Detection of COVID-19 from Chest X-rays Using Convolutional Neural Networks

Garrett Yoon Yifan Li

December 7, 2021

1 Purpose

To compare the performance of different methodologies featuring machine learning and deep convolutional neural networks (CNN) to correctly detect COVID-19 pneumonia from chest x-ray images.

2 Introduction

The novel coronavirus is one of these most challenging problems for the modern world. Although often individuals infected with COVID-19 experience mild-to-moderate respiratory illness, some patients develop life-threatening complications (Wiersinga et al., 2020). An estimated 3 million people have died from COVID-19. Chest x-rays (CXR) are one of the most important and essential tools to detect visual responses to SARS-COV-2 infection, including pneumonia. Due to the limited number of expert radiologists, the usage of deep neural networks models to aid the diagnosis process have been studied to help ease the burden on the healthcare infrastructure. Researchers from around the world have built different models to label COVID-19 in CXR images, using convolutional neural networks (CNN). This approach allows for transfer learning to be successfully incorporated in many applications, especially where large amounts of data can be hard to find (Christodoulidis et al., 2017) and may reduce the time required to develop and train deep learning model from scratch.

3 Related Work

Currently many biomedical health problems and complications are using artificial intelligence-based solutions (Esteva et al., 2019). Especially, convolutional neural networks (CNN) have been proven to be effective in feature extraction and later applied into machine learning. (Bluche et al., 2013). For example, CNNs have enhanced the image quality of medical imaging and aid the diagnosis of pulmonary nodules and pediatric pneumonia (Choe et al., 2019). Transfer learning technique has drastically eased the process by allowing quickly training a deep learning framework with limited number of images. Rajpurkar et al reported a 121-layer CNN (CheXNet) on chest X-rays to detect 14 different pathologies, including pneumonia using an ensemble of different networks (Rajpurkar et al., 2017). Additionally, a pre-trained convolutional neural networks have been used as feature extraction techniques in conjunction with machine learning models to correctly identify chest x-ray pathology (Ho and Gwak, 2019). Recently, several groups have reported deep machine learning techniques using X-ray images for detecting COVID-19 pneumonia. For COVID-19 detection, the following CNNs were most commonly used: ResNet-18, ResNet-50, ResNet-101 , DenseNet-201 (Huang et al., 2017), ChexNet/DenseNet-121 (Rajpurkar et al., 2017), and InceptionV3 (Szegedy et al., 2016).

4 Hypothesis

Several works report the ability of transfer learning in the detection of COVID-19 from chest x-ray imaging. Ioannis et al. (B.Sundaram, 2017) reported transfer learning approach for classifying dataset of 1427 X-ray images containing 224 COVID-19, 700 Bacterial Pneumonia and 504 Normal x-ray images with accuracy, sensitivity, and specificity of 96.78%, 98.66%, and 96.46% respectively. CoroNet, based on the residual network (ResNet) architecture, achieved a high accuracy predicting COVID-19 in a 4-class classification problem. Since publication, this dataset has had additional images and an added class, so we would like to investigate how deep learning models perform on it compared to the original study. We predict that a classification model can be built with high performance to delineate COVID-19 from other similar lung pathologist using deep CNNs. We expect that transfer learning will perform better than using models with randomly initialized weights.

5 Dataset

We used the Kaggle COVID-19 Radiography Database compiled by a team of researchers from University of Qatar and University of Dhaka consisting of 21,885 frontal-view X-ray images. COVID-19 chest x-ray images were compiled from multiple sources including The Italian Society of Medical Interventional Radiology (SIRM) COVID-19 Database, The Novel Coronavirus 2019 Dataset, and multiple publications. At the time of writing, the database consisted of COVID-19 images, normal chest x-rays, lung opacity/non-COVID lung infection images, and viral pneumonia images. The number of images are shown in Table 1. Images were resized to 224 x 224 pixels and rendered in Portal Network Graphics file format. Data was split in a ratio of 60% training, 20% validation, and 20% test.

