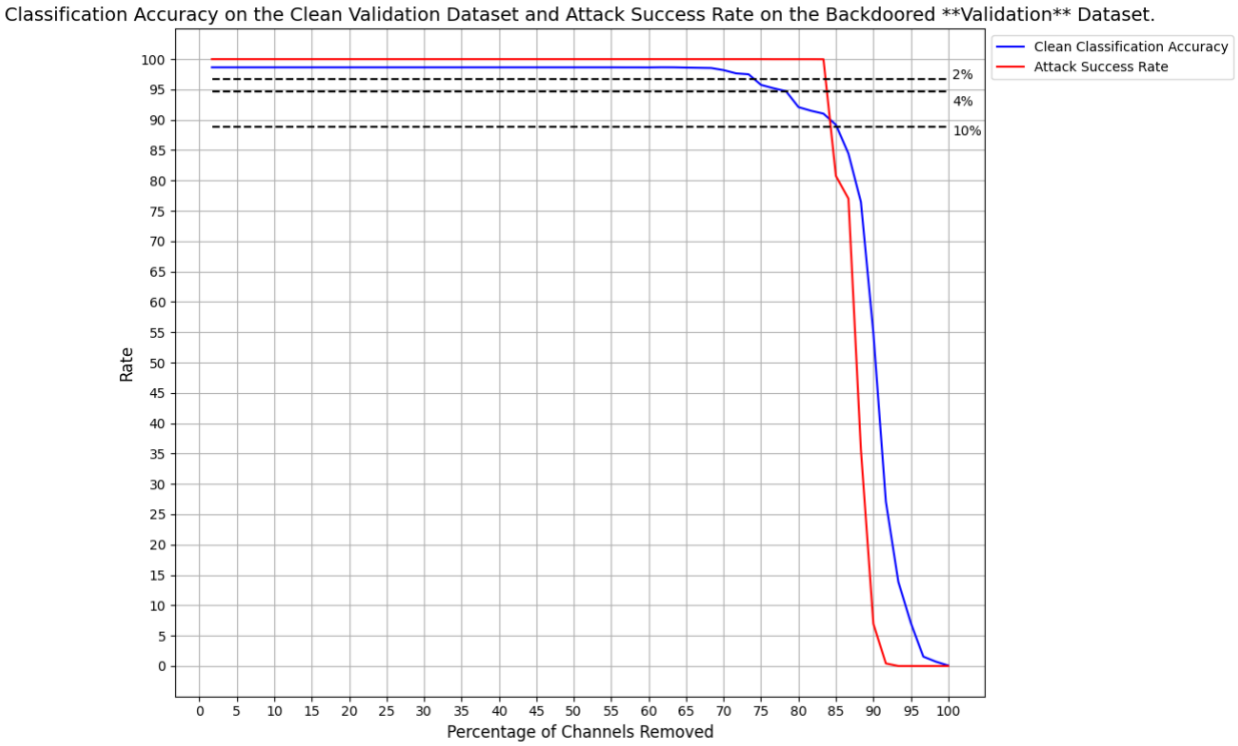
# Lab 4: Backdoor Attacks

| Iteration | Channels Removed | Percentage Channels Removed | Clean Accuracy | Accuracy Difference | Attack Success Rate |
|---|---|---|---|---|---|
| 0 | 0 | 1.67% | 98.65% | 0.00% | 100.00% |
| 1 | 26 | 3.33% | 98.65% | 0.00% | 100.00% |
| 2 | 27 | 5.00% | 98.65% | 0.00% | 100.00% |
| 3 | 30 | 6.67% | 98.65% | 0.00% | 100.00% |
| 4 | 31 | 8.33% | 98.65% | 0.00% | 100.00% |
| 5 | 33 | 10.00% | 98.65% | 0.00% | 100.00% |
| 6 | 34 | 11.67% | 98.65% | 0.00% | 100.00% |
| 7 | 36 | 13.33% | 98.65% | 0.00% | 100.00% |
| 8 | 37 | 15.00% | 98.65% | 0.00% | 100.00% |
| 9 | 38 | 16.67% | 98.65% | 0.00% | 100.00% |
| 10 | 25 | 18.33% | 98.65% | 0.00% | 100.00% |
| 11 | 39 | 20.00% | 98.65% | 0.00% | 100.00% |
| 12 | 41 | 21.67% | 98.65% | 0.00% | 100.00% |
| 13 | 44 | 23.33% | 98.65% | 0.00% | 100.00% |
| 14 | 45 | 25.00% | 98.65% | 0.00% | 100.00% |
| 15 | 47 | 26.67% | 98.65% | 0.00% | 100.00% |
| 16 | 48 | 28.33% | 98.65% | 0.00% | 100.00% |
| 17 | 49 | 30.00% | 98.65% | 0.00% | 100.00% |
| 18 | 50 | 31.67% | 98.65% | 0.00% | 100.00% |
| 19 | 53 | 33.33% | 98.65% | 0.00% | 100.00% |
| 20 | 55 | 35.00% | 98.65% | 0.00% | 100.00% |
| 21 | 40 | 36.67% | 98.65% | 0.00% | 100.00% |
| 22 | 24 | 38.33% | 98.65% | 0.00% | 100.00% |
| 23 | 59 | 40.00% | 98.65% | 0.00% | 100.00% |
| 24 | 9 | 41.67% | 98.65% | 0.00% | 100.00% |
| 25 | 2 | 43.33% | 98.65% | 0.00% | 100.00% |
| 26 | 12 | 45.00% | 98.65% | 0.00% | 100.00% |
| 27 | 13 | 46.67% | 98.65% | 0.00% | 100.00% |
| 28 | 17 | 48.33% | 98.65% | 0.00% | 100.00% |
| 29 | 14 | 50.00% | 98.65% | 0.00% | 100.00% |
| 30 | 15 | 51.67% | 98.65% | 0.00% | 100.00% |
| 31 | 23 | 53.33% | 98.65% | 0.00% | 100.00% |
| 32 | 6 | 55.00% | 98.65% | 0.00% | 100.00% |
| 33 | 51 | 56.67% | 98.64% | 0.01% | 100.00% |
| 34 | 32 | 58.33% | 98.64% | 0.01% | 100.00% |
| 35 | 22 | 60.00% | 98.63% | 0.02% | 100.00% |
| 36 | 21 | 61.67% | 98.66% | -0.01% | 100.00% |
| 37 | 20 | 63.33% | 98.65% | 0.00% | 100.00% |
| 38 | 19 | 65.00% | 98.61% | 0.04% | 100.00% |
| 39 | 43 | 66.67% | 98.57% | 0.08% | 100.00% |
| 40 | 58 | 68.33% | 98.54% | 0.11% | 100.00% |
