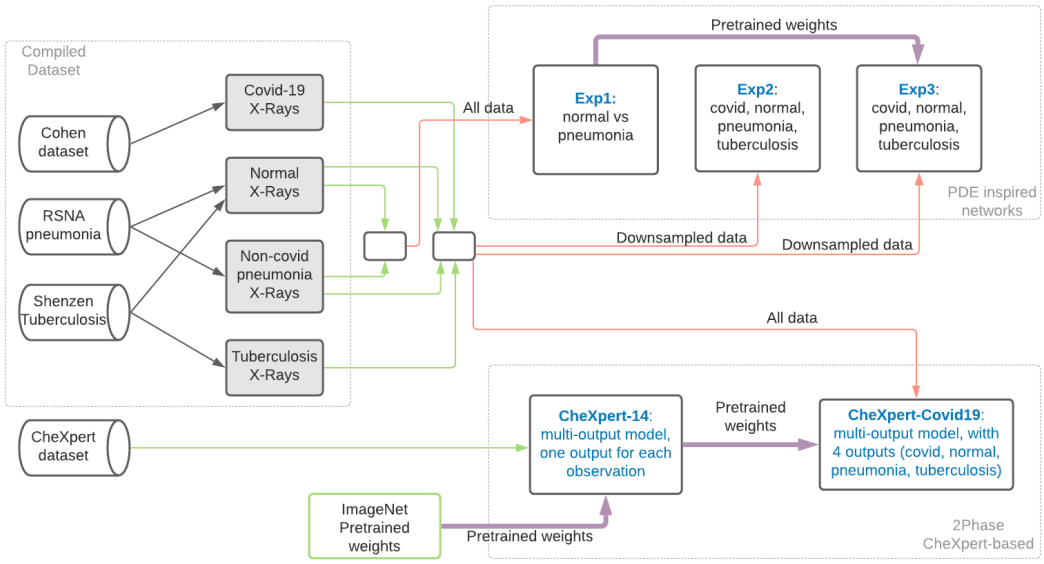


Figure 2. The proposed method for learning: compiling the dataset (**left**), employing partial differential equation (PDE)-inspired networks (**top-right**), and using CheXpert dataset in 2-phase transfer learning (**bottom-right**).

3. Experiments and Results

First, we designed three experiments with parabolic and hyperbolic CNNs. Second, we used the classical DenseNets.

3.1. Experiments with PDE Inspired CNNs

In order to benefit from the stability of PDEs, we conducted several experiments for training PDEs inspired networks for the classification of X-rays. For PDE-inspired CNNs, we use the Meganet [1,15]. We employ one architecture with different dynamics, corresponding to Parabolic, second-order, and Hamiltonian CNNs. For each dynamic, we use four ResNet blocks with 32, 64, 112, 112 channels. The number of the learned parameters for the Hamiltonian CNN is 263,428, while for the parabolic CNN and for the second-order CNN is 501,892.

The input images are RGB and their size is 192×192. Random flip augmentation was applied. The split between train/test sets is done randomly, with 80% examples in the training set.

Because of the imbalance in the dataset, we designed three types of experiments:

Exp_1—train only for the two classes (i.e., normal and pneumonia) for which the number of images is larger than 1000.

Exp_2—train for all four classes, but with downsampling of the normal and pneumonia classes.

Exp_3—similar to Exp_2, but instead of starting from the random weights, we start from the weights obtained in Exp_1.

3.1.1. Exp_1: Distinguishing between Normal and Pneumonia X-ray Images

Figure 3 shows the learning curves of accuracy for parabolic, second-order, and Hamiltonian CNNs. For training, all the three networks obtained values bigger than 98%, while for validation values higher than 93%. The Hamiltonian network reaches 93.6% accuracy on the validation set, with 94.4% precision and 95.1% recall for the normal class, respectively 92.4% precision and 91.4% recall for pneumonia class. The dataset is not completely balanced: in the validation set, there are 1,836 images for the normal class, and 1203 for pneumonia.

These results are comparable with the ones reported in literature for the detection of pneumonia from this dataset. Pan et al. [16], the winners of the RSNA challenge, have combined the detection of the pneumonia with the localization. Differently from us, Pan et al. have used during training the information about the area where the consolidation is present. Ensemble learning has been used to combine 10 models for the classification (into normal vs. pneumonia), and 50 models for object detection (for localization of the consolidation) [16]. The used architectures including Inception ResNetV2, Xception and DenseNet169, while transfer learning was based on ImageNet pre-trained weights.

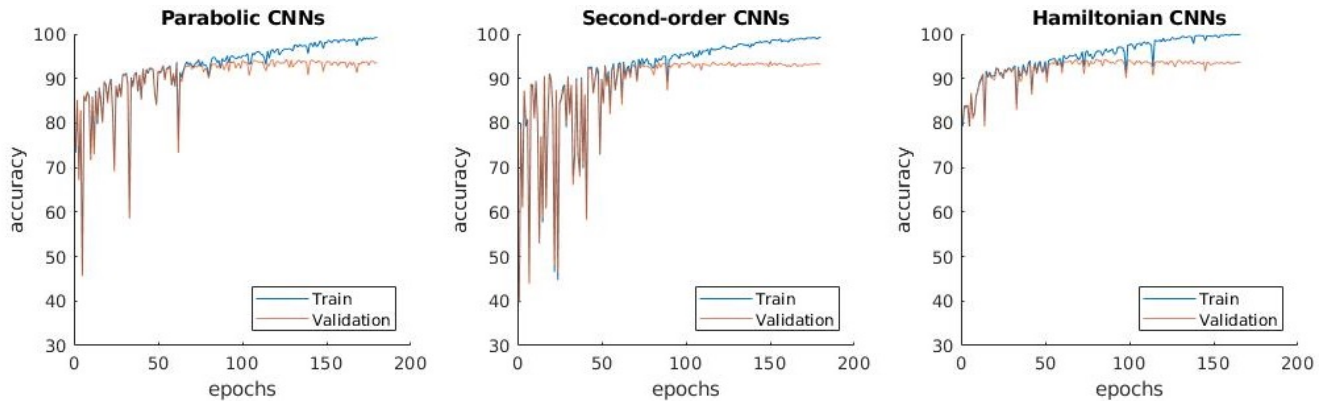


Figure 3. Exp_1: Accuracy on normal vs. pneumonia classes (train: 12,150, validation: 3039 examples).

Ruthotto and Haber have argued in [1] that parabolic CNNs are better suited for tasks not affected by smoothing the images, while the hyperbolic CNNs are suited for tasks which require preservation of edge information. Both the learning curve and the reached accuracy on training for the Hamiltonian CNNs seem to confirm this hypothesis. The reversibility property of these networks ensures that nothing is lost from the images, and they seem to be more appropriate for the classification of X-ray images. One justification might be that the image includes the structural elements of the chest area, and it is of most importance to know where abnormalities occur.

