Shallow Neural Networks beat Deep Neural Networks trained with transfer learning

Shallow Neural Networks beat Transfer Learning

A Use Case based on training Neural Networks to identify Covid-19 in chest X-ray images

DIMITRIOS MANOLAKIS

University of Macedonia, Department of Applied Informatics, dmanolakis@uom.edu.com

GEORGIOS SPANOS

University of Macedonia, Department of Applied Informatics, gspanos@uom.edu.com

IOANNIS REFANIDIS

University of Macedonia, Department of Applied Informatics, yrefanid@uom.edu.com

Since the start of the covid-19 health crisis, there have been many studies on the application of deep learning models in order to detect the virus on chest X-ray images. Training large neural networks on big data sets is a computationally intensive task, consuming a lot of power and needing a lot of time. Thus, usually only researchers in large institutions or companies have the necessary resources to bring the task to fruition. Other researchers employ transfer learning, a technique that is based on using pre-trained deep neural networks that have been trained on a similar dataset and retrain only their last neuron layers. However, using deep neural networks with transfer learning is not always the best option; in some cases, training a shallow neural network from scratch achieves better results. In this paper we compare training from scratch, shallow neural networks to transfer learning from deep neural models. Our experiments have been conducted on a publicly available dataset containing chest X-ray images concerning covid-19 patients, as well as non-covid-19 ones. Surprisingly enough, training from scratch shallow neural networks produced significantly better results in terms of both specificity and sensitivity. The results of the models' evaluation showed that the three shallow neural networks achieved specificity rates higher than 98%, while having a sensitivity rate of 98%, exceeding the best performing pre-trained model, the DenseNet121, which achieved a specificity rate of 91.3%, while having a sensitivity rate of 98%.

CCS CONCEPTS: • **Computing methodologies** → **Neural networks; Computer vision;** *Supervised learning.*

Additional Keywords and Phrases: Deep Learning, Transfer Learning, Convolutional Neural Networks

1 INTRODUCTION

Diagnosing covid-19 from a chest X-ray or a computed tomography (CT) is considered a difficult task for radiologists since the abnormalities these images contain, may look very similar for many other lung problems. It is also quite possible for patients with covid-19 to have normal chest X-rays or chest CTs, eliminating therefore, any chance of correctly diagnosing the virus. For these reasons, considering as well, that the virus is

highly contagious, most experts and medical societies advise against the use of an imaging test alone to diagnose or rule out covid-19.

On the other hand, convolutional neural networks (CNNs) are especially good at dealing with visual data, being the dominant approach currently in the scientific field of computer vision, which aims to extract information from images and videos. Up to this point, there have been some attempts to use CNNs on the problem of predicting covid-19 from chest X-ray images, by applying transfer learning on pre-trained models of this type [1, 13, 14]. This technique is generally considered to be quite effective, as it is possible to leverage pre-trained models of many layers and complicated architectures. In this paper, the architectures used for the pre-trained models are: VGG16 [5], ResNet50 [6], ResNet101 [6], ResNet152 [6] and DenseNet121 [7], with the number of layers ranging from 16 to 152 layers. For the shallow CNNs three architectures are used: CNN5, CNN6 and CNN7, composed by 5, up to 7 layers.

Considering the importance of data reliability and validity in medical image analysis, locating a trustworthy and valid dataset was prioritized, in order to correctly identify the best solution for the particular problem. The publicly available Covid – X-ray – 5k dataset (5000 X-rays) was chosen, which was created by examining already labeled chest X-rays from relevant databases, with the help of a board-certified radiologist, as part of a published research paper [1] that uses transfer learning. Using techniques such as data augmentation [10], we were able to effectively expand the dataset's size while in order to optimize the architecture for the various neural networks that were created, k-fold cross-validation was used. Furthermore, we experimented with many different metrics such as precision – recall curves, ROC curves, average precision score [18], specificity and sensitivity rates, in order to determine the validity of each metric in this highly imbalanced dataset. Additionally, we calculated confidence intervals for each metric, by applying normal approximation of the binomial distribution, with the aim of deriving a more reliable estimation of the true values.

