DOI: 10.48011/asba.v2i1.991

On Transfer Learning for Classifying COVID-19 in Chest X-Rays Images

Elilson Santos * **Lúcio Flavio de Jesus Silva** *
Omar Andres Carmona Cortes **

* *Programa de Pós-Graduação em Engenharia da Computação e Sistemas – Universidade Estadual do Maranhão (UEMA), São Luis, MA, Brazil (e-mail: elilson.java@gmail.com , lucioslv@hotmail.com.br).*
** *Departamento de Computação – Instituto Federal de Educação, Ciência e Tecnologia do Maranhão (IFMA), São Luis, MA, Brazil (e-mail: omar@ifma.edu.br)*

Abstract: COVID-19 is an infectious disease exceptionally caused by the severe acute respiratory syndrome. The disease has spread worldwide quickly and can lead to death in just a few days. In this context, investigating rapid forms of detection that assist doctors in the decision-making process is essential to saving lives. This paper investigates fourteen Convolutional Neural Network architectures using Transfer Learning. We used a database composed of 2,902 radiographs divided into three classes: Normal, COVID-19, and Viral Pneumonia. The results showed that DenseNet201 presented the best results regarding the classification reaching an average Recall of 92.1% and an F1_Score of 95.1%.

Keywords: Deep Learning, Transfer Learning, Classification, COVID-19.

1. INTRODUCTION

Coronavirus disease 2019 (COVID-19) is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (Tsang et al., 2020). The illness has led to millions of infected and thousands of deaths around the world. The World Health Organization (WHO) released the Situation Report (World Health Organization, 2020) informing 18,142,718 cases and 691,013 deaths worldwide in August-04th-2020. Almost half of the patients are only in the Americas, counting 9,741,727 cases and 365,334 deceased.

Additionally to the high degree of contagiousness, the disease evolves fast. According to Wilson et al. (2020), the mortality in terms of days is a median of 13 days passed from pneumonia confirmation to death. Therefore, it is essential to provide tools that help diagnose the disease as fast as possible, since early detection can give proper treatment time.

Thus, this work investigates using deep neural networks to help the diagnostic of COVID-19 by analyzing lung x-ray images. The proposal fits a new field called Medicine 4.0 (Wolf and Scholze, 2017), in which one of the focus application includes a combination of innovative artificial intelligence technologies, including deep learning algorithms, to develop Clinical Decision Support Systems.

The decision support system relates to procedures for improving clinical decisions by providing evidence-based medical information at the time of the doctor-patient contact or the treatment decision (Schnurr et al., 2018).

Because COVID can lead the patient to death in a short amount of time, as previously mentioned, and aiming slowing down the contagiousness, it is essential to investigate the manners of improving and fastening the diagnosis task. Also, attending the new trends of Medicine 4.0, in this work, we investigate the performance of fourteen convolutional neural networks (CNNs), also known as deep neural networks.

Usually, deep neural networks, especially the CNNs that deal with images, are devised by many layers and connections, making them computationally heavy to training from scratch. Thus, a solution called transfer learning can be applied to overcome this issue. The idea is to use previous knowledge to solve related problems, as humans use to do. Thus, transfer learning is a method of reusing a model or knowledge for another related task (Sarkar et al., 2018).

Regarding CNNs, some pre-trained models can be used to leverage image-based tasks. In this context, we investigate fourteen architectures for detecting COVID-19 as follows: DenseNet169, DenseNet201, Resnet50, Mobilenet, VGG16, Mobilenet_v2, DenseNet121, Nasnet_Large, VGG19, Xception, Inception_Resnet_v2, Inception_v3, Nasnet_Mobile, and Resnet50_V2. Moreover, we use a database composed of 2,902 x-ray images, which is the most common way of detecting lung diseases, consequently being useful to detect COVID-19 infection.

The common features observed in the X-Ray images of patients with COVID-19 are patchy infiltrates or opacities that bear similarities to other viral pneumonia features (Horry et al., 2020); thus, CNNs seem to be fitable to the task, since its primary purpose is extracting features from images.

For this sake, this paper is divided as follows: Section 2 presents related works; Section 3 illustrates how CNN and transfer learning work; Section 4 describes the dataset, metrics and shows the results of the experiment; finally, Section 6 presents the conclusions of this work.

2. RELATED WORK

Regular neural networks, such as Multi-Layer Perceptron networks, are usually trained from scratch, relying only on guidance training data. Then, in 1991, Pratt et al. (1991) suggested that the training of a neural network could be "recycled" in order to speed up the learning process, calling this approach of transfer learning.

