

RELIABLE COVID-19 DETECTION USING CHEST X-RAY IMAGES

Aysen Degerli[†], Mete Ahishali[†], Serkan Kiranyaz^{*}, Muhammad E. H. Chowdhury^{*}, and Moncef Gabbouj[†]

[†] Faculty of Information Technology and Communication Sciences, Tampere University, Tampere, Finland

^{*} Department of Electrical Engineering, Qatar University, Doha, Qatar

ABSTRACT

Coronavirus disease 2019 (COVID-19) has emerged the need for computer-aided diagnosis with automatic, accurate, and fast algorithms. Recent studies have applied Machine Learning algorithms for COVID-19 diagnosis over chest X-ray (CXR) images. However, the data scarcity in these studies prevents a reliable evaluation with the potential of overfitting and limits the performance of deep networks. Moreover, these networks can discriminate COVID-19 pneumonia usually from healthy subjects only or occasionally, from limited pneumonia types. Thus, there is a need for a robust and accurate COVID-19 detector evaluated over a large CXR dataset. To address this need, in this study, we propose a reliable COVID-19 detection network: ReCovNet, which can discriminate COVID-19 pneumonia from 14 different thoracic diseases and healthy subjects. To accomplish this, we have compiled the largest COVID-19 CXR dataset: QaTa-COV19 with 124,616 images including 4603 COVID-19 samples. The proposed ReCovNet achieved a detection performance with 98.57% sensitivity and 99.77% specificity.

Index Terms— SARS-CoV-2, COVID-19 Detection, Machine Learning, Deep Learning

1. INTRODUCTION

Coronavirus disease 2019 (COVID-19), caused by severe acute respiratory syndrome Coronavirus-2 (SARs-CoV-2), was declared a pandemic by the World Health Organization in March 2020. The disease affects seriously people in high-risk groups (especially the elderly) leading to hospitalization, intubation, and even death [1]. In order to prevent the spread of the disease, detection, and isolation of infected patients have the utmost importance. However, the diagnosis of COVID-19 is challenging due to its similar symptoms with other viral infections such as fever, cough, fatigue, and breathlessness [2]. Therefore, reliable detection of the disease has significant importance.

Recent diagnostic tools to detect COVID-19 are nucleic acid detection with real-time polymerase chain reaction (RT-PCR), computed tomography (CT), and chest X-ray (CXR) imaging. RT-PCR has become the gold standard for COVID-19 diagnosis. However, RT-PCR tests suffer from instability and high false alarm rate [3]. On the other hand, CT imaging has higher sensitivity compared to RT-PCR test; thus, recommended for the suspected cases [4]. However, the performance of CT imaging in the early COVID-19 cases has limited sensitivity [4]. Thus, CXR imaging is widely used for the diagnosis of COVID-19 mainly because of its advantages that are faster acquisition, less radiation exposure, and easy accessibility compared to the aforementioned tools [5].

Many studies utilized Deep Learning (DL) algorithms for COVID-19 detection [?, 6, 7]. However, the reliability of these models is under question due to their hidden decision-making process.

In fact, the activation maps of the deep models reveal the unreliability of their decision-making process, where irrelevant areas on the CXRs, outside of the lung area such as bones, background, or text, affect the decision of the network. Therefore, several studies [8–10] attempted to prevent deep models to learn from these irrelevant areas on the CXRs with a two-staged approach for COVID-19 detection by processing only the lung areas with lung segmentation as their first stage. At the second stage, only the segmented lung area on the CXRs are given to the deep models as the input. Although these studies have achieved good performance for COVID-19 detection, data scarcity is the main drawback that can yield overfitting and hinders an accurate evaluation. Moreover, the datasets used in these studies encapsulate none or limited thoracic diseases, i.e., viral and bacterial pneumonia against COVID-19 pneumonia that makes them unreliable in real-case scenarios for COVID-19 diagnosis.

