# Backdoor Attacks – Report

**Name: Jash Rathod  Net ID: jsr10000**

## GitHub Link: https://github.com/jashrathod/backdoor-attacks

**Backdoor attacks:** Backdoor attacks on machine learning systems are a sort of cyber danger. During training, a hidden damaging behavior known as a 'backdoor' is discreetly inserted to a neural network in these assaults. This modified network, often known as a 'BadNet,' operates normally with regular inputs but behaves maliciously when particular inputs specified by the attacker are utilized. Key Points of Backdoor Attacks:

1. Trigger-Based Activation (distinctive form, mark, or item in a picture)
2. Hidden Malicious conduct
3. Wrong Output on Purpose

**Pruning Defense:** Pruning defense is a method of combating backdoor assaults. It seeks to close the backdoor while preserving the neural network's ability to fulfill its primary functions. Steps in Pruning Defense:

1. Focusing on the Last Pooling Layer and removing channels based on activity
2. Pruning and checking are repeated until accuracy falls below a certain threshold
3. Making the Fixed Network (G) (determines if original (B) and pruned (B') networks provide same results)

**Table with the accuracy on clean test data and the attack success rate (on backdoored test data) as a function of the fraction of channels pruned (X):**

| Pruning progress percentage | Clean data accuracy of Modified Model | Attack Success Rate (ASR) |
|---|---|---|
| **0.0 to 0.45** | 98.64899974019230 | 100.0 |

| Pruning progress percentage | Clean data accuracy of Modified Model | Attack Success Rate (ASR) |
|---|---|---|
| **0.45** | 98.64899974019230 | 100.0 |
| **0.4666666666666670** | 98.64899974019230 | 100.0 |
| **0.48333333333333300** | 98.64899974019230 | 100.0 |
| **0.5** | 98.64899974019230 | 100.0 |
| **0.5166666666666670** | 98.64899974019230 | 100.0 |
| **0.5333333333333330** | 98.64899974019230 | 100.0 |
| **0.55** | 98.64033948211660 | 100.0 |
| **0.5666666666666670** | 98.64033948211660 | 100.0 |
| **0.5833333333333330** | 98.63167922404090 | 100.0 |
| **0.6** | 98.63167922404090 | 100.0 |
| **0.6166666666666670** | 98.62301896596520 | 100.0 |
| **0.6333333333333330** | 98.57105741751100 | 100.0 |
| **0.65** | 98.47579457867850 | 100.0 |
| **0.6666666666666670** | 98.44115354637570 | 100.0 |
| **0.6833333333333330** | 98.08608296527240 | 100.0 |

| | | |
|---|---|---|
| **0.7** | 97.54914696457960 | 100.0 |
| **0.7166666666666670** | 97.39326231921710 | 100.0 |
| **0.7333333333333330** | 95.61790941370050 | 100.0 |
| **0.75** | 95.02901186455360 | 99.9913397419243 |
| **0.7666666666666670** | 94.49207586386080 | 99.9913397419243 |
| **0.7833333333333330** | 91.85935740885080 | 99.9913397419243 |
| **0.8** | 91.27045985970380 | 99.9913397419243 |
| **0.8166666666666670** | 90.80280592361650 | 99.98267948384860 |
| **0.8333333333333330** | 88.94951069541870 | 80.73958603966400 |
| **0.85** | 84.24699056031870 | 77.015675067117 |
| **0.8666666666666670** | 76.29687364683470 | 35.71490430414830 |
| **0.8833333333333330** | 54.75015155451630 | 6.954187234779600 |
| **0.9** | 26.994024421927800 | 0.4243526457088420 |
| **0.9166666666666670** | 13.813111630726600 | 0.0 |
| **0.9333333333333330** | 7.066770589763580 | 0.0 |
| **0.95** | 1.5501861955486300 | 0.0 |
| **0.9666666666666670** | 0.7188014202823240 | 0.0 |
| **0.9833333333333330** | 0.0779423226812159 | 0.0 |

**Evaluating the Repaired Models:**

| Threshold | Clean test accuracy of Refined Model | Attack Success Rate (ASR) |
|---|---|---|
| **2** | 95.744349 | 100.0 |
| **4** | 92.127825 | 99.991340 |
| **10** | 84.333593 | 77.015675 |

**Evaluating the GoodNet G models (`ModifiedModel`):**

| Threshold | Clean test accuracy of Goodnet Model | Attack Success Rate (ASR) |
|---|---|---|
| **2** | 95.90023382696803 | 100.0 |
| **4** | 92.29150428682775 | 99.98441153546376 |
| **10** | 84.54403741231489 | 77.20966484801247 |