

ML for Cyber Security – Lab 4 Report

Ankit Rajvanshi (ar7996)

Github Repo: <https://github.com/sudoDollar/NYUMLSecLab4>

Accuracy and Attack Success Rates

| Threshold (X %) | 0 | 2 | | 4 | | 10 | |
|-------------------------|---------|---------|----------|---------|----------|---------|----------|
| Network Model | BadNets | B_prime | GoodNets | B_prime | GoodNets | B_prime | GoodNets |
| Clean Test Accuracy (%) | 98.62 | 95.90 | 95.74 | 92.29 | 92.12 | 84.54 | 84.33 |
| Attack Success Rate (%) | 100 | 100 | 100 | 99.98 | 99.98 | 77.21 | 77.21 |

BadNets (B): Original Backdoored Model

B_Prime (B'): Network generated after pruning last pooling layer till X%

GoodNets (G): Network with N+1 output classes.