In this study, to address the aforementioned issues we propose ReCovNet: a **reliable COVID-19 detection network**, which is an end-to-end network solution. Instead of detecting COVID-19 directly from the CXR image or the segmented lung area on the CXR, we embed this information into the ReCovNet model by transfer learning from a segmentation network. For this purpose, we initially train the segmentation network and detach its encoder block to reconstruct the ReCovNet model for COVID-19 detection. Additionally, in this work, we extend the QaTa-COV19 dataset that was introduced in our previous study [11]. The extended version of QaTa-COV19 is the largest COVID-19 dataset with 124,616 images including 4603 COVID-19 samples. The control group CXRs consists of 14 different thoracic diseases and healthy subjects. Moreover, the QaTa-COV19 consists of a subset of 1065 early COVID-19 cases showing no or limited sign of COVID-19 pneumonia, which makes the diagnosis more challenging. Accordingly, the proposed ReCovNet trained over the largest QaTa-COV19 dataset has an outstanding performance with a reliable diagnosis compared to state-of-the-art deep models. Lastly, the benchmark QaTa-COV19 dataset is publicly shared with the research community¹.

The rest of the paper is organized as follows. In Section 2, we introduce the QaTa-COV19 dataset and give the details of our proposed ReCovNet model along with the state-of-the-art deep models. In Section 3, we report the experimental results, and we conclude the paper in Section 4.

2. MATERIALS AND METHODOLOGY

In this section, first we introduce the benchmark QaTa-COV19 dataset. Then, the state-of-the-art deep models are introduced for COVID-19 diagnosis. Lastly, we propose the ReCovNet model for reliable COVID-19 detection.

¹The benchmark QaTa-COV19 is publicly shared at the repository <https://www.kaggle.com/aysendegerli/qatacov19-dataset>.

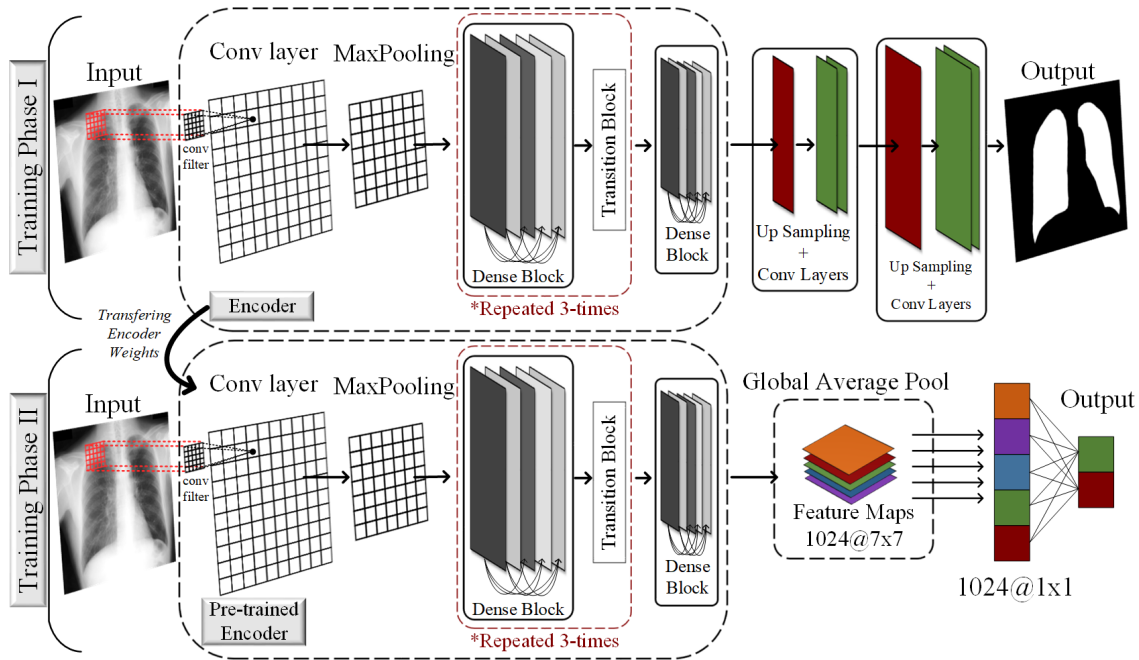


Fig. 1: The proposed ReCovNet, where the transfer learning is performed from the segmentation network initially trained on CXRs for lung segmentation.

