CS482/682 Final Project Report Group 6
Predictive Modeling of Covid-19 with Chest X-rays

Keefer Chern/kchern1, Yao Zixuan/zyao5, Ziyi Wang/zwang223, Mou Zhang/mzhan106

1 Introduction

Background Covid-19 is an extremely distressing ongoing pandemic that has negatively impacted the lives of all people globally. One significant issue surrounding the virus is difficulty in getting a diagnosis for the virus, especially in poorer countries without any medical resources. We wish to create an algorithm that will quickly determine whether an individual has been infected by the Covid-19 virus with only an chest X-ray. Development of a working model would help facilitate decision support and help reduce physician error when treating patients.

Related Work Some prior work has been done working utilizing a VGG architecture to model Covid-19 X-rays. They were able to achieve over 80% accuracy while working with three classes of healthy, bacterial, and Covid-19 images. What we hope to accomplish is to utilize different architectures to model the X-ray data while looking at whether we could improve on their results. In addition, the data points they possessed was even smaller than our and did not utilize any methods to compensate for the lack of data.

2 Methods

Dataset The dataset utilized in this project originated from two sources: COVID-19 image data collection and RSNA Pneumonia Detection Challenge on Kaggle. From the data, lateral chest X-rays were filtered out and only the frontal chest X-rays were kept. This was done for the sake of consistency in the dataset when training and eliminate some possible confounding variables. We had in total 3 tasks: Task 1 (100 normal and 100 COVID 2 classes classification), Task 2 (75 normal, 75 COVID and 75 other pneumonia 3 classes classification), Task 3 (75 normal, 75 COVID, 75 bacterial pneumonia and 75 viral pneumonia 4 classes classification). For each task, we manually filter and use different datasets, because we have different experimental purposes. It is also summarized in Table 1: Ablation Table. In addition image augmentation is performed by having the images rotated when training.

Setup, Training and Evaluation We utilized three architectures in this project. The first architecture was VGG-16. It takes a set of X-ray images (batch size = 8) as input and generates a classification prediction of the X-Ray images. It has 13 convolutional layers, 5 max pooling layers and 3 dense layers. The weights of the first 13 convolutional layers are fixed using transfer learning technique. Then, the convolutional layer features from each slice of the X-rays are combined by an average-pooling operation and fed to the last 3 dense layers, which are trained on our own Chest X-Ray dataset.
The second architecture was ResNet 101. The reason why we decided to utilize this architecture was the inspiration of the related work and how it could be improved upon. It also takes a series of X-rays as input and generates a classification prediction of the X-Ray image. The convolutional layer features from each slice of the X-rays are combined by an average-pooling operation and the resulting feature map is fed to the fully connected layers to generate a probability score for each class. Both the VGG and ResNet have been pretrained on ImageNet. Since we are using transfer learning, a new top layer was constructed and trained on our Chest X-Ray dataset, referring to Figure 1. In addition, we would be able to analyze how VGG and ResNet would learn the X-ray images differently.

