Self-supervised deep convolutional neural network for chest X-ray classification

Matej Gazda, Jakub Gazda, Jan Plavka, Peter Drotar

***Abstract*—Chest radiography is a relatively cheap, widely available medical procedure that conveys key information for making diagnostic decisions. Chest X-rays are almost always used in the diagnosis of respiratory diseases such as pneumonia or the recent COVID-19. In this paper, we propose a self-supervised deep neural network that is pretrained on an unlabeled chest X-ray dataset. The learned representations are transferred to downstream task – the classification of respiratory diseases. The results obtained on four public datasets show that our approach yields competitive results without requiring large amounts of labeled training data.**

***Index Terms*—self-supervised learning, contrastive learning, deep learning, convolutional neural network, chest X-ray, COVID-19.**

I. Introduction

The medical imaging utilization has undergone a rapid increase in recent decades, increasing by more than 50% for some modalities [1]. Even though the rate of increase has slowed down in recent years [2], medical imaging is still considered as a significant diagnostic source.

Among all medical imaging modalities, radiography is cost effective and is frequently employed by hospitals, emergency services, and other medical facilities. Chest radiography (or chest X-ray (CXR)) is a painless, noninvasive, and powerful investigatory method that conveys crucial respiratory disease information. For these types of diseases, CXR is a basal diagnostic tool. In CXR, pulmonary opacification ("the white lung field") represents the result of a decrease in the ratio of gas to soft tissue in the lung. Pulmonary opacification has several likely causes, including atelectasis, bronchogenic carcinoma, pleural effusion, and tuberculosis, as well as both bacterial and viral pneumonia.

A diagnosis of pneumonia is usually made after considering a combination of the clinical symptoms (cough, fever, pathological respiratory sounds), laboratory results (white blood cell count, C-reactive protein and procalcitonin levels, blood gas analysis, and sputum culture), and the presence of pulmonary opacification on CXR. Although the diagnosis and treatment of pneumonia are straightforward in most cases, rapid and accurate diagnosis is specifically required in uncertain cases because complications resulting from an initial misdiagnosis may lead to prolonged hospitalization and resource-draining medical care. Currently, pneumonia is one of the most frequent causes of death for patients of all ages [3]; moreover, pneumonia accounts for a significant number of hospital admissions [4].

The recent outbreak of coronavirus disease 2019 (COVID-19) has ushered in unprecedented challenges for most medical facilities. The enormous number of infections calls not only for prevention but also for early diagnosis followed by effective treatment. In this scenario, chest radiography has proven to be one of the most time- and cost-effective tools for COVID-19 diagnosis [5]. In CXR, patients who suffer from COVID-19 pneumonia present a combination of different multifocal patterns of pulmonary opacification. However, compared to community-acquired bacterial pneumonia, these changes are frequently bilateral. Furthermore, while the distribution of these changes is initially peripheral, during the course of the disease, they usually spread to other parts of the lung parenchyma as well. A recent study [6] showed that the correctly diagnosing mild and moderate COVID-19 from CXR is challenging, even for experienced radiologists. A shortage of medical personnel, the need for large numbers of diagnostic decisions even under unfavorable conditions, and the need for quick and accurate medical decisions all mean that the need for computer-aided diagnostics is now greater than ever before.

In this study, we addressed the two most pronounced use cases of the CXR classification. Pneumonia is one of the most frequent and serious inflammatory conditions, and COVID-19 is currently the disease with a devastating impact on healthcare services and even economies in many states worldwide.

This study makes several contributions. First, we trained a deep convolutional neural network (CNN) in a self-supervised fashion on a large dataset of unlabeled CXR images. Second, we proposed utilizing this pretrained CNN as a feature extractor for several downstream tasks aimed at CXR classification. Third, while the proposed CXR classification network does not require large amounts of labeled data, it still achieves performance levels comparable to those of its supervised counterparts. The extensive experiments on COVID-19 and pneumonia detection tasks validated that the proposed model obtains very reasonable CXR feature representations; thus it enables accurate CXR classifications on the four evaluated datasets.

Manuscript submitted on March 04, 2021. This work was supported by the Slovak Research and Development Agency under contract No. APVV-16-0211 and by the Scientific Grant Agency of the Ministry of Education, Science, Research and Sport of the Slovak Republic and the Slovak Academy of Sciences under contract VEGA 1/0327/20.

M. Gazda and P. Drotár are with the Intelligent Information Systems Lab, Faculty of Electrical Engineering and Informatics, Technical University of Kosice, Letna 9, 04201 Kosice, Slovak Republic (e-mail: matej.gazda@tuke.sk; peter.drotar@tuke.sk).

J. Gazda is with the 2nd Department of Internal Medicine, Pavol Jozef Safarik University and Louis Pasteur University Hospital, Trieda SNP 1, 04011, Kosice, Slovak Republic (e-mail: jakub.gazda@upjs.sk).

J. Plavka is with the Department of mathematics and theoretical informatics, Faculty of Electrical Engineering and Informatics, Technical University of Kosice, Letna 9, 04201 Kosice, Slovak Republic (e-mail: jan.plavka@tuke.sk).

The remainder of this paper is organized as follows. In Section II, we provide a brief overview of the related works about CNN utilization for CXR classification and contrastive learning. In Section III, we introduce the proposed approach for a self-supervised CNN and its architecture. In Sections IV and V we describe the datasets used in this study, present the experiments and report their results. Finally, after providing a discussion in Section VI, we draw conclusions in Section VII.

II. Related works

In this section, we present the related works in terms of CXR classification and the methods involved in our work.

A. Convolutional neural networks for CXR classification

Two main enablers have occurred in the CXR classification domain in recent years. The most frequent limitation for successful pattern recognition in the biomedical imaging domain has always been (and still is) a scarcity of data. This problem was partially overcome with the introduction of transfer learning for convolutional neural networks (CNNs). Fine-tuned CNNs have shown enormous potential and have even outperformed fully trained CNNs in many applications [7]. The second enabler consisted of data: several large chest X-ray datasets were made publicly available that have made it possible to utilize many new methodologies for chest X-ray classification [8], [9]. Most of the forthcoming works have taken advantage of one or both of these enablers.

CXR classification has drawn attention from the research community for several years, but the arrival of the COVID-19 pandemic boosted interest in this topic, and the number of works has increased substantially. The majority of recent works on CXR classification focus solely or partially on COVID-19 classification.

Given the recent findings, the most straightforward approach, to diagnose COVID-19 from CXR images is to use existing CNN architectures pretrained on ImageNet and fine-tune them on a target COVID-19 dataset. This was the approach employed by the authors of [10]. They fine-tuned four state-of-the-art convolutional networks (ResNet18, ResNet50, SqueezeNet, and DenseNet-121) to identify COVID-19. Similarly, Apostolopoulos et al. [11] evaluated five other CNN architectures pretrained on ImageNet and found that the most promising results were achieved by the VGG-19 architecture and the compact MobileNet network.

