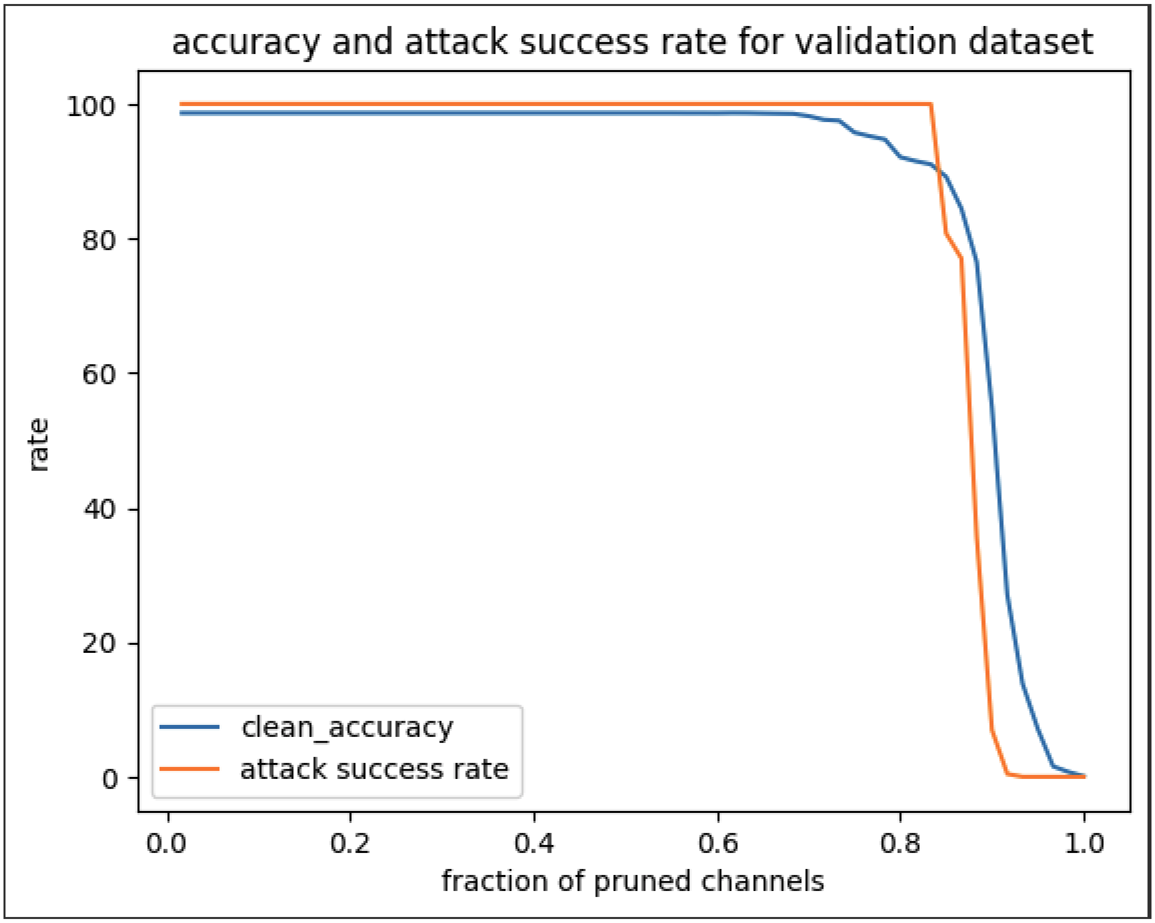
Lab 4 Report

Github: https://github.com/kaushik-42/ML_CyberSec_NYU_lab
Name: Kaushik Tummalapalli
NetId: kt2651

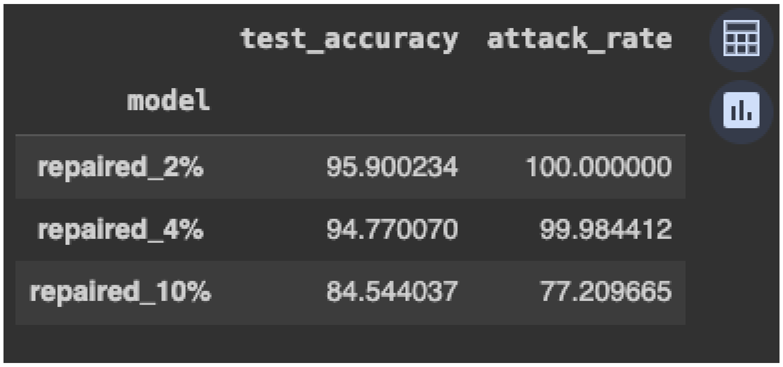
I have received various test accuracies while testing on the clean datasets with the “pruned” version of the bad nets which were given from the 2020 hacks git repository. We have used the pruning technique to remove the affected layers in the given bad neural networks which changes the behavior for some particular output classes. Pruning is an important technique in deep learning which will also help in reducing the complexity in terms of the time and space complexity which mainly reduces the number of parameters because we will be dealing with so many parameters once it reaches the final layer of the neural network.