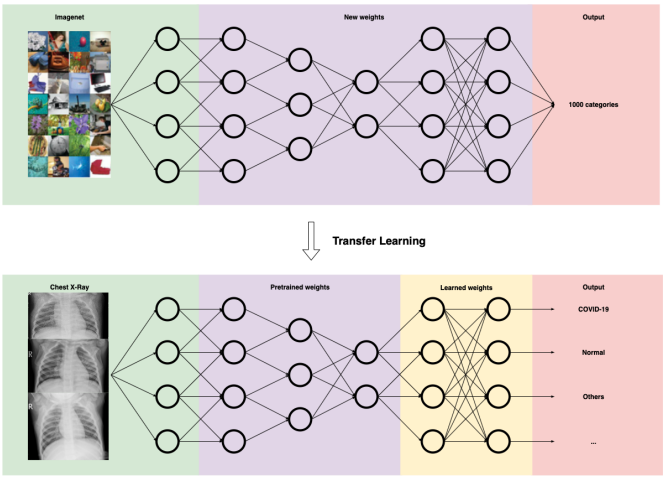

Figure 1: Transfer Learning

Model	Task & Dataset	Accuracy(80/20)	Cross Validation
ResNet 101 (transfer learning)	Task 1 (100 normal : 100 COVID)	90%	89.5%
VGG 16 (transfer learning)	Task 1 (100 normal : 100 COVID)	97.5%	97.5%
CNN	Task 1 (100 normal : 100 COVID)	100%	95.0%
ResNet 101 (transfer learning)	Task 2 (75 normal : 75 COVID : 75 other pneumonia)	65%	63.1%
VGG 16 (transfer learning)	Task 2 (75 normal : 75 COVID : 75 other pneumonia)	93%	91.5%
CNN	Task 2 (75 normal : 75 COVID : 75 other pneumonia)	91%	87.1%
ResNet 101 (transfer learning)	Task 3 (75 normal : 75 COVID : 75 bacterial : 75 viral)	65%	48.7%
VGG 16 (transfer learning)	Task 3 (75 normal : 75 COVID : 75 bacterial : 75 viral)	79%	80.3%
CNN	Task 3 (75 normal : 75 COVID : 75 bacterial : 75 viral)	73%	80.3%

Table 1: Ablation Table

The last architecture is a self-constructed CNN. The main motivation for this architecture was to see whether X-ray images can be modeled in a simpler architecture compared to VGG and ResNet. After multiple test, we arrived at an architecture that provided great training speed and accuracy. Our optimal architecture has 3 convolutional layers, 3 max pooling layers and a dropout, uses ReLu as activation function.

Initially, we have a 80/20 split for training and testing the models. In addition, due to the lack of data, we perform a 5-fold cross validation to compensate. To be specific, we divided all the image data into 5 folds and run the experiment for 5 times. Each time we use a different fold as validation set (the same as test set here) and other 4 folds as training sets to get an accuracy. After conducting all 5 experiments, we calculate the average accuracy of these 5 experiments as the final accuracy.

3 Results

Our results are organized in Table 1: Ablation Table. The table contains the accuracy from the 80/20 split as well as the cross validation runs. In addition, we provide detailed statistics and graphs for the 80/20 split training in Appendix, to show detailed performance and how accuracy change through training process. Overall, the accuracy we obtained were much higher than initially expected.

In the 2 classes 80/20 split, our self-constructed CNN performed the best (100%), followed by VGG (97.5%), and ResNet performed the worst (90%). In the 3 classes 80/20 split, ResNet continually performed the worst (65%), while VGG performed the best (93%), and our self-constructed CNN performed in between (91%). Lastly in the hardest 4 classes case, VGG performed the best (79%), followed by our self-constructed CNN (73%), and followed by ResNet (65%). Our cross-validation presented hopeful results, where the accuracy were not drastically changed. Details are shown in the Ablation Table.

Notice that the recall rate of both VGG and CNN are extremely high (93% or above, shown in Appendix) in all tasks even on this small dataset, which means if the patient does get infected by coronavirus, he/she is very likely to be diagnosed according to our neural network.

4 Discussion

We decided to pursue multiple classes in this project to understand what effects they may have on the model. This is in order to reflect real world settings where pneumonia may not necessarily be from Covid-19, but other bacterial or viral infections. Our results presented an anticipated results where accuracy decreased when we had more classes. This result is probably due to the fact that these Covid-19 is a viral disease and could have certain features that are similar to other pneumonias.

In addition, it is quite apparent that ResNet does not perform well on the provided data. Both ResNet and VGG were trained on ImageNet, which means that the having pre-trained models are not likely the source of errors, since VGG performs so much better. One possible explanation for the lower performance of ResNet may be the lack of data. Both self-build CNN and VGG have relatively simple structure (3 layers and 16 layers respectively) while ResNet has 101 layers. Utilizing cross-validation is a method to compensate for the lack of data in a model. We utilized cross validation to check how our models performed. Overall, the accuracies from the cross-validation runs were quite close to the accuracies from the 80/20 split runs. As such, this result indicates that our models may not be heavily hurt by the lack of data present.

