Pneumonia Detection and Classification from X-ray Images – a Deep Learning Approach

Jelena Bozickovic, Ivan Lazic, *Member, IEEE,* and Tatjana Loncar Turukalo, *Member, IEEE*

***Abstract*— Chest X-rays are one of the first medical imaging tools used for correctly assessing different causes of pneumonia. With the recent spread of the SARS-CoV-2 virus, fast diagnostics and differentiation of the COVID-19 disease from other causes (bacterial or viral) is important. In this study we evaluated four different neural network architectures and applied transfer learning in order to try to detect and classify pneumonia in patient images in a 4-class problem (normal, viral, bacterial and COVID-19). We applied data augmentation on the outnumbered COVID-19 class and compared the effects of single end-to-end network training to a two-stage variant. The best results were obtained using the ResNet50 model with an average cross-validation accuracy of 89.97%. Across all models the COVID-19 and normal X-ray images showed very high precision and sensitivity scores.**

***Index Terms*—Chest X-ray, Deep Learning, Convolutional Neural Networks, COVID-19, SARS-CoV-2, Pneumonia**

I. Introduction

The increase and spread of registered cases of COVID-19 has accentuated the importance and accelerated development of efficient and reliable methods for diagnosis of this viral disease [1]. The fast commonly used diagnostic technique is real-time reverse transcription-polymerase chain reaction (RT-PCR) [2]. However, low sensitivity of RT-PCR test (60%–70%), and the deficit of the tests in developing countries emphasize the role of chest radiology (computed tomography and X-ray) and blood analysis in diagnostics and timely treatment [3]. The main features of pneumonia caused by SARS-CoV-2 on the X-ray images are peripheral and lower lobe predominant rounded airspace opacities and multifocal rounded opacities and nodules [4] with multifocal non-peripheral airspace opacities being less pronounced as a feature.

The urgency created by the COVID-19 pandemic requires fast diagnostics and differentiation of the COVID-19 caused pneumonia cases from other such as bacterial or viral). Early detection of pneumonia development is another goal, as it largely influences the flow and consequences of the disease. Even before the COVID-19 cases, pneumonia was one of the leading causes of death among children under 5 years old and lower respiratory tract infections (LRTI) responsible for 2.8 million deaths annually [5]. As one of the first examinations when COVID-19 or LRTI are suspected, both X-ray or CT scans provide useful resources for a machine learning approach to pneumonia detection and classification. The abundance of new cases facilitate data gathering purposes and support further improvements towards development of an automatic diagnostic support model.

Deep learning techniques have already been applied to the problem of pneumonia detection in X-ray images [6] showing that an accurate deep learning model can assist in diagnosis, especially when medical expertise or experience are insufficient. COVID-19 has been an incentive to speed up the progress in this area. The literature review from 2020 exhibits the performance of many well-known pre-trained architectures in pneumonia detection and classification tasks. ResNet50, InceptionV3 and InceptionResNetV2 pre-trained models with ImageNet [7] data were used in [8] with an average accuracy of 98%, 97% and 87% for the three models respectively for a binary classification problem between normal and COVID-19 cases with only 50 samples per class. In [9] the pretrained 121-DenseNet model was used, as in [6], achieving an accuracy of 87.2% for a 4-class classification problem between normal chest X-ray scans and 3 different pneumonia cases caused by bacteria, viruses or COVID-19. However, the dataset used in [9] was heavily imbalanced with 155 sample images of the COVID-19 class whereas the images for other classes were taken from [10]. Regarding the same 4-class classification problem with a balanced dataset of around 300 samples per class, authors in [11] used the Xception model pre-trained on the ImageNet data, where they achieved an accuracy of 89.5%. In these works, a small set of COVID-19 image data was available, both [8] and [11] focused mainly on end-to-end training of the used pre-trained models without resorting to data augmentation techniques.

In this work we evaluate several ImageNet pretrained models on the 4-class classification task using an expanded chest X-ray dataset from healthy patients and patients with bacterial, viral and COVID-19 induced pneumonia. Additionally, we compare the results obtained from end-to-end training of the models and a two-stage training process associated with transfer learning and fine-tuning.