In 1998, (LeCun et al., 1998) proposed an extension of Neural Networks, called CNN, that is devised by several layers aiming image classification; thus, it is a natural evolution that this new architecture also uses the transfer learning technique. Particularly, these kinds of neural networks have the innate ability to detect patterns into images, making them attractive for biomedical image processing.

Since then, many applications helping to diagnose many diseases have been proposed. For example, Gulshan et al. (2016) tries to detect diabetic retinopathy in retinal photographs. Ehteshami Bejnordi et al. (2017)'s work helps to detect metastasis in lymph nodes. Ismail and Sovuthy (2019) uses two CNNs, called VGG16 and ResNet50, to identify breast cancer in IRMA's x-ray dataset. Šarić et al. (2019) investigates the classification of lung cancer in histopathological images, also using VGG16 and ResNet50.

Recently, due to the 2020's COVID pandemic, efforts have been driven to help the diagnose of COVID-19 because it is a fast illness; therefore, the faster the diagnostic, the better the patient's chances. Wang et al. (2020) investigates the use of three CNN architectures named ResNet50, ResNet101, and ResNet152 in an x-ray images database.

Waheed et al. (2020) proposes a model for augmenting an x-ray dataset to improve the classification of the VGG16 CNN. Horry et al. (2020) studies seven CNN models for classifying COVID in three different datasets.

In our work, we investigate the performance of fourteen CNNs called: DenseNet169, DenseNet201, Resnet50, Mobilenet, VGG16, Mobilenet_V2, DenseNet121, Nasnet_Large, VGG19, Xception, Inception_Resnet_V2, Inception_V3, Nasnet_Mobile, and Resnet50_V2. These fourteen CNNs were already made available in the Keras package directly; therefore, it is unnecessary to go after these CNNs.

3. CNN AND TRANSFER LEARNING

A CNN architecture is typically devised by three components: convolutional layer, polling layer, and a fully connected one. The convolutional layers use filters that scan over a portion of the image and extract features from it. These features are usually colors, shapes, and edges that ultimately define a specific image (Beysolow II, 2017). CNNs can have as many convolutional layers as necessary. The more convolutional layers, the more features are extracted.

After convolutional layers, we add pooling layers that are responsible for pooling the feature maps into an image (Beysolow II, 2017), providing a dimensionality reduction. The reduction is obtained by applying a single maximum or average of the values inside a box produced by the convolutional layer. Thus, the fully connected layers, a regular MLP network, receive a smaller image than that one presented to the CNN input. Figure 1 represents a sketch of a CNN.

As several layers devise a CNN, we can quickly found pretrained ones to "recycle" the previous knowledge. Usually, those CNNs were trained by an accessible image dataset called ImageNet (imagenet, 2020) (Deng et al., 2009), which is composed of 14,197,122 images; however, other datasets can also have been used depending on the application. Thus, the idea of transfer learning is to update the CNN parameters using a new set of images. In this context, Torrey and Shavlik (2009) defines transfer learning as the improvement of learning in a new task through the transfer of knowledge from a related task that has already been learned.

All in all, transfer learning can present three main benefits (Silva and Cortes, 2020):

- We can use models that were carefully designed by experts;
- Because experts created those models, we do not need to worry about what architecture or layers to use or include;
- Due to their careful design, they tend to perform well in image detection.

Next, we show the experiments we carried out in this paper.

4. EXPERIMENTS

4.1 Environment and Setup

We implemented the application in Python Python (2020) 3.0 and the Neural Networks using Keras (Keras, 2020). The code and the training step have been done in Google Colaboratory Google (2020), also known as Google Colab. This system was essential to this work because we can use GPU computing for training the ANN.

The virtual machine is a two CPUs Intel® Xeon 2.30 GHz, 14 GB of RAM, and 37.11 GB of HD. Even though we used GPUs on the training step, we could not choose what type of GPU we could connect. Usually, Colab includes Nvidia K80s, T4s, and P4s. Using GPUs, the training step of each fold takes about 40 minutes. Each CNN was trained using ten epochs and k-fold with k=5; therefore, the training and test proportion were 80/20. The parameter optimization was done using Adam algorithm Kingma and Ba (2014) with learning rate = 0.001, $\beta_1 = 0.9$, and $\beta_1 = 0.999$

4.2 Metrics

Firstly, it is necessary to define the meaning of true positives, true negatives, false positives, and false negatives.