Plot representing accuracy vs attack success rate for the validation dataset:



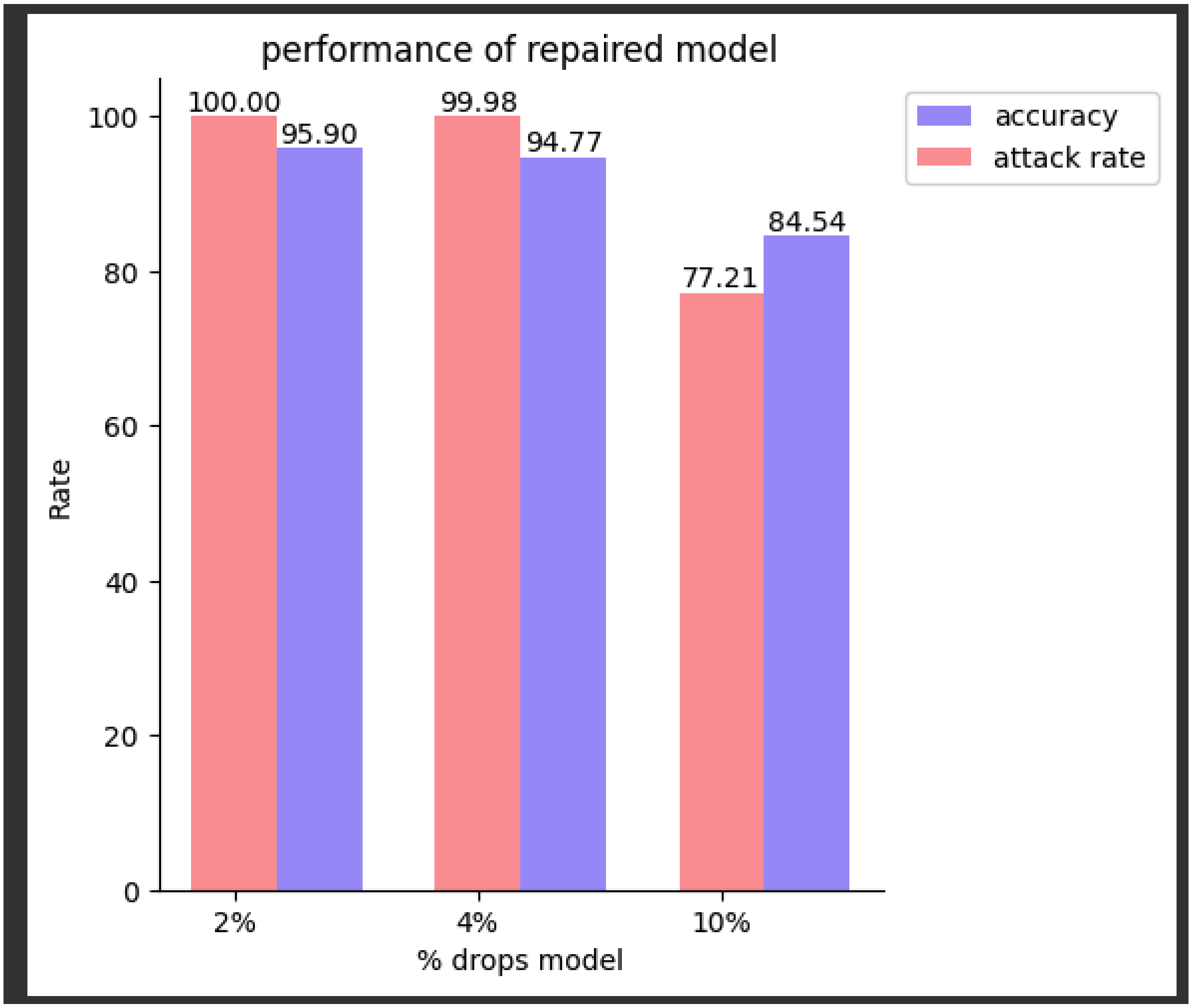
The Attack Success rate when the accuracy drops at least 30% is 6.954.

For the repaired net, which is by pruning out the bad net model by 2%, 4% and 10%, we received test accuracies as well as the attack rate for the following 3 repaired net models we have trained:



	test_accuracy	attack_rate
model		
repaired_2%	95.900234	100.000000
repaired_4%	94.770070	99.984412
repaired_10%	84.544037	77.209665

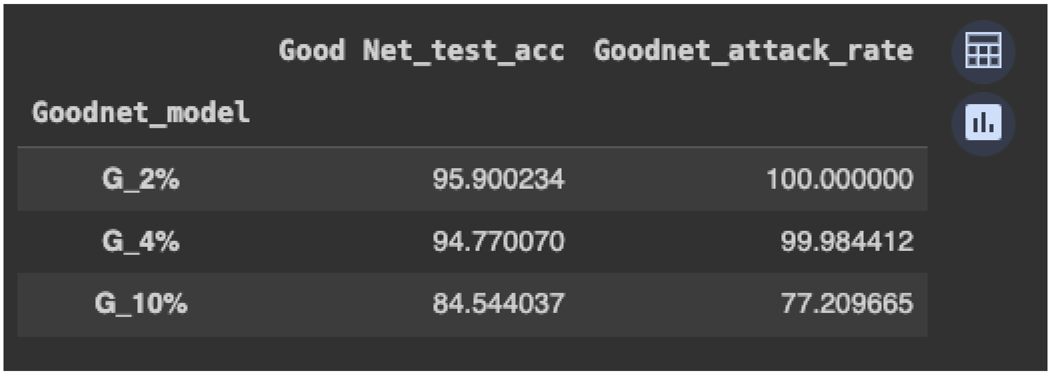
Performance of the Repaired Net:



For combined Good net models:
Combined 2% drops model, the clean test data Classification accuracy: 95.90023382696803

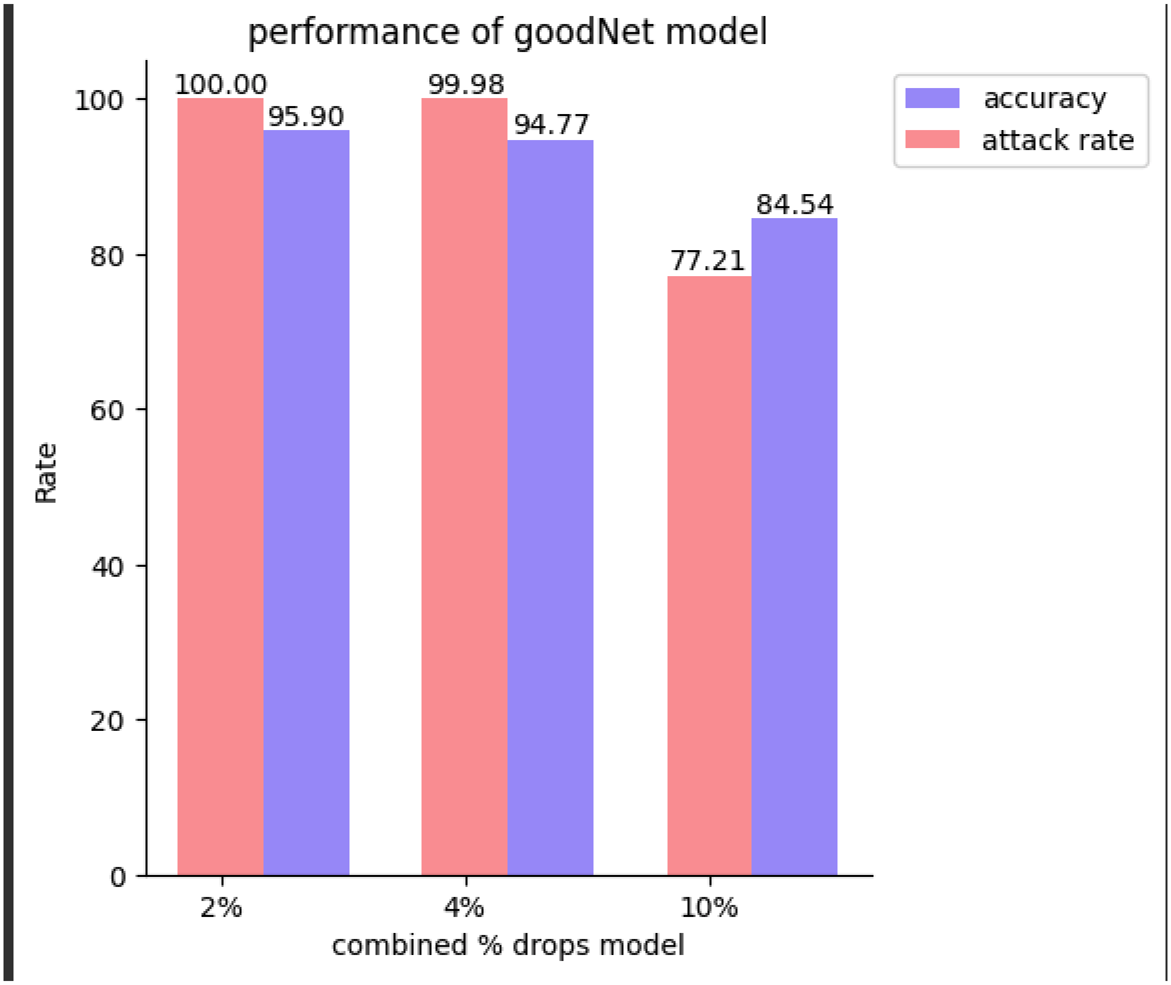
Combined 2% drops model, Attack Success Rate: 100.0
Combined 4% drops model, the clean test data Classification accuracy: 94.77007014809041
Combined 4% drops model, Attack Success Rate: 99.98441153546376
Combined 10% drops model, the clean test data Classification accuracy: 84.54403741231489
Combined 10% drops model, Attack Success Rate: 77.20966484801247

Test Accuracies and attack rates for models combined:



	Good Net_test_acc	Goodnet_attack_rate
Goodnet_model		
G_2%	95.900234	100.000000
G_4%	94.770070	99.984412
G_10%	84.544037	77.209665

Performance of the GoodNet Model:


performance of goodNet model
Rate
combined % drops model
accuracy
attack rate