There are many possible future improvements and applications that we can have for this project. The most important improvement is having more data to train on. Although we have attempted to compensate for the late of data with cross-validation, having real data will further help the credibility and accuracy of the models we have developed. In addition, testing more different types of architecture will further facilitate our understanding of applying deep learning into medical diagnosis. Our proposed approach is to develop a interface that would allow both researchers and medical personnel to upload and use our models for prediction. Not only will it help increase data points, but also assist individuals in their task. With the current accuracies that we have achieved with our models, they can be utilized for decision support in the medical field in help determining whether an individual has Covid-19 with a chest X-ray.

References

[1] Detecting COVID-19 induced Pneumonia from Chest X-rays with Transfer Learning. Available: https://towardsdatascience.com/detecting-covid-19-induced-pneumonia-from-chest-x-rays-with-transfer-learning-an-implementation-311484e6afc1

[2] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

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

[4] Narin, A., Kaya, C. and Pamuk, Z., 2020. Automatic Detection of Coronavirus Disease (COVID-19) Using X-ray Images and Deep Convolutional Neural Networks. arXiv preprint arXiv:2003.10849.

[5] Szegedy, C., Ioffe, S., Vanhoucke, V. and Alemi, A.A., 2017, February. Inception-v4, inception-resnet and the impact of residual connections on learning. In Thirty-first AAAI conference on artificial intelligence.

[6] Khan, A., Sohail, A., Zahoora, U. and Qureshi, A.S., 2019. A survey of the recent architectures of deep convolutional neural networks. arXiv preprint arXiv:1901.06032.

[7] Torrey, L. and Shavlik, J., 2010. Transfer learning. In Handbook of research on machine learning applications and trends: algorithms, methods, and techniques (pp. 242-264). IGI Global.

[8] Pan, S.J. and Yang, Q., 2009. A survey on transfer learning. IEEE Transactions on knowledge and data engineering, 22(10), pp.1345-1359.

[9] LeCun, Y., Bengio, Y. and Hinton, G., 2015. Deep learning. nature, 521(7553), pp.436-444.

[10] LeCun, Y., Boser, B.E., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W.E. and Jackel, L.D., 1990. Handwritten digit recognition with a back-propagation network. In Advances in neural information processing systems (pp. 396-404).

[11] Chen, T., Li, M., Li, Y., Lin, M., Wang, N., Wang, M., Xiao, T., Xu, B., Zhang, C. and Zhang, Z., 2015. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. arXiv preprint arXiv:1512.01274.

[12] Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K. and Lungren, M.P., 2017. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:1711.05225.

[13] Zech, J.R., Badgeley, M.A., Liu, M., Costa, A.B., Titano, J.J. and Oermann, E.K., 2018. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS medicine, 15(11).

[14] Team, T.D., 2018. Pneumonia detection in chest radiographs. arXiv preprint arXiv:1811.08939.

[15] Hemdan, E.E.D., Shouman, M.A. and Karar, M.E., 2020. Covidx-net: A framework of deep learning classifiers to diagnose covid-19 in x-ray images. arXiv preprint arXiv:2003.11055.

[16] Li, L., Qin, L., Xu, Z., Yin, Y., Wang, X., Kong, B., Bai, J., Lu, Y., Fang, Z., Song, Q. and Cao, K., 2020. Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT. Radiology, p.200905.

Appendix

COVID-19 image data collection
(https://github.com/ieee8023/covid-chestxray-dataset)
RSNA Pneumonia Detection Challenge
(https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/data)

