Research Article
iMedPub Journals
www.imedpub.com

American Journal of Computer Science and Information Technology
ISSN 2349-3917

2022
Vol.10 No.1:113

COVID-19 Detection on Chest X-ray Using an Enhanced Neural Network Model: Impact of Network Complexity, Data Augmentation and Transfer Learning

Himal Bamzai-Wokhlu
Buchholz, 5510 NW 27th Avenue, Gainesville, FL 32607

Received date: January 31, 2022, Manuscript No. IPACSIT-22-14853; **Editor assigned date:** February 02, 2022, PreQC No. IPACSIT-22-14853(PQ);
Reviewed date: February 12, 2022, QC No IPACSIT-22-14853 **Revised date:** March 17, 2022, Manuscript No. IPACSIT-22-14853(R);
Published date: March 25, 2022, DOI:10.36648/2349-3917.10.1.22
Citation: Himal Bamzai-Wokhlu (2022) COVID-19 Detection on Chest X-ray Using an Enhanced Neural Network Model: Impact of Network Complexity, Data Augmentation and Transfer Learning. Am J Compft Scfi Infform Technofl Vofl .10 Iss No. 1: 113

Abstract

Machine learning (ML) algorithms have potential to rapidly screen COVID-19 from chest x-ray (CXR). Current deep convolutional neural network (DCNN) models for COVID-19 detection are limited by small datasets and overfitting. We hypothesized that less network complexity, heavy data augmentation, and transfer learning would result in the best model. A COVID-19 detection model was developed using the COVIDx public dataset of 16,352 de-identified CXRs associated with known COVID-19 status by reverse transcriptase polymerase chain reaction (RT-PCR). Twenty-four pre-trained DCNNs with various enhancement features were compared using 80/20 split for testing and validation. Among 5 pretrained DCNN's, the low complexity but deep ResNet18 architecture performed best. Data augmentation using horizontal flip (HF), Gaussian blur (GB), and cutout (CO) improved ResNet18 performance- with the ResNet18-CO/GB model performi¬¬ng best at 1,000 iterations. Although transfer learning using an extrinsic pneumonia detection model did not boost performance, transfer learning from tuberculosis (TB) detection models enhanced performance of ResNet18-HF and ResNet18-CO/HF/GB models. Comparing the top models at 10K iterations, the best model was ResNet18-GB/CO without transfer learning with sensitivity 82.0%, specificity 96.5%, and accuracy 94.5%. Our findings suggest utility for automated COVID-19 detection by CXR using DCNN's enhanced by data augmentation more so than transfer learning.

Keywords: Artificial intelligence, COVID-19, Diagnosis, Machine learning, Transfer learning

1. Introduction

COVID-19 is a respiratory viral illness that has erupted into a global pandemic, affecting over 220 million people worldwide [1]. Diagnostic errors or delays in detection can impact timely identification and management of high-risk patients. The gold standard in diagnostic testing, reverse transcriptase polymerase chain reaction (RT-PCR) is not available everywhere and turnaround times vary. Alternatively, point-of-care antigen testing has lower sensitivity ranging from 60 to 85% [2]. Thus, there is strong impetus to detect COVID-19 by CXR using artificial intelligence (AI).

Based on the three-dimensional neural pattern inspired by the visual cortex of animals, deep convolutional neural networks (DCNN) are particularly suited to identify patterns in CXR imaging. Several models targeting diagnosis of pneumonia, tuberculosis, atelectasis, and lung cancer have been developed [3][4]. Furthermore, DCNN's are not prone to human factors like fatigue or distraction that decay performance. Jones et al. showed that CNN models matched or outperformed radiologists in detection of 11 out of 14 different pathologies from CXR [5]. When it comes to detecting COVID-19 by CXR, radiologists are not particularly accurate with one series reporting sensitivity of <50 % [6]. It remains to be seen how machine learning (ML) detection of COVID-19 by CXR compares to clinicians and existing standards for other rapid assays, like antigen testing.

Various deep convolutional neural network (DCNN) models have been proposed for COVID-19 detection, with accuracies ranging from 82% to 98% [7, 8, 9, 10]. For example, Ozturk et al. reported an accuracy of 98% for binary detection of COVID-19 using DarkCovidNet [7]. Another group reported comparable success using a three-way classification, discriminating COVID from other forms of viral pneumonia as well as healthy patients [8]. Nonetheless, the main limitation of early models is that they were derived from small datasets (<500 COVID-19 patients) making them prone to excess validation loss and overfitting [9].

