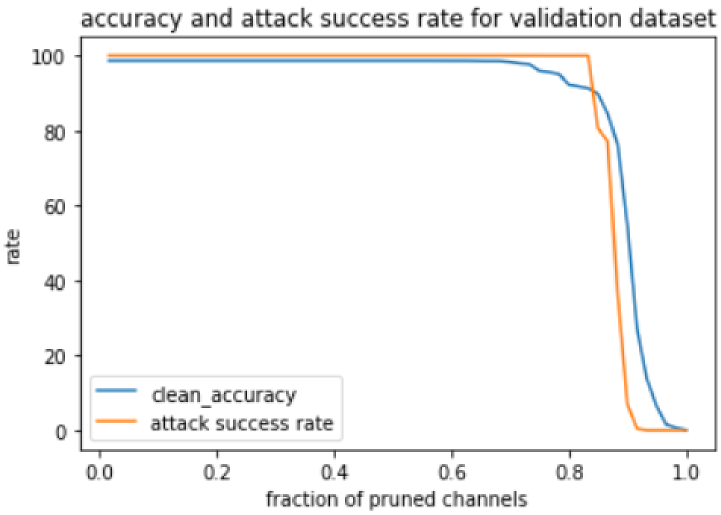
Lab2 Report

The Github link is https://github.com/RickyLCT/ML-in-Cyber-Security-Lab2.

In this lab, I implemented the prune defense that we learnt from the class. I pruned the conv layer based on the last pooling average across the while validation and we needed to pruned “conv_3”. The code was improved based on the instruction demo given by the TA. The codes were clear in the ipynb file. If you wanted to run the codes on your own, all you needed to do was change some file path which were located differently in your Google Drive. According to the instruction, we should save the model when is accuracy dropped at least 2%, 4%, and 10%. The models were stored in the model directory in the Github repo. The following table indicated the accuracy on clean test data and the attack success rate of each pruned model.

	accuracy	Attack success rate
2%	95.90%	100.00%
4%	92.29%	99.98%
10%	85.54%	77.21%
2% repaired	95.74%	100.00%
4% repaired	92.13%	99.98%
10% repaired	84.33%	77.21%

And the following figure states that the prune defense was not that successful since the attack success rate did not drop too much. When the accuracy dropped by 30%, the success rate dropped to 0.70, which was ok but not good.


accuracy and attack success rate for validation dataset
rate
fraction of pruned channels
clean_accuracy
attack success rate