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 [1]. 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. The Covid-19 and diseased X-ray images originated from COVID-19 image data collection [2]. The healthy chest X-ray images originated from the RSNA Pneumonia Detection Challenge on Kaggle [3]. 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 data set when training and eliminate some possible confounding variables. 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 series of X-rays as input and generates a classification prediction of the X-Ray image. 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, which we decided to use to have more variety of transfer learning models. The reason we decided to utilize this architecture was due to 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. The last architecture is a self constructed CNN, the main motivation for this architecture was to see whether

X-ray image data can be modeled in a simpler architecture compared to VGG and ResNet. After multiple test, our optimal architecture has 3 convolutional layers, max pooling, dropout, and a ReLu activation function. The self constructed CNN utilized custom layer design. After multiple tests and changes, we arrived at an architecture that provided great training speed and accuracy. 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.

3 Results

Our results are organized in the following ablation tables. The tables contain the accuracy from the 80/20 split as well as the cross validation runs. In addition we provide graphs for the 80/20 split training, to see how accuracy changes through the process of training. The accuracy we obtained from our models were much higher than initially expected. In the 2 class 80/20 split, our self-constructed CNN performed the best(100%), followed by VGG(97.5%), and ResNet performed the worst (90%). The the 3 class 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 class 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 our accuracies were not drastically changed. In the case of our self-constructed CNN, the 2 class(95%), 3 class(87.1%), and 4 class(80.3%) models had accuracies quite close to the 80/20 split. For VGG, the 2 class(97.5%), 3 class(91.5%), and 4 class(80.3%) models had accuracies quite close to the 80/20 split. For ResNet, the 2 class(89.5%), 3 class(63.1%), and 4 class(48.7%) models had accuracies quite close to the 80/20 split. Notice that the recall rate of both VGG and CNN are extremely high (93% or above) 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 the more classes we had. 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 pneumonia diseases. 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.