In this study, we sought to compare the relative impact of DCNN enhancement features, specifically network complexity, data augmentation, and transfer learning on model performance and on the problems of validation loss and overfitting. All three features have potential to impact the performance of such models. In the case of architectural complexity, there is conflicting data on the impact of network complexity on COVID-19 detection models. Comparing the performance of various neural network architectures, Ismael and Sengul reported that Wide ResNet50 and VGG16 were both more accurate than the less complex

ResNet18 model [10]. However, in direct comparisons for non-COVID image recognition, such as the ImageNet Large Scale Visual Recognition Challenges, less complex deep learning architectures perform better [11]. Identifying optimal architectures is highly relevant to future model development.

The issue of data deficit also can be mitigated by data augmentation or transfer learning strategies. By applying various image transformations on the dataset, data augmentation can artificially "increase" the amount of and variation of data being read into the model. For instance, Ismael and Sengul reported an accuracy of 94.7% using a data-augmented ResNet50-enhanced standard vector (SVM) with the linear kernel function [10]. Another underexplored strategy to circumvent the data deficit is instance-based transfer learning [8, 9]. Transfer learning involves a multi-step training plan—by first pre-training the model on a relevant pathology using a distinct dataset, such that it then facilitates downstream recognition of pathology of interest during the subsequent training phase on the dataset of interest. For instance, Apostolopoulos and Mpesiana reported 96.8% accuracy and 98.9% sensitivity using a transfer learning approach to train their deep-learning model to categorize X-ray images as either common bacterial pneumonia, COVID-19, and otherwise healthy patients [8]. Although promising, there are few studies that measure the impact of transfer learning on model accuracy.

In May 2021, de-identified open-source COVID-NET was made available containing over 16,000 CXR's associated with confirmed COVID-19 RT-PCR status, including 2358 X-rays associated with COVID infection. Using this dataset, we sought to develop an optimized CNN model for COVID-19 detection by CXR, using a binary classification schema, with a minimum sensitivity of ≥80% and a target specificity of ≥97%, consistent World Health Organization standards for antigen assays [12]. We hypothesized that the combination of low architectural complexity, heavy data augmentation, and transfer learning would result in a high-performance COVID-19 CXR detection model with less validation loss and overfitting.

2.Materials And Methods

2.1 Dataset Description and Pre-Processing

The CXR images used to train and test our model were obtained from a publicly available and de-identified COVID-19 CXR image dataset collected by the COVID-Net Open Initiative, COVIDx CXR-2 [13]. The dataset, originally made available through GitHub, is a combination of five different publicly available datasets and currently represents the largest open-source collection of COVID-19 CXRs. The dataset contains 16,352 CXR images from 15,100 patients. Images were pre-classified according to known COVID-19 status by RT-PCR, as shown in Table I. COVID-negative cases included patients with both normal CXR as well as patients with non-COVID pneumonia. In addition, two additional datasets containing tuberculosis (TB) and pneumonia (PN) data were utilized to inform detection models that were used for transfer learning. The TB data set is an open-source dataset containing 3500 Normal CXRs and 700 tuberculosis CXRs collected from several sources, acquired through Kaggle [14]. The pneumonia dataset contains 4273 pneumonia CXRs and 2709 normal CXRs [15]. All datasets used in this study were publicly available, open access, and de-identified, and therefore did not require institutional review. As part of pre-processing, all images were resized to 600 x 700 pixels for consistency in convolutional layers.

Table 1. X-ray Image Counts According to Known COVID-19 RT-PCR status

	Covid-19 positive	Covid-19 negative
Test	1,886	11,194
Validation	472	2,799
Total	**2,358**	**13,993**

2.2 Experimental Design

To train models to detect COVID-19 via CXR, for each version of the model, we randomly allotted 80% and 20% of the images into the training and validation subsets, respectively. Performance characteristics that were assessed included sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F-score. We also assessed validation losses, training losses, and their differences to assess overfitting using a Stochastic Gradient Descent Function. For the final model contenders, we also measured accuracy.

2.3 Common Model Features to Minimize Validation Loss

The Adam optimization algorithm was used to minimize training loss for all models. The Adam optimizer, a leading optimization algorithm in CNN image classification, is an extension of Stochastic Gradient Descent and uses both momentum and adaptive learning rates [16]. The hyperparameters used for training included batch size = 5, learning rate = 0.05, number of epochs = 20, weight of COVID-19 images in loss function = 13993/2358, weight of COVID-19 (-) images in loss function = 1).

2.4 Selection of a Pre-Trained Neural Network Architecture

To select an architectural backbone for our DCNN, we compared five existing pre-trained architectures that are regarded as a gold standard for image detection ML. These included AlexNet [17], VGG16 [18], ResNet18 [19], WideResNet [19][20], and DenseNet-161 [21].

Model Feature 1. Architectural Complexity

For all the pre-trained CNN's, we measured complexity by finding the number of trainable parameters in each network. To assess the impact of parameter number on validation loss, training loss, and validation loss-training loss (as an indicator of overtraining or overfitting), least square regression line (LSRL) t-tests were performed and compared across the models.