Taking into account the results in Figure 3, we argue that the stability property of the three networks facilitates learning from X-ray, even with a small number of parameters (i.e., much smaller than a DenseNet network), and without any transfer learning.

3.1.2. Exp_2: Distinguishing between Normal, Pneumonia, Tuberculosis and COVID-19

Since the dataset is highly imbalanced, mainly due to the extremely small number of COVID-19 images, we applied downsampling. For the training, we kept 280 images for the normal and the pneumonia classes. The final structure of the training set is 96 COVID-19, 280 normal, 280 pneumonia, 268 tuberculosis, while the structure for the test set is: 25 COVID-19, 70 normal, 70 pneumonia, 67 tuberculosis. The learning curves for the accuracy appear in Figure 4. Over-fitting is reached around the 100th epoch, and even though the accuracy on training reaches values above 95%, the accuracy of the test set is around 88%.

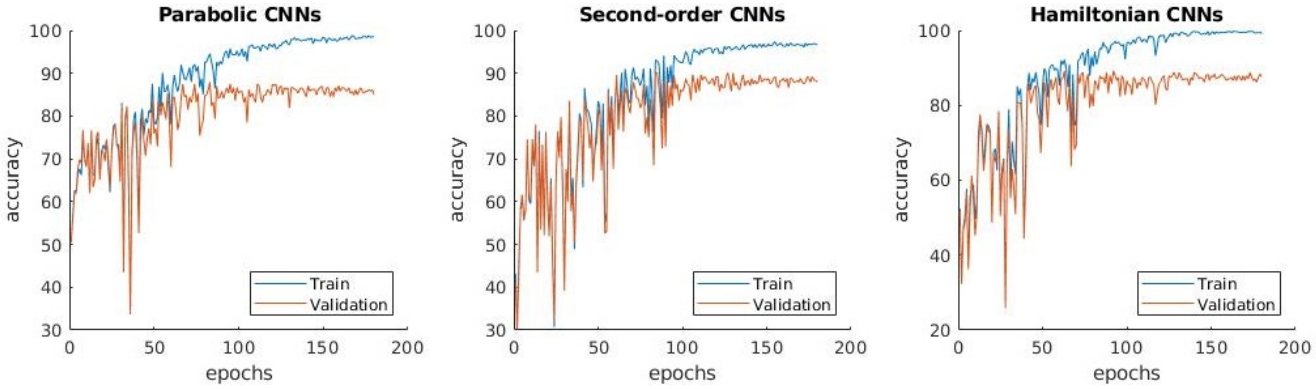


Figure 4. Exp_2: Accuracy on the downsampled COVID-19 dataset (train: 924, validation: 232 examples).

The confusion matrix for the Hamiltonian network from Figure 5a shows that tuberculosis is almost perfectly learned, while the other classes have several problems. The right column stands for the precision on each class, while the last row for the recall. The overall accuracy is given in the lower right corner. The classes are in the following order: COVID-19, normal, (non-COVID) pneumonia, tuberculosis.

A number of 14 normal images (out of 70) are wrongly classified as pneumonia, therefore the recall for the normal class is only 0.786. The recall for pneumonia is 0.90: 4 pneumonia images are wrongly considered normal, and 3 are considered COVID-19. Note that the precision and recall metrics for normal and pneumonia classes are smaller than in Exp_1. This is expected, on one hand due to the fact that the training set is very small compared to Exp_1, and on the other hand, due to the fact that two other classes are introduced, with COVID-19 as the most difficult to detect class.

For COVID-19, 4 out of 25 images are wrongly considered as pneumonia. There is also a 1 image wrongly classified as normal and another 1 as tuberculosis. The precision for COVID-19 class is 0.818 and the recall 0.72.

3.1.3. Exp_3: Distinguishing between Four Classes with Transfer Learning

This experiment augments Exp_2 with transfer learning, where the weights are taken from the results of Exp_1. One question which appeared after running Exp_1 and Exp_2

was: How would the two-class Hamiltonian CNN from Exp_1, which was not trained on COVID-19 samples, classify the images from the COVID-19 class? Exp_1 Hamiltonian CNN classifies 95 of them as pneumonia and 26 as normal. This is not surprising, since COVID-19 has the clinical signs and radiological features of a pneumonia.

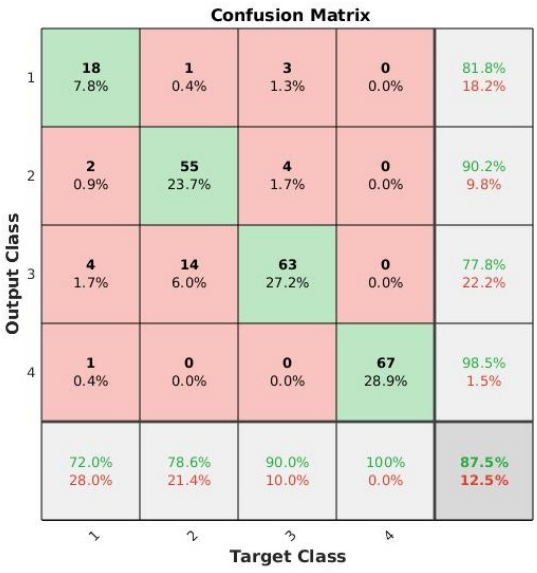
So, a first observation is that the networks from Exp_1 identified some features which could be correlated to the COVID-19 class.

The second observation is that the number of images in the classes normal and pneumonia is much larger than the number of images in the classes COVID-19 and tuberculosis.

Based on these two observations, we decided to apply transfer learning from Exp_1 into training on the downsampled dataset. Hence, we train the networks for the four classes on the downsampled dataset built for Exp_2, but the initial weights are set to the ones learned for the classification into normal and pneumonia (Exp_1).

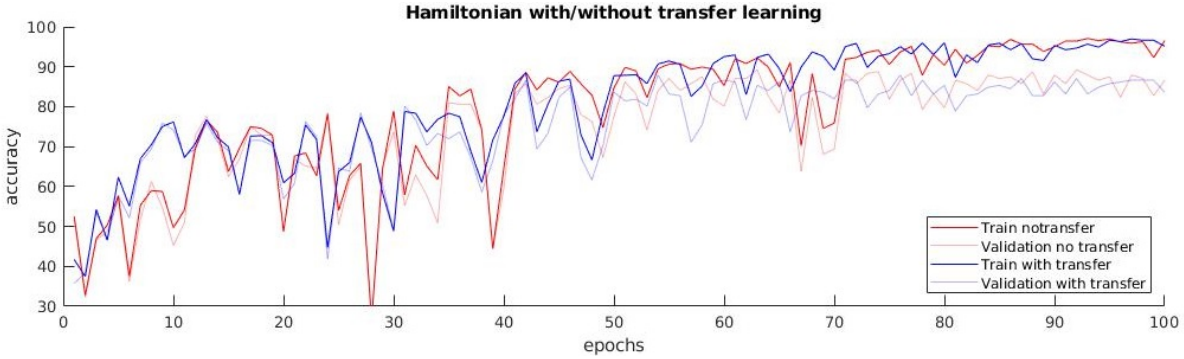
On running Exp_3, we made the following observations:

- The confusion matrix in Figure 5b shows that the normal and pneumonia classes are better predicted in Exp_3 than in Exp_2, even though the used dataset is the same. Differently, COVID-19 is a little bit worse in Exp_3.
- The learning curves in Figure 5c show that the accuracy in Exp_3 does not have the high variations from one epoch to another, which is present in Exp_2.
- The difference between training and validation accuracy is smaller in Exp_3 compared to Exp_2.