Class	Number of Images
Normal	10912
Lung Opacity	6012
COVID-19	3616
Viral Pneumonia	1345

Table 1: **Number of images per class label.**

6 Materials and Methods

The detection problem is a multiclass classification problem where the input is a front-view chest x-ray and the output is a label of one of the following: Normal, COVID-19, Lung Opacity, or Viral Pneumonia. First, we tried using the pre-trained neural networks as feature extraction mappings in combination with machine models, specifically. XGBoost and Random Forest. We then used established neural network architectures both with pre-trained weights and training from scratch. We then visualized selected COVID-19 images using class activation maps (Zhou et al., 2016).

6.1 Feature Extraction and Machine Learning

Two well known tree-based machine learning methods used are XGBoost and Random Forest. XGBoost stands for eXtreme Gradient Boosting, which is an implemention of gradient boosted decision trees designed for speed and performance that has dominated competitive machine learning (Brownlee, 2021). Random forest has a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest outputs a class prediction and the class with the most votes becomes the final prediction (Yiu, 2019).

We used pre-trained ResNet-18, ResNet-50, and DenseNet-121 convolutional neural network models as feature extractors. We intialized the convolutional part of each model up to the last fully-connected layer. We then run the model on our training and validation data images once, recording the output of the last pooling layer as "bottleneck features". The bottleneck features are a low-dimensional vector that can act as a visual representation of each image. Using these feature vectors as input to a machine learning model can significantly reduce training time compared to retraining and fine-tuning CNN models. Using the bottleneck features of each model, we then trained a model using machine learning classifiers Random Forest and XGBoost to correctly predict the image as one of the four classes.

6.2 CNN Model Architecture

Three different CNN models were validated and tested in this study: ResNet-18, ResNet-50, and DenseNet-121.

Residual neural networks (ResNet) employ skip-connections to allow gradient information to flow through the network directly to avoid the problem of vanishing gradients. ResNet-18 and ResNet-50 are similar in architecture with 16 and 48 layer of convolutional layer respectively along with 1 MaxPool and 1 Average Pool layer.

DenseNets improve the flow of information of gradients through the network through concatenation of subsequent outputs while allowing for feature reuse along with parameter sharing in a feed forward fashion. For each layer, the feature-maps of all preceding layers are used as inputs, and its own feature-maps are used as inputs into the following layers. DenseNets have several compelling advantages: they alleviate the vanishing-gradient problem, strengthen feature propagation, encourage feature reuse, and substantially reduce the number of parameters.

6.3 Class Activation Map

Class activation maps (CAM) are a powerful technique used in computer vision for classification tasks. It allows the scientist to inspect the image to be categorized and understand which parts/pixels of that image have contributed more to the final output of the model. A class activation map for a particular category indicates the discriminative image regions used by the CNN to identify that category (Zhou et al., 2016). It is generated using global average pooling and we use CAM to detect features that might be of importance for classification of chest x-rays. We visualized class activation for COVID-19 images using one of our best performing models.

6.4 Hyperparameter Search and Selection

For each of the CNN models, we performed a hyperparameter search for the following: initial learning rate, optimizer, and drop-out rate. We used a subset of the original dataset for training. We incorporated reduction of learning rate if validation loss did not decrease for a set amount of epochs. Among the initial learning rates from 1E-5 to 1E-3, 1E-3 has the lowest validation loss in the DenseNet models and 1E-4 had the lowest validation loss for the ResNet models after a set number of epochs. We also repeated this for varying batch sizes from 4 to 64. A batch size of 64 resulted in lowest validation loss for all architectures. Stochastic gradient descent (SGD) performed slightly better than Adam. Varying the dropout rate did not affect performance significantly. Based on this search, we trained the networks with a learning rate 1E-3 for DenseNet and 1E-4 for ResNet, using the SGD optimizer with momentum.

6.5 Image Augmentation

Images were imported from the COVID-19 Radiography Dataset. Data transformations included randomly cropping and randomly horizontal flipping. Lastly, the images were normalized based on the means and standard deviation of the ImageNet database.