Other authors have tried to optimize the performance by designing a CNN architecture tailored for CXR classification. These models are inspired by existing architectures such as CoroNet [12], which was inspired by the Xception design, DarkCovidNet [13], which was based on the DarkNet [14] CNN, and COVID-CAPS [15], which capitalizes on the capsule networks that preserve the spatial information between images. Instead of using an existing architecture, Wang et al. [16] employed generative synthesis to develop COVID-Net—a machine-designed deep CNN. Other authors have proposed methods to preprocess X-ray images [17] or extract specific features [18] to boost the classification performance.

Many of the above approaches have shown very promising results and achieved high classification accuracies. However, these methods must be considered with caution because several additional aspects must be considered before accepting a particular design as a production-ready solution. First, many of the aforementioned studies combine the COVID-19 dataset for experiments with other publicly available datasets to create a dataset used for model training and testing. This increases the chances that the model then provides an output not only based on disease-related features but also based on some dataset-specific aspects, such as contrast and saturation. Second, some studies, such as [15] and [16], utilize only simple hold-out model validation. The criticisms of previous studies are detailed in [19], where the authors propose a more robust solution called COVID-SDNet and utilize a new dataset for model validation.

Some of the authors have considered other practical aspects of CXR classification, such as the limited available datasets. Oh et al. [20] proposed a solution for overcoming the lack of sufficient training datasets based on a pretrained ResNet-18, which processed CXR images through smaller patches. The authors of [21] approached viral pneumonia detection as an anomaly detection problem. Using this approach, they were able to avoid having to train the model on large numbers of different pneumonia cases and focus solely on viral pneumonia. Recently, Luo et al. [22] proposed a framework to integrate the knowledge from different datasets and effectively trained a neural network to classify thoracic diseases.

To date, all the previous approaches have relied on backbone networks pretrained on ImageNet. Transfer learning makes CNNs trained on large-scale natural images suitable for medical images as well. However, the disparities between natural images and X-ray images is quite significant. Training a CNN from scratch on a large X-ray dataset can further boost the performance. Some papers already exist [23], [24], confirming that this is a viable approach.

B. Contrastive learning of visual representations

Self-supervised neural networks provide unprecedented performance in computer vision tasks. Generative models operate mostly in the pixel space, which is computationally expensive and unsustainable on larger scales. On the other hand, contrastive discriminative methods operate on the augmented views of the same image, thus avoiding the computationally costly generation of the pixel space. In addition, contrastive discriminative methods currently achieve state-of-the-art performances on self-supervised learning tasks [25] [26]. Various approaches exist regarding to model training in a self-supervised way. The main paradigm has shifted towards instance discriminative models, where similar contrastive learning (SimCLR) [27], momentum contrast for unsupervised visual representation learning (MoCo) [28] and bootstrap your own latent architecture (BYOL) [29] have demonstrated as-yet-untapped potential. The representations learned by these architectures are on par with those of their supervised counterparts [30] [31].

From the point of view of pretext task selection, contrastive learning can be divided into context-instance contrast and

context-context contrast [32]. The former tries to find relations between the local features and the global representation of an instance (i.e., wheels and windows to a car). We believe that the learned local features help to distinguish between the target classes. Some examples of pretext tasks working in the context-instance principle are a jigsaw puzzle [33] and rotation angle detection [34].

The Context-context contrast architectures focus on the relationships between the global representations of different samples. CMC [25], MoCo [28], and SimCLR [27] contrasts between the positive and negative pairs, where the positive pairs constitute the same image augmented in different ways while the negative pairs constitute all remaining images. The number of negative and positive pairs depends solely on the type of self-supervised architecture.

SimCLR and MoCo share the idea of using positive and negative pairs, but they differ in how the pairs are handled. In SimCLR, [28] negative pairs are processed within the batch; thus, SimCLR requires a larger batch size. MoCo's representations of negative keys are maintained in a separate queue encoded by a momentum encoder. BYOL claims to achieve better results than SimCLR and MoCo without using negative samples in its loss function. Different from SimCLR and MoCo, BYOL employs an L_2 error loss function instead of contrastive loss while using a principle similar to the momentum encoder introduced in MoCo.

BYOL takes advantage of two neural networks called "online" and "target" networks that learn by interactions between each other. BYOL initializes the optimization step by including one augmented view of a single image. It teaches the online network to correctly predict the representation of a differently augmented view of the same image produced by the target network.

III. Methods

A. Self-supervised Learning

Self-supervised learning (SSL) is a subset of unsupervised learning methods that aim to learn meaningful representations from unlabeled data. The representations can then be reused for downstream (target) tasks as either a base for fine-tuning or as a fixed feature extractor for models such as logistic regression, SVM, and many others. Because manually annotated labels are not available in the training data, the SSL's first step is to generate pseudolabels automatically through carefully selected pretext tasks.

Formally, self-supervised learning can be defined as the minimization of an objective function $J(\boldsymbol{\theta})$ parameterized by parameters $\boldsymbol{\theta} \in \mathbb{R}^d$, that represents the mean loss over all training samples:

$$J(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{x} \sim \hat{p}_{data}} \mathcal{L}(m(\boldsymbol{x}; \boldsymbol{\theta}), \pi(\boldsymbol{x})), \tag{1}$$

where $\hat{p}_{data}$ is an empirical distribution, $\mathcal{L}$ is the per-example loss function, $m(\cdot, \cdot)$ is the model prediction when the input is $\boldsymbol{x}$, and $\pi(.)$ is a function that returns pseudolabel for input $\boldsymbol{x}$ based on the pretext task.

Optimization of such a neural network is accomplished similarly to supervised learning – by updating the parameters $\boldsymbol{\theta}$ in the direction of the antigradient using methods based on stochastic gradient descent:

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \eta \frac{1}{B} \sum_{i=Bt+1}^{B(t+1)} \frac{\partial \mathcal{L}(m(\boldsymbol{x}_i; \boldsymbol{\theta}), \pi(\boldsymbol{x}_i))}{\partial \boldsymbol{\theta}}, \tag{2}$$

where $\mathcal{L}$ is a loss function of the i-th example from the batch sampled at time t, B stands for the batch size, and η stands for a hyperparameter called the learning rate.

Modern SSL designs decouple the neural network architecture from the downstream tasks, which makes the transfer of knowledge more straightforward. State-of-the art SSLs such as SimCLR [27] and MoCo [28] use the ResNet50 [35] architecture on datasets such as CIFAR-10 [36] and ImageNet [37] just as their supervised counterparts do.

