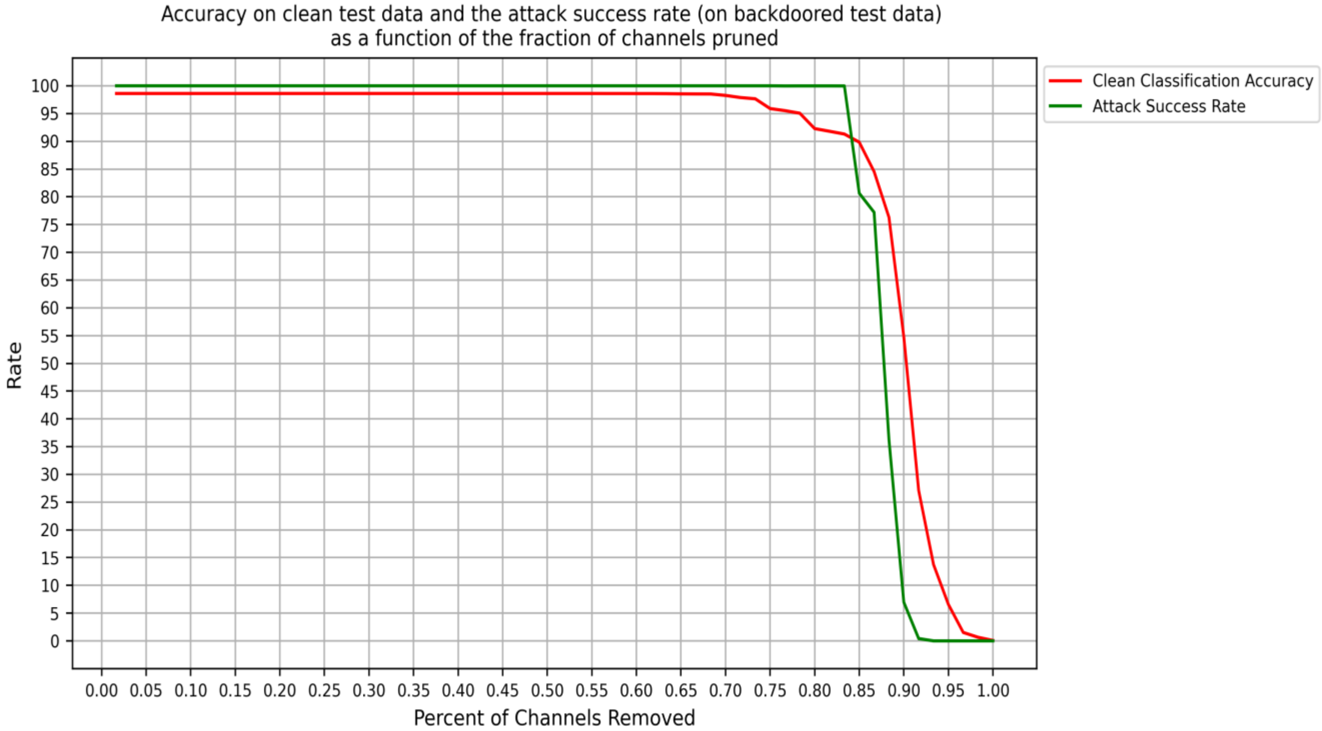
EL-GY-9163: Machine Learning for Cyber-security
Lab 4 Report

To construct the backdoor detector G, we first need to repair the badnet B itself. The repair process involves pruning the weights of the last convolution layer based on activations from clean validation data (obtained from valid.h5), which are derived from the last pooling layer before the fully connected (FC) layers of badnet B. The activations are averaged over all samples in the validation set, and then the first three dimensions are averaged to obtain a vector of activations for each of the 60 channels (neurons). These channels are arranged in increasing order according to the average activations. Subsequently, we use this arrangement to prune one channel at a time, comparing the validation accuracy with the original badnet accuracy. Pruning stops as soon as the validation accuracy drops by at least X% (2%, 4%, 10%) below the original accuracy, and the repaired network B' is saved.