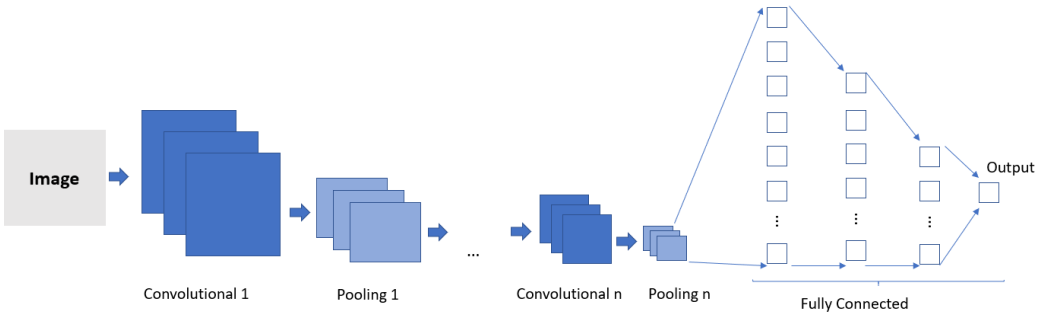


Fig. 1. Representation of a CNN

True positives and true negatives are correct classifications, *i.e.*, normal, pneumonia, and COVID-19 classified correctly. In contrast, false positive and false negative are wrong classifications. Having said that, we can define our metrics.

Equation 1 present the first metric called precision. This metric is the ratio between true positives and true positive plus false positives. Thus, a low precision indicates that the number of correct classifications is too low, or the number of false positives is too high.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

The next metric is the accuracy, as presented in Equation 2, which is the percentage of correct classifications. A low accuracy could indicate that the number of wrong classifications (false positives and false negatives) is high.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{2}$$

The recall presented in Equation 3 is the ratio between the true positives and true positives plus false positives. This metric indicates that the algorithm is performing well in classifying true positives. However, if this metric is low, it can mean that a high number of misclassification is going on. Thus, this metric is essential to minimize the number of false negatives, which produce the patient's worst scenario.

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

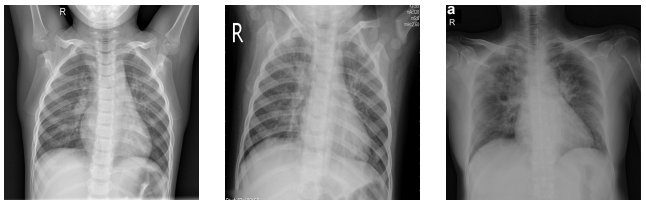
Finally, the F1 Score presented in Equation 4 is the harmonic mean between precision and recall. In this context, F1 ends up being a big picture of the performance because precision takes into account false positives, and recall takes into account false negatives. Thus, F1_Score gives an idea of whether the classifying algorithm is providing too many incorrect classifications.

$$F1\_score = \frac{2 \times precision \times recall}{precision + recall} \tag{4}$$

4.3 X-Ray dataset

The dataset is from Kaggle composed of 2902 images in three classes: 1340 (normal), 218 (COVID-19), and 1344 (Viral Pneumonia). Figure 2 present an excerpt of 6 images with their respective classification.

(a) Normal (b) Pneumonia (c) COVID-19



(d) Normal (e) Pneumonia (f) COVID-19

Fig. 2. Classes Example: normal, viral pneumonia, and COVID-19

All original images have the resolution 1024 x 1024, and they have to be converted into images with resolution of 224 x 224. Also, we performed the following operations in order to avoid the overfitting of the neural network: rotation_range = 50, width_shift_range = 0.2, height_shift_ranged = 0.2, zoom_range = 0.1, horizontal_flip = True, and vertical_flip = True.

5. RESULTS

Table 1 shows the mean of the fourteen CNNs according to the metrics presented previously. As we can see, Resnet50 reaches the best accuracy with 94.3%. Mobilenet achieved the best precision. DenseNet201 got the best Recall and F1_Score. And, ResNet50_V2 presented the worst results. Next, we detail the results of DenseNet201, Resnet50, and Mobilinet, showing the confusion matrix and training curves.

As previously presented, DenseNet201 presented the best Recall and F1_Score. Table 2 shows its the confusion matrix; Figures 3 and 4 illustrates the training and validation behavioral. As we can see, DenseNet201 misclassified only 25 cases in the average. The remarkable fact is that only 2 cases of Pneumonia were wrongly classified as COVID-19, and no Normal x-rays were improperly classified as COVID-19.