Jelena Božičković – Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, 21102, Novi Sad, Serbia (e-mail: jelena.bozickovic@live.com).

Ivan Lazić – Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, 21102, Novi Sad, Serbia (e-mail: ivan.lazic@uns.ac.rs).

Tatjana Lončar Turukalo - Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, 21102, Novi Sad, Serbia (e-mail: turukalo@uns.ac.rs).

II. Materials and Methods

A. Database

In this study two publicly available databases were used. Images representing classes of normal X-rays, viral X-rays and bacterial X-rays were collected from the Chest X-Ray Images (Pneumonia) database [10], while the COVID-19 images class were collected from the COVID-19 Radiography Database [12].

The databases contain a total of 5669 samples, of which: 1575 samples from the class of normal X-rays, 2530 samples from the class of bacterial images and 1345 from the class of viral images. The COVID-19 class, however, only had 219 samples which were taken from 137 patients. The initial database used in this research was designed so that 15% of the patients with COVID-19 were moved to the validation and test set with only 1 unique image per patient. The same number of X-ray images are taken from other classes for the validation and test set as well. The rest of the image samples were used as the training set. The number of samples by class is shown in Table I. As the number of samples of the COVID-19 disease class was significantly smaller than the number of samples in other classes, the COVID-19 samples were augmented using random rotation by 15º, zooming in the range from 0.8 to 1.2 pixels, image rotation around the vertical axis and translation by up to 0.1 fraction of total width and height of the image. This resulted in a database with approximately balanced classes. All of the input images were scaled to 224x224 pixels. A sample image for each class is presented in Fig. 1.

B. Baseline deep learning models

Deep learning is a powerful framework for supervised learning which benefits from adding more layers and more units to achieve excellent performance in modeling complex functions, given sufficiently large labeled dataset [13]. Large datasets facilitate the use of larger models, offer better generalization, with the burden placed on the training process in terms of time and computational power, which is balanced by advances in hardware, software and parallelization [14].
For smaller data sets, overfitting can be prevented using pre-trained network models [15], which are usually trained on very large datasets, such as the ImageNet database, and used for feature extraction.

In this study, 4 different pretrained models were evaluated: ResNet50 [16], InceptionV3 [17], InceptionResNetV2 [18] and Xception [19].

TABLE I
NUMBER OF SAMPLES PER CLASS FOR THE INITIAL DATABASE

Set / Class	Training	Test	Valid.
Bacterial	1308	21	21
COVID19 augmented	1304	21	21
Normal	1310	21	21
Viral	1290	21	21

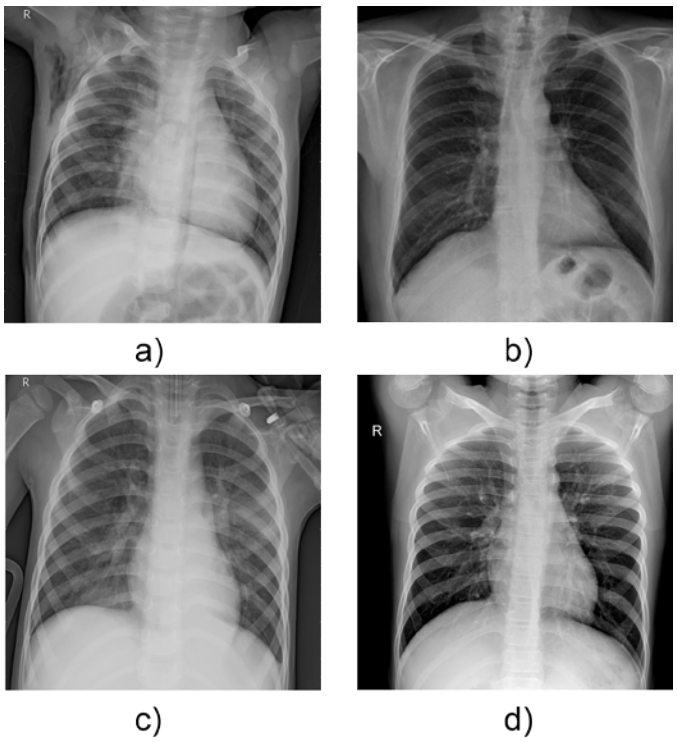

Fig. 1. Sample input from a) bacterial, b) COVID-19, c) viral and d) normal X-ray images.