2.1. The Benchmark QaTa-COV19 Dataset

The benchmark QaTa-COV19 dataset, compiled by researchers of Qatar University and Tampere University is so far the largest COVID-19 dataset including 4603 COVID-19 and 120,013 control group CXRs. The detection task on this dataset is especially challenging since QaTa-COV19 consists of 1065 samples from early COVID-19 cases that show no or limited sign of COVID-19 pneumonia. COVID-19 samples have been collected from publicly available datasets and repositories [9, 12–17], and were pre-processed by excluding low-quality images and any duplication. The control group images were collected from several datasets: ChestX-ray14 [18], X-rays from pediatric patients [19], and Chest X-rays (Indiana University) [20]. We have only used the bacterial and viral pneumonia CXRs from pediatric patients to increase pneumonia samples for a challenging diagnosis. Additionally, we included only the lateral-view CXRs from Chest X-rays (Indiana University) dataset since all other samples in the control group are from frontal-view CXRs, whereas COVID-19 samples include CXRs both from lateral and frontal views.

Table 1: Details of QaTa-COV19 dataset.

Data	Training Samples	Augmented	Augmented Training Samples	Test Samples
ChestX-ray14	86,524	✗	86,524	25,596
Bacterial Pneumonia	2130	✓	5000	630
Chest X-rays (Indiana University)	2816	✓	5000	832
Viral Pneumonia	1146	✓	5000	339
COVID-19	3553	✓	10,000	1050
Total	96,169		111,524	28,447

Table 1 shows the number of samples in the QaTa-COV19 dataset. COVID-19 detection is performed against the control group images, which consists of 14 different thoracic diseases and healthy subjects. Therefore, we perform a binary classification problem. Since the train and test sets of the ChestX-ray14 dataset are predefined, we have randomly split Chest X-rays (Indiana University), bacterial and viral pneumonia, and COVID-19 CXRs with the same train/ test ratio as in [18]. The CXRs in the dataset are resized to 224×224 pixels. We have augmented the images except for ChestX-ray14 samples using the Image Data Generator in Keras. The images are 10% randomly shifted both horizontally and vertically, and in a 10-degree range randomly rotated. Lastly, the 'nearest' mode is selected to fill the blank sections.

2.2. COVID-19 Detection with Deep Models

DL algorithms achieved state-of-the-art results on many computer vision tasks, including COVID-19 detection. Especially during the pandemic, recent studies concluded that DL algorithms with Convolutional Neural Networks can achieve outstanding performance for COVID-19 diagnosis. Nevertheless, the major issue in DL is that supervised deep models require a large amount of data to generalize well over unseen data. Thus, when subjected to data scarcity, such models fail in the testing phase due to overfitting. In this study, our first objective is to investigate the performances of state-of-the-art deep models by transfer learning on the largest COVID-19 dataset: QaTa-COV19. The state-of-the-art networks are selected as follows:

- **DenseNet-121** [21] is a 121-layer deep network that achieves a maximum information flow by connecting the layers with additional input nodes.
- **ResNet-50** [22] is a deep network with 50-layers that introduces residual blocks to prevent gradient vanishing in deep model structures by shortcut connections that merge input and

output through the stacked layers.

- **Inception-v3** [23] is a deep network with low computational complexity compared to other state-of-the-art deep models. The reduced complexity is ensured by pruning and factorizing operations inside the network.
- **Inception-ResNet-v2** [24] unites the structure of the inception model [23] with residual blocks [22] to achieve state-of-the-art results in computer vision tasks with a less computational cost.

In order to utilize the deep models in the COVID-19 detection task, we modify their output layers by inserting a global average pooling layer, a fully connected layer with 2-neurons, and a softmax activation function. The transfer learning is performed on the models by initializing their weights with the ImageNet weights.