Transfer Learning

Despite recent advancements in deep learning and hardware accessibility, neural network training still tends to be slow and resource intensive. Thus, the transfer of knowledge from one domain to another reduces these burdens and has proven to be effective in numerous applications [38] [39].

Transfer learning is applicable to tasks with different degrees of label availability. The knowledge extracted from a base domain can be acquired in an unsupervised, semisupervised, or supervised fashion. For unsupervised pretraining, transfer learning is defined as follows. Let $\mathcal{D}_S = (\mathcal{X}_S, \mathcal{P}_S)$ be a pretext dataset consisting of a set of samples $\mathcal{X}_S$ with corresponding pseudolabels $\mathcal{P}_S$ generated by the underlying pretext task and a downstream dataset $\mathcal{D}_T = (\mathcal{X}_T, \mathcal{Y}_T)$, where $\mathcal{X}_T$ denotes the set of training samples and $\mathcal{Y}_T$ denotes the set of true labels. Given example source datasets $\mathcal{D}_S$, pretext tasks $\mathcal{T}_S$, downstream datasets $\mathcal{D}_T$, and downstream tasks $\mathcal{T}_T$, transfer learning aims to reduce the loss function $\mathcal{L}$ of the model used for downstream tasks ($\mathcal{T}_T$) using the knowledge acquired from pretext tasks $\mathcal{T}_S$, where $\mathcal{D}_S \neq \mathcal{D}_T$ or $\mathcal{T}_S \neq \mathcal{T}_T$.

B. Proposed approach

For the discriminative pretext task, we chose a proven contrastive approach similar to that used in SimCLR and MoCo [40] [27].

The contrasting method learns by comparing representations of differently augmented input images. The input images are formed into positive and negative pairs. A positive pair is formed by two images that were augmented differently from a common source image. Conversely, two images are denoted as a negative pair when they were augmented from different source images. The neural network learns to discriminate between the positive pairs and negative pairs by maximizing the agreement between two differently augmented views of the same data example.

The representations are compared by cosine similarity, which is defined for two vectors $\boldsymbol{u}, \boldsymbol{v}$ as follows:

$$\mathrm{sim}(\boldsymbol{u}, \boldsymbol{v}) = \frac{\boldsymbol{u}^T \boldsymbol{v}}{||\boldsymbol{u}||_2 \cdot ||\boldsymbol{v}||_2}. \tag{3}$$

TABLE I
PROJECTION HEAD $n(.)$

Layer	Size	Bias
Global Average Pooling Layer	-	-
Dense Layer	2048	True
Batch Normalization Layer & ReLU	-	-
Dense Layer	128	False

Other similarity functions such as Euclidean distance or dot product could also be employed.

The proposed learning architecture consists of three parts: a backbone neural network $m(.)$, a projection head $n(.)$, and a stochastic data augmentation module $\mathcal{A}$. The backbone network $m(.)$ is a ResNet50 Wide network that extracts representations from the augmented data examples. The projection head $n(.)$ (see Table I) transforms the output from the backbone network into a latent space where contrastive loss is applied. The size of the output vector is a hyperparameter allowing the final size to be adjusted to properly reflect the size of the original image. The data augmentation module $\mathcal{A}$ is a module that returns random augmentations as follows: resized crop, horizontal flip, rotation, Gaussian blur and color-jitter. A resized crop involves a random crop of the image followed by a resize back up to the original image size. The entire learning architecture is depicted in Fig. 1.

The optimization step is performed as follows. First, for a minibatch sampled at time t and consisting of images $\mathcal{X}_t = \{\boldsymbol{x}_{Bt+1}, \boldsymbol{x}_{Bt+2}, \ldots, \boldsymbol{x}_{Bt+B}\}$ of size B is drawn from the dataset samples $\mathcal{X}$, $\mathcal{X}_t \subseteq \mathcal{X}$ similarly to supervised learning. Then, for each image in the minibatch, a positive pair is formed by augmenting the image twice with random augmentations, from which $2B$ images are obtained. The images are then encoded via the backbone network $m(.)$ to obtain the representation vectors. The representations are passed through the projection head $n(.)$ to obtain projection vectors. The set of projection vectors is denoted as $\mathcal{Z}_t$. To calculate the model error, we apply the NT-Xent loss (*normalized temperature-scaled cross entropy loss*) introduced in [41]. For a positive pair $(\boldsymbol{z}_i, \boldsymbol{z}_j)$ drawn from the set of projections of augmented images, $\mathcal{Z}_t$ is loss calculated as follows:

$$l(\boldsymbol{z}_i, \boldsymbol{z}_j) = -\log \frac{e^{\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau}}{\sum\limits_{z_k \in \mathcal{Z}_t - \{\boldsymbol{z}_i\}} e^{\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau}}, \qquad (4)$$

where sim is a similarity function, $\mathcal{Z}_t - \{\boldsymbol{z}_i\}$ are the $2B$ projections of augmented images (with the exception of the projection $\boldsymbol{z}_i$), and τ is a hyperparameter called temperature.

After the loss is calculated, we backpropagate the errors to optimize the weights of the backbone neural network $m(.)$ and the projection head $n(.)$. At the end of the training process, we extract the features from the last layer of the backbone neural network $m(.)$ and discard the projection head $n(.)$.

IV. DATA

In this study, we utilized several chest X-ray datasets. First, a large-scale dataset is required for network pretraining. Therefore, to formulate the pretext task, we utilized the CheXpert dataset [8], which contains 224,316 chest radiographs from 65,420 patients. The samples were labeled by extracting data from radiology reports from October 2002 to July 2017 for 14 commonly observed conditions such as pneumonia, pneumothorax, and cardiomegaly. It should be noted that even though labels are available, we do not use these during pretraining because the proposed model is unsupervised.

TABLE II
DATASETS USED IN THIS STUDY

Dataset	# samples	# class 0	# class 1	source
pretext				
CheXpert	224,316	na	na	[8]
target				
Cell	5,323	3,883	1,349	[42]
ChestX-ray-14	2,706	1,353	1,353	[9]
C19-Cohen	418	564	243	[43]
COVIDGR	852	426	426	[19]

To evaluate the transferability of the model to an external target dataset, we acquired four public datasets. First, the Cell dataset [42] includes 5,323 X-ray images from children, including 3,883 cases of viral and bacterial pneumonia and 1,349 normal images. The labels were provided by two expert physicians and verified by a third physician. The second dataset is the ChestX-ray14 [9] dataset, which comprises 112,120 X-ray images with eight disease labels from 30,805 patients. We used only a subset of this dataset by selecting only patients with pneumonia and a matched number of healthy controls. The other two datasets were compiled only recently and were intended for COVID-19 detection. The C19-Cohen dataset [43] is a collection of different types of pneumonia (viral, bacterial, and fungal). We selected two classes: 304 patients with COVID-19 and 114 patients with other types of pneumonia. Finally, we also evaluated the proposed model on the COVIDGR dataset [19], which contains 426 CXR images of COVID-19 patients with four different severity levels and the same number of control subjects. Note that while 76 out of these 426 COVID-19 patients were diagnosed as positive by PCR, but their CXRs were evaluated as normal, making the classification task more challenging. A brief summary of all the datasets utilized in this study is presented in Tab. II.