| 41 | 3 | 70.00% | 98.19% | 0.46% | 100.00% |
| 42 | 42 | 71.67% | 97.65% | 1.00% | 100.00% |
| 43 | 1 | 73.33% | 97.51% | 1.14% | 100.00% |
| 44 | 29 | 75.00% | 95.76% | 2.89% | 100.00% |
| 45 | 16 | 76.67% | 95.20% | 3.45% | 99.99% |
| 46 | 56 | 78.33% | 94.72% | 3.93% | 99.99% |
| 47 | 46 | 80.00% | 92.09% | 6.56% | 99.99% |
| 48 | 5 | 81.67% | 91.50% | 7.15% | 99.99% |
| 49 | 8 | 83.33% | 91.02% | 7.63% | 99.98% |
| 50 | 11 | 85.00% | 89.17% | 9.47% | 80.74% |
| 51 | 54 | 86.67% | 84.44% | 14.21% | 77.02% |
| 52 | 10 | 88.33% | 76.49% | 22.16% | 35.71% |
| 53 | 28 | 90.00% | 54.86% | 43.79% | 6.95% |

| | | | | | |
|---|---|---|---|---|---|
| 54 | 35 | 91.67% | 27.09% | 71.56% | 0.42% |
| 55 | 18 | 93.33% | 13.87% | 84.78% | 0.00% |
| 56 | 4 | 95.00% | 7.10% | 91.55% | 0.00% |
| 57 | 7 | 96.67% | 1.55% | 97.10% | 0.00% |
| 58 | 52 | 98.33% | 0.72% | 97.93% | 0.00% |
| 59 | 57 | 100.00% | 0.08% | 98.57% | 0.00% |




From these visualizations, we can observe a significant decrease in the backdoor attack success rate when a large portion of neurons is pruned. Initially, the attack success rate hovers around 100%, while the clean classification accuracy remains stable. This can be explained as follows: initially, we prune neurons that are either all zeros or poorly activated, making them irrelevant to both a genuine network and a malicious BadNet.

As the number of channels removed exceeds 70% but remains below 83% of their initial quantity, we observe a decline in clean classification accuracy. This suggests that we are now pruning neurons responsible for classifying clean inputs but not those activated by malicious inputs. Beyond 83% of all neurons removed, both the attack success rate and clean classification accuracy experience a drop. This indicates that we are now removing neurons activated by both clean and malicious inputs.

It's noteworthy that complete elimination of the backdoor attack is challenging since doing so would lead to a decline in clean classification accuracy. For instance, reducing the attack success rate to 6% by disabling 90% of neurons results in a significant decrease in clean classification accuracy to almost 50%.