The pre-trained models have been trained on the well-known image database ImageNet [9] which contains random objects like balloons or chairs, thus having zero similarity with the radiography images we are trying to analyze. As a result, our main objective is to examine and compare both the performance of transfer learning on various pre-trained models compared to that of shallow CNNs with a small number of layers, and also, to examine the capabilities these neural models have on the difficult task of chest X-ray image classification.

The remainder of the paper is organized as follows. Section 2 contains a review of the background literature. Section 3 provides a summary on the techniques used as part of the data preprocessing that occurred, as well as a full description of the two proposed frameworks. Section 4 presents the experimental results, the analysis of each framework's performance, as well as comparisons with other works. Finally, Section 5 concluded the paper and poses future directions.

2 RELATED WORKS

Even before the covid-19 health crisis, deep learning techniques where widely used on chest X-ray image classification. For example, there have been many transfer learning applications on pneumonia detection in chest X-ray images. As shown in [11], various pre-trained models such as AlexNet, DenseNet121, ResNet18, InceptionV3 and GoogleNet, which were trained and tested on over 5232 images, were able to achieve high levels of performance, with ResNet18 exceeding the other models, with a test accuracy of 94,23%. Moreover, a shallow CNN of 6 layers proved to be quite effective on the same task [12]. Trained on a dataset of 5,856 chest X-ray images the model was able to achieve a validation accuracy of approximately 94%.

In the case of predicting covid-19 on chest X-ray images, there have been numerous attempts to apply transfer learning, aiming to optimize and fine-tune various pre-trained models. As shown in [13], where five pre-trained CNN (ResNet50, ResNet101, ResNet152, InceptionV3 and Inception-ResNetV2) were trained using 5-fold cross validation on three different binary classifications with four classes (covid-19, healthy, viral pneumonia and bacterial pneumonia), ResNet50 achieved the highest overall classification performance of 98.43%, while in [14] transfer learning is applied on four pre-trained models (VGG-16, VGG-19, MobileNet and InceptionResNetV2) using a small dataset of 545 chest X-rays (181 covid-19 samples) with MobileNet achieving a testing accuracy of 96.8%. The published paper [1] from which the dataset we used was extracted from, applied transfer learning on four pre-trained models (ResNet18, ResNet50, SqueezeNet, DenseNet121). SqueezeNet achieved the best performance, maintaining a specificity rate of 92.9%, while having a sensitivity rate of 98%.

Although in [15] a shallow CNN of 4 layers, trained on an imbalanced dataset of 5856 chest X-rays (1583 covid-19 samples), managed to achieve an accuracy of 99,69%, highlighting the capabilities of these simples architectures, there are no studies that examine extensively the performance of both techniques on the same dataset, using multiple architectures of pre-trained and shallow neural networks on the task of predicting covid-19 on chest X-ray images. Most performance comparisons on transfer learning, focus on the technique of using the pre-trained models as feature extractors [22] against fully training these models. As shown in [16] on the task of facial expression recognition, applying transfer learning on VGG16 provided higher general classification accuracy than the training from scratch method while in [17], similar results were demonstrated, when comparing the two techniques on the task of predicting breast cancer on histology images.

According to the aforementioned related works and to the best of our knowledge, the present research work is the first attempt of comparison between transfer learning and shallow neural networks in the field of covid-19 prediction from chest X-rays.

3 DATASET & METHODOLOGY

The dataset is composed from two other datasets (Covid-chestxray-Dataset [2] & Chex-Pert [3]). In total, there are 184 covid-19 samples (84 for training and 100 for testing) while for the Negative class, there are 5000 samples (2000 for training and 3000 for testing).

3.1 K-Fold Cross Validation

Using k-Fold cross validation on the training set, we were able to create multiple combinations of validation and training sets, which were later used for training the models multiple times, while selecting for each architecture, the model with the highest performance. Since the dataset is highly imbalanced, with the Positive class being composed of just 100 samples, using the average testing accuracy as a metric to judge the models' performance would be invalid. As a result, average precision score [18] is preferred since it combines precision and recall, two metrics that take into consideration the models' performance with respect to the true and false positives, emphasizing additionally in the performance of the Positive class.