Model Feature 2. Data Augmentation using Image Transformation

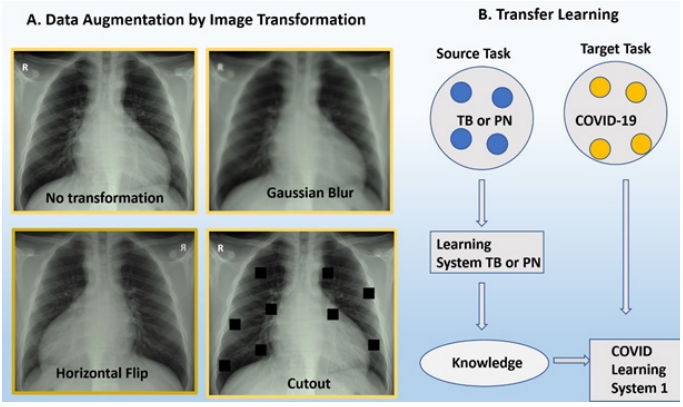
Using the best performing pre-trained model, data augmentation image transformation strategies, as shown in Figure 1A, were implemented either alone or in combination to artificially "increase" the amount of and variation of data being read into the model. To address some of our questions about data augmentation, we had to collect data from over six combinations of Random Horizontal Flip (HF), Gaussian Blur (GB), and Cutout (CO).

a) Random Horizontal Flip – inverts images by flipping across their horizontal axis. 50% of the images in our model underwent this transformation.

b) Gaussian Blur - a blurring technique whose visual effect is image smoothing in order to reduce visual noise based on a Gaussian distribution with the parameter of kernel size. For this study, we used a kernel size of 0.5.

c) Cutout – randomly chooses squares of a specified size remove from the image. For this study, 10 holes of dimensions 20 pixel*20 pixel were cutout.

Figure 1. Schematic of Data Augmentation and Transfer Learning Strategies



A. Image Transformation.

B. Transfer learning with TB or pneumonia datasets.

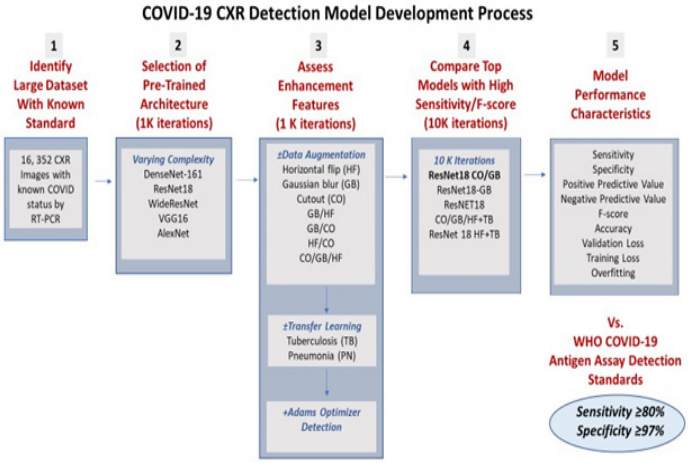
Model Feature 3. Instance-Based Transfer Learning

The impact of instance-based transfer learning on model performance for prediction of COVID-19 was applied to the best pre-trained model using the TB or PN datasets. These datasets were used to train various ResNet18 models first before retraining the transfer learning enhanced models on the COVIDx-CXR dataset.

2.5 Comparing the Top Models to Identify an Optimal Final Model

We selected the CNN with the optimal combination of performance characteristics, focusing on sensitivity and F1 score to identify the best performing models. For the initial models, statistics reported are an average over 1,000 iterations. To select the optimal final model, the top four performing models based on sensitivity and f-score were run for 10,000 iterations. Figure 2 summarizes the development process for our tailored DCNN.

Figure 2: Development Processes for COVID-19 CXR Detection Models



The model development process included 1) identifying a large publicly available CXR dataset 2) selecting from neural network architectures of varying complexity, 3) assessing impact of data-augmentation and transfer learning 4) reassessing the top models at 10,000 iterations, (top model shown in **bold**) 5) and comparing performance characteristics to pre-specified World Health Organization Criteria for COVID-19 antigen testing.

2.6 Software, Code Availability, and Statistics

The five pre-trained open-source architectures already exist in the PyTorch library. We used PyTorch 1.6 to develop and train various iterations of the DCNN models. PyTorch is an open-source ML library based on the Torch library, released under the Modified BSD license [2]. Model performance characteristics were visualized and recorded through Tensor Board. Code can be found at: https://github.com/mynameishimal/ML-project.git. The confusion matrix package was used to calculate performance characteristics in PyTorch. Least square regression line t-tests

were performed using Excel. P-value <0.05 was considered significant.