Table 1. Results in all fourteen CNNs: mean of accuracy, precision, recall, F1_Score and Loss

Model	Accuracy	Precision	Recall	F1_score	Loss folds
Resnet50	**0.943**	0.983	0.915	0.946	0.035
Mobilenet	0.942	**0.989**	0.888	0.932	0.037
DenseNet169	0.939	0.984	0.892	0.933	0.040
Vgg16	0.934	0.988	0.869	0.921	0.033
DenseNet121	0.934	0.988	0.890	0.930	0.035
DenseNet201	0.933	0.987	**0.921**	**0.951**	0.031
Mobilenet_v2	0.932	0.983	0.870	0.919	0.038
Vgg19	0.931	0.980	0.875	0.922	0.037
Xception	0.917	0.972	0.840	0.897	0.037
Nasnet_Large	0.916	0.979	0.792	0.871	0.032
Nasnet_Mobile	0.914	0.972	0.816	0.880	**0.024**
Inception_v3	0.908	0.974	0.844	0.900	0.035
Inception_Resnet_v2	0.905	0.974	0.858	0.908	0.033
Resnet50_v2	0.678	0.749	0.151	0.214	0.256

Table 2. Best Confusion Matrix - DenseNet201

	COVID-19	Normal	Pneumonia
COVID-19	**42**	0	2
Normal	0	**267**	20
Pneumonia	2	1	**247**

Concerning the training and validation mean, curves of accuracy tend to converge. On the other hand, the loss tends to decrease until the fifth epoch; then the validation curve tends to go up. Maybe because the overfitting started, even though the difference increases only 0.05 in the loss.

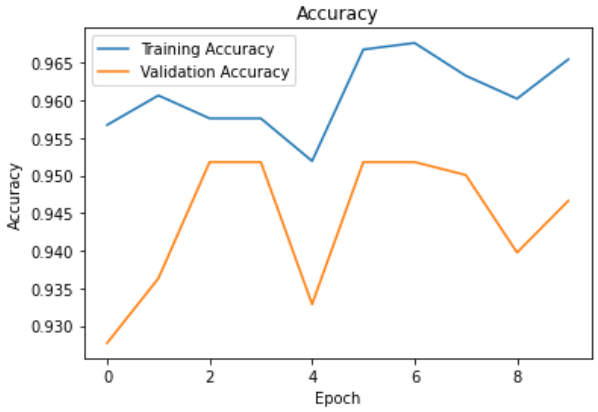


Fig. 3. Accuracy of training and validation in DenseNet201

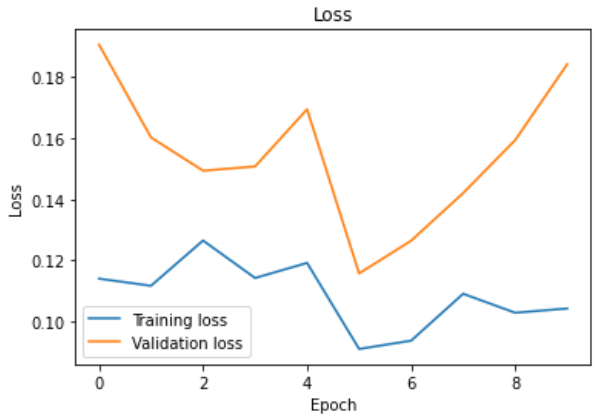


Fig. 4. Loss of training and validation in DenseNet201

Considering that Resnet50 presented the best accuracy, Table 3 shows its confusion matrix, in which we can see that the Resnet50 was incorrect only in 26 cases (all of them involving viral Pneumonia), being very similar to DenseNet201.

Table 3. Best Confusion Matrix - Resnet50

	COVID-19	Normal	Pneumonia
COVID-19	**42**	0	3
Normal	1	**267**	20
Pneumonia	1	1	**246**

Figures 5 and 6 presents the mean of accuracy and loss as epochs increase in ResNet50. As we can see, the accuracy is going up toward one (100%). On the other hand, the validation went up in the sixth epoch going down in the next ones, which is a behavior we expect.