In the case of the pre-trained models, we applied 4-fold cross validation, that is, in each iteration 75% of the original training set was used for training, while the validation set used the remaining 25% of the original training set. Regarding the simple CNNs, 10-fold cross validation was used, that is, 90% of the original training set was used on the new ones. Creating slightly bigger training datasets is needed in the case of these network

structures since they are trained from scratch, which makes them susceptible to overfitting when trained on a very small amount of data.

3.2 Data Augmentation

The main idea behind this technique, which is widely used in many previous similar studies [1, 14, 17], is to apply a small transformation on an image, which for example can be either a shift to the right or to the left, or a zoom, and then feed the image to the neural network. The number of images stays the same, but in each training epoch different transformations of the original dataset are used, making the model more robust and accurate as it is trained on different variations of the same image, whereas it also prevents overfitting [10]. It is important to note, that these transformations are applied exclusively to the training dataset in order to avoid manipulating the test and validation data beforehand. Another very important fact is that in the case of medical imaging, augmentation may affect the resulting classification [20]. For that reason, significant transformations such as flipping the image horizontally or vertically were avoided, minimizing as much as possible the effect of the transformations which were applied.

An additional problem concerns the wide variation in the resolution of images in the dataset, which contains some low – resolution images (below 400x400) and some high – resolution ones (more than 1900x1400). With the aim of solving this problem, three more useful operations are applied on every dataset (train, validation and test):

1. Resizing the images in order for all of them to have the same size (224x224). This size was selected since the pre-trained models, were trained on images of that size.
2. Convert the single-channel X-ray images (grayscale) to a three-channel format by repeating the values in the image across all channels. When training simple Convolutional Networks a single channel can be used, but when pre-trained models are used for transfer learning, three channel inputs (RGB) are required.
3. Rescaling (Normalizing) the images. Rescaling the scales array of the original image pixel values to be between [0, 1] achieves a balanced contribution to the overall loss and greatly improves the convergence of the optimization algorithm [19].

3.3 Transfer Learning

Transfer learning is a machine learning method where a model which was built for a specific task is adapted into solving another task. The most popular version of transfer learning makes use of the pre-trained models as feature extractors [16, 17, 22]. Only the fully connected layers, which create the final predictions using the features that were extracted from the previous layers, are adapted on the new data. These layers are completely removed from each pre-trained model and replaced from new fully connected layers that correspond to the specific problem (binary classification in our case). This technique is proposed when working with a small amount of data and for that reason it is applied on this project as well. Since this is the only part of the network that gets trained, four different architectures were used for each pre-trained model and optimized using 4-fold cross-validation, choosing in each case the best performing fully connected layers combination for each model. Figure 1 shows these four architectures with increasing complexity (parameter-wise), from left to right.

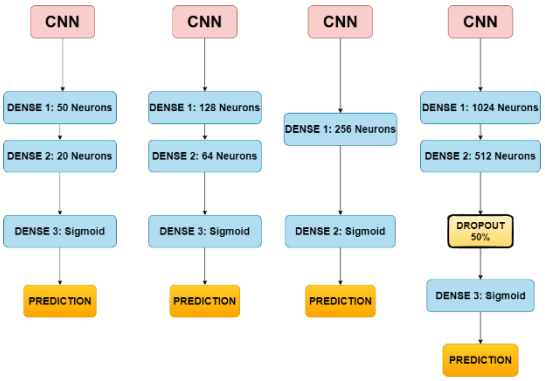

Figure 1: Architectures of the fully connected layers

3.4 Shallow Convolutional Neural Networks

The architecture used on these structures is based on the simplicity of the VGG16 architecture, since both architectures use convolutional layers with 3x3 filters and stride set to 1, as well as, max pooling layers with 2x2 filters and stride set to 2. In order to optimize the performance of these networks we employed 10-fold cross-validation and, we experimented with three different architectures for the convolutional and the pooling layers (number of layers and number of filter in each layer). The exact architectures used in each shallow CNN are displayed in Figure 2.