3. Results

3.1 Backbone CNN Architectures with Fewer Trainable Parameters Perform Better

Table 2 summarizes the performance characteristics for various architecture backbones, based on increasing complexity, measured by number of trainable parameters. Sensitivity was highest for the relatively dense AlexNet at 79.2%, but other performance metrics were poor. The smallest CNN, DenseNet-161, had the greatest F-score. However, the slightly larger but still lightweight ResNet18 performed the best overall, with a high sensitivity (72.2%) and the highest specificity (79.3%). WideResNet50, despite having a ResNet foundation with ~50 million (M) more parameters, did not perform as well.

Table 2. Performance Metrics for Pre-Trained CNN Architectures of Varying Complexity

Pre-Trained CNN Model	No. Parameters (M)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	F-score (%)
DenseNet-161	7	67.9	66.1	56.1	76.5	67.2
ResNet18	**11**	**72.2**	**79.3**	**66.5**	**87.3**	**62.0**
WideResNet50	61	65.7	42.3	35.6	59.0	46.8
AlexNet	66	79.2	11.7	30.8	10.8	38.7
VGG16	134	71.4	22.5	31.1	22.0	36.9

Validation set statistics are averaged over 1,000 iterations. M=million. Top performing model is highlighted in **bold**

3.2 Validation Loss Correlates Positively with Network Complexity

Of the pre-trained architectures, ResNet18 had the lowest validation loss without overfitting (loss difference of -0.05), as shown in Supplementary Table 1.

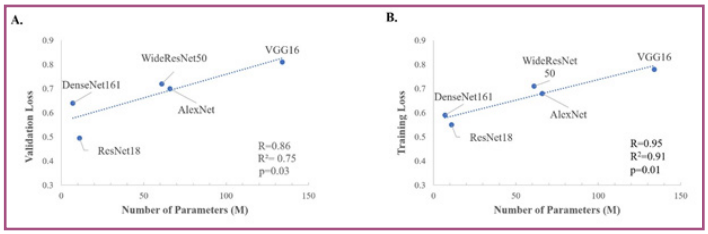
Supplementary Table 1. Losses for CNN Architectures of Varying Complexity

Pre-trained CNN Model	No Parameters (M)	Validation Loss	Training Loss	Loss Difference	Over fitting
DenseNet161	7	0.64	0.59	0.05	Yes
ResNet18	11	0.50	0.55	-0.05	No
AlexNet	61	0.72	0.71	0.01	Yes
WideResNet50	66	0.70	0.68	0.02	No
VGG-16	134	0.81	0.78	0.03	No

Loss was averaged over 1,000 iterations. Loss Difference is a degree of overfitting.

The other models demonstrated higher validation and training losses and were more prone to overfitting. In addition, a significant correlation was observed between validation loss and complexity (LSRL T-Test, R=0.86, p=0.03) across architectures (Figure 3).

Figure 3. Impact of Complexity on Validation and Training Losses



Figures **3A** and **3B**, respectively show the positive correlation between then number of parameters in a pre-trained neural network architecture in millions (M), i.e., complexity and validation loss and training loss, respectively, at 1,000 iterations

3.3 Multiple Data Augmentation Boost Sensitivity and F1-Score for COVID-19 Detection

Performance characteristics adding on various image transformations are summarized in Table 3. Both HF and GB augmented the baseline ResNet18 model, whereas CO alone was associated with worsening performance parameters except for specificity. Applying multiple image transformations in combination boosted model performance; the best performing transfer learning model was ResNet18-CO/HF/GB+TB with a sensitivity of 83.0%, specificity of 97.2%, and overall accuracy of 93.0% for detecting COVID-19 from CXRs.

Table 3. Performance Characteristics of Data Augmentation by Image Transformation

Model	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	F-score (%)
ResNet18	72.2	79.3	66.5	87.3	62.0
ResNet18-HF	75.3	80.0	69.8	76.0	65.3
ResNet18-GB	75.0	88.6	73.8	91.9	72.6
ResNet18-CO*	65.7	85.5	47.4	85.3	50.2
ResNet18-HF/GB	75.6	86.7	66.6	90.4	65.6
ResNet18-CO/GB*	**83.4**	**95.1**	**83.1**	**95.4**	**80.3**
ResNet18-CO/HF	71.5	84.4	64.9	76.9	62.5
ResNet18-CO/HF/GB	74.4	91.5	63.7	80.0	64.5

Validation set statistics are an average of over 1,000 iterations. CO=cutout, HF=horizontal flip, GB=Gaussian blur. Top performing model is in **bold**. *selected for further study

3.4 Validation Loss and Fitting Are Optimal with Multiple Image Transformations

Validation loss was lowest for the ResNet-CO/GB model at 0.17 – much lower than ResNet18 alone (0.05) and without overfitting (loss difference -0.01) (Supplementary Table 2). Other models had much higher validation loss.

