1 Team

Team Members	Rodrigue de Schaetzen, Youssef Farag, Philippe Solodov
Kaggle Team Name	SailCuresCOVID

2 Introduction

The COVID-19 X-ray classification problem consists of training an image classifier with a highly limited sample size. Whereas in the MNIST classifier several thousand images were made available for training, this problem comprises of a mere 70 labeled images. Additionally, the distribution of the image classes is highly skewed which means the baseline model outputs an accuracy of 82%.

3 Method

3.1 Data Pre-processing

To reduce some of the computational load, the 3 channel images were converted to single-channel with values in the range of [0, 1]. Next, we applied a series of transformations to augment our dataset and improve the robustness of the model. Specifically, during training, a random transformation from a fixed set is selected and applied to input images. During validation, every transformation from this set is applied to the input such that the final prediction is an average across all transformations of this image. This distinction between the validation and training augmentations is found in our *dataset_classes.py*.

3.2 Pipeline

The pretrained image classifier **resnet18** was used as our base model. It should be noted that we modified the first and final layer of the network to tweak the model to this task. As a result, it is uncertain how useful the pretrained network was considering some of the weights were lost due to this modification. Nevertheless, our results suggested this network was a suitable model architecture for the COVID-19 X-ray classification problem.

To overcome the lack of data, we implemented the leave-one-out cross validation technique. For each of the 70 training images, a different model was trained whose validation set consisted of only that image. Recall, one random transformation is applied to each training image and all 12 transformations are applied to the validation set. Following training, all 70 models became part of our aggregated model. The ensemble model is evaluated by calculating the average accuracy of the 70 models tested on their validation set. Finally, predictions on the test set are generated by outputting the mean score of the aggregated model.

4 Experiments

The pipeline discussed in the previous section became our first submission on kaggle and achieved a score of 0.80000. Approaches to improving the classifier were built off of this base model.

After implementing loss and accuracy monitoring to the pipeline, we were quickly able to optimize the learning rate. A learning rate scheduler monitoring validation loss was added, though this seemed to have a negligible affect on experiments. In addition, we modified the training phase to return the model with the lowest validation loss instead of the model outputted by the last epoch. This allowed us to significantly increase the number of epochs without having to worry about overfitting, while also returning the best model from the training regime. Various experiments strongly suggested our pipeline had considerably improved following this modification. Given that benign examples are underrepresented, we also decided to weight the loss to disproportionately penalize misclassifying benign examples. This decision was also motivated by the fact that the submission result would be determined by a weighted mean of precision and recall. We found a loss weight of 0.4 using the Pytorch BinaryCrossEntropy loss to produce the best result. All techniques discussed above helped achieve a kaggle score of 0.96551.

Our final approach consisted of running a series of experiments testing the loss weight and the effect of a learning rate scheduler. All experiments produced the same final predictions as our best submission. Thus, we were unable to find what we believe to be the single misclassified image in our test set predictions.

5 Results

Model	β F-Score
Ensemble X-RaysNet	0.96551

6 Conclusion

The COVID-19 classification problem exposed our team to the challenges of training a model with a small, imbalanced dataset. The leave-one-out cross validation technique along with data augmentations proved to be highly effective for this task. We were able to leverage what we learned through this semester as well as our own prior experience with neural networks to develop a robust classifier. Given more time, we would have conducted ablation studies on each addition to our pipeline, allowing us to confirm the usefulness of our various components.

We believe a solution to this task would have had greater value given a few modifications to the problem, and the data. For instance, a larger training dataset, with a distribution of labels matching the real world would have produced a more reliable model for real medical purposes. Additionally, patient metadata such as age, medical history, and progression through the illness could provide very valuable information to the model, improving its accuracy and robustness. Even providing the particular diagnoses of the 'non-COVID' images would be useful, and help better distinguish COVID from other illnesses or infections.