The ResNet50 [16] model is based on a residual training mode to simplify the learning of deep neural networks. The network architecture involves reformulating the layers so that they learn the residual functions depending on the input layers. The depth of the residual network is 8 times larger than the VGG [20] network, but it is less complex.

The InceptionV3 [17] model allows for an expansion of depth and width of deep neural networks in a way that does not require more computing power. The model generates features on several levels using 1x1, 3x3 and 5x5 convolution filters.

InceptionResNetV2 [18] is a model that combines Inception models and residual models. It has been shown that training with residual connections significantly speeds up training compared to the Inception model itself. It has also been proven that the combination of these two models gives better results compared to the individual models.

The Xception [19] model represents such an architecture of a convolutional neural network in which the convolutional layers are completely separated. Specifically, the hypothesis behind the Xception model architecture is as follows: mapping correlations between channels and spatial correlations in feature maps can be completely separated. Network architecture consists of linearly arranged separable convolutional layers with residual connections.

All of the models show exceptional results on the ImageNet dataset classification problem, making them powerful feature extractors and classifiers. Using the stored model weights as the learned knowledge, the networks can be applied on a new set of data using transfer learning by detaching the original classification layers and training only the specific classification layers needed for the required 4-class problem.

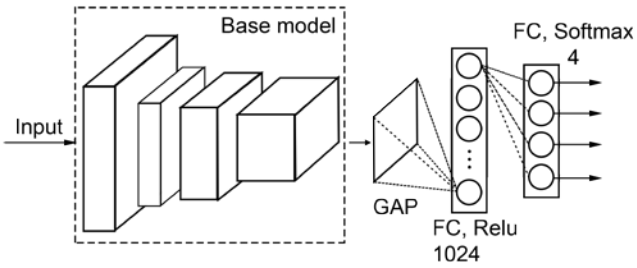

Fig. 2. Deep network architecture used for the experiments.

The models' performance can then be additionally improved by fine-tuning the original baseline model weights to the input data, as the original set of classes didn't include X-ray images. This allows the network to adjust its original weights to the new data, thus enabling it to better suit the new classification problem. This is usually done with a lower learning rate to prevent the model from completely forgetting the valuable knowledge initially learned on the large dataset.

C. Network architecture and model training

The network architecture used in this work is presented in Fig 2. After one of the mentioned base models, a classification layer is added which consists of a global average pooling (GAP) and two fully connected (FC) layers. The models were constructed using the Tensorflow 2.0 library.

Three different strategies were used to train the classifier with each baseline model:

- the first approach involved unlocking all layers of the pre-trained model.
- the second involved a two-step training procedure: in the first step, the layers of the base model were frozen, while in the second step, these layers were fully trainable.
- the third additionally introduced a dropout layer before the classification layers and performed two-step training as in the second experiment.

All models were trained using all three strategies. In the first experiment for all baseline models, the parameters for the Adam [21] optimizer, the learning speed and the epsilon parameter were set to 10^{-5} and 0.1, determined empirically. For the two-step training methods, the initial training of classification layers was done with a default learning rate of 0.001, while the second training was done with a learning rate of 10^{-5} and epsilon of 0.1. The size of the batch was 32. In the training process the early stopping method was used, which monitored the validation loss and stopped the training if the validation loss didn't improve for 15 epochs. With the obtained optimal epoch number, new models and datasets were constructed from the original images in a cross-validation fashion. The number of folds was 5 and the split between the new training and test sets was 80%-20% with data augmentation being applied to the COVID-19 class in a similar way as with the initial training. In this case, there wasn't a need for a validation set as model parameters were already determined in advance.