Supplementary Table 2. Validation and Training Loss for Various Data Augmentations

Pre-trained CNN Model

	Validation Loss	Training Loss	Loss Difference	Overfitting
ResNet18	0.50	0.55	-0.05	No
ResNet18-HF	0.53	0.53	0	No
ResNet18-GB	0.47	0.50	-0.03	No
ResNet18-CO	0.61	0.60	0.01	Yes
ResNet18-HF/GB	0.60	0.58	0.02	Yes
ResNet18-CO/GB	0.17	0.18	-0.01	No
ResNet18-CO/HF	0.53	0.51	-0.02	No
ResNet18-CO/HF/GB	0.49	0.78	-0.29	No

CO=cutout, GB=Gaussian blur, HF= horizontal flip

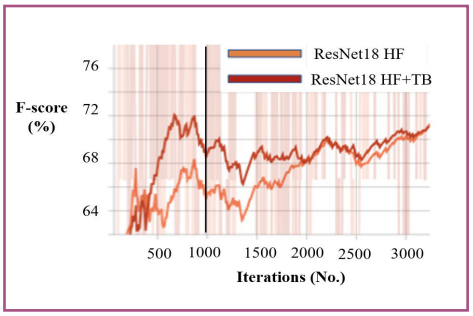
3.5 Transfer Learning with Data Augmentation Improves Model Performance

Results from instance-based transfer learning approaches to CNN based COVID-19 detection using TB and PN datasets are summarized in Table 4. Model performance with the PN transfer learning was poor overall, with no significant improvement seen. In contrast, when performing TB transfer, benefits were seen but only when data augmentation was also incorporated in the model. For instance, adding TB transfer learning to the ResNet18 alone decreased F-score from 62.0 to 54.5%. In contrast, adding the TB transfer learning to the ResNet18-HF model resulted in an overall F-score improvement due to markedly increased specificity from 80.0% to 87.7%. As Supplementary Figure 1 illustrates, the performance gain achieved early in the transfer learning process over certain data augmented models diminished with progressive iterations.

Table 4. Performance Characteristics with Transfer Learning

Model	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	F-score (%)
ResNet18-TB	69.9	77.3	51.6	60.3	54.5
ResNet18-HF+TB*	74.5	87.7	73.5	87.5	68.3
ResNet18-GB+TB	71.7	79.6	56.7	79.4	58.1
ResNet18-HF/GB+TB	70.1	88.3	64.2	77.7	63.4
ResNet18-CO/GB+TB	71.9	87.7	68.1	82.7	66.5
ResNet18-CO/HF/GB+TB*	**83.0**	**97.2**	**81.2**	**93.1**	**79.8**
ResNet18-PN	69.7	81.9	53.7	66.3	50.0
ResNet18-HF+PN	64.4	76.4	54.3	63.9	54.2
ResNet18-GB+PN	74.7	80.0	59.3	67.1	59.2
ResNet18-HF/GB+PN	73.0	82.2	63.4	73.6	63.2
ResNet18-CO/GB+PN	68.7	77.4	52.1	61.3	53.8
ResNet18-CO/HF/GB+PN	63.7	77.4	48.6	58.9	53.8

Supplementary Figure 1. Data Augmentation Models with or without Transfer Learning



Supplementary Figure 1. F-score gains achieved early in the transfer learning process over the HF-augmented model appear to diminish with progressive iterations. HF=Horizontal Flip, TB=Tuberculosis

3.6 Validation Loss Varies with Transfer Learning

The addition of transfer learning to data augmented models impacted validation loss variably (Supplementary Table 3). In general, validation loss was high for transfer learning models. However, for the top transfer learning model, ResNet18-CO/HF/GB +TB, validation loss and training loss was low but still had overfitting.

