# Title: Analysis of Channel Pruning as a Defense Against Backdoor Attacks in Neural Networks

Prathyusha Kadali, pk2669

Git: https://github.com/pk6797/MLCyber-BackdoorAttacks

## Introduction

This report analyzes the findings from the GitHub repository pk6797/MLCyber-BackdoorAttacks which explores the effectiveness of channel pruning in neural networks as a defense mechanism against backdoor attacks. The study is significant in the context of increasing security concerns in machine learning models, particularly in scenarios where models are susceptible to adversarial manipulations.

## Methodology

The study involved pruning channels in the neural network, known as BadNet, to observe the impact on the accuracy of clean test data and the attack success rate (ASR) on backdoored test data. The channels were pruned in increasing order of activation values of the last pooling layer of BadNet. The pruning process was stopped once the validation accuracy dropped by a pre-determined percentage below the original accuracy.

## Results

The original accuracy of BadNet on clean test data was 98.62%, with an ASR of 100%. The results after pruning were as follows:

- 2% Drop (45 channels pruned): Clean data accuracy was 95.75%, and ASR remained at 100%.
- 4% Drop (48 channels pruned): Clean data accuracy was 92.09%, ASR slightly reduced to 99.99%.
- 10% Drop (52 channels pruned): Clean data accuracy fell to 84.44%, ASR significantly dropped to 77.21%.
- 30% Drop (54 channels pruned): Clean data accuracy drastically decreased to 54.86%, ASR reduced to 6.96%.

## Table

| **Fraction of Channels Pruned** | **Clean Test Data Accuracy (%)** | **Attack Success Rate (%)** |
|---|---|---|
| Original (0% pruned) | 98.62 | 100.00 |
| 2% Drop (45 channels pruned) | 95.75 | 100.00 |
| 4% Drop (48 channels pruned) | 92.09 | 99.99 |
| 10% Drop (52 channels pruned) | 84.44 | 77.21 |
| 30% Drop (54 channels pruned) | 54.86 | 6.96 |

## Discussion

The results demonstrate a clear trade-off between maintaining high accuracy on clean test data and reducing the ASR through channel pruning. While aggressive pruning significantly lowers the ASR, it also leads to a substantial decrease in model accuracy. These findings highlight the challenges in neural network security, where defenses against backdoor attacks must be balanced against the overall performance of the model.

## Implications and Conclusion

The study underscores the need for robust security measures in machine learning models, especially in applications where model integrity is crucial. While channel pruning emerges as an effective strategy to mitigate backdoor attacks, its impact on model accuracy cannot be overlooked. Future research should focus on optimizing pruning strategies to achieve a balance between security and performance.