(**a**) Exp_2: Confusion matrix for the Hamiltonian convolutional neural networks (CNNs) on downsampled COVID Dataset.

(**b**) Exp_3: Confusion matrix for the Hamiltonian CNNs on downsampled COVID Dataset with transfer learning.



(**c**) Comparison between learning curves for training with/without transfer (Exp_2 and Exp_3).

Figure 5. Confusion matrix and learning curve for the Hamiltonian networks trained in Experiment 2 and Experiment 3.

Based on the above observations, we conclude that, for parabolic and hyperbolic networks, learning first from a larger dataset for 2 classes and after that moving to 4 classes with the downsampled dataset is more reliable, even though from the numerical point of view of COVID-19 class, Exp_2 seems better than Exp_3. Further investigation is needed here, and a method other than downsampling could be used for dealing with the imbalanced data, like class weights, such that all the available data is used. Furthermore, visualization of the activation maps or filtered images could get some insights into what the networks learned.

3.2. 2-Phases Learning with CheXpert as X-ray Embedding

The second approach relies on using the classical DenseNets. It is a common practice in image classification to use transfer learning from ImageNet [17]. Since we classify medical images, we expect to have limited benefits in doing this transfer. Therefore, the proposed method exploits the benefits of the transfer learning in two phases: first from ImageNet, and then from a general and large dataset of chest X-rays, i.e., CheXpert. The first phase is building models for the detection of the 14 observations from the CheXpert dataset.

We call these models CheXpert-14 and we start their training from ImageNet pretrained weights. The second phase is to build models for the detection of the three diseases: COVID-19, pneumonia, tuberculosis, together with the normal aspect from our custom-made dataset. We name these models CheXpert-COVID19, since we start learning from the CheXpert-14 weights as pretrained weights (see Figure 2). CheXpert [18] is a large public dataset for chest radiograph interpretation. It consists of 224,316 chest radiography from 66,240 patients, collected from Stanford Hospital for patients in 2002–2017. For each patient, there are 14 observations that can be positive, negative or uncertain, except for NoFinding, which can only be positive or negative (see Table 2). Even though the data are real, since the annotations are extracted with an automated rule-based labeler from the radiology reports, the data are not error-free.

Table 2. CheXpert labeler performance on the 14 observations [18].

	Positive	**Negative**	**Uncertain**
Atelectasis	0.998	0.833	0.936
Cardiomegaly	0.973	0.909	0.727
Consolidation	0.999	0.981	0.924
Edema	0.993	0.962	0.796
Pleural Effusion	0.996	0.971	0.707
Pneumonia	0.992	0.750	0.817
Pneumothorax	1.000	0.977	0.762
Enlarged Cardiom.	0.935	0.959	0.854
Lung Lesion	0.896	0.900	0.857
Lung Opacity	0.966	0.914	**0.286**
Pleural Other	0.850	1.000	0.769
Fracture	0.975	0.807	0.800
Support Devices	0.933	0.720	-
No Finding	0.769	-	-

Table 2 lists the performance (F1) of the labeler on all 14 observations [18]. Note here that the performance varies among the observations. For instance, the observations support device and NoFinding have F1 values lower than 0.8 for one of the possible values (positive, negative, uncertain), while the observations atelectasis, consolidation, cardiomegaly, edema, pleural effusion, pneumothorax have values higher than 0.8 for positive and negative values. This suggests that the annotations for the later observations are more reliable than the former mentioned ones.

3.2.1. First Phase: Classifying the 14 Observations from CheXpert

The architecture of the network for CheXpert-14 models is standard: a DenseNet121 followed by a Dense layer and the output layer (Figure 6—truncated due to the lack of space).

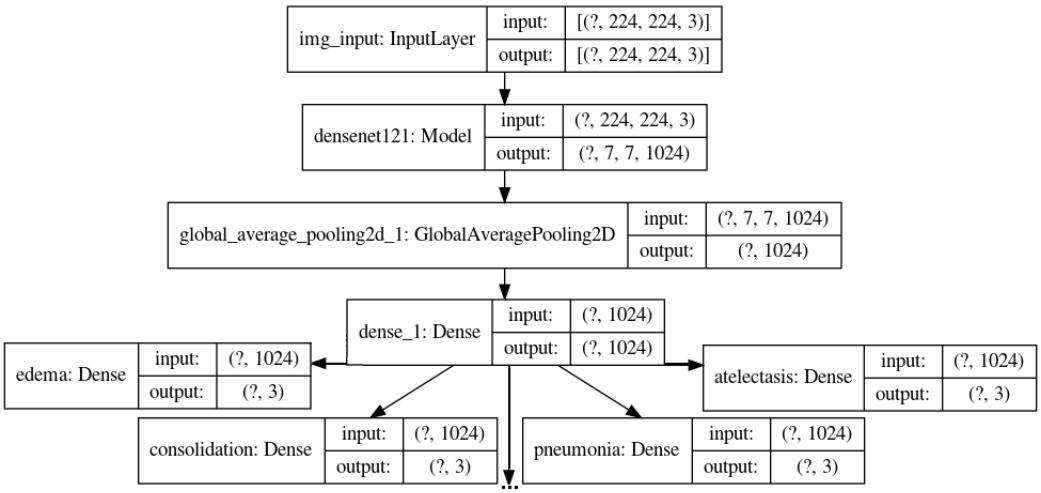


Figure 6. CheXpert-14 models: using DenseNet121 for training with 14 multi-outputs and 2 or 3 classes for each output.