2.3. ReCovNet: Reliable COVID-19 Detection Network

DL algorithms are often considered as black-box since their decision-making process is latent. In order to reveal their mystery in the decision-making process, the authors in [25] proposed Grad-CAM method that computes activation maps indicating the areas on the input image considered by the deep model during the classification task. In the COVID-19 detection task, our observations on the activation maps with the Grad-CAM approach show that the state-of-the-art deep models tend to learn and perform the classification from irrelevant areas on the CXRs, such as bones, background, or text. Therefore, the decisions of these models may be considered unreliable for COVID-19 detection. In order to overcome the unreliability issue, this study proposes ReCovNet: an end-to-end network for reliable COVID-19 detection.

ReCovNet is a deep network that considers the lung areas on the input CXR images to detect COVID-19 pneumonia. The structure of the proposed ReCovNet is given in Fig. 1. Accordingly, to construct ReCovNet, a segmentation network is trained in phase-I. The structure of the lung segmentation network is a convolutional autoencoder that maps the input image, \mathbf{X} to its corresponding output mask, \mathbf{M} : $\mathbf{M} \leftarrow P_{\theta, \phi}(\mathbf{X})$. Any deep model can be used as the encoder block of the network, ε_{θ} . On the other hand, the decoder block of the segmentation network is similar to the U-Net [26] model except for its u-shaped architecture, where the low-level features at the encoder block are concatenated with the high-level features at the decoder level. The u-shaped architecture is excluded by removing the skip connections, which performs the concatenation operation. The reason for constructing an encoder-decoder network without skip connections is that the contributions from the initial layers are avoided; therefore, the network can make decisions from the high-level features that are closer to segmentation mapping of the input image. Based on our observations, this approach improves the performance of ReCovNet in terms of reliability observed in the activation maps. The decoder block of the segmentation network consists of $\phi \in \{b_j, w_j\}_{j=1}^L$ with L number of layers composed of five stages. Each stage consists of an upsampling layer by $\times 2$, and sequentially two times of the convolutional layer, batch normalization, and Rectified Linear Unit (ReLU) activation function. The output of the last stage is connected to a convolutional layer with a sigmoid activation function to reconstruct the segmentation mask at the output. In order, the number of convolutional layer filters are $\{256, 128, 64, 32, 16, 1\}$ with kernel of size of $k = (3 \times 3)$. Lastly, training is performed over N number of samples $\{x_{s,train}^j, M^j\}_{j=1}^N$, where x_s and M are the training data, and ground-truth segmentation masks, respectively. The

loss function used in training is a hybrid function, which is the summation of the binary focal and dice loss functions.

During phase-II of the training, we construct the convolutional layers of ReCovNet by ε_{θ} that generates the latent features $f \leftarrow \varepsilon_{\theta}(\cdot)$. Then, f is vectorized and downsampled by attaching a global average pooling layer and a fully connected layer with 2-neurons using softmax activation function. We perform the classification task with categorical cross-entropy loss function by training ReCovNet over N number of samples $\{x_{train}^j, y_{train}^j\}_{j=1}^N$, where x and y are the training data and ground-truth labels, respectively. During this training phase, ε_{θ} is not frozen; therefore, the latent features f are further adjusted to the benchmark QaTa-COV19 dataset. Overall, during the inference, ReCovNet does not require prior lung segmentation to provide reliable COVID-19 detection. Finally, we propose two versions of the proposed model: ReCovNet-v1 is formed by DenseNet-121 encoder due to its good performance in the COVID-19 detection task, and ReCovNet-v2 is formed by ResNet-50 encoder.

3. EXPERIMENTAL EVALUATION

In this section, the experimental setup is presented. Then, the experimental results are given on the benchmark QaTa-COV19 dataset.

3.1. Experimental Setup

The performance metrics are calculated on the test (unseen) set of the QaTa-COV19 dataset. We consider COVID-19 CXRs as positive-class, whereas control group samples as negative-class. Accordingly, we form the confusion matrix (CM) elements as follows: true positive is the number of correctly classified COVID-19 samples, false positive is the number of misclassified control group samples as the positive class member, true negative is the number of correctly detected control group samples, and false negative is the number of misclassified COVID-19 samples as the negative class members. The performance metrics are defined as follows: *sensitivity* is the rate of correctly detected COVID-19 samples in all positive samples, *specificity* is the ratio of correctly classified control group samples in all negative samples, *precision* is the rate of correctly classified positive samples among all the members detected as positive class members, *accuracy* is the rate of correctly detected samples among all the data. Moreover, we define the *F-score* as follows:

$$F(\beta) = (1 + \beta^2) \frac{(\text{precision} \times \text{sensitivity})}{\beta^2 \times \text{precision} + \text{sensitivity}} \quad (1)$$

where the harmonic average between *precision* and *sensitivity* is defined as *F1-score* as $\beta = 1$. On the other hand, to minimize the effect of false negatives over false positives, *F2-Score* is defined as $\beta = 2$. The major performance metric in COVID-19 detection is *sensitivity* since any misdetection of the disease threatens global health. Hence minimizing the false alarm ($1 - \text{specificity}$) is our target.

The networks are implemented using Tensorflow library on NVidia @ GeForce RTX 2080 Ti GPU card. The optimizer choice is Adam with its default momentum parameters. ReCovNet models are trained with 15-epochs, the learning rate of $\alpha = 10^{-5}$, and a batch size of 64. The segmentation networks are trained with 15-epochs, the learning rate of $\alpha = 10^{-4}$, and a batch size of 32. We have utilized Montgomery County X-ray Set [27] and Japanese Society of Radiological Technology (JSRT) [28] datasets to train the segmentation models. All the images are frontal-view CXRs and have their corresponding ground-truths except for the JSRT. Thus, the segmentation masks provided by [29] are used as ground-truths

Table 2: COVID-19 detection performance results (%) computed over the test (unseen data) set of QaTa-COV19 dataset using four state-of-the-art and the proposed ReCovNet deep models.

Model	Sensitivity	Specificity	Precision	F1-Score	F2-Score	Accuracy
ResNet-50	96.571	99.953	98.734	97.641	96.996	99.828
Inception-v3	94.762	99.821	95.307	95.033	94.870	99.634
Inception-ResNet-v2	94.286	99.803	94.828	94.556	94.394	99.599
DenseNet-121	97.429	99.974	99.320	98.365	97.801	99.880
ReCovNet-v1	97.810	99.901	97.438	97.624	97.735	99.824
ReCovNet-v2	98.571	99.770	94.262	96.369	97.678	99.726

for JSRT [28]. Overall, the number of CXRs is 385 in the lung segmentation dataset. For the performance evaluation, we split this data with a ratio of 80% training to 20% test sets. Then, training samples are augmented up to 1000 samples.

3.2. Experimental Results

In this section, the performance of segmentation networks is first investigated. Over the test set of segmentation dataset, the segmentation model with DenseNet-121 encoder has achieved 96.12% sensitivity and 98.59% specificity, and with ResNet-50 encoder 97.12% sensitivity and 98.22% specificity for the lung segmentation task.

The COVID-19 detection performance results of the state-of-the-art and ReCovNet models are presented in Table 2. For each model, we have observed that their performance on COVID-19 detection is successful with > 94% sensitivity. The best model from the state-of-the-art deep models is DenseNet-121 with 97.43% sensitivity and 99.97% specificity. The performance of ReCovNet-v1 is very close to DenseNet-121. However, the best sensitivity in COVID-19 detection is achieved by the ReCovNet-v2 by 98.57%, which is an outstanding performance for the diagnosis on the largest COVID-19 dataset. Moreover, ReCovNet-v2 also holds a high specificity level of 99.77%. Table 3 shows the confusion matrices of the best performing models, which are DenseNet-121 from state-of-the-art deep models and ReCovNet-v2 from the proposed networks. The best detection (sensitivity) rate is achieved by ReCovNet-v2, which misses only 15 COVID-19 samples among 1050 images.

The results on the largest COVID-19 dataset, which includes many CXR images from different thoracic diseases, shows that deep models can achieve elegant COVID-19 detection performance. However, the activation maps extracted by Grad-CAM [25] approach reveal the contribution of the irrelevant regions and this is

Table 3: Confusion matrices of the best performing DenseNet-121 and proposed ReCovNet-v2 model for COVID-19 detection.