Model	Learning Rate	Batch Size
ResNet-18	1e-4	64%
ResNet-50	1e-4	64 %
DenseNet-121	1e-3	64%

Table 2: **Best hyperparameters for each model architecture.**

6.6 Model Training

We investigated two different strategies for weight initialization for ResNet-18, ResNet-50, and DenseNet-121. First, we randomly initialized the model weights and thus the model is trained from scratch. Second, we initialize the network with pre-trained weights from ImageNet, where the knowledge is transferred from a different domain and task. We employed the fine-tuning methodology, allowing one or more layers to be retrained. In our experiment, we allowed all the layers of each network to be trained. The loss function we used for the single label classification is the cross-entropy loss function in Pytorch.

The two models were trained using SGD (stochastic gradient descent) with momentum optimizer ($\beta = 0.9$) and mini-batch size of 32. Initial learning rates were set to 0.0001 for ResNet models and 0.001 for DenseNet models. We reduce the learning rate by a factor of 10 if the validation accuracy does not improve after a training epoch. Performance was monitored over training epochs by computing validation loss and accuracy after each epoch. Early stopping of training would occur if validation loss did not decrease for 3 consecutive epochs of training. Maximum number of epochs for training was set to 50. The model with the best validation accuracy in a given epoch was selected as the final model for testing, as a way of a empirical risk minimization.

7 Results

Overall accuracy, precision, recall and F1-score were computed for each methodology. The multi-class classification result was recorded for each combination of neural network architecture and machine learning model. The performance of using pre-trained CNNs in conjunction with traditional machine learning models is reported in Table 1. The best performing model was DenseNet with XGBoost. For any set architecture, XGBoost achieved better prediction metrics compared to random forest.

Model	Accuracy	Precision	Recall	F1-Score
ResNet-18 + RF	73.44	78.86	59.52	64.18
ResNet-18 + XGBoost	74.86	75.84	65.71	69.31
ResNet-50 + RF	74.47	79.17	64.20	68.36
ResNet-50 + XGBoost	80.70	81.85	75.94	78.38
DenseNet-121 + RF	76.28	80.55	66.42	70.96
DenseNet-121 + XGBoost	**82.83**	**84.79**	**79.49**	**81.79**

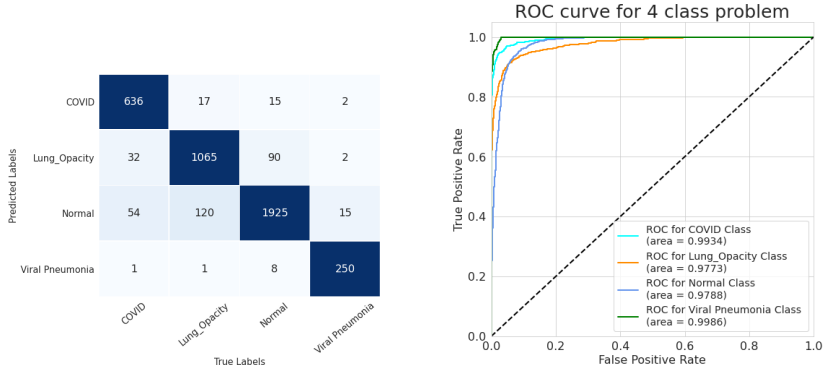
Table 3: **Macro average performance of using feature extraction with machine learning.**

Performance of each trained model via empirical risk minimization is reported in Table 2. Pre-trained models achieved significantly higher performance compared to random weight initialization. All pre-trained models achieved similar performance, with respective macro average F1-Scores of 90.79%, 91.90%, and 91.10% for ResNet-18, ResNet-50, and DenseNet-121 respectively.

The confusion matrix on the test set for the 4-class problem along with per-class ROC curves are shown in Figure 1 for the pre-trained DenseNet-121 model. A sample of COVID-19 images along with its sample class activation map are shown in Figure 2.

Model	Accuracy	Precision	Recall	F1-Score
ResNet-18 + Scratch	78.27	71.91	81.72	75.14
ResNet-50 + Scratch	85.71	84.48	85.33	84.88
DenseNet-121 + Scratch	80.37	73.97	82.15	77.21
ResNet-18 + Pre-Trained	90.72	89.68	92.02	90.79
ResNet-50 + Pre-Trained	91.57	90.67	**92.93**	**91.90**
DenseNet-121 + Pre-trained	**91.80**	**91.10**	92.42	91.74

Table 4: **Comparison of performances across models trained by empirical risk minimization.**



(a) Confusion Matrix (b) ROC Curve

Figure 1: **Performance of pre-trained ResNet-50 model on test set images.** a) Number of correctly identified images for each class is displayed on the diagonal. b) Per-class ROC curves with ROAUC reported.