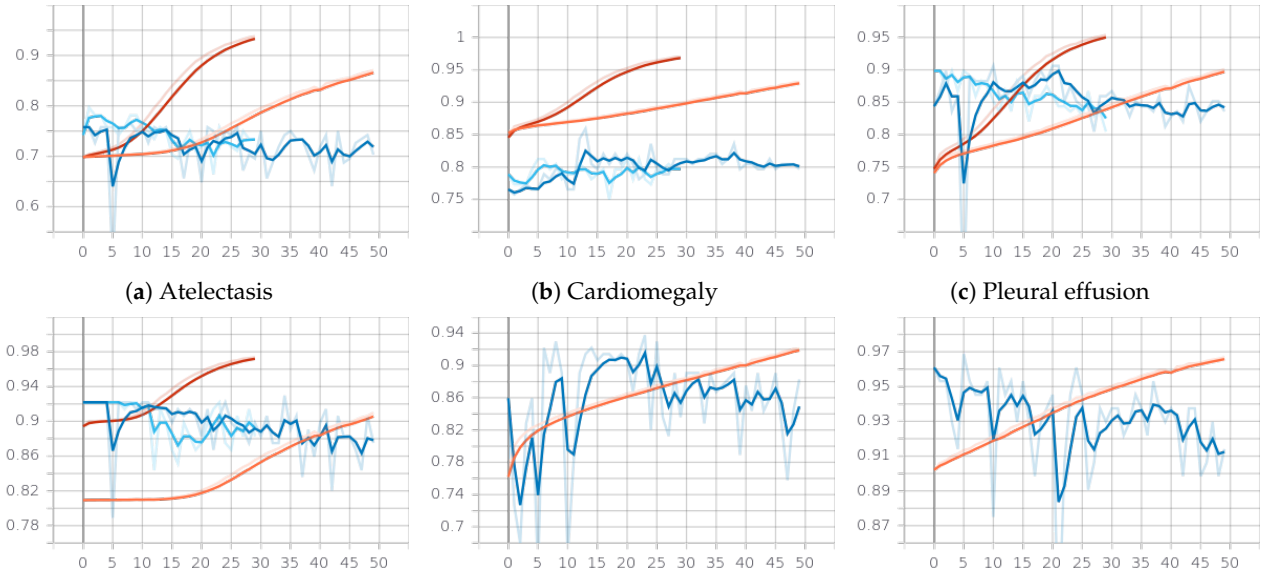
For each of the 14 observations, there is an output as a softmax layer with 3 neurons (or 2 neurons for no finding class), meaning that we have a 3-class classification for each of the 13 observations and a 2-class classification for no finding class.

CheXpert dataset is imbalanced [18], and the difference between the number of examples for positive, negative and uncertain classes also varies among observations. For example, for cardiomegaly, there are 12.26% positive samples, 3.52% uncertain, and 84.23% in the negative class. Lung opacity has 49.39% positive, 2.31% uncertain and 48.3% for the uncertain class. Therefore, in addition to fine-tuning parameters of the optimizer, we tested different combinations of the following elements:

1. Class weights: for each of the 14 observations, the weights of the 3 classes is computed according to the number of examples in each class (except for the no finding class, where the weight is computed only for 2 classes since there is no example in the uncertain class).
2. Reliable classes: training is done only for the the first 5 classes for which F1 of the labeler on positive and negative classes are all very high: edema, consolidation, atelectasis, cardiomegaly, pleural effusion.
3. Value of the weight for the uncertain class: the weight of the uncertain class for some outputs is halved such that the errors on this class to have reduced impact on the loss. The reason for doing this is the smaller quality of the labeler on this class.
4. Replacing the uncertain class: the samples from the class uncertain for some observations are considered from either positive or negative class.

Figure 7 shows the learning curves for six observations in two similar experiments: in $Exp14\_1$, the learning is done on all the 14 observations for 50 epochs, with the class weight computed according to the number of the examples in each class, while in $Exp14\_2$, the learning is done only on the 5 best annotated classes, with the halved weight for the uncertain class, for 25 epochs. The orange and dark blue curves are for the training, respectively, the validation set for $Exp14\_1$, while red and light blue are used for $Exp14\_2$.

Both experiments reveal that: (i) the overfitting is reached at different epochs for different observations; (ii) cardiomegaly seems to be underfit due to the difference between values on the validation and training sets; (iii) the value for the validation accuracy ranges from 75% for atelectasis and 94% for the pneumothorax. The observations support device and pneumothorax were included as outputs only for the experiment $Exp14\_1$.



(**a**) Atelectasis (**b**) Cardiomegaly (**c**) Pleural effusion

(**d**) Consolidation (**e**) Support device (**f**) Pneumothorax

Figure 7. Learning curve for the accuracy of 6 observations for *CheXpert-14* models. Two experiments ($Exp14\_1$ and $Exp14\_2$).

We underline again that we built the CheXpert-14 models in order to use the transfer knowledge approach for avoiding the lack of generality specific to learning from a small dataset. We considered that the features identified by CheXpert-14 models are considerably more valuable for learning from chest X-rays than features identified by training only on ImageNet.

3.2.2. Second Phase: Classifying COVID-19, Pneumonia, Tuberculosis and Normal with CheXpert Pretrained Weights

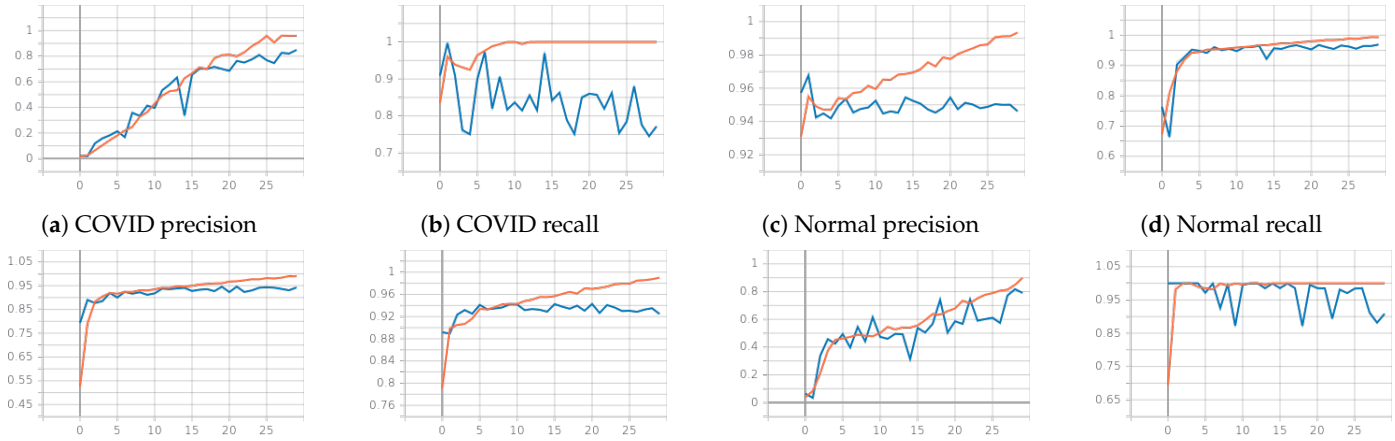
Our final aim is to build models which can distinguish among COVID-19, pneumonia, tuberculosis and normal classes. Since we build on top of the CheXpert-14 models, we named these models CheXpert-COVID19. We used the same architecture as the one for CheXpert-14 (Figure 6), with changes in the output layer, since there are 4 outputs, each with 2 classes (positive and negative). The network relies on all the layers from the CheXpert model, except the top layer. We decreased the learning rate in order to take small steps in adapting previously learned features on the new classification problem.