Supplementary Table 3. Losses for Various Transfer Learning Models

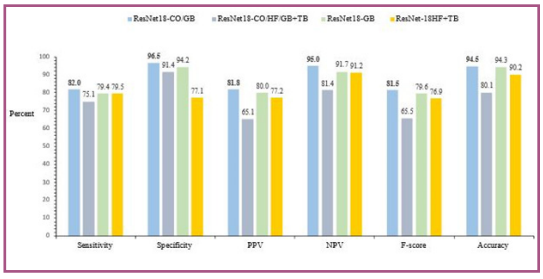
Model	Validation Loss	Training Loss	Loss Difference	Overfitting
ResNet18-TB	0.59	0.61	-0.02	No
ResNet18-HF+TB	0.44	0.42	0.02	Yes
ResNet18-GB+TB	0.56	0.58	-0.02	No
ResNet18-HF/GB+TB	0.57	0.60	-0.03	No
ResNet18-CO/GB+TB	0.47	0.40	0.07	Yes
ResNet18-CO/HF/GB+TB	0.27	0.17	0.10	Yes
ResNet18-PN	0.63	0.59	0.04	Yes
ResNet18-HF+PN	0.63	0.56	0.07	Yes
ResNet18-GB+PN	0.39	0.39	0.00	No
ResNet18-HF/GB+PN	0.56	0.54	0.02	Yes
ResNet18-CO/GB+PN	0.62	0.61	0.01	Yes
ResNet18-CO/HF/GB+PN	0.64	0.61	0.03	Yes

CO=cutout, GB= Gaussian Blur, PN=Pneumonia. TB= tuberculosis

3.7 Final Model Optimization and Performance Metrics

At 1,000 iterations, the best overall performing model was ResNet18-CO/GB and the second-best model was ResNet18-CO/HF/GB+TB. Figure 4 demonstrates the performance characteristics of the top models at 10,000 iterations. ResNet18-CO/GB remained the best performing model at 10,000 iterations. For the final model, sensitivity was 82.0%, specificity was 96.5%, and accuracy was 94.5%. ResNet18-GB was the second-best model with a sensitivity of 79.4%, specificity 94.2%, and accuracy of 94.3%. The best transfer learning model at 10,000 iterations, ResNet18-HF+TB had a similar F-score and accuracy but its performance was adversely impacted by low specificity (77.1%).

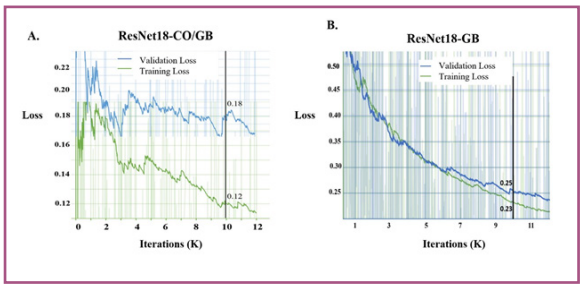
Figure 4. Performance Characteristics of Best Performing Models at 10,000 iterations



Sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), F-score, and accuracy are shown for the validation set, with top performer by category in **bold**

All four models demonstrated mild overfitting at 10,000 iterations, ranging from 0.02 to 0.07, with relatively low validation loss ≤0.25. Figure 5 demonstrates the loss characteristics for the top two models. Validation and training losses were least for the ResNet18-CO/GB model at 0.18 and 0.12 respectively, with loss difference of 0.06 consistent with mild overfitting. The second-best performing model, ResNet18-GB demonstrated higher relative validation and training losses of 0.25 and 0.23, respectively with minimal overfitting (loss difference of 0.02).

Figure 5. Validation and Training Loss Characteristics for the Top 2 Models

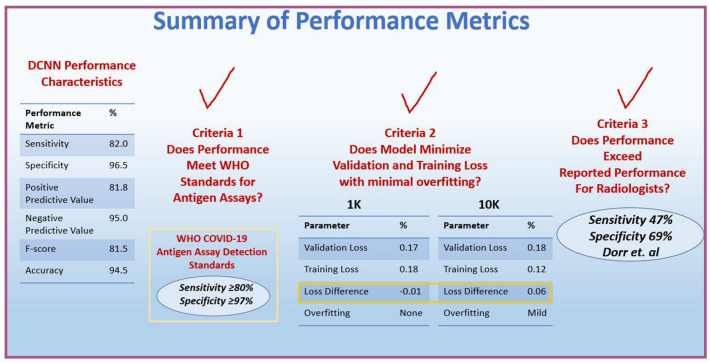


A. Top performing Model B. Runner-up Model at 10K iterations. Validation loss is shown in blue. Training loss is shown in green

4. Discussion

In this study, we developed a robust deep learning CNN model for detecting COVID-19 by CXR that meets a variety of metrics, as outlined in the central diagram in Supplementary Figure 2-- including World Health Organization target standards for COVID antigen tests. After testing 24 model variations, our final one, ResNet18-CO/GB demonstrated robust performance characteristics with a sensitivity 82.0% and accuracy of 94%, both of which are comparable to existing COVID-19 antigen assays in clinical use. Although ours may not supersede some previously published COVID-19 CXR DCNN detection models, this study has the advantage of using the largest open access CXR datasets available- underscoring the potential clinical relevance of our model. Importantly, this model demonstrated performance characteristics that likely exceeds the ability of clinicians to detect COVID-19 by CXR alone without assistance, previously reported at sensitivity 47% and specificity 79% [6].