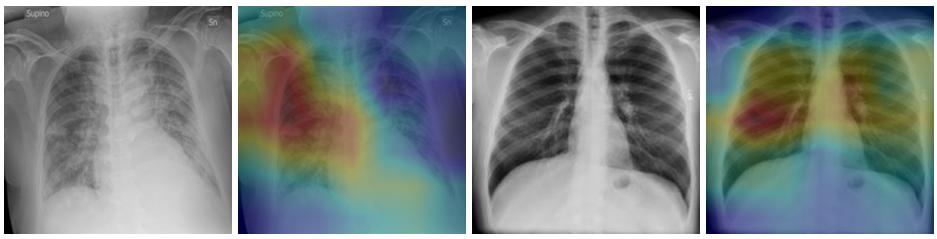


Figure 2: **COVID-19 CXR sample images with accompanying class activation maps.** The maps highlight discriminative image regions used by the DenseNet-121 model for classification. Images with increased artifact may have disrupted class activation maps.

8 Discussion

In this study, we assessed the performance of one method utilizing machine learning and pre-trained deep convolutional neural networks (CNNs) against conventional transfer learning to correctly classify COVID-19 pneumonia from other chest X-rays. We expect that CNNs would perform better than machine learning methods and transfer learning would help the model achieve higher performance.

XGBoost performed on feature vectors extracted from the pre-trained DenseNet-121 model has the highest performance with 82.23% accuracy, 84.79% precision, 79.49% recall and 81.79% F1-score compared to other combinations of CNNs and random forests. There could be a combined advantage of using gradient boosting trees with the concatenated outputs from DenseNets. Together, this method can achieve a good representation of bottleneck features used for the classification of chest x-ray images. Although the performance was much lower than another study which used this method (Wang et al., 2020), we note that the dataset used in that study was much smaller in size than ours and the task was only a

binary classification. Furthermore, each ground-truth label class of their data only came from one source, while our data originated from multiple sources, which may overestimated their performance based on image data distribution.

Transfer learning has been prominently used in the field of medical imaging classification (Rajpurkar et al., 2017). After fine-tuning, pre-trained models classified images in the studied dataset more accurately than models with random weight initialization. Notably, DenseNet-121 and Resnet-50 achieved the best performance metrics. This was expected due to our dataset size and the limited information we had on COVID-19 CXR. Architectures with larger model capacity could capture the image features better than a less deep ResNet while still maintaining gradients. The pre-trained DenseNet-121 achieved the highest accuracy at 91.80% and precision at 91.10%, while pre-trained ResNet50 has the highest recall at 92.93% and a F1-score at 91.90%. Khan et al. performed a similar experiment using transfer learning on a multi-class classification problem, with their 4-class predictive model achieving an accuracy of 89.6% and precision of 90% (Khan et al., 2020). Recall is especially important in medical predictive models, with the low number of false negatives in our study to be reassuring. Of note, the extremely high per-class AUROC of the ResNet model may suggest that model is successful at ranking the class patterns well and differentiating the positive labels, but poorly selects the threshold and has a lower number of true negatives as the accuracy is relatively lower.

Of note, a previous study classified a subset of this database (Chowdhury et al., 2020). Although this study had higher performance that our results, we identify some issues with the work. The authors reported the performance classifying all the data, and did not utilize a correct testing set, resulting in artificially high performance metrics. The authors also used image augmentation to generate new images for a more labelled dataset, however due to the small dataset size, this can result in over-fitting to repeated images. CNNs overall achieved significantly better performance than the machine learning methods. This is expected because CNNs is fully connected feed forward neural networks, which give them the ability to reduce the number of parameters without losing on the quality of the model. CXR have high dimensionalites which is suited for CNNs. We initially thought that the using machine learning methods on the feature vectors extracted from the last layer of CNNs would maintain the dimensionalities of the images and reduce the computational capacity. However, based on the performance differences, the machine learning methods still could not achieve similar performance compared with CNNs.