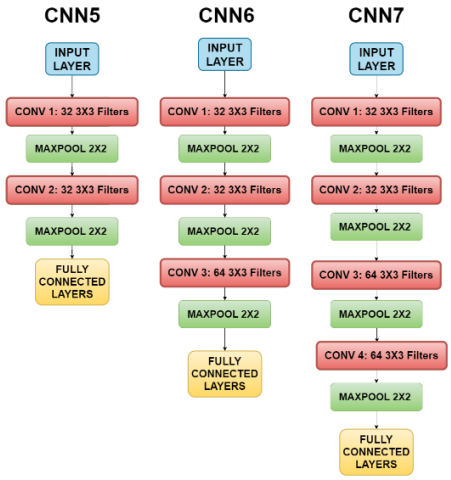

Figure 2: Architectures of the shallow convolutional neural networks

4 EXPERIMENTAL RESULTS

4.1 Threshold Selection and Confidence Intervals

After selecting the architecture with the best average performance for each pre-trained model and shallow network, using average precision score and k-fold cross-validation, it was crucial to select a specific threshold for the predictions each model outputs through the sigmoid neuron in the last layer. As mentioned earlier, it is important to take into consideration the high importance of the predictions about the positive class. For that reason, we selected for each model the threshold that produces a sensitivity rate of 97% and compared their performance on the corresponding specificity rate.

The testing dataset contains a very limited number of samples, especially for the positive class and as a result, the values for the sensitivity and specificity rates may be unreliable. In order to address this issue, as also conducted in [1], we estimated the 95% confidence intervals for each metric by applying normal approximation of the binomial distribution [21].

Table 1 contains the confidence intervals for the sensitivity and specificity metrics and for each of the 8 models. The width of the intervals for the sensitivity rates is quite large since there is a limited number of positive samples. On the other hand, for the specificity rates the intervals are quite small since for the negative class, the testing dataset contains 3000 samples. It is also clear, that the shallow neural networks are able to achieve a considerably greater performance than the pre-trained models since while keeping the sensitivity rate of ***97%***, the best performing pre-trained model DenseNet121 achieves a specificity rate of ***95%*** whereas the best performing shallow network CNN6 achieves a specificity rate of ***99.84%***.

The published paper [1] from which the dataset was extracted from, also used transfer learning on pre-trained models. The threshold value for each model was selected targeting to achieve 98% sensitivity rate. Table 2 provides a comparison of the performance of our shallow neural networks against the paper's pre-trained models' performance, highlighting that in this case as well, the proposed framework is overperforming.

Table 1: Confidence Intervals

Model	sensitivity	specificity
VGG16	97.00 ± 3.34%	86.07 ± 1.23%
ResNet50	97.00 ± 3.34%	68.07 ± 1.66%
ResNet101	97.00 ± 3.34%	71.05 ± 1.62%
ResNet152	97.00 ± 3.34%	72.70 ± 1.59%
DenseNet121	97.00 ± 3.34%	95.00 ± 0.77%
CNN5	97.00 ± 3.34%	99.44 ± 0.26%
CNN6	97.00 ± 3.34%	99.84 ± 0.14%
CNN7	97.00 ± 3.34%	99.70 ± 0.19%

4.2 Precision – Recall and ROC Curves

In order to examine the general performance of each model, we need to go through the comparison for all possible threshold values. precision recall and ROC curves illustrate the diagnostic ability of the classifiers, as we can see in Figure 3, which also displays the Area Under the Curve (AUC) metric. All models have a similar performance according to the ROC AUC values, even though it is clear from the values reported in Table 1, that this is not the case. On the contrary, the precision recall AUC values display an accurate overall

performance for the models accordingly to our previous results, emphasizing the fact that these plots are appropriate for problems with extremely imbalanced datasets and a high priority on the predictions about the positive class [8].