Note that the problem framing is different here than in the PDE-inspired networks (Exp_2 and Exp_3). In parabolic and hyperbolic networks, one of the COVID-19, pneumonia, tuberculosis or normal classes is mandatory for each image, since the networks predict 4 probabilities whose sum is 1. Differently, in CheXpert-COVID19, it is possible for an image to be positive for more than one of the COVID-19, pneumonia, tuberculosis or normal classes, or to be negative for all.

We considered that this design choice gives more flexibility to the network, and it is more appropriate to the medical perspective. That is, because a patient can have none of the three diseases, but at the same time, he or she does not have a normal aspect on the chest X-ray.

Differently to PDEs networks, for CheXpert-COVID19 model, the training is done on the complete dataset (Table 1). The imbalancing effect is tackled here with different class weights for each of the 4 outputs. The imbalancing effect is increased by the choice of having 4 outputs, since for instance, for COVID-19 examples, the negative examples include all the examples from normal, pneumonia, and tuberculosis.

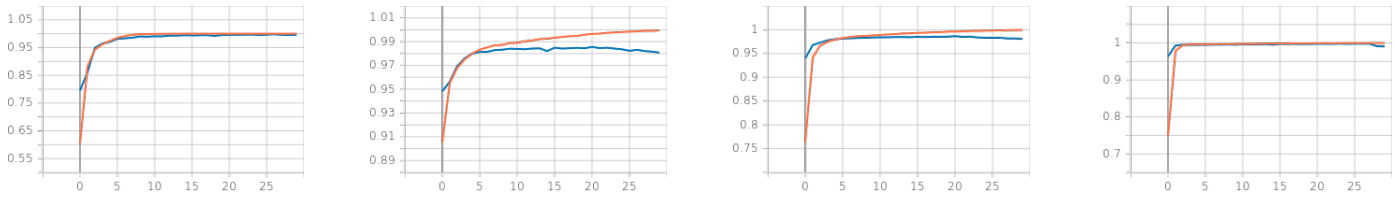
Given that the dataset is imbalanced, the accuracy is not so informative, even though we obtain values ≥ 0.97 for all the 4 outputs. Figure 8 presents the precision and the recall for all the 4 outputs, while Figure 9 includes the AUC for all four outputs.



(**a**) COVID precision (**b**) COVID recall (**c**) Normal precision (**d**) Normal recall

(**e**) Pneumonia precision (**f**) Pneumonia recall (**g**) Tuberculosis precision (**h**) Tuberculosis recall

Figure 8. Learning curve for CheXpert-COVID19 model for COVID-19, normal, pneumonia, tuberculosis (orange—training, blue—validation).



(**a**) COVID AUC (**b**) Normal AUC (**c**) Pneumonia AUC (**d**) Tuberculosis AUC

Figure 9. AUC for CheXpert-COVID19 model for COVID-19, normal, pneumonia, tuberculosis (orange—training, blue—validation).

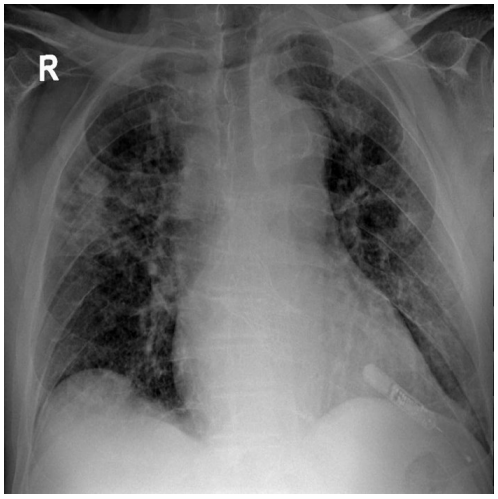
We can observe that the precision for COVID-19 increases, while the recall tends to decrease after several epochs. At the epoch 24, precision is 0.90 for the training set and 0.81 for the validation set, while the recall is 0.99 for the training set and 0.75 for the validation (see Table 3). At the same time, at epoch 26, the precision on validation is 0.74 and the recall 0.88. The effect of having a very small number of samples for COVID-19 is still present. At the same epoch 24, the precision for the normal output is 0.98 on the training set, 0.94 on the validation, while the recall is 0.98 on training and 0.96 on validation. We remind the reader that the number of normal images is much higher than the number of COVID-19 images. Tuberculosis is well recognized when present, with high values for recall on both training and validation sets, but the precision is small. This is due to the fact that several COVID-19 images are wrongly included in the positive class for the tuberculosis output. We underline again that with the CheXpert-COVID19 architecture, an image can have positive class for more than one outputs (i.e. to have a positive class for both tuberculosis and COVID-19). The experiments for CheXpert-14 and CheXpert-COVID19 models were run on Nvidia Tesla V100 with TensorFlow 2.0 (https://www.tensorflow.org/, accessed on 9 January 2021).

Table 3. CheXpert-COVID19 model: precision and recall on training and validation sets for the epoch 24 (and for epoch 26 for COVID-19.

Output	Training		Validation	
	Precision (%)	Recall (%)	Precision (%)	Recall (%)
COVID	90	99	82 (74)	77 (88)
Normal	98	98	94	96
Pneumonia	97	97	94	92
Tuberculosis	77	99	60	97

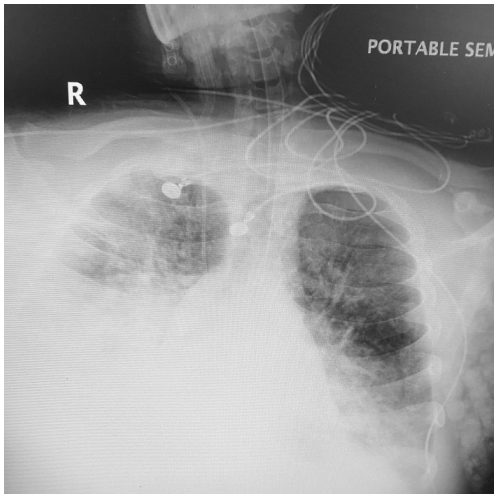
3.2.3. Qualitative Analysis

To go beyond the numerical evaluation, we applied our models on a set of X-rays from a different set, COVID-Net, and we analyzed the results together with two radiologists. We detail here the 14 observations predicted by the CheXpert-14 model, together with the predictions of the model CheXpert-COVID19 for the 4 class (Figure 10).