9 Conclusion

Our results confirmed our hypothesis that CNNs do have the ability to classifying COVID-19 pneumonia from other lung pathologies with well-annotated data. Potentially, deep neural networks can help more non-radiologist identify and diagnose COVID-related symptoms and alleviate the current crisis. One limitation is the requirement for well-annotated data which is hard to come by in medicine. We would like to explore more on the self-supervised learning which is a form of unsupervised learning where the data provides the supervision without the annotation. We would like to try self-supervised learning utilizing contrastive loss on our dataset without the labels to see whether we can produce similar results.

10 Contributions

Garrett Yoon and Yifan Li contributed equally to the development of the CNN models, using transfer learning and machine learning frameworks. Yifan Li wrote the introduction, related works, hypothesis, and methods of this paper. Garrett Yoon generated the results figures and wrote the discussion of this paper.

References

W. Joost Wiersinga, Andrew Rhodes, Allen C. Cheng, Sharon J. Peacock, and Hallie C. Prescott. Pathophysiology, transmission, diagnosis, and treatment of coronavirus disease 2019 (covid-19). *JAMA*, 324(8):782, 2020. doi: 10.1001/jama.2020.12839.

Stergios Christodoulidis, Marios Anthimopoulos, Lukas Ebner, Andreas Christe, and Stavroula Mougiakakou. Multisource transfer learning with convolutional neural networks for lung pattern analysis. *IEEE Journal of Biomedical and Health Informatics*, 21(1):76–84, 2017. doi: 10.1109/JBHI.2016.2636929.

Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, Jeff Dean, and et al. A guide to deep learning in healthcare. *Nature Medicine*, 25(1):24–29, 2019. doi: 10.1038/s41591-018-0316-z.

Théodore Bluche, Hermann Ney, and Christopher Kermorvant. Feature extraction with convolutional neural networks for handwritten word recognition. In *2013 12th International Conference on Document Analysis and Recognition*, pages 285–289, 2013. doi: 10.1109/ICDAR.2013.64.

Jooae Choe, Sang Min Lee, Kyung-Hyun Do, Gaeun Lee, June-Goo Lee, Sang Min Lee, and Joon Beom Seo. Deep learning–based image conversion of ct reconstruction kernels improves radiomics reproducibility for pulmonary nodules or masses. *Radiology*, 292(2):365–373, 2019. doi: 10.1148/radiol.2019181960.

Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning, 2017.

T. K. Ho and J. Gwak. Multiple feature integration for classification of thoracic disease in chest radiography. *Appl. Sci*, 2019.

G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

P.Lakhaniand B.Sundaram. Deep learning at chest radiography:auto- mated classification of pulmonary tuberculosis by using convolutional neural networks,. *Radiology*, 2017.

B. Zhou, A. Khosla, Lapedriza. A., A. Oliva, and A. Torralba. Learning Deep Features for Discriminative Localization. *CVPR*, 2016.

Jason Brownlee. A gentle introduction to xgboost for applied machine learning, Feb 2021. URL `https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/`.

Tony Yiu. Understanding random forest, Aug 2019. URL `https://towardsdatascience.com`.

Dingding Wang, Jiaqing Mo, Gang Zhou, Liang Xu, and Yajun Liu. An efficient mixture of deep and machine learning models for covid-19 diagnosis in chest x-ray images. *Plos One*, 15(11), 2020. doi: 10.1371/journal.pone.0242535.

Asif Iqbal Khan, Junaid Latief Shah, and Mohammad Mudasir Bhat. Coronet: A deep neural network for detection and diagnosis of covid-19 from chest x-ray images. *Computer Methods and Programs in Biomedicine*, 196:105581, 2020. doi: 10.1016/j.cmpb.2020.105581.

Muhammad Chowdhury, Tawsifur Rahman, Amith Khandakar, Rashid Mazhar, Muhammad Kadir, Zaid Mahbub, Khandakar Islam, Muhammad Khan, Atif Iqbal, Nasser Emadi, and et al. Can ai help in screening viral and covid-19 pneumonia? *IEEE Access*, 8:132665–132676, 2020. doi: 10.1109/access.2020.3010287.