V. EXPERIMENTS AND RESULTS

In this section, we analyze some aspects of the network pretraining task and report the experimental results.

To demonstrate the generalizability of our approach, we formulated four classification tasks on four publicly available CXR datasets. We avoided the combination of datasets for a particular classification task and used only one dataset for a specific classification task. In this manner, we try to avoid the bias and criticism outlined in [19]. The datasets details are presented in Tab. II. For the Cell dataset and ChestX-ray-14, we classified subjects as having pneumonia or being healthy. Similarly, for COVIDGR, we discriminated between patients with and without COVID-19 disease. Finally, because C19-Cohen dataset does not include healthy controls, we discriminated between COVID-19 and other types of pneumonia.

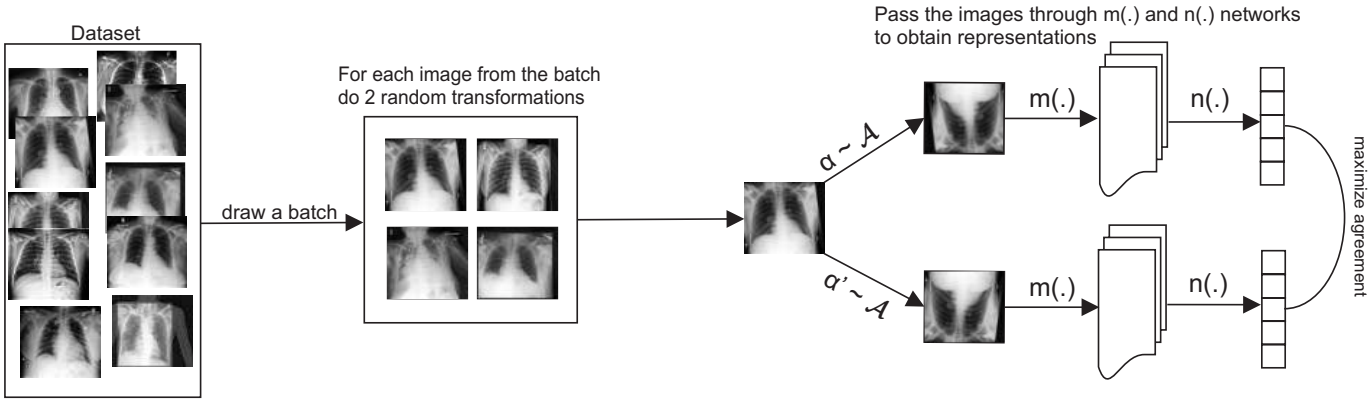

Fig. 1. Self-supervised architecture

A. Network training and feature extraction

We pretrained a ResNet-50 Wide model in the self-supervised task-agnostic way on the large CheXpert dataset of CXR images. The effective batch size was set to 128, and the temperature value was 0.5. We used the Adam optimizer [44] with a learning rate of 0.0005. The training process convergence is depicted in Fig. 2 (a). The loss on the pretext task plateaus at approximately the 25th epoch and does not decrease further but instead oscillates around some value. This result raises the question of whether it is necessary to train the model beyond the 25th epoch. However, the relationship between the loss on the pretext task and the model's performance on the target task has not yet been established. Here, we analyze the performance in terms of the prediction accuracy on the two datasets (the Cell dataset and the COVIDGR dataset). The prediction accuracy for each epoch is depicted in Fig. 2 (b) and Fig. 2 (c). Although the loss on the pretext task does not improve significantly beyond the 25–30th epochs, the accuracy achieved for predictions on the Cell dataset increases when we employ models that have been trained with a larger number of epochs. This indicates an interesting phenomenon. During pretraining, even after the loss on the pretext task is no longer improving, the model is still learning. In contrast, to the prediction accuracy obtained on the Cell dataset, for the COVIDGR dataset, the accuracy did not gradually improve for models trained beyond the 25th epoch. However, this can be explained by the composition of the images in the dataset. As explained before, COVIDGR also contains CXR images from COVID-19 patients in which expert radiologists did not find any pathological changes. The highest reported result so far on this dataset is 76.18% [19]. Our model achieved this accuracy level quite early; thus we hypothesize that there was no room for further improvement.

To visualize the learned representations, we chose models from four different checkpoints and extracted features. The models were trained for 10, 25, 50 and 100 epochs. Fig. 3 shows t-SNE visualizations of features extracted from the Cell and COVIDGR datasets by these four models. While the Cell dataset exhibits a noticeable but slight improvement in the separability of two classes, the two classes for the COVIDGR dataset seems to be interlaced through all four images.

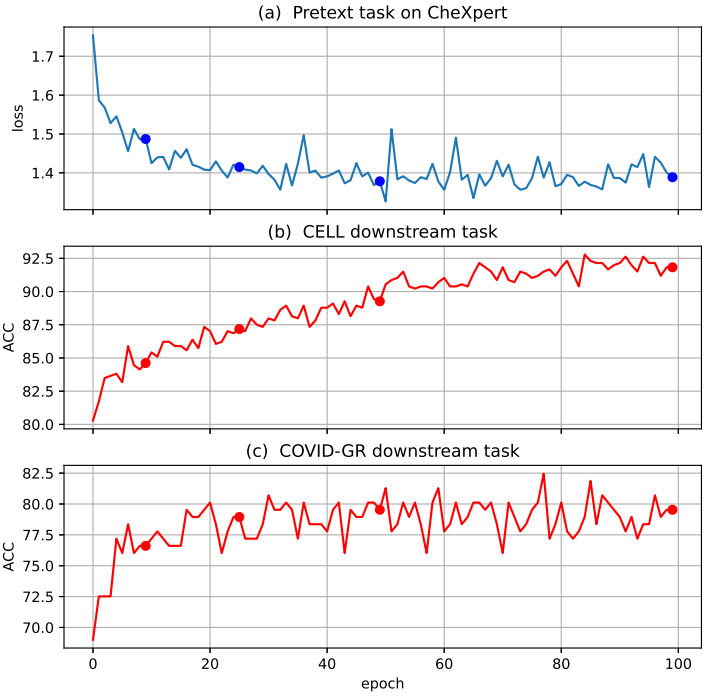

Fig. 2. Loss and accuracy in each epoch. (a) Loss on the pretext task and accuracy on the (b) Cell dataset and (b) COVIDGR dataset.