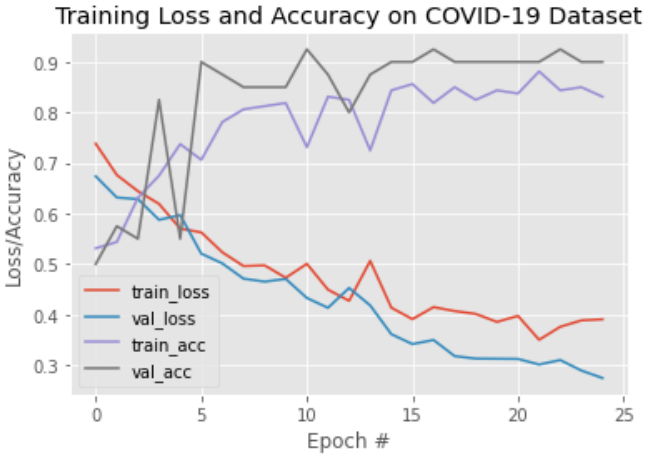

Figure 2: 2 classes ResNet

2 classes ResNet	precision	recall	f1-score	support
covid	0.86	0.95	0.90	20
normal	0.94	0.85	0.89	20
accuracy			0.90	40
macro avg	0.90	0.90	0.90	40
weighted avg	0.90	0.90	0.90	40

Table 2: 2 classes ResNet Evaluation Table

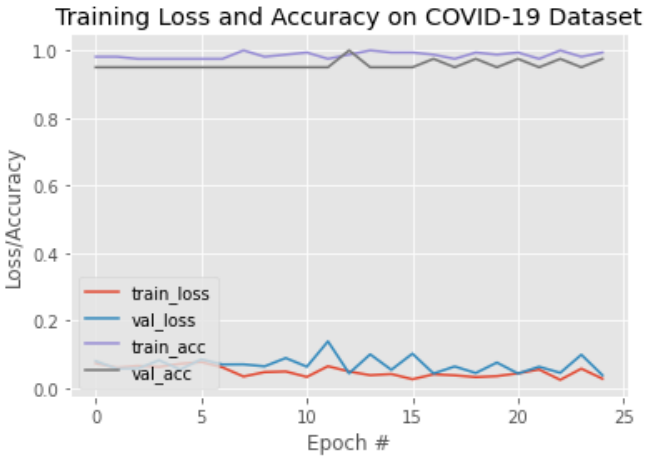

Figure 3: 2 classes VGG

2 classes VGG	precision	recall	f1-score	support
covid	0.95	1.00	0.98	20
normal	1.00	0.95	0.97	20
accuracy			0.97	40
macro avg	0.98	0.97	0.97	40
weighted avg	0.98	0.97	0.97	40

Table 3: 2 classes VGG Evaluation Table

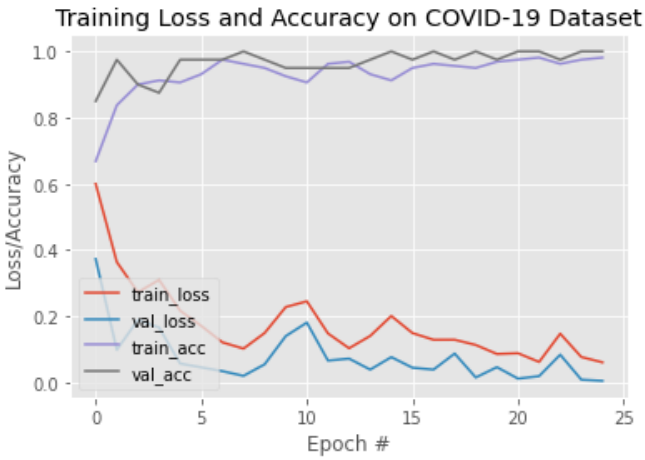

Figure 4: 2 classes CNN

2 classes CNN	precision	recall	f1-score	support
covid	1.00	1.00	1.00	20
normal	1.00	1.00	1.00	20
accuracy			1.00	40
macro avg	1.00	1.00	1.00	40
weighted avg	1.00	1.00	1.00	40