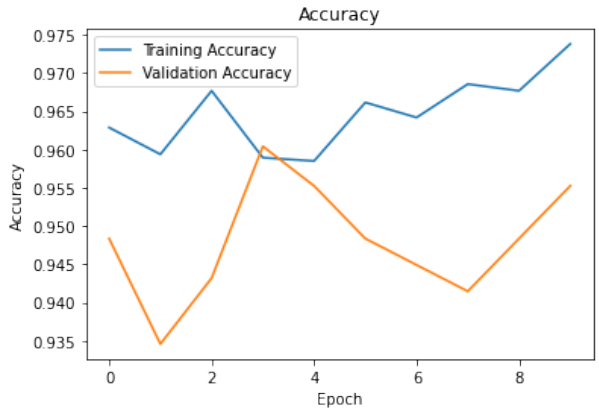


Fig. 5. Accuracy of training and validation in Resnet50

Regarding Mobilenet, we could see in Table 1 that the referred CNN reaches the best precision; thus, Table 4 shows its confusion matrix, in which the number of misclassification is higher than DenseNet201 and ResNet50.

Table 4. Confusion Matrix - Mobilinet

	COVID-19	Normal	Pneumonia
COVID-19	**39**	0	2
Normal	4	**263**	22
Pneumonia	1	5	**245**

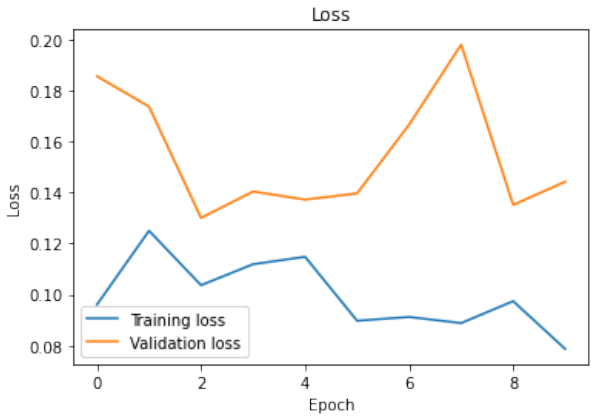


Fig. 6. Loss of training and validation in Resnet50

Moreover, Figures 7 and 8 shows that the accuracy and loss curves are not converging to the same place, especially considering the loss curve.

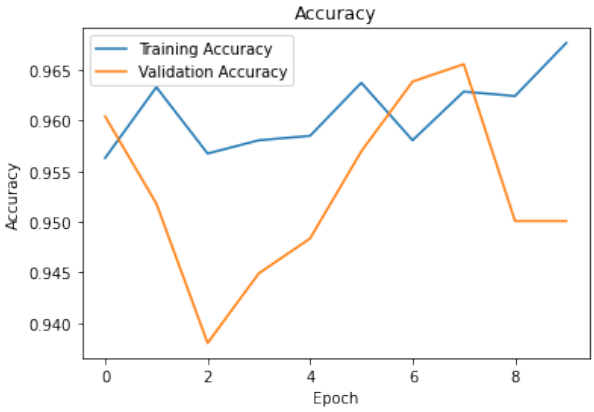


Fig. 7. Accuracy of training and validation in Mobilenet

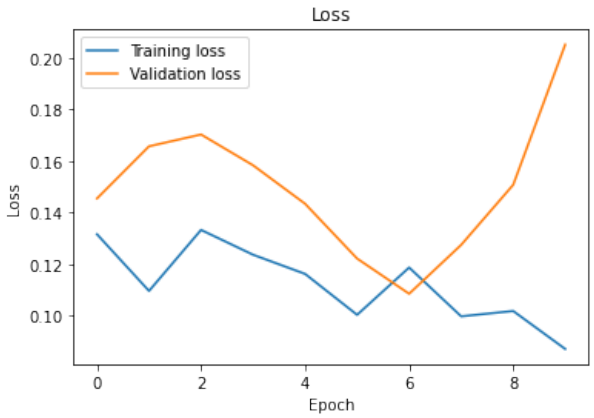


Fig. 8. Loss of training and validation in Mobilinet

6. CONCLUSION

This paper presented an investigation of fourteen CNNs trained by using transfer learning on classifying lung infections based on x-rays, particularly COVID-19 and viral pneumonia.

Results have shown that DenseNet201 reached the best results in general because it got the best results in Recall and F1_Score that are essential metrics based on false negatives, *i.e.*, this CNN produces less false negatives whose the worst scenario is the patient's death.

Mobilenet got the best accuracy. On the other hand, it presented a higher number of misclassifications than DenseNet201 and ResNet50.