TABLE II
AVERAGE MODEL EVALUATION METRICS

		Experiment 1				Experiment 2				Experiment 3			
Model	Class	Accuracy [%]	Precision [%]	Sensitivity [%]	F1	Accuracy [%]	Precision [%]	Sensitivity [%]	F1	Accuracy [%]	Precision [%]	Sensitivity [%]	F1
ResNet50	0	87.75	71.53	83.49	0.77	87.41	74.37	82.62	0.78	89.97	80.26	84.90	0.82
	1		98.52	97.26	0.98		97.04	98.62	0.98		97.78	99.31	0.99
	2		96.32	97.22	0.97		96.38	93.07	0.95		99.26	95.19	0.97
	3		84.63	75.31	0.79		81.85	77.12	0.79		82.57	81.42	0.82
InceptionV3	0	87.96	75.21	81.92	0.78	86.07	72.46	82.37	0.77	87.43	76.64	83.77	0.80
	1		96.30	97.88	0.97		96.32	97.83	0.97		96.32	98.60	0.97
	2		96.32	93.17	0.95		94.89	91.98	0.93		95.61	91.81	0.94
	3		84.02	79.49	0.81		80.32	74.21	0.77		81.14	77.15	0.79
Inception ResNetV2	0	87.41	73.02	85.82	0.79	85.40	73.76	77.52	0.76	87.96	74.42	85.49	0.79
	1		95.61	97.83	0.97		95.58	98.51	0.97		95.58	99.31	0.97
	2		95.58	93.74	0.95		94.07	92.91	0.93		94.84	94.23	0.94
	3		85.42	74.59	0.80		78.17	73.84	0.76		87.01	76.32	0.81
Xception	0	89.48	78.23	87.77	0.82	87.86	76.80	86.42	0.81	87.51	73.76	87.13	0.80
	1		97.78	97.11	0.97		94.84	98.60	0.97		96.30	98.62	0.97
	2		95.66	93.91	0.95		97.12	91.17	0.94		98.52	89.96	0.94
	3		86.27	80.80	0.83		82.67	78.03	0.80		81.77	76.19	0.79

A new model is then trained for each variant of the baseline network and for each of the three strategies and for each testing fold of the cross-validation.

III. Results

Table II presents the average evaluation results for different architectures and training strategies over different folds using accuracy, precision, sensitivity and F1 measure. The classes numbers from 0 to 3, correspond to bacterial, COVID-19, normal and viral class respectively. In the case of the ResNet50 and InceptionResNetV2 models, the performance of the models indicated the improved performance when the two-step training strategy was used. The pre-trained ResNet50 model has achieved the overall best results with an accuracy of 89,97%, whereas the InceptionResNetV2 had an accuracy of 87,96%. In most cases, adding the dropout layer before the classifier improved the two-step training process. The Xception and InceptionV3 models achieved their best score of 89,48% and 87,96% using the first training strategy. Overall, the best precision and sensitivity is obtained on the COVID-19 class, with the normal class following it. The most difficult task proved to be distinguishing between the bacterial and viral classes.

IV. Conclusion

From an engineering point of view, the research showed that it is possible to use pre-trained architectures in order to detect and classify different types of pneumonia, including COVID-19. In general, the two-stage training strategy provided better results in almost all of the baseline models. The analysis of confusion matrices indicated COVID-19 X-ray scans can successfully be differentiated from images of viral pneumonia. As the network is trained and tested on a limited input of COVID-19 cases, improvements are expected by including more patient images to the current database. Additionally, further testing can be done on other non-explored pretrained deep neural models such as the EfficientNet series, NASNet or DenseNet models, as well as trying out simpler models.

Acknowledgment

This research has been supported by the Ministry of Education, Science and Technological Development through the project no. 451-03-68/2020-14/200156: “Innovative scientific and artistic research from the FTS domain” and COST Actions CA15120 and CA19136.

References

[1] “Coronavirus Disease (COVID-19) – events as they happen”, www.who.int, 2020. [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/events-as-they-happen

[2] Z.Y. Zu, M.D. Jiang, P.P. Xu, W. Chen, Q.Q. Ni, G.M. Lu, L.J. Zhang, “Coronavirus Disease 2019 (COVID-19): A Perspective from China,” *Radiology*, vol. 296, no. 2, pp. E15–E25, Feb. 2020, doi: 10.1148/radiol.2020200490.