B. Numerical results

To examine the predictive performance of the proposed approach, we employ transfer learning and use the pretrained network as a fixed feature extractor. We adopted logistic regression as a classifier and evaluated the model on four different CXR datasets. To ensure the model's generalizability and to avoid overfitting, we used stratified k-fold cross-validation. The datasets were divided into training, validation, and testing subsets. If the original paper that introduced the datasets also provided the specification of train/validation/test subsets of data, we used that division to achieve fair comparisons

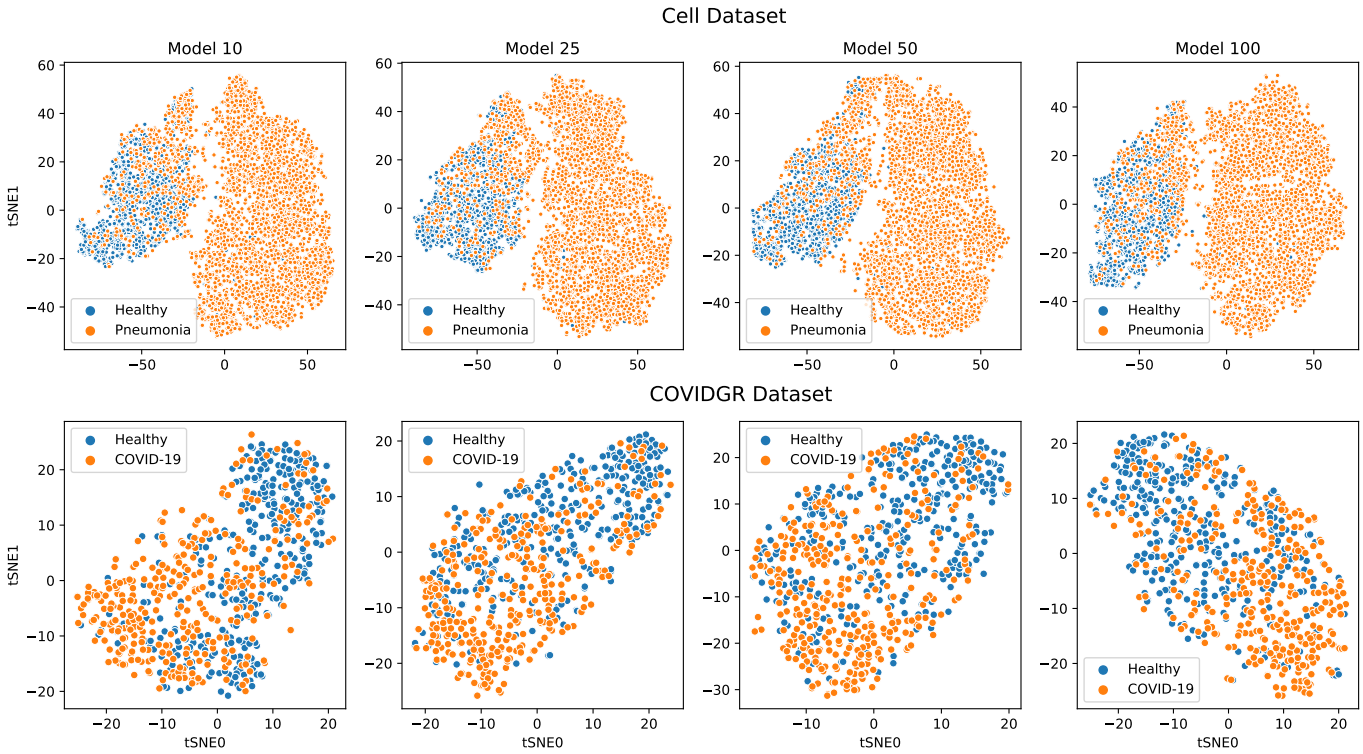

Fig. 3. t-SNE visualization of the features extracted by models selected in different stages of the pretraining process (10th, 25th, 50th and 100th epochs). The features shown are from the Cell and COVIDGR datasets.

of model accuracy with the published results (COVIDGR, ChestX-ray-14 and Cell). Otherwise, we divided the data as follows: 70% as training samples, 10% for validation and 20% as test data (C19-Cohen). Furthermore, we ensured that the CXR images of a particular patient were present only within the same data subset to prevent data leakage that would cause positive bias.

To determine the optimal logistic regression parameters, we searched through the parameter space $C = \{0.01, 0.05, 0.1, 0.2, 0.5, 1\}$. The logistic regression weights were automatically adjusted to be inversely proportional to the class frequencies in the input data. The other parameters were set to their default values. The best model was adopted after a grid search based on the area under the receiver operating characteristic curve (AUC) metric.

We also evaluated the amount of data required for the pretext task to correctly identify relevant features that are beneficial for the downstream task by testing with three different data fractions: 1%, 10% and 100%.

The results of the trained models are depicted in Table III. To provide a better overview of model performance, we calculated several metrics: accuracy (ACC), AUC, sensitivity (SEN), specificity (SPE) and F1-score.

Our first observation is that the prediction accuracy differs between the datasets. The model prediction accuracy varies more in the pneumonia classification task (Cell, ChestX-ray-14) than in the COVID-19 classification (C19-Cohen, COVIDGR). This variation is caused by the different dataset compositions. The datasets were compiled from different sources and acquired by different devices that influence the image characteristics. However, more importantly, the Cell dataset is composed of CXR children aged four to six years, making classification a specific type of task.

On the Cell dataset, the AUC and ACC tend to increase as the dataset fraction size increases for the pretext task. The highest AUC 97.7% of our model is higher than the 96.6% reported in [42], which used transfer learning based on ImageNet. This result shows that representations learned in a self-supervised fashion on smaller datasets with semantically closer domains are more beneficial than is supervised pretraining on large but semantically very different datasets such as ImageNet.

The model evaluated on the ChestX-ray-14 dataset yielded significantly lower results than the model evaluated on the Cell dataset. In this case, it is independent of the fraction of dataset used for training the pretext task. Tab. III shows that the proposed model achieves a higher score than the results published in [9]. However, the comparison is not entirely fair because the authors of [9] were solving a multiclass problem, which could have had a negative impact on the model accuracy compared to a binary classification.

Models trained on the COVIDGR and Cohen-19 datasets achieved comparable results. One model trained on C19-COHEN achieved AUCs up to 91.5% when using a 10% fraction of the dataset in the pretext task. Surprisingly, it outperformed another model trained on the pretext task with the entire dataset. We hypothesize that this may have been caused by the better performance of the logistic regression model due to the hyperparameters found in the grid search. Some of the previously published papers combined the C19-

Cohen dataset with CXRs of healthy controls obtained from other datasets. We intentionally avoided such combinations, and to the best of the authors' knowledge, no published paper has conducted classification only on the C19-Cohen dataset; thus, we cannot directly compare the performance of our model with those of others on the C19-Cohen dataset.