Supplementary Figure 2. Summary Diagram



Performance characteristics of the model to key metrics [6,12].

In the process of model development, several key findings emerged. First, low complexity ResNet18 was the optimal architectural prototype for COVID-19 CXR classification, with better performance characteristics than more complex CNN's. Second, data augmentation was critical to augmenting DCNN performance. All three image transformations, random HF, GB, and CO, enhanced model performance in various combinations. Stacking image transformations provided the most effective enhancement strategy- yielding ResNet18+CO/GB as the optimal model. Finally, transfer learning was of little benefit as a standalone CNN feature, and was best used in combination with data augmentation. Importantly, although transfer learning appeared promising when models were run with fewer iterations, these models exhibited declining performance at 10,000 iterations and ultimately, did not perform as well as data augmented models.

4.1 Backbone Architecture Impacts Model Performance

Using ResNet18 as architectural underpinning was an important feature of this model. Previous studies reported Wideset and VGG-16 as having better performance characteristics than ResNet18 for COVID CXR detection models [25]. However, we found that the less complex CNN's like Densenet and ResNet18 had higher specificity and F-score. There was also a strong correlation between CNN complexity and validation loss in COVID-19 detection, with ResNet having the least validation loss and no overfitting to suggest an overtrained model. Structurally, whereas the back propagation technique can cause VGG16 and AlexNet's stagnate with progressive iterations, ResNet resolves the so-called vanishing gradient problem and has the advantage of other architectural solutions that lend depth, like shortcut or skip connections.

4.2 Data Augmentation Boosts Model Performance

This study supports the longstanding theory that data augmentation is critical to drive ML performance and translates this to CXR detection of COVID-19. All three image transformations- GB, HF, and CO- either alone or in combination improved the ResNet18 model. GB, in particular was effective standalone feature; this observation in a COVID-19 model extends previous findings that GB improved deep-learning image classification accuracy of non-COVID pathologies by CXR's by 0.05% [26]. The present study found that data augmentation methods were not necessarily additive in their effects. For example, the combined use of HF/GB improved did not improve on the ResNet-GB model even though each feature individually boosted ResNet18 performance. Similarly, CO alone did not boost ResNet18 performance, whereas the CO/GB combination yielded the most robust model. This observation underscores the importance of experimenting with different combinations of image transformations.

4.3 Transfer Learning Variable Effects on CNN Performance

Another unique aspect of this study is that the experimental design teases out the use of different types of instance-based transfer learning and the relative impact of stacking data augmentation effects on transfer learning models. For example, tuberculosis transfer learning improved multiple models, whereas using a pneumonia transfer learning dataset did not. Interestingly, TB transfer learning, as a standalone feature, did not boost the performance of the ResNet18 model. Instead, it worked best in combination with data augmentation features. The best performing transfer learning model resulting from combining all three image transformations (CO, GB, and HF) prior to transfer learning. This is consistent with the previous finding of Zhang et al. who reported good results in their COVID-19 CXR detection model, using a ResNet 34 model, with multiple image transformations (random resized crop, rotation, horizontal flip and vertical flip) in combination with transfer learning [9].

Ultimately, in this study, however, the best transfer learning models did not perform as well as the best data augmented models. It is possible the architectural backbone selected may impact the value of transfer learning. In a comparison of 15 different CNN architectures, Rahaman et al. concluded that VGG-19 works best with transfer learning for COVID-19 detection [27]. Another contributor might be that transfer learning models appeared to demonstrate declining performance with progressive iterations beyond 1,000 suggesting that these models may be prone to overfitting. Successful application of transfer learning in COVID CXR detection may be a function of multiple variables,

e.g., a CNN size, supervision features, the relevance of the transfer learning dataset, and the size of the datasets involved; further investigations are required to determine under what circumstances it would enhance future COVID-19 CXR models.

4.4 Minimizing Validation Loss, Training Loss, and Overfitting

In addition to identifying a robust COVID-19 CXR detection model, this study uniquely focused on the problem of validation loss and overfitting that are common to small datasets. Deep learning models perform best with training by large data sets. The final model ResNet-CO/GB had the lowest validation loss of all the models consistent with its robustness. However, we did identify some mild overfitting when the model was run for 10,000 iterations. This can be addressed in a few ways including terminating the model earlier, enlarging the training dataset, or adding other data augmentation techniques. Nonetheless, this model provides a foundation for future model development by addressing the question of which enhancements work best

4.5 Limitations

This study has its limitations. First, the model is based on open-source data, so the methodology of chest X-ray acquisition and the clinical stage of the COVID is not known. Second, the model is designed to assess for COVID-19 in a binary fashion. In the situation where a patient has pathology that is not COVID-19, additional assessment may be required. Such binary decision models still benefit clinical practice by providing a meaningful screen but may require physician overread or incorporation into a more complex CXR interpretation model, such as CHEXNet which is designed to detect other pathologies [28]. It remains to be seen how this model compares to unassisted radiologists or impacts clinical workflow in prospective study.