Future work includes: (i) testing these fourteen CNNs in other image datasets such as computerized tomographies and ultrasound; and (ii) embedding the DenseNet201, Resnet50, and Mobilenet in a mobile application to help physicians with no access to computers in the moment of the diagnosis.

REFERENCES

Šarić, M., Russo, M., Stella, M., and Sikora, M. (2019). Cnn-based method for lung cancer detection in whole slide histopathology images. In *2019 4th International Conference on Smart and Sustainable Technologies (SpliTech)*, 1–4.

Beysolow II, T. (2017). *Introduction to Deep Learning Using R: a step-by-step guide to learning and implementing Deep Learning Models Using R*. Apress.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., and Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE Computer Vision and Pattern Recognition*.

Ehteshami Bejnordi, B., Veta, M., van Diest, P.J., van Ginneken, B., Karssemeijer, N., Litjens, G., van der Laak, J.A.W.M., and the CAMELYON16 Consortium (2017). Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. *JAMA*, 318(22), 2199–.

Google (2020). *Google Colaboratory*. `https://colab.research.google.com/notebooks/intro.ipynb`, Visit on 06-15-2020.

Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., Kim, R., Raman, R., Nelson, P.C., Mega, J.L., and Webster, D.R. (2016). Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA*, 316(22), 2402–2410.

Horry, M.J., Chakraborty, S., Paul, M., Ulhaq, A., Pradhan, B., Saha, M., and Shukla, N. (2020). Covid-19 detection through transfer learning using multimodal imaging data. *IEEE Access*, 8, 149808–149824.

imagenet (2020). *ImageNet*. `http://www.image-net.org/`, Visit on 05-31-2020.

Ismail, N.S. and Sovuthy, C. (2019). Breast cancer detection based on deep learning technique. In *2019 International UNIMAS STEM 12th Engineering Conference (EnCon)*, 89–92.

Keras (2020). *Keras Page*. https://keras.io/, Visit on 06-15-2020.

Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2324.

Pratt, L.Y., Mostow, J., and Kamm, C.A. (1991). Direct transfer of learned information among neural networks.

In *Ninth National Conference on Artificial Intelligence*, 584–589.

Python (2020). *Python Page*. https://www.python.org/, Visit on 06-15-2020.

Sarkar, D., Bali, R., and Ghosh, T. (2018). *Transfer Learning with Python*. Pack Publishing.

Schnurr, H.P., Aronsky, D., and Wenke, D. (2018). *MEDICINE 4.0—Interplay of Intelligent Systems and Medical Experts*, 51–63. Springer International Publishing, Cham.

Silva, D.C.S.e. and Cortes, O.A.C. (2020). On convolutional neural networks and transfer learning for classifying breast cancer on histopathological images using gpu. In *Congresso Brasileiro de Engenharia Biomédica*, 1–6.

Torrey, L. and Shavlik, J. (2009). Transfer learning. In E.S. Olivas, J.D.M. Guerrero, M.M. Sober, J.R.M. Benedito, and A.J.S. Lopez (eds.), *Handbook Of Research On Machine Learning Applications and Trends: Algorithms, Methods and Techniques*, chapter 11, 242–264. Information Science Reference, , Hershey, New York.

Tsang, T., Wu, P., Lin, Y., Lau, E., Leung, G., and Cowling, B. (2020). Effect of changing case definitions for covid-19 on the epidemic curve and transmission parameters in mainland china: a modelling study. *The Lancet Public Health*, 5.

Waheed, A., Goyal, M., Gupta, D., Khanna, A., Al-Turjman, F., and Pinheiro, P.R. (2020). Covidgan: Data augmentation using auxiliary classifier gan for improved covid-19 detection. *IEEE Access*, 8, 91916–91923.

Wang, N., Liu, H., and Xu, C. (2020). Deep learning for the detection of covid-19 using transfer learning and model integration. In *2020 IEEE 10th International Conference on Electronics Information and Emergency Communication (ICEIEC)*, 281–284.

Wilson, N., Kvalsvig, A., Barnard, L., and Baker, M.G. (2020). Case-fatality risk estimates for covid-19 calculated by using a lag time for fatality. *Emerging Infectious Diseases*, 26(6), 1339–1441.

Wolf, B. and Scholze, C. (2017). "medicine 4.0". *Current Directions in Biomedical Engineering*, 3.

World Health Organization (2020). *Situation Report - 158*. `https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200626-covid-19-sitrep-158.pdf?sfvrsn=1d1aae8a_2`, Visit on 06-26-2020.