[3] J. P. Kanne, B. P. Little, J. H. Chung, B. M. Elicker, and L. H. Ketai, “Essentials for Radiologists on COVID-19: An Update—Radiology Scientific Expert Panel,” *Radiology*, vol. 296, no. 2, pp. E113–E114, Feb. 2020, doi: 10.1148/radiol.2020200527.

[4] UCLA Radiology, “COVID-19 Chest X-Ray Guideline”, Los Angeles, Westwood, Manhattan Beach, Santa Monica, CA. Available: https://www.uclahealth.org/radiology/covid-19-chest-x-ray-guideline.

[5] GBD 2015 Mortality and Causes of Death Collaborators, “Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: a systematic analysis for the Global Burden of Disease Study 2015”, Lancet 2016; 388: 1459–1544

[6] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. P. Lungren, A. Y. Ng, “CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning,” *arXiv:1711.05225 [cs, stat]*, Dec. 2017, [Online]. Available: http://arxiv.org/abs/1711.05225.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2009, pp. 248–255, doi: 10.1109/CVPR.2009.5206848.

[8] A. Narin, C. Kaya, and Z. Pamuk, “Automatic Detection of Coronavirus Disease (COVID-19) Using X-ray Images and Deep Convolutional Neural Networks,” *arXiv:2003.10849 [cs, eess]*, Jul. 2020, [Online]. Available: http://arxiv.org/abs/2003.10849.

[9] A. Mangal, S. Kalia, H Rajgopal, K. Rangarajan, V. Namboodiri, S. Banerjee, C. Arora, “CovidAID: COVID-19 Detection Using Chest X-Ray,” *arXiv:2004.09803 [cs, eess]*, Apr. 2020, [Online]. Available: http://arxiv.org/abs/2004.09803.

[10] D. S. Kermany, K. Zhang, and M. H. Goldbaum, “Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification,” 2018, doi: 10.17632/RSCBJBR9SJ.2.

[11] 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, Nov. 2020, doi: 10.1016/j.cmpb.2020.105581.

[12] M. E. H. Chowdhury, T. Rahman, A. Khandakar, R. Mazhar, M. Abdul Kadir, Z. Bin Mahbub, K. R. Islam, M. S. Khan, A. Iqbal, N. Al-Emadi, M. Bin I. Reaz, T. I. Islam, “Can AI help in screening Viral and COVID-19 pneumonia?,” *arXiv:2003.13145 [cs]*, Jun. 2020, [Online]. Available: http://arxiv.org/abs/2003.13145.

[13] I. Goodfellow, Y. Bengio, and A. Courville, “Deep learning”, MIT Press, Nov. 2016.

[14] Y. LeCun, Y. Bengio and G. Hinton, "Deep learning", *Nature*, vol. 521, no. 7553, pp. 436-444, 2015. Available: 10.1038/nature14539

[15] Y. Bengio, A. Courville, and P. Vincent, “Representation Learning: A Review and New Perspectives,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013, doi: 10.1109/TPAMI.2013.50.

[16] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” *arXiv:1512.03385 [cs]*, Dec. 2015, [Online]. Available: http://arxiv.org/abs/1512.03385.

[17] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the Inception Architecture for Computer Vision,” *arXiv:1512.00567 [cs]*, Dec. 2015, [Online]. Available: http://arxiv.org/abs/1512.00567.

[18] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, “Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning,” *arXiv:1602.07261 [cs]*, Aug. 2016, Available: http://arxiv.org/abs/1602.07261.

[19] F. Chollet, “Xception: Deep Learning with Depthwise Separable Convolutions,” *arXiv:1610.02357 [cs]*, Apr. 2017, Available: http://arxiv.org/abs/1610.02357.

[20] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” *arXiv:1409.1556 [cs]*, Apr. 2015, [Online]. Available: http://arxiv.org/abs/1409.1556.

[21] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” *arXiv:1412.6980 [cs]*, Jan. 2017, Available: http://arxiv.org/abs/1412.6980.