Table 4: 2 classes CNN Evaluation Table

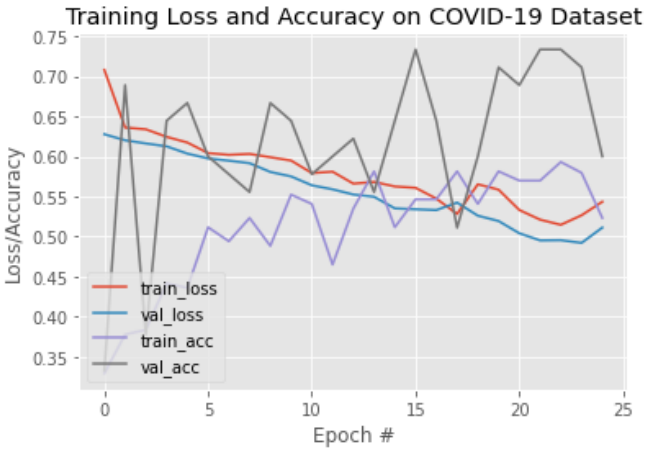

Figure 5: 3 classes ResNet

3 classes ResNet	precision	recall	f1-score	support
covid	0.52	1.00	0.68	15
normal	1.00	0.60	0.75	15
other pneumonia	0.43	0.20	0.27	15
accuracy			0.60	45
macro avg	0.65	0.60	0.57	45
weighted avg	0.65	0.60	0.57	45

Table 5: 3 classes ResNet Evaluation Table

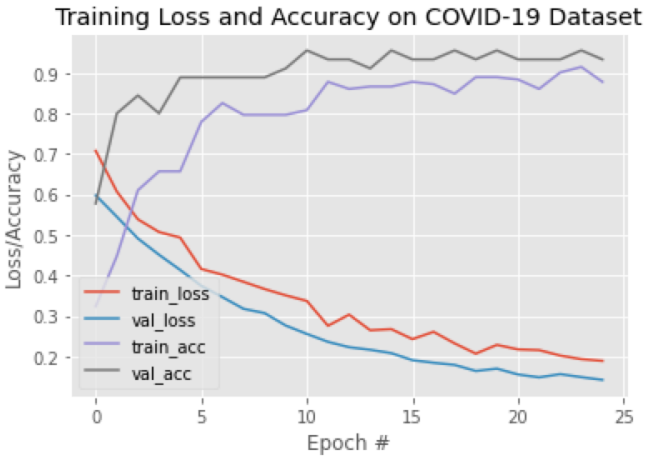

Figure 6: 3 classes VGG

3 classes VGG	precision	recall	f1-score	support
covid	1.00	0.93	0.97	15
normal	0.88	1.00	0.94	15
other pneumonia	0.93	0.87	0.90	15
accuracy			0.93	45
macro avg	0.94	0.93	0.93	45
weighted avg	0.94	0.93	0.93	45

Table 6: 3 classes VGG Evaluation Table

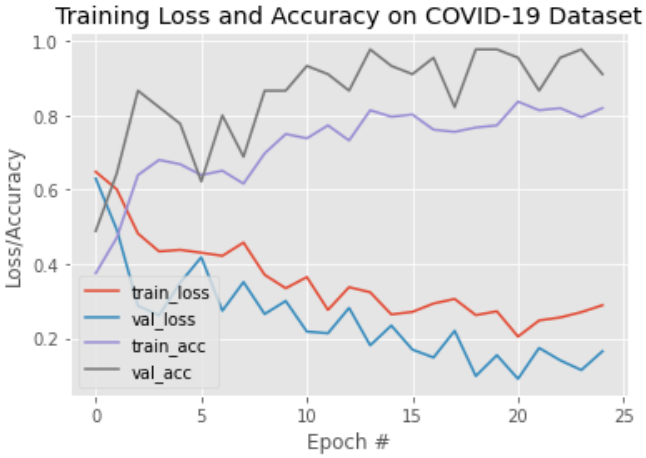

Figure 7: 3 classes CNN

3 classes CNN	precision	recall	f1-score	support
covid	1.00	0.93	0.97	15
normal	0.87	0.87	0.87	15
other pneumonia	0.88	0.93	0.90	15
accuracy			0.91	45
macro avg	0.91	0.91	0.91	45
weighted avg	0.91	0.91	0.91	45