Model	Description	Accuracy	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

Appendix

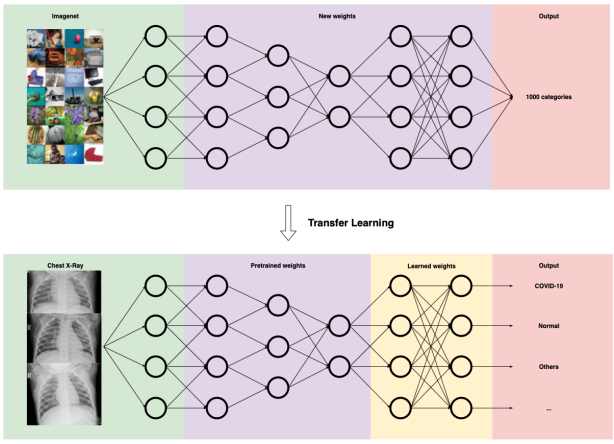

Figure 1: Transfer Learning

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

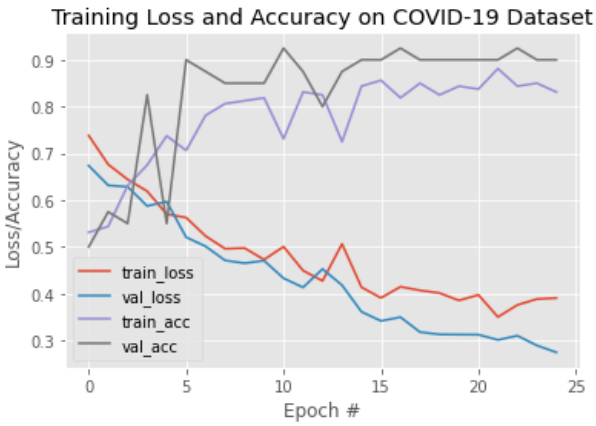

Figure 2: 2 classes ResNet

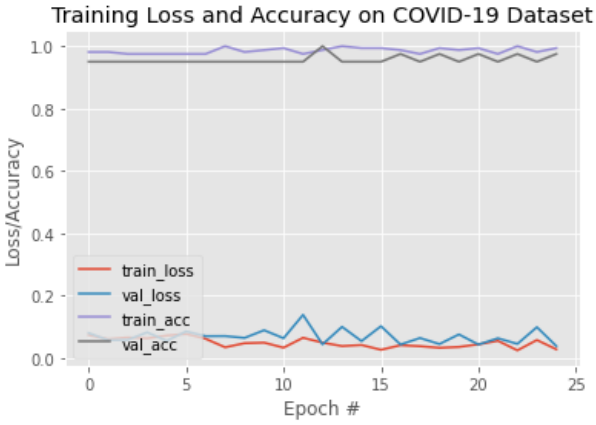

Figure 3: 2 classes VGG

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

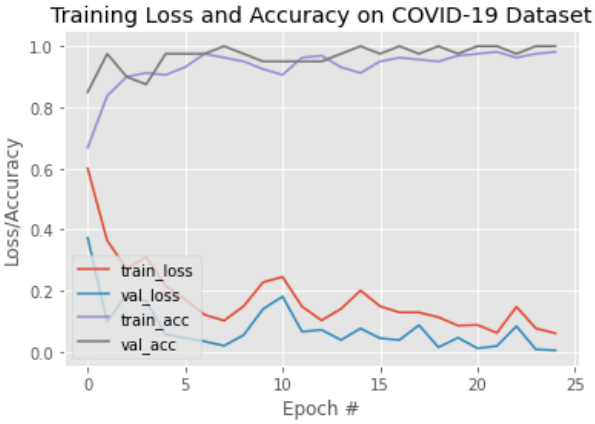

Figure 4: 2 classes CNN

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

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

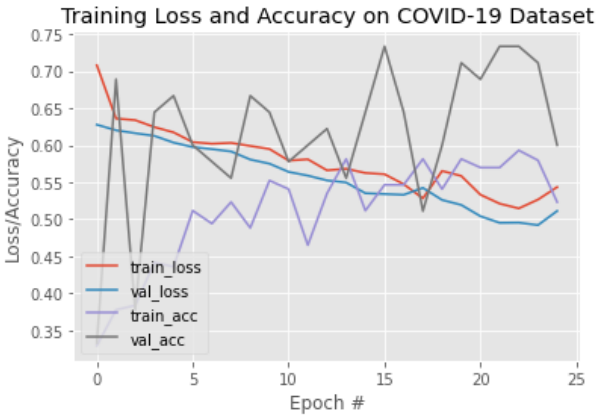

Figure 5: 3 classes ResNet

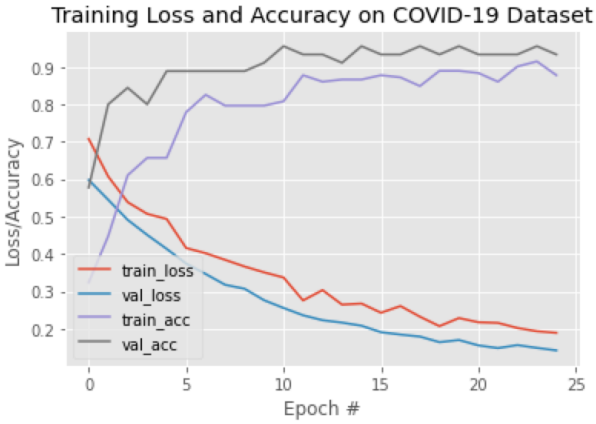

Figure 6: 3 classes VGG

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

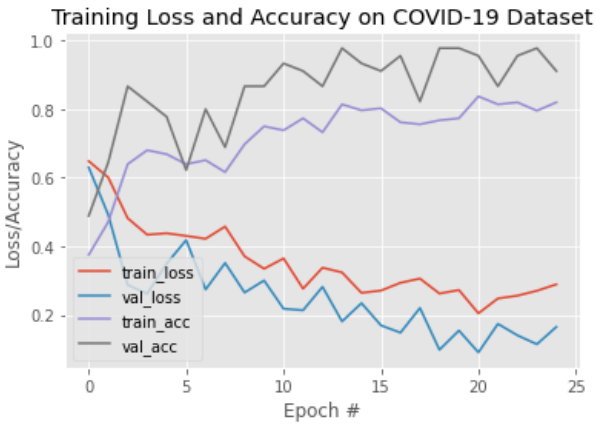

Figure 7: 3 classes CNN

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

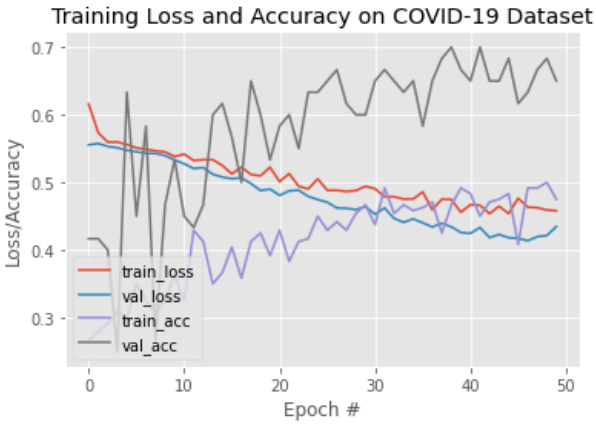

Figure 8: 4 classes ResNet

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

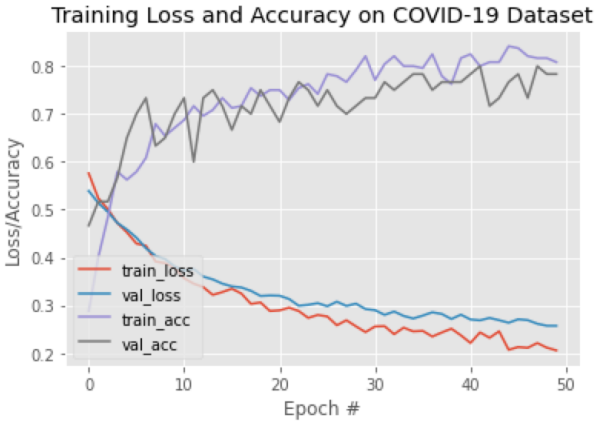

Figure 9: 4 classes VGG

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

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

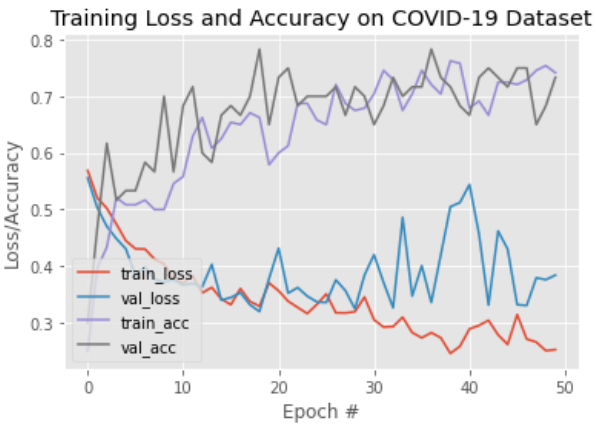

Figure 10: 4 classes CNN