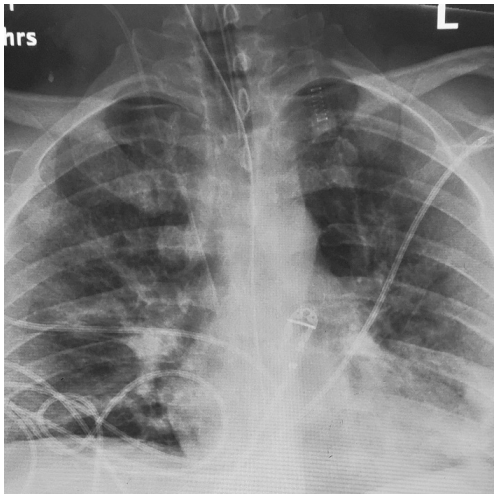
Chexpert	Absent	Present	Uncertain	Decision(max)/>0.5
No Finding	1.000	0.000	-1.000	False
Enlarged Cardiomediastinum	0.968	0.021	0.011	Absent
Cardiomegaly	0.978	0.020	0.002	Absent
Lung Opacity	0.130	0.861	0.009	Present
Lung Lesion	1.000	0.000	0.000	Absent
Edema	0.959	0.032	0.010	Absent
Consolidation	0.886	0.108	0.006	Absent
Pneumonia	0.955	0.004	0.041	Absent
Atelectasis	0.964	0.018	0.017	Absent
Pneumothorax	0.556	0.406	0.038	Absent
Pleural Effusion	0.880	0.099	0.021	Absent
Pleural Other	0.999	0.001	0.000	Absent
Fracture	0.011	0.983	0.006	Present
Support Devices	0.911	0.088	0.001	Absent

Covid Set	probability to be	Decision>0.5
Covid	1.000	Present
Normal	0.000	Absent
Pneumonia	0.110	Absent
Tuberculosis	0.002	Absent



Chexpert	Absent	Present	Uncertain	Decision(max)/>0.5
No Finding	0.997	0.003	-1.000	False
Enlarged Cardiomediastinum	0.980	0.008	0.012	Absent
Cardiomegaly	0.995	0.003	0.002	Absent
Lung Opacity	0.734	0.251	0.015	Absent
Lung Lesion	0.986	0.012	0.001	Absent
Edema	0.983	0.015	0.002	Absent
Consolidation	0.443	0.551	0.006	Present
Pneumonia	0.975	0.009	0.015	Absent
Atelectasis	0.923	0.068	0.009	Absent
Pneumothorax	0.995	0.002	0.003	Absent
Pleural Effusion	0.163	0.806	0.031	Present
Pleural Other	0.994	0.005	0.000	Absent
Fracture	0.971	0.028	0.001	Absent
Support Devices	0.311	0.687	0.002	Present

Covid Set	probability to be	Decision>0.5
Covid	0.919	Present
Normal	0.000	Absent
Pneumonia	0.968	Present
Tuberculosis	0.036	Absent



Chexpert	Absent	Present	Uncertain	Decision(max)/>0.5
No Finding	0.482	0.518	-1.000	True
Enlarged Cardiomediastinum	0.999	0.001	0.000	Absent
Cardiomegaly	0.999	0.000	0.000	Absent
Lung Opacity	0.456	0.524	0.020	Present
Lung Lesion	1.000	0.000	0.000	Absent
Edema	0.892	0.104	0.004	Absent
Consolidation	0.788	0.176	0.036	Absent
Pneumonia	0.967	0.009	0.024	Absent
Atelectasis	0.875	0.065	0.059	Absent
Pneumothorax	0.935	0.052	0.013	Absent
Pleural Effusion	0.961	0.038	0.001	Absent
Pleural Other	1.000	0.000	0.000	Absent
Fracture	1.000	0.000	0.000	Absent
Support Devices	0.342	0.656	0.001	Present

Covid Set	probability to be	Decision>0.5
Covid	0.990	Present
Normal	0.000	Absent
Pneumonia	0.396	Absent
Tuberculosis	0.934	Present

Figure 10. Three predictions of CheXpert-14 and CheXpert-COVID19 models on representative X-rays of COVID-19 patients; public available images from COVID-net dataset [19].

In the first example from Figure 10, all the predictions are correct, except the one for the fracture.

In the second example, even though the X-ray images are over-saturated at the bottom, the presence of consolidation and pleural effusion are correct, together with the decision for COVID-19. The prediction for pneumonia is wrong if we consider pneumonia to be the non-COVID pneumonia.

The third example is correctly classified as COVID-19, but the probability of having tuberculosis is too large (0.797). The probability of "No finding" is 0.52, which wrongly asserts the image to be without abnormalities (even though it is quite close to the 0.5 threshold). The model correctly identifies the presence of the support device and the lung opacity.

According to the qualitative analysis, even though the quality of the predictions improved after using the 2-phase training with CheXpert compared to a training without CheXpert, there is still room for improvement. The most important limitation remains the small number of samples for COVID-19-positive.

4. Discussion and Related Work

4.1. PDE-Inspired Networks

From our knowledge, this is the first approach of using PDE-inspired networks on chest X-rays. The performance of these networks on large datasets, as normal vs. pneumonia, is extremely good, even though the number of the trained parameters is small and the resolution of the images is 192×192. Precision and recall values (%) are 92.4/91.4 for non-COVID pneumonia, and 94.4/95.1 for normal class (obtained in Exp_1). For traditional CNNs, these values are common when non-COVID pneumonia is detected. The novelty of our results comes from using PDE-inspired CNNs.

Hyperbolic networks are also competitive with traditional CNNs for detection of tuberculosis. In [20], an accuracy of 84.4 is reported for the Shenzhen dataset. We underline that in the experiments Exp_2 and Exp_3, the downsampled dataset includes all the examples for tuberculosis class from the Shenzhen dataset. The precision and recall values for tuberculosis obtained in Exp_2 and Exp_3 are 98.5/100, respectively 98.5/95.5.

We obtained considerably smaller values for the detection of COVID-19. The values for precision and recall on the downsampled dataset are 81.8/72 without transfer learning, respectively 77.3/68.0 with transfer learning from Exp_1 (pneumonia vs. normal). In the next subsection, we compare these values to the ones obtained by CAD4COVID-XRAY [21]. We consider that CAD4COVID-XRAY stands out among the current results for COVID-19 identification from X-rays: not only a dataset extended with data from two hospitals is used, but also a comparison of the system performance to the performance of several radiologists was conducted.

Since discretizations of stable PDEs are used, we expected, and it indeed happened on our data, that once the objective function is learned, even though the training continues for several epochs, the value of the objective function does not increase (as it happens in the case of classical CNNs at some epochs after the overfitting point).

4.2. 2-Phase CheXpert-Based Networks

We compare CheXpert based networks with other existing approaches for COVID-19 identification in terms of: the considered classes and data sources, the used networks and the training method.

COVID-Net [19] was among the first efforts towards learning from X-ray images for COVID-19 and explaining the predictions through visualization. The used data included normal and pneumonia cases from RSNA pneumonia challenge, while for COVID-19, they combined several datasets, including Cohen dataset and Kaggle COVID-19 Radiography Database [9]. Wang et al. have used VGG-19, ResNet-50 and their own COVID-Net architecture. The best results, obtained with the COVID-Net architecture, are: accuracy 93.3%, recall for COVID-19 91%, respectively 95% for normal and 94% for non-COVID pneumonia.