Table 7: 3 classes CNN Evaluation Table

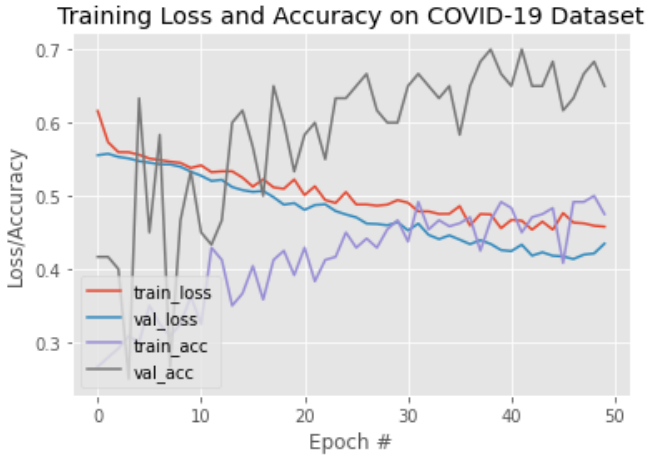

Figure 8: 4 classes ResNet

4 classes ResNet	precision	recall	f1-score	support
covid	0.91	0.67	0.77	15
normal	0.62	1.00	0.77	15
other bacteria	0.40	0.27	0.32	15
other viral	0.67	0.67	0.67	15
accuracy			0.65	60
macro avg	0.65	0.65	0.63	60
weighted avg	0.65	0.65	0.63	60

Table 8: 4 classes ResNet Evaluation Table

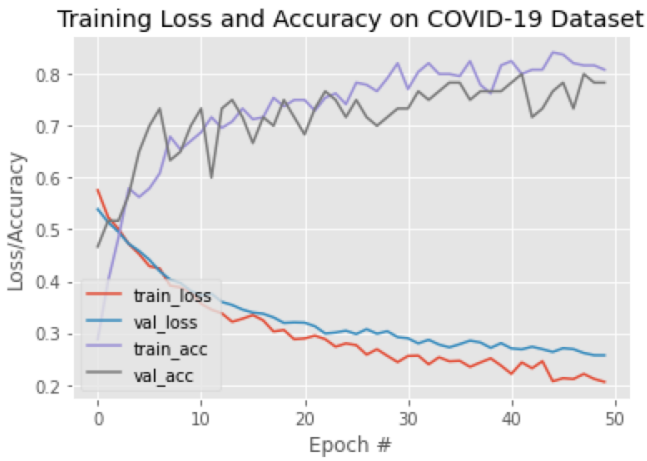

Figure 9: 4 classes VGG

4 classes VGG	precision	recall	f1-score	support
covid	0.83	1.00	0.91	15
normal	0.70	0.93	0.80	15
other bacteria	0.82	0.60	0.69	15
other viral	0.82	0.60	0.69	15
accuracy			0.78	60
macro avg	0.79	0.78	0.77	60
weighted avg	0.79	0.78	0.77	60

Table 9: 4 classes VGG Evaluation Table

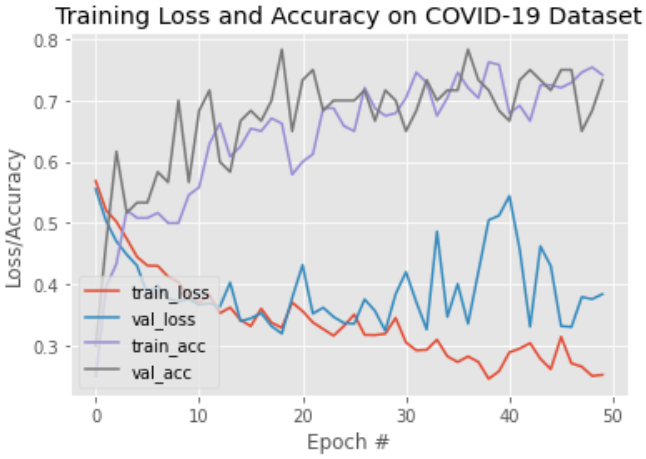

Figure 10: 4 classes CNN

4 classes CNN	precision	recall	f1-score	support
covid	0.82	0.93	0.87	15
normal	0.80	0.80	0.80	15
other bacteria	0.69	0.60	0.64	15
other viral	0.60	0.60	0.60	15
accuracy			0.73	60
macro avg	0.73	0.73	0.73	60
weighted avg	0.73	0.73	0.73	60

Table 10: 4 classes CNN Evaluation Table