Encouraging results were achieved on the COVIDGR dataset in differentiating between healthy and COVID-19 CXRs. Our CNN pretrained in a self-supervised fashion was able to outperform the supervised COVID-SDNet [19] model by a few percent. Although this difference is not large, it should be noted that self-supervised learning does not require a large, labeled training dataset, which can save a substantial human resources.

C. Explaining CNN decisions

To shed some light on the CNN decisions, we employ Grad-CAM to highlight the important regions of the CXR image corresponding to a decision. Fig. 4 shows CXRs of six different patients correctly classified as pneumonia cases. Images of both ground-glass opacities and consolidations are present together with air bronchograms. An air bronchogram is a dark radiographic appearance of an air-filled bronchus (dark thread-like line) made visible by the opacification of the surrounding alveoli ("white lung field"). An air bronchogram is another pathological sign frequently associated with pneumonia. In this instance, the areas in the lungs highlighted by the attention map cover the regions with visible pulmonary opacification and air bronchograms, which provides the grounds for a correct diagnosis. Some other areas highlighted by the attention map exist outside the lung region. These areas cannot be linked to any pathology caused by pneumonia and probably reflect zones incorrectly evaluated by the visualization algorithm or by the CNN itself.

VI. Discussion

We proposed an approach based on a self-supervised convolutional network and evaluated it on four datasets. To avoid the critiques presented by [19] and others [45], we did not combine existing datasets, which clearly limits the available training and testing data size. On the other hand, it helps ensure that the model learns only parameters related to specific aspects of the disease pathology, not the differences instigated by different devices or acquisition procedures. To increase the confidence of our approach, we evaluated the trained model on four different datasets focused on two different diseases. We differentiated between CXR with no findings and CXR-containing pathologies and also between CXR from patients with different pneumonia types and COVID-19-induced pneumonia.

One issue we touched on briefly in Section V-A is the selection of the optimal model for transfer learning. Fig. 2 clearly shows that the relationship between model performance on the pretext task and a subsequent downstream task is not straightforward. The question that arises is how to select the optimal checkpoint for the pretrained model. Based on the traditional training view, the user may be tempted to stop the training at approximately the 40th epoch, based on the loss curve in Fig. 2(a), because the loss is no longer improving. However, at this point, the model is far from optimal in the sense of the prediction accuracy on the downstream task (at least for the Cell dataset 2(b)). Further research is needed to establish the relationship between model performance on pretext tasks and downstream tasks.

We provide decision support for the classification of CXR images; however, the output should be taken with caution. The final diagnosis should always be based on a combination of the clinical symptoms (cough, fever, pathological respiratory sounds) and laboratory results. However, in critical pandemic situations such as the one we are experiencing currently, medical staff are extremely busy, and a solution that provides rapid automated information could help to reduce the burden on medical personnel.

Clearly there is a strong need for further validation and detailed assessment led by transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) [46] before an automated approach can be used in clinical practice. Additional datasets from different types of devices need to be included in testing and evaluation. However, this study has proven that it is possible to train the model in a self-supervised fashion and apply it successfully to medical imaging tasks without the need for large amounts of labeled data. This may open new research horizons because a similar approaches can be examined for other types of medical imaging, such as computer tomography and retinal imaging.

The proposed approach for CXR classification is based on a deep CNN, and a fundamental principle of these models is that they work in a black-box manner. The ability to explain the decisions of deep learning models is still in its early stages and is a hot research topic. We provide some explanations based on the learned attention regions in Fig. 4, but these are not able to explain all the aspects and peculiarities included in the final decision. Because explainability is of crucial importance for medical applications, this and other limitations will be addressed in future work.

A. Fault analysis

We also investigated some misclassifications to obtain a better understanding of CNN decisions. Fig. 5 shows four CXRs that were incorrectly classified. Both cases (a) and (b) were misclassified as pneumonia. The patient in Fig. 5(a) has a small consolidation-like area ("white lung field") in the (right) middle lobe, and Case (b) shows a distinct diffuse reticular interstitial pattern. Both patterns may resemble pneumonia findings, which may have led the model to incorrect classification. It is likely that a radiology specialist would make the same mistake if the only available information was the radiograph. The true cause of these patterns would have to be determined by the patient's clinical history and additional radiological examinations (such as computed tomography scan). On the other hand, the CXR images in 5(c) and 5(d) were also misclassified as pneumonia but these do not present any structural changes that could be associated with the incorrect classification and should have been classified as healthy patients. It is difficult

TABLE III
PREDICTION PERFORMANCE OF THE PROPOSED APPROACH ON FOUR CXR DATASETS. WE PROVIDE PREVIOUS PUBLISHED RESULTS FOR COMPARISON PURPOSES. * THERE IS NO AVAILABLE RESULT FOR COMPARISON BECAUSE PREVIOUSLY PUBLISHED RESULTS COMBINE THE C19-COHEN DATASET WITH SOME OTHER CXR DATASET. MOREOVER, DATA ARE CONTINUOUSLY ADDED TO THE 19-COHEN DATASET. ** RESULT OF PNEUMONIA PREDICTION AS PART OF A MULTI-CLASS CLASSIFICATION.

Dataset	Fraction of dataset															Published result
	1%					10%					100%					ACC/AUC
	ACC	AUC	SEN	SPE	F1	ACC	AUC	SEN	SPE	F1	ACC	AUC	SEN	SPE	F1	
Cell	85.6	96.6	99.5	62.4	76.4	86.9	96.9	99.2	66.2	79.1	91.5	97.7	98.7	79.5	87.5	92.8/96.8 [42]
C19-Cohen	84.9	89.2	90.6	73.6	76.5	81.8	91.5	85.8	73.6	72.9	81.1	88.2	83	77.4	73.2	na*
COVIDGR	79.5	86.6	83.5	75.6	78.8	77.8	86.0	80	75.6	77.4	78.4	87.1	83.5	73.3	77.3	76.16/na [19]
ChestX-ray-14	71.5	79.1	72.8	71.4	83	71.2	78.1	72.0	71.2	82.9	71.4	78.4	71.3	71.5	83.0	na/65.8** [9]