(a) Confusion Matrix DenseNet-121

DenseNet-121		Predicted	
		Control Group	COVID-19
Ground Truth	Control Group	27390	7
	COVID-19	27	1023

(b) Confusion Matrix ReCovNet-v2

ReCovNet-v2		Predicted	
		Control Group	COVID-19
Ground Truth	Control Group	27334	63
	COVID-19	15	1035

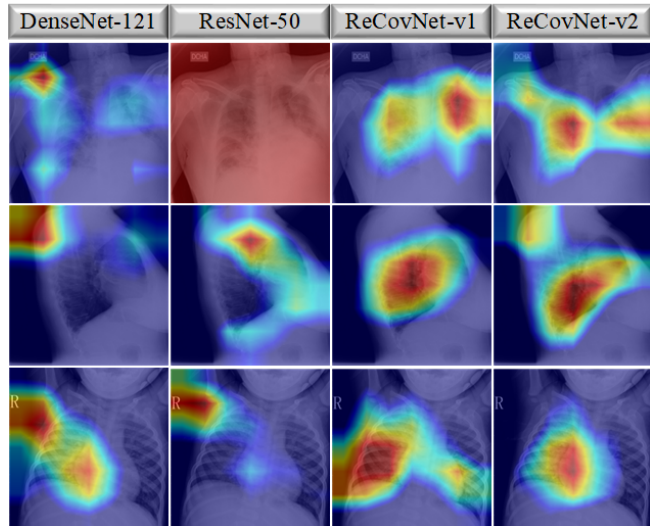


Fig. 2: The activation maps are extracted by Grad-CAM [25] approach for the models. The top two rows are COVID-19 samples, whereas the bottom row is a CXR from the control group images.

a major issue of these models in COVID-19 diagnosis. To exemplify this issue, we have compared the proposed ReCovNet-v1 and ReCovNet-v2 models with deep models as shown in Fig. 2. The activation maps show that DenseNet-121 and ResNet-50 models obviously get the information from irrelevant regions on the CXRs while the proposed models focus on the relevant regions.

4. CONCLUSIONS

The diagnosis of COVID-19 is a crucial task to prevent the further spread of the disease. This study investigates the limitations of the state-of-the-art deep models that are trained for COVID-19 detection directly from CXRs. To address these problems, we propose an end-to-end reliable COVID-19 detection network with pre-trained convolutional layers. We have compiled and publicly shared the largest COVID-19 dataset: QaTa-COV19, which includes 4603 COVID-19 samples, and 120,013 CXRs from 14 different thoracic diseases and normal samples. The experimental results over this benchmark dataset have shown that the proposed approach has achieved the highest sensitivity level compared to competing methods. We also demonstrated how the proposed models properly focus their analysis in the relevant region of the CXR instead of irrelevant activation observed in the competing models. In our future work, more CXR images will be used to train the lung segmentation models to further increase the reliability of our approach in COVID-19 detection.