Our performance for the detection of the normal and the non-COVID pneumonia classes are comparable with the ones obtained by COVID-Net. For the detection of COVID-19, we have smaller values, but here we have a possible explanation related to the used dataset. In our view, the [9] is a problematic dataset, since it includes pediatric X-ray images for normal and pneumonia classes.

When training DenseNet, with transfer learning directly from ImageNet, on the set from [9] (which is included in the COVID-net dataset), we also obtained good precision and recall (above 90%) for COVID-19 class. However, these results are not included here because of the previously mentioned pitfalls (recall Figure 1).

The fact that the classical deep neural network architectures work on chest X-ray images is also shown by Bressem et al. [22]. Densenet121, ResNet, SqueezeNet and VGG-19 are compared on CheXpert dataset and on a set for COVID-19, similar to the one used in COVID-net. We also trained on the CheXpert dataset but for different reasons, i.e. transfer the weights trained on this large dataset to the training on the small 4 classes dataset.

CAD4COVID-XRAY [21] is another approach based on the classical CNNs. Different to us, Murphy et al. have a significantly larger number of COVID-19 positive X-rays, 994, gathered from public sources, but also from Nerbhoven Hospital and Radboud University Medical center. Still, 994 images remains a small number, but significantly larger than what we used. Murphy et al. have pretrained the system on RSNA pneumonia challenge in order to tackle the issue of having too few samples for COVID-19. This approach is similar to our Exp_3 from PDEs inspired networks and with CheXpert-COVID19 model. The differences are: (i) in the case of the PDEs inspired networks, we used a downsampled dataset for the final training. (ii) in the case of the CheXpert-COVID19, we transfer from CheXpert instead of RSNA. Murphy et al. have considered the following classes: normal, non-COVID pneumonia, COVID-19 pneumonia, and other abnormalities inconsistent with pneumonia. They obtained AUC ROC = 0.81: at a 60% sensitivity; the specificity was 85%, at a 75% sensitivity, the specificity was 78%, and at 85% sensitivity, the specificity was 61%. Even though these values are not above 90%, Murphy et al. concluded that the AI system is comparable or even better than the human. They reached this conclusion after comparing the results of the AI system with 6 radiologists with experience. These sensitivity/specificity values for detecting COVID-19 pneumonia are comparable to the ones obtained by both PDE-inspired networks and the CheXpert-COVID19 model (see the confusion matrices for the Hamiltonian CNNs—Figure 5a,b, and for CheXpert-COVID19, recall the learning curves for precision and recall from Figure 8). Our CheXpert-COVID19 obtained better values on the validation set, but for a real comparison between methods only, we need to extend the set for COVID-19 positive X-rays.

From a different perspective, CT scans include significantly more information compared to X-ray images. The interested reader can explore the recent work of Yang et al. [23] or Zhou et al. [24] for screening COVID-19 or for the prediction of the disease severity/monitoring. The ImagingCOVID19AI European initiative, which gathers several hospitals from Europe, is of much interest, given the current technological difficulties, due to the lack of COVID-19 images.

4.3. Comparison between PDE-Inspired Networks and CheXpert-Based Network

We compare parabolic and hyperbolic networks with CheXpert-based network along several dimensions: (i) the size of the models (ii) the impact of the transfer learning; (iii) the quality of learning of the control class, tuberculosis.

The number of parameters is much smaller for the parabolic and hyperbolic networks: 8*e*6 for each of the models for CheXpert and the CheXpert-based COVID-model, while for the parabolic network 501,892, and hyperbolic network even less, 263,428. The fact that the small models of parabolic and hyperbolic networks successfully learned to differentiate pneumonia from normal images is inline with the conclusion suggested in [25] that "large models designed for ImageNet might be too overparameterized for the very small data regime".

Transfer learning: in both approaches, PDE-based and classical DenseNets approach, pretrained weights improve the final performance for the identification of the diseases.

In case of Exp_3 for PDE-inspired networks, the data on which we build the pretrained weights and the data for which we apply the transfer contain common classes (normal and pneumonia). The model built with transfer (Exp_3) is better at the level of normal and pneumonia than the model built without transfer (Exp_2) when only a small part of the data for normal and pneumonia is used.

For CheXpert-COVID19 model, the generality of the network is increased when CheXpert pretrained is used. Classification based on wrong relevant areas in the image, particularly to learning from a very small number of COVID-19 images, is avoided when using the pretrained weights on CheXpert. Nevertheless, in order to prove that the generality is increased in both cases when the transfer is used, further work is needed, which we believe must involve a group of radiologists.

Learning the control class—tuberculosis: in all the experiments with PDEs inspired networks on the four-class dataset, tuberculosis is well learned. In the case of the CheXpert-COVID19 model, images with tuberculosis are well classified, but sometimes, COVID-19 images are wrongly classified as tuberculosis. We could assume that this happens due to the stability property of the PDEs networks, but this needs further investigation.

5. Code and Data Availability

The data stored as MATLAB files, the logs and models for the parabolic and hyperbolic networks, together with trained weights for CheXpert-14 and CheXpert-COVID19 models and some running examples, are publicly available at GitLab.utcluj.ro/Anca.Marginean/XRay, accessed on 9 January 2021.

6. Conclusions

In the context of the actual pandemic, in certain areas with an increased number of cases and a lack of trained specialists (radiologists), the neural network's interpretation of chest X-rays from COVID-19 patients can act as a triage instrument, being a valuable tool, even if only to raise suspicion where the seen aspect is not normal, leading to more specific testing.

We framed the problem of identification of COVID-19 from X-rays as a multi-label classification into COVID-19, non-COVID pneumonia, tuberculosis and normal aspect. The contributions of this article are: (1) applying parabolic and hyperbolic CNNs on chest X-rays, on large and balanced dataset for normal and non-COVID pneumonia classes, and on the downsampled dataset for COVID-19, normal, pneumonia and tuberculosis classes; (2) using CheXpert dataset for a domain-specific transfer learning for differentiation of COVID-19 from pneumonia, tuberculosis and normal findings; (3) compiling a dataset with COVID-19, pneumonia, and tuberculosis, along with normal chest X-ray images. Our results have shown that PDE-inspired networks are competitive for learning from chest X-rays, even though the models are much smaller compared to DenseNets, and even when they are trained on an extremely small dataset. However, in order to have reliable models for the detection of COVID-19 from a very small number of images, domain-specific transfer learning helps in the case of both PDE-inspired networks and traditional DenseNets.

One observation which emerged during this work is that trained models with a high performance when analyzing X-rays should be treated cautiously and the radiologists should be involved in the design and learning process.

As future work, we plan to: (i) extend the set of COVID-19 images; (ii) do an extensive qualitative analysis of the results obtained by the PDE-inspired networks; (iii) train parabolic and hyperbolic networks on CheXpert dataset and explore the impact of the stability and reversibility properties on the learned features; (iv) consider the severity and the evolution of the disease.