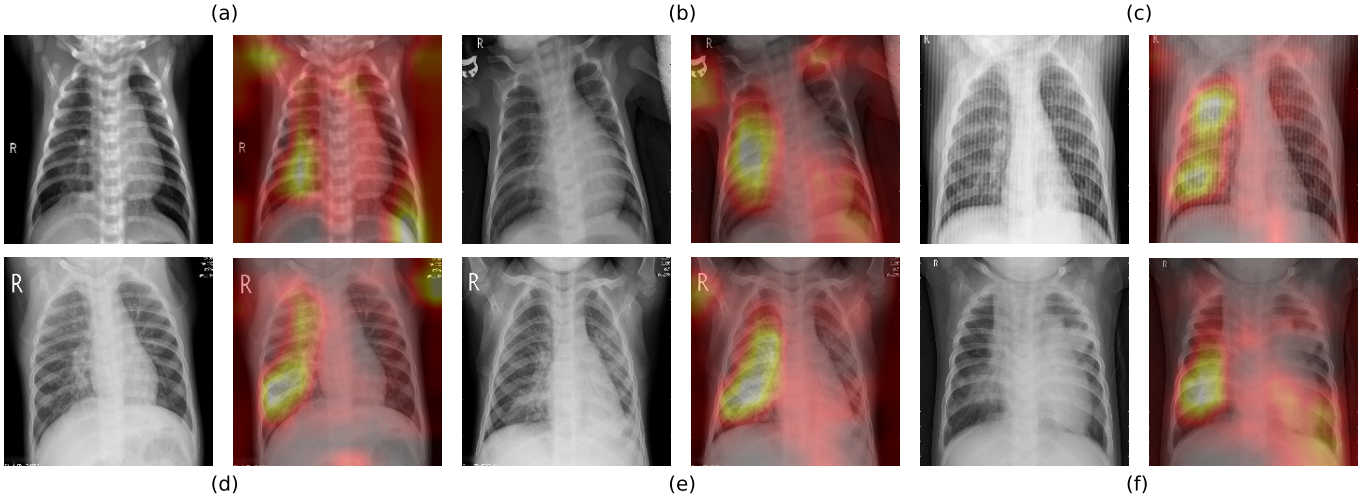

Fig. 4. Visualizations of 6 CXR images from the Cell dataset and their Grad-CAM attention maps. These images were taken of patients with pneumonia and were diagnosed as pneumonia by our model.

to determine the cause of these particular misinterpretations. This demonstrates the disadvantage of a CNN behaving as a black box.

VII. CONCLUSIONS

The current pandemic further highlights the need for including diagnostic decision support systems in clinical decision making. The successful incorporation of these systems into contemporary medical devices could automate certain tasks and reduce medical personnel workloads. Automated solutions also contribute strongly during noncritical times by providing medical specialists with more time for tasks and duties that require a more careful or specific approach.

As a contribution to medical expert systems, we introduced a solution that classifies CXR images. The proposed approach utilizes a CNN pretrained on an unlabeled dataset of CXR images. By avoiding the need for labeled data, which are both scarce and expensive in the medical domain, our approach offers new possibilities for CNN utilization by demonstrating that CNN networks do not need to be trained on only natural images (such as the ImageNet dataset), as in the majority of approaches today; they can instead be trained on images that are semantically closer to the target task. In our case, a network pretrained on the ChestXpert dataset was able to learn meaningful representations and extract relevant features for pneumonia and COVID-19 detection. The obtained results of our unsupervised model are competitive with their supervised counterparts. Considering that self-supervised contrastive learning for visual representations is a very new topic,

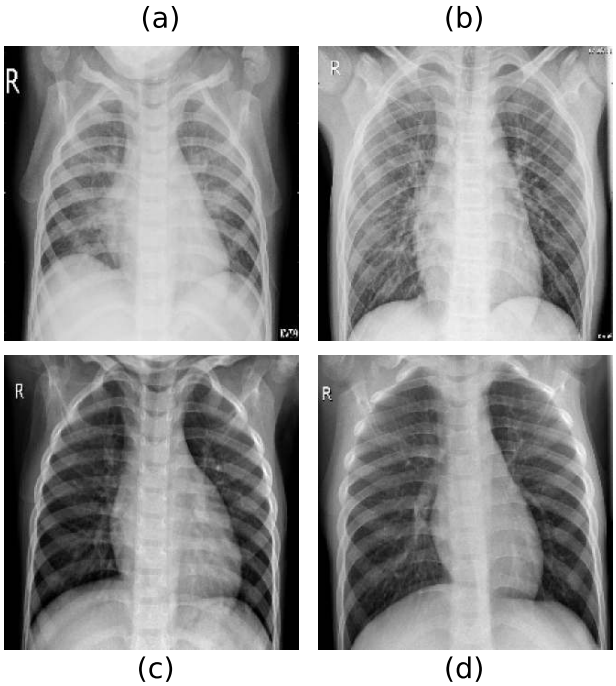

Fig. 5. CXR images of four healthy subjects classified as pneumonia.

this approach represents a huge potential. Later methodological improvements may further boost the performance.

ACKNOWLEDGMENT

We would like to thank Dr. Ján Buša and David Hubáček, MD. for valuable comments.

REFERENCES

[1] R. Smith-Bindman, D. L. Miglioretti, E. Johnson, C. Lee, H. S. Feigelson, M. Flynn, R. T. Greenlee, R. L. Kruger, M. C. Hornbrook, D. Roblin, L. I. Solberg, N. Vanneman, S. Weinmann, and A. E. Williams, "Use of Diagnostic Imaging Studies and Associated Radiation Exposure for Patients Enrolled in Large Integrated Health Care Systems, 1996-2010," *JAMA*, vol. 307, no. 22, pp. 2400–2409, 06 2012. [Online]. Available: https://doi.org/10.1001/jama.2012.5960

[2] A. S. Hong, D. Levin, L. Parker, V. M. Rao, D. Ross-Degnan, and J. F. Wharam, "Trends in diagnostic imaging utilization among medicare and commercially insured adults from 2003 through 2016," *Radiology*, vol. 294, no. 2, pp. 342–350, 2020, pMID: 31891320. [Online]. Available: https://doi.org/10.1148/radiol.2019191116

[3] O. Ruuskanen, E. Lahti, L. C. Jennings, and D. R. Murdoch, "Viral pneumonia," *The Lancet*, vol. 377, no. 9773, pp. 1264 – 1275, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0140673610614596

[4] J. S. Brown, "Community-acquired pneumonia," *Clinical Medicine*, vol. 12, no. 6, pp. 538–543, 2012. [Online]. Available: https://www.rcpjournals.org/content/12/6/538

[5] 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. W.-H. Chung *et al.*, "Frequency and distribution of chest radiographic findings in patients positive for covid-19," *Radiology*, vol. 296, no. 2, pp. E72–E78, 2020.