Table 2: Comparison with Related Work

Model	sensitivity	specificity
ResNet18	98.00 ± 2.7%	90.7 ± 1.1%
ResNet50	98.00 ± 2.7%	89.6 ± 1.1%
SqueezeNet	98.00 ± 2.7%	92.9 ± 0.9%
DenseNet121	98.00 ± 2.7%	75.1 ± 1.5%
CNN5	98.00 ± 2.7%	98.9 ± 0.3%
CNN6	98.00 ± 2.7%	98.3 ± 0.4%
CNN7	98.00 ± 2.7%	98.3 ± 0.4%

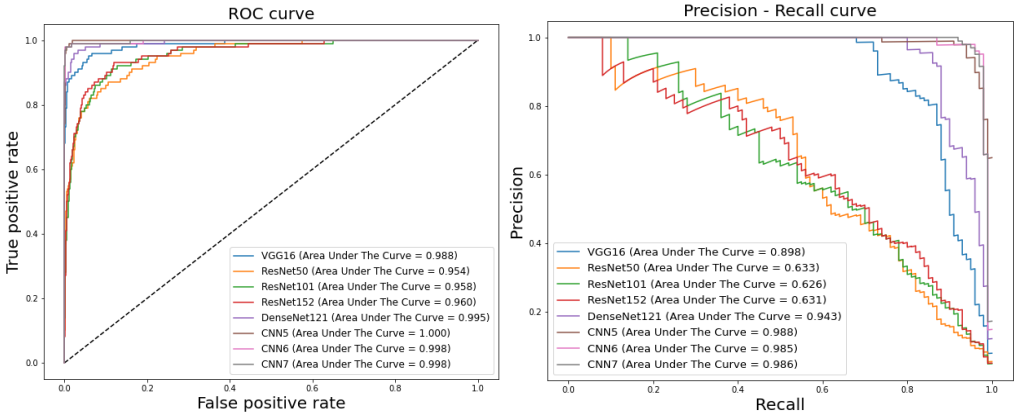

Figure 3: ROC curves (left) and precision – recall curves (right)

5 CONCLUSIONS

This paper examines the capabilities of neural networks on the task of predicting covid-19 on chest X-rays and reports a performance comparison of transfer learning on deep CNNs against training from scratch shallow CNNs. According to the results of the models' evaluation, it is clear that even though both pre-trained models and shallow neural networks are able to produce a promising performance in terms of sensitivity and specificity rates, the latter framework managed to beat the transfer learning technique used on 8 different architectures of pre-trained models. As mentioned earlier, it is quite possible that the dissimilarity between the ImageNet images and the radiography images is hindering the performance of pre-trained models on the covid-19 binary classification task. We also confirmed that precision recall curves is the appropriate metric [8] in order to evaluate the general performance of models trained on imbalanced datasets, while also, solving a problem with a high priority on the predictions of the positive class. Moreover, we were able to essentially upgrade the publicly available dataset, creating two separate versions composed of 4 and 10 folds of training and validation sets.

Even though the results are quite promising, the use of insufficient number of covid-19 chest X-ray images is a limitation in our study and as such, further experiments using ideally a larger dataset are needed in order to correctly estimate the actual performance of neural networks on this problem. It is also worth mentioning that small and less diverse datasets could possibly create biased models that rely heavily on the source dataset instead of the relevant medical information [23]. Creating a large dataset of (covid positive) chest X-ray images from various sources will be a very important future work in order to further investigate if convolutional neural networks can aid in the diagnosis of covid-19. Additionally, there is a large collection of pre-trained models that are not yet used for this task, which can possibly achieve better performance if used in future projects. Finally, the augmented dataset allows for further experimentation with the fully connected layers of the pre-trained models, as well as, with shallow neural networks of different amount of layer, different amount of filters and filter sizes in each layer or even different optimizers, activation and loss functions.

REFERENCES

[1] Minaee, S., Kafieh, R., Sonka, M., Yazdani, S. and Jamalipour Soufi, G., 2020. Deep-covid: Predicting covid-19 from chest X-ray images using deep transfer learning. Medical Image Analysis, 65, p.101794.

[2] Cohen, Joseph Paul, Paul Morrison, and Lan Dao. "covid-19 image data collection." arXiv preprint arXiv:2003.11597, 2020

[3] Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D., Halabi, S., Sandberg, J., Jones, R., Larson, D., Langlotz, C., Patel, B., Lungren, M. and Ng, A., 2019. CheXpert: A Large chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. Proceedings of the AAAI Conference on Artificial Intelligence, 33, pp.590-597.