Author Contributions: Conceptualization, A.N.M., R.R.S. and C.-M.P.; methodology, A.N.M., D.D.M. and A.G.; software, A.N.M., G.A.M. and C.L.T.; validation, D.D.M., G.A.M., A.P., G.Z.M., C.O.M. and M.M.C.; formal analysis, A.N.M., R.R.S., C.L.T. and C.-M.P.; data curation, D.D.M., G.A.M., A.P., G.Z.M. and M.M.C.; writing—original draft preparation, A.N.M., D.D.M., G.A.M., A.P., A.G. and C.-M.P.; writing—review and editing, A.N.M., D.D.M., G.A.M., A.P., A.G. and C.-M.P.; visualization, A.N.M. and D.D.M.; project administration, A.N.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ruthotto, L.; Haber, E. Deep Neural Networks Motivated by Partial Differential Equations. *J. Math. Imaging Vis.* **2020**, *62*, 352–364. [CrossRef]
2. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
3. Luján-García, J.E.; Moreno-Ibarra, M.A.; Villuendas-Rey, Y.; Yáñez-Márquez, C. Fast COVID-19 and Pneumonia Classification Using Chest X-ray Images. *Mathematics* **2020**, *8*, 1423. [CrossRef]
4. Sciacca, F.; Bell, D.J. COVID-19. 2020. Available online: https://radiopaedia.org/articles/covid-19-4 (accessed on 9 January 2021).
5. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269.
6. Cruz, B.G.S.; Sölter, J.; Bossa, M.N.; Husch, A.D. On the Composition and Limitations of Publicly Available COVID-19 X-ray Imaging Datasets. *arXiv* **2020**, arXiv:abs/2008.11572.
7. Cohen, J.P.; Morrison, P.; Dao, L. COVID-19 image data collection. *arXiv* **2020**, arXiv:abs/2003.11597.
8. Winther, H.B.; Laser, H.; Gerbel, S.; Maschke, S.K.; B. Hinrichs, J.; Vogel-Claussen, J.; Wacker, F.K.; Höper, M.M.; Meyer, B.C. COVID-19 Image Repository, 2020. Available online: https://www.rsna.org/news/2020/march/covid-19-imaging-data-repository (accessed on 20 December 2020). [CrossRef]
9. Chowdhury, M.E.H.; Rahman, T.; Khandakar, A.; Mazhar, R.; Kadir, M.A.; Mahbub, Z.B.; Islam, K.R.; Khan, M.S.; Iqbal, A.; Emadi, N.A.; et al. Can AI Help in Screening Viral and COVID-19 Pneumonia? *IEEE Access* **2020**, *8*, 132665–132676. [CrossRef]
10. Kermany, D.; Zhang, K.; Goldbaum, M. Labeled Optical Coherence Tomography (OCT) and Chest X-ray Images for Classification. Mendeley Data. 2018. Available online: https://data.mendeley.com/datasets/rscbjbr9sj/2 (accessed on 20 December 2020). [CrossRef]
11. Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; Summers, R.M. ChestX-ray8: Hospital-Scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3462–3471.
12. Jaeger, S.; Candemir, S.; Antani, S.; Wáng, Y.X.J.; Lu, P.X.; Thoma, G. Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. *Quant. Imaging Med. Surg.* **2014**, *4*, 475–477. [CrossRef] [PubMed]
13. Tian, Y.; Wang, Y.; Krishnan, D.; Tenenbaum, J.B.; Isola, P. Rethinking Few-Shot Image Classification: A Good Embedding Is All You Need? *Lect. Notes Comput. Sci.* **2020**, *12359*, 266–282. [CrossRef]
14. Martínez-García, M.; Zhang, Y.; Wan, J.; McGinty, J. Visually Interpretable Profile Extraction with an Autoencoder for Health Monitoring of Industrial Systems. In Proceedings of the 2019 IEEE 4th International Conference on Advanced Robotics and Mechatronics (ICARM), Toyonaka, Japan, 3–5 July 2019; pp. 649–654.
15. XtractOpen. Meganet.jl: A Fresh Approach to Deep Learning Written in Julia. 2018. Available online: https://github.com/XtractOpen/Meganet.jl (accessed on 20 December 2020).
16. Pan, I.; Cadrin-Chênevert, A.; Cheng, P.M. Tackling the Radiological Society of North America Pneumonia Detection Challenge. *Am. J. Roentgenol.* **2019**, *213*, 568–574. [CrossRef] [PubMed]
17. Ovalle-Magallanes, E.; Avina-Cervantes, J.G.; Cruz-Aceves, I.; Ruiz-Pinales, J. Transfer Learning for Stenosis Detection in X-ray Coronary Angiography. *Mathematics* **2020**, *8*, 1510. [CrossRef]
18. 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. *Proc. AAAI Conf. Artif. Intell.* **2019**, *33*, 590–597. [CrossRef]
19. 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. *Sci. Rep.* **2020**, *10*, 19549. [CrossRef] [PubMed]

20. Pasa, F.; Golkov, V.; Pfeiffer, F.; Cremers, D.; Pfeiffer, D. Efficient Deep Network Architectures for Fast Chest X-ray Tuberculosis Screening and Visualization. *Sci. Rep.* **2019**, *9*. [CrossRef] [PubMed]
21. Murphy, K.; Smits, H.; Knoops, A.J.; Korst, M.B.; Samson, T.; Scholten, E.T.; Schalekamp, S.; Schaefer-Prokop, C.M.; Philipsen, R.H.; Meijers, A.; et al. COVID-19 on the Chest Radiograph: A Multi-Reader Evaluation of an AI System. *Radiology* **2020**, *296*, 166–172. [CrossRef] [PubMed]
22. Bressem, K.K.; Adams, L.C.; Erxleben, C.; Hamm, B.; Niehues, S.M.; Vahldiek, J.L. Comparing different deep learning architectures for classification of chest radiographs. *Sci. Rep.* **2020**, *10*. [CrossRef]
23. Yang, W.; Cao, Q.; Qin, L.; Wang, X.; Cheng, Z.; Pan, A.; Dai, J.; Sun, Q.; Zhao, F.; Qu, J.; et al. Clinical characteristics and imaging manifestations of the 2019 novel coronavirus disease (COVID-19):A multi-center study in Wenzhou city, Zhejiang, China. *J. Infect.* **2020**, *80*, 388–393. [CrossRef] [PubMed]
24. Zhou, T.; Lu, H.; Yang, Z.; Qiu, S.; Huo, B.; Dong, Y. The ensemble deep learning model for novel COVID-19 on CT images. *Appl. Soft Comput.* **2021**, *98*, 106885. [CrossRef] [PubMed]
25. Raghu, M.; Zhang, C.; Kleinberg, J.; Bengio, S. Transfusion: Understanding Transfer Learning for Medical Imaging. *arXiv* **2019**, arXiv:abs/1902.07208.