ML for Cyber Security - Lab 4

Name: Shashank Shekhar
netID: ss16116
GitHub Link: https://shorturl.at/twBD4

Objective:

Use methods like layer pruning, accuracy-based model saving, and vulnerability assessment to refine a machine-learning model, with the final output being a repaired BadNet.

Methodology:

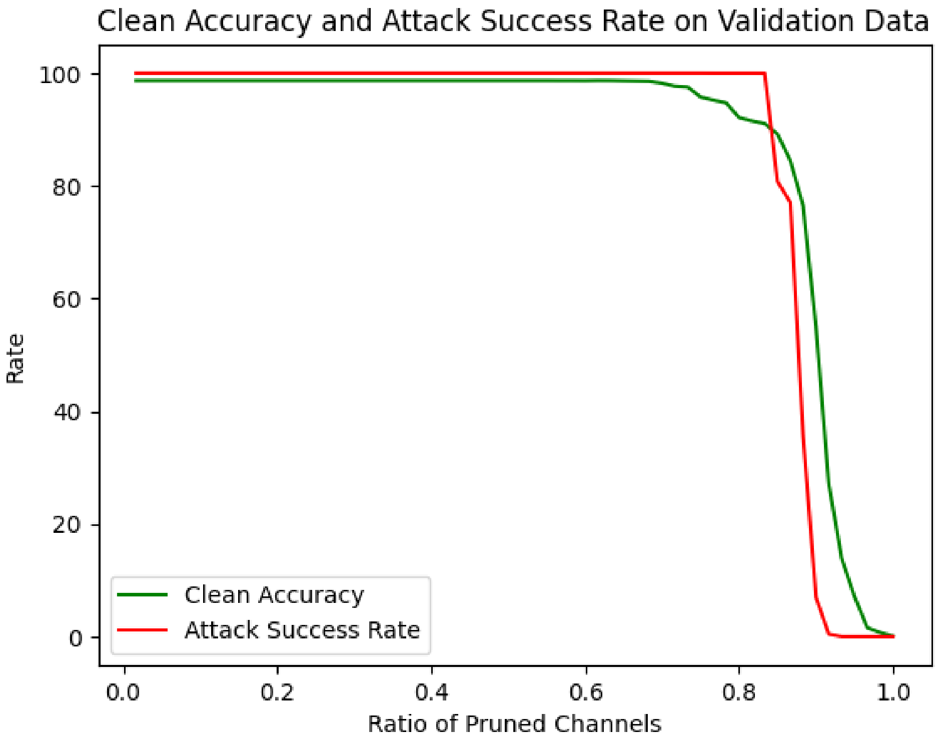
Pruning: The conv_3 layer is pruned based on the decreasing order of the average activation from the last pooling operation, with intermediate models being saved at pre-defined thresholds for an accuracy drop of 2%, 4%, and 10%.

The model's attack success rate was 6.95 when the validation accuracy dropped by 30%.

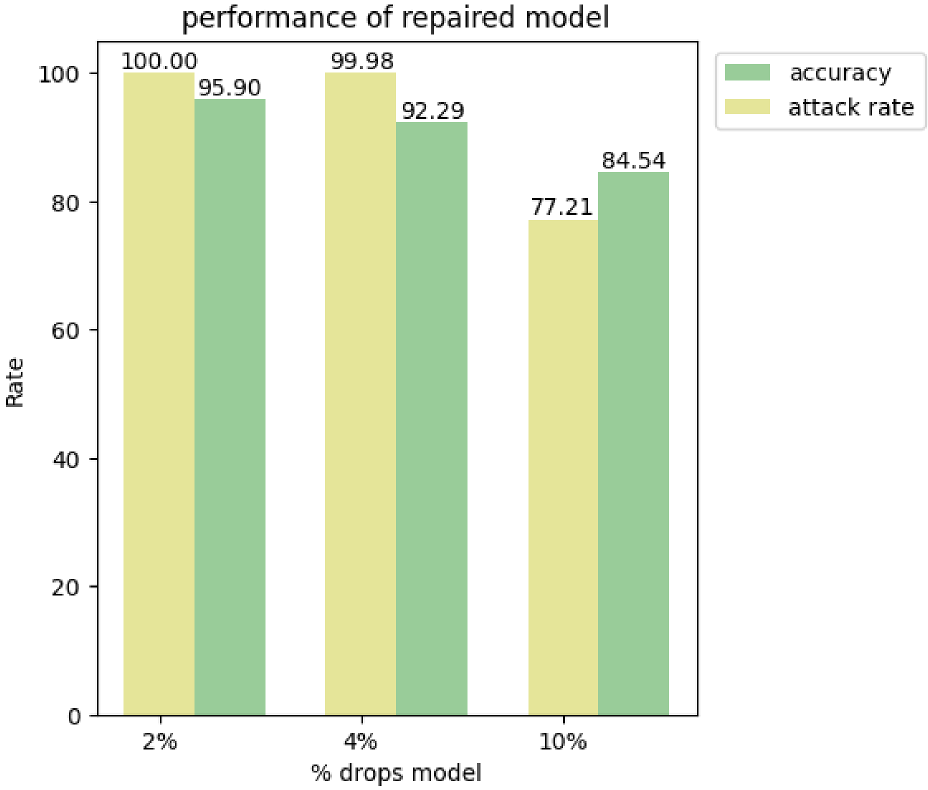
Model Integration: The original BadNet was combined with the repaired model, resulting in a GoogNet to enhance the model's performance.

Results:

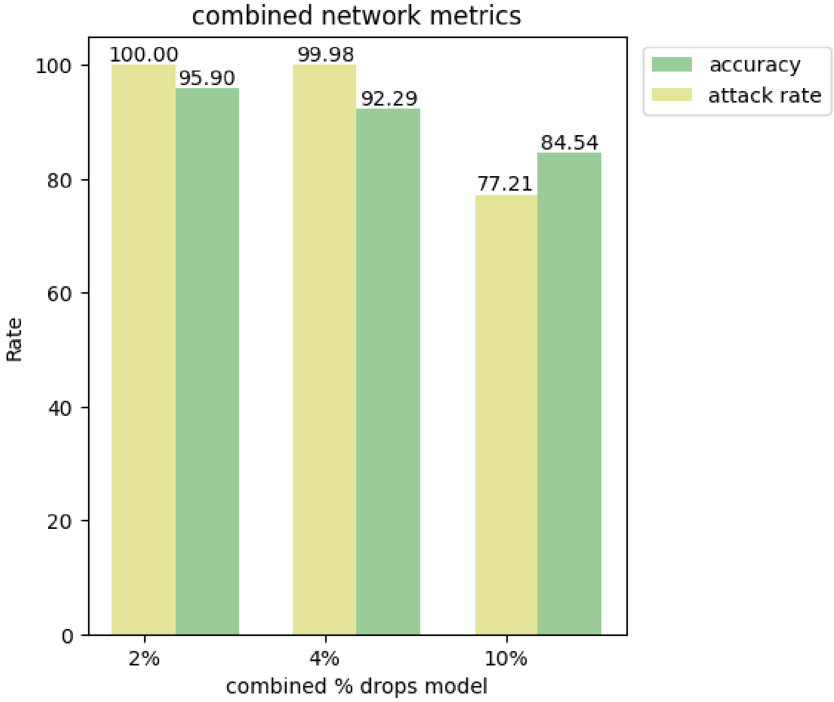
The attack success rate as a function of the fraction of channels pruned:



Performance metrics for the repaired network:



Performance metrics for the combined model:


combined network metrics
accuracy
attack rate
Rate
combined % drops model