5. REFERENCES

- [1] World Health Organization, “Coronavirus disease 2019 (covid-19): situation report, 88,” 2020.
- [2] T. Singhal, “A review of coronavirus disease-2019 (covid-19),” *The Indian Journal of Pediatrics*, vol. 87, no. 4, pp. 281–286, 2020.
- [3] A. Tahamtan and A. Ardebili, “Real-time rt-pcr in covid-19 detection: issues affecting the results,” *Expert Review of Molecular Diagnostics*, vol. 20, no. 5, 2020.
- [4] A. Bernheim, X. Mei, M. Huang, Y. Yang, Z. A. Fayad, N. Zhang, K. Diao, B. Lin, X. Zhu, K. Li, et al., “Chest ct findings in coronavirus disease-19 (covid-19): relationship to duration of infection,” *Radiology*, vol. 295, no. 3, pp. 200463, 2020.
- [5] D. J. Brenner and E. J. Hall, “Computed tomography—an increasing source of radiation exposure,” *New England Journal of Medicine*, vol. 357, no. 22, pp. 2277–2284, 2007.
- [6] T. D. Pham, “Classification of covid-19 chest x-rays with deep learning: new models or fine tuning?,” *Health Information Science and Systems*, vol. 9, no. 1, pp. 1–11, 2020.
- [7] M. E. H. Chowdhury, T. Rahman, A. Khandakar, R. Mazhar, M. A. Kadir, Z. B. Mahbub, K. R. Islam, M. S. Khan, A. Iqbal, N. A. Emadi, M. B. I. Reaz, and M. T. Islam, “Can ai help in screening viral and covid-19 pneumonia?,” *IEEE Access*, vol. 8, pp. 132665–132676, 2020.
- [8] M. Z. Alom, M. Rahman, M. S. Nasrin, T. M. Taha, and V. K. Asari, “Covid_mtnet: Covid-19 detection with multi-task deep learning approaches,” *arXiv preprint arXiv:2004.03747*, 2020.
- [9] A. Haghani, M. M. Majdabadi, and S. Ko, “Covid-cxnet: Detecting covid-19 in frontal chest x-ray images using deep learning,” *arXiv preprint arXiv:2006.13807*, 2020.
- [10] E. Goldstein, D. Keidar, D. Yaron, Y. Shachar, A. Blass, L. Charbinsky, I. Aharony, L. Lifshitz, D. Lumelsky, Z. Nee-man, et al., “Covid-19 classification of x-ray images using deep neural networks,” *arXiv preprint arXiv:2010.01362*, 2020.
- [11] A. Degerli, M. Ahishali, M. Yamac, S. Kiranyaz, M. E. H. Chowdhury, K. Hameed, T. Hamid, R. Mazhar, and M. Gabbouj, “Covid-19 infection map generation and detection from chest x-ray images,” *arXiv preprint arXiv:2009.12698*, 2020.
- [12] M. I. Vayá, J. M. Saborit, J. A. Montell, A. Pertusa, A. Bustos, M. Cazorla, J. Galant, X. Barber, D. Orozco-Beltrán, F. Garcia, et al., “Bimcv covid-19+: a large annotated dataset of rx and ct images from covid-19 patients,” *arXiv preprint arXiv:2006.01174*, 2020.
- [13] “COVID-19 Image Repository,” 2020, <https://github.com/ml-workgroup/covid-19-image-repository>. [Accessed on 12-January-2021].
- [14] “COVID-19 DATABASE,” 2020, <https://www.sirm.org/category/senza-categoria/covid-19/>. [Accessed on 12-January-2021].
- [15] J. P. Cohen, P. Morrison, and L. Dao, “Covid-19 image data collection,” *arXiv preprint arXiv:2003.11597*, 2020.
- [16] “COVID-19 Radiography Database,” 2020, <https://www.kaggle.com/tawsifurrahman/covid19-radiography-database>. [Accessed on 12-January-2021].
- [17] “Chest Imaging,” 2020, <https://www.eurorad.org/>. [Accessed on 12-January-2021].
- [18] 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.
- [19] 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.
- [20] “Chest X-rays (Indiana University),” 2020, https://www.kaggle.com/raddar/chest-xrays-indiana-university?select=indiana_reports.csv. [Accessed on 12-January-2021].
- [21] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2261–2269.
- [22] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
- [23] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826.
- [24] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, “Inception-v4, inception-resnet and the impact of residual connections on learning,” in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 2017, p. 4278–4284.
- [25] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 618–626.
- [26] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical Image Computing and Computer-assisted Intervention*. Springer, 2015, pp. 234–241.
- [27] S. Jaeger, S. Candemir, S. Antani, Y.-X. J. Wang, P.-X. Lu, and G. Thoma, “Two public chest x-ray datasets for computer-aided screening of pulmonary diseases,” *Quantitative Imaging in Medicine and Surgery*, vol. 4, no. 6, pp. 475, 2014.
- [28] J. Shiraishi, S. Katsuragawa, J. Ikezoe, T. Matsumoto, T. Kobayashi, K.-i. Komatsu, M. Matsui, H. Fujita, Y. Kodera, and K. Doi, “Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists’ detection of pulmonary nodules,” *American Journal of Roentgenology*, vol. 174, no. 1, pp. 71–74, 2000.
- [29] B. V. Ginneken, M. B. Stegmann, and M. Loog, “Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database,” *Medical Image Analysis*, vol. 10, no. 1, pp. 19–40, 2006.