[4] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Rafal Jozefowicz, Yangqing Jia, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Mike Schuster, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[5] Simonyan, K., & Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[6] He, K., Zhang, X., Ren, S., & Sun, J. 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp.770-778.

[7] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. 2017. Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700-4708.

[8] Davis, J., and Goadrich, M. 2006. The Relationship between precision-recall and ROC Curves. In Proceedings of the 23rd International Conference on Machine Learning (pp. 233–240). Association for Computing Machinery.

[9] Russakovsky O., Deng J., Su H., Krause J., Satheesh J., Ma S., Huang Z., Karpathy A., Khosla A., Bernstein M., Berg A., Fei-Fei L. 2015. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV); 115(3):pp.211-252

[10] Shorten, C., Khoshgoftaar, T.M. A survey on Image Data Augmentation for Deep Learning. J Big Data 6, 60 (2019). https://doi.org/10.1186/s40537-019-0197-0

[11] Chouhan, V., Singh, S., Khamparia, A., Gupta, D., Tiwari, P., Moreira, C., Damaševičius, R. and de Albuquerque, V., 2020. A Novel transfer learning Based Approach for Pneumonia Detection in chest X-ray Images. *Applied Sciences*, 10(2), p.559.

[12] Stephen, O., Sain, M., Maduh, U. and Jeong, D., 2019. An Efficient Deep Learning Approach to Pneumonia Classification in Healthcare. *Journal of Healthcare Engineering*, 2019, pp.1-7.

[13] Narin, A., Kaya, C. and Pamuk, Z., 2021. Automatic detection of coronavirus disease (covid-19) using X-ray images and deep Convolutional Neural Networks. *Pattern Analysis and Applications*, 24(3), pp.1207-1220.

[14] Mohammadi, R., 2020. transfer learning-Based Automatic Detection of Coronavirus Disease 2019 (covid-19) from chest X-ray Images. *Journal of Biomedical Physics and Engineering*, 10(5).

[15] Mukherjee, H., Ghosh, S., Dhar, A., Obaidullah, S., Santosh, K. and Roy, K., 2021. Shallow Convolutional Neural Network for covid-19 Outbreak Screening Using chest X-rays. *Cognitive Computation*.

[16] Oztel, I., Yolcu, G. and Oz, C., 2019. Performance Comparison of transfer learning and Training from Scratch Approaches for Deep Facial Expression Recognition. *2019 4th International Conference on Computer Science and Engineering (UBMK)*.

[17] Shallu and Mehra, R., 2018. Breast cancer histology images classification: Training from scratch or transfer learning?. *ICT Express*, 4(4), pp.247-254.

[18] Zhang, E. and Zhang, Y., 2009. average precision. *Encyclopedia of Database Systems*, pp.192-193.

[19] LeCun, Y., Bottou, L., Orr, G. and Müller, K., 2012. Efficient BackProp. *Lecture Notes in Computer Science*, pp.9-48.

[20] Glickman, C., 2021. *Data Augmentation in Medical Images*. [online] Medium. Available at: <https://towardsdatascience.com/data-augmentation-in-medical-images-95c774e6eaae> [Accessed 4 October 2021].

[21] Wallis, S., 2013. Binomial Confidence Intervals and Contingency Tests: Mathematical Fundamentals and the Evaluation of Alternative Methods. *Journal of Quantitative Linguistics*, 20(3), pp.178-208.

[22] Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C. and Liu, C., 2018. A Survey on Deep transfer learning. *Artificial Neural Networks and Machine Learning – ICANN 2018*, pp.270-279.

[23] Gianluca Maguolo and Loris Nanni. 2020. A critic evaluation of methods for COVID-19 automatic detection from X-ray images.Artificial NeuralNetworks and Machine Learning – ICANN 2018 Lecture Notes in Computer Science(2020), 270–279. https://doi.org/10.1007/978-3-030-01424-7_27