[6] J. Russell, A. Echenique, S. R. Daugherty, and M. Weinstock, "Chest x-ray findings among urgent care patients with covid-19 are not affected by patient age or gender: A retrospective cohort study of 636 ambulatory patients," *The Journal of Urgent Care Medicine*, pp. 13–18, 2020.

[7] 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.

[8] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya *et al.*, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 590–597.

[9] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2097–2106.

[10] S. Minaee, R. Kafieh, M. Sonka, S. Yazdani, and G. Jamalipour Soufi, "Deep-covid: Predicting covid-19 from chest x-ray images using deep transfer learning," *Medical Image Analysis*, vol. 65, p. 101794, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1361841520301584

[11] I. D. Apostolopoulos and T. A. Mpesiana, "Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks," *Physical and Engineering Sciences in Medicine*, vol. 43, no. 2, pp. 635–640, 2020. [Online]. Available: https://doi.org/10.1007/s13246-020-00865-4

[12] A. I. Khan, J. L. Shah, and M. M. Bhat, "Coronet: A deep neural network for detection and diagnosis of covid-19 from chest x-ray images," *Computer Methods and Programs in Biomedicine*, vol. 196, p. 105581, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0169260720314140

[13] T. Ozturk, M. Talo, E. A. Yildirim, U. B. Baloglu, O. Yildirim, and U. Rajendra Acharya, "Automated detection of covid-19 cases using deep neural networks with x-ray images," *Computers in Biology and Medicine*, vol. 121, p. 103792, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0010482520301621

[14] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 7263–7271.

[15] P. Afshar, S. Heidarian, F. Naderkhani, A. Oikonomou, K. N. Plataniotis, and A. Mohammadi, "Covid-caps: A capsule network-based framework for identification of covid-19 cases from x-ray images," *Pattern Recognition Letters*, vol. 138, pp. 638 – 643, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167865520303512

[16] L. Wang, Z. Q. Lin, and A. Wong, "Covid-net: a tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images," *Scientific Reports*, vol. 10, no. 1, p. 19549, 2020. [Online]. Available: https://doi.org/10.1038/s41598-020-76550-z

[17] M. Heidari, S. Mirniaharikandehei, A. Z. Khuzani, G. Danala, Y. Qiu, and B. Zheng, "Improving the performance of cnn to predict the likelihood of covid-19 using chest x-ray images with preprocessing algorithms," *International Journal of Medical Informatics*, vol. 144, p. 104284, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S138650562030959X

[18] S. Varela-Santos and P. Melin, "A new approach for classifying coronavirus covid-19 based on its manifestation on chest x-rays using texture features and neural networks," *Information Sciences*, vol. 545, pp. 403 – 414, 2021. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0020025520309531

[19] S. Tabik, A. Gómez-Ríos, J. L. Martín-Rodríguez, I. Sevillano-García, M. Rey-Area, D. Charte, E. Guirado, J. L. Suárez, J. Luengo, M. A. Valero-González, P. García-Villanova, E. Olmedo-Sánchez, and F. Herrera, "Covidgr dataset and covid-sdnet methodology for predicting covid-19 based on chest x-ray images," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 12, pp. 3595–3605, 2020.

[20] Y. Oh, S. Park, and J. C. Ye, "Deep learning covid-19 features on cxr using limited training data sets," *IEEE Transactions on Medical Imaging*, vol. 39, no. 8, pp. 2688–2700, 2020.

[21] J. Zhang, Y. Xie, G. Pang, Z. Liao, J. Verjans, W. Li, Z. Sun, J. He, Y. Li, C. Shen, and Y. Xia, "Viral pneumonia screening on chest x-rays using confidence-aware anomaly detection," *IEEE Transactions on Medical Imaging*, pp. 1–1, 2020.

[22] L. Luo, L. Yu, H. Chen, Q. Liu, X. Wang, J. Xu, and P. A. Heng, "Deep mining external imperfect data for chest x-ray disease screening," *IEEE Transactions on Medical Imaging*, vol. 39, no. 11, pp. 3583–3594, 2020.

[23] H. Sowrirajan, J. Yang, A. Y. Ng, and P. Rajpurkar, "Moco pretraining improves representation and transferability of chest x-ray models," 2020.

[24] Z. Wang, Y. Xiao, Y. Li, J. Zhang, F. Lu, M. Hou, and X. Liu, "Automatically discriminating and localizing covid-19 from community-acquired pneumonia on chest x-rays," *Pattern Recognition*, vol. 110, p. 107613, 2021. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320320304167

[25] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[26] O. Henaff, "Data-efficient image recognition with contrastive predictive coding," in *International Conference on Machine Learning*. PMLR, 2020, pp. 4182–4192.

[27] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," *arXiv preprint arXiv:2002.05709*, 2020.

[28] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.

[29] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar *et al.*, "Bootstrap your own latent: A new approach to self-supervised learning," *arXiv preprint arXiv:2006.07733*, 2020.

[30] I. Misra and L. v. d. Maaten, "Self-supervised learning of pretext-invariant representations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6707–6717.

[31] A. Newell and J. Deng, "How useful is self-supervised pretraining for visual tasks?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7345–7354.

[32] X. Liu, F. Zhang, Z. Hou, Z. Wang, L. Mian, J. Zhang, and J. Tang, "Self-supervised learning: Generative or contrastive," *arXiv preprint arXiv:2006.08218*, vol. 1, no. 2, 2020.

[33] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *European Conference on Computer Vision*. Springer, 2016, pp. 69–84.

[34] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," *arXiv preprint arXiv:1803.07728*, 2018.

[35] 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.
[36] A. Krizhevsky, V. Nair, and G. Hinton, "Cifar-10 (canadian institute for advanced research)." [Online]. Available: http://www.cs.toronto.edu/~kriz/cifar.html
[37] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.
[38] A. Van Opbroek, M. A. Ikram, M. W. Vernooij, and M. De Bruijne, "Transfer learning improves supervised image segmentation across imaging protocols," *IEEE transactions on medical imaging*, vol. 34, no. 5, pp. 1018–1030, 2014.
[39] 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.
[40] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3733–3742.
[41] K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 2016, pp. 1857–1865.
[42] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018.
[43] J. P. Cohen, P. Morrison, L. Dao, K. Roth, T. Q. Duong, and M. Ghassemi, "Covid-19 image data collection: Prospective predictions are the future," *arXiv 2006.11988*, 2020. [Online]. Available: https://github.com/ieee8023/covid-chestxray-dataset
[44] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[45] G. Maguolo and L. Nanni, "A critic evaluation of methods for covid-19 automatic detection from x-ray images," *arXiv preprint arXiv:2004.12823*, 2020.
[46] G. S. Collins and K. G. Moons, "Reporting of artificial intelligence prediction models," *The Lancet*, vol. 393, no. 10181, pp. 1577–1579, 2019.