5. Conclusion

In this study, we developed a robust, relatively straightforward model of COVID-19 CXR detection with sensitivity 82%, specificity 96.5%, accuracy 94.5%. These performance characteristics exceed previously reported physician's ability to detect COVID-19 by CXR without AI and are comparable to existing World Health Organization standards for COVID-19 antigen assays in clinical use. This study added value to existing literature by exploring the impact of various CNN facets and enhancement in terms of performance and validation loss. Future directions would be to further assess other data augmentation strategies, to consider the use of different architectural backbones with transfer learning, and to assess this model performance relative to clinicians.

6. Acknowledgements, Source of Funding and Disclosures

I would like to thank Dr. Parsa Akbari for his valuable mentorship in supervising this research and in manuscript review. This independent research was performed without a funding source. I have no disclosures. Conflicts of interest: none to report.

References

1. Ensheng Dong, Hongru Du, and Lauren Gardner. An interactive web-based dashboard to track COVID-19 in realtime. Lancet Infect. Dis., 20(5):533–534, May 2020.

2. Options for the use of rapid antigen tests for COVID-19 in the EU/EEA and the UK.https://www.ecdc.europa.eu/en/publications-data/options-use-rapid-antigen-tests-covid-19-eueea-and-uk, November 2020. Accessed: 2021-10-9.

3. Sergii Stirenko, Yuriy Kochura, Oleg Alienin, Oleksandr Rokovyi, Yuri Gordienko, Peng Gang, and Wei Zeng.Chest X-Ray analysis of tuberculosis by deep learning with segmentation and augmentation. In2018 IEEE 38thInternational Conference on Electronics and Nanotechnology (ELNANO), pages 422–428, April 2018.

4. as4401s.GitHub - as4401s/COVID-19-X_ray-image classification. https://github.com/as4401s/COVID-19-X_ray-image-classification. Accessed: 2021-9-26.

5. Catherine M Jones, Quinlan D Buchlak, Luke Oakden-Rayner, Michael Milne, Jarrel Seah, Nazanin Esmaili, et al. Chest radiographs and machine learning - past, present and future.J. Med. Imaging Radiat. Oncol.,65(5):538–544, August 2021.

6. Francisco Dorr, Hernán Chaves, María Mercedes Serra, Andrés Ramirez, Martín Elías Costa, Joaquín Seia, et al. Covid-19 pneumonia accurately detected on chest radiographs with artificial intelligence. Intelligence-based medicine, 3:100014, 2020.

7. Tulin Ozturk, Muhammed Talo, Eylul Azra Yildirim, Ulas Baran Baloglu, Ozal Yildirim, and U Rajendra Acharya.Automated detection of COVID-19 cases using deep neural networks with x-ray images. Comput. Biol. Med.,121:103792, June 2020.

8. Ioannis D Apostolopoulos and Tzani A Mpesiana. Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks. Australas. Phys. Eng. Sci. Med., 43(2):635–640, June 2020.

9. Elene Firmest Ohata, Gabriel Maia Bezerra, Joao Victor Souza das Chagas, Aloisio Vieira Lira Neto, Adri-ano Bessa Albuquerque, et al. Automatic detectionof COVID-19 infection using chest x-ray images through transfer learning. IEEE/CAA Journal of AutomaticaSinica, 8(1):239–248, January 2021.

10. Aras M Ismael and Abdulkadir ¸Sengür. Deep learning approaches for COVID-19 detection based on chest x-ray images. Expert Syst. Appl.,

164:114054, February 2021.

11. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. International journal of computer vision, 115(3), 211-252.

12. World Health Organization. Antigen-detection in the diagnosis of sars-cov-2 infection. October 2021

13. Linda Wang, Zhong Qiu Lin, and Alexander Wong. COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest x-ray images.Sci. Rep., 10(1):19549, November 2020.

14. Tawsifur Rahman. Tuberculosis (TB) chest x-ray database.

15. Daniel S, Michael Goldbaum, Wenjia Cai, Carolina C S Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, Justin Dong, et al. Identifying medical diagnoses and treatable diseases by Image-Based deep learning. Cell, 172(5):1122–1131.e9, February 2018.

16. Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. December 2014.

17. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neuralnetworks. Commun. ACM, 60(6):84–90, May 2017.

18. Qassim, H., Feinzimer, D., and Verma, A. (2017). Residual squeeze vgg16. arXiv preprint arXiv:1705.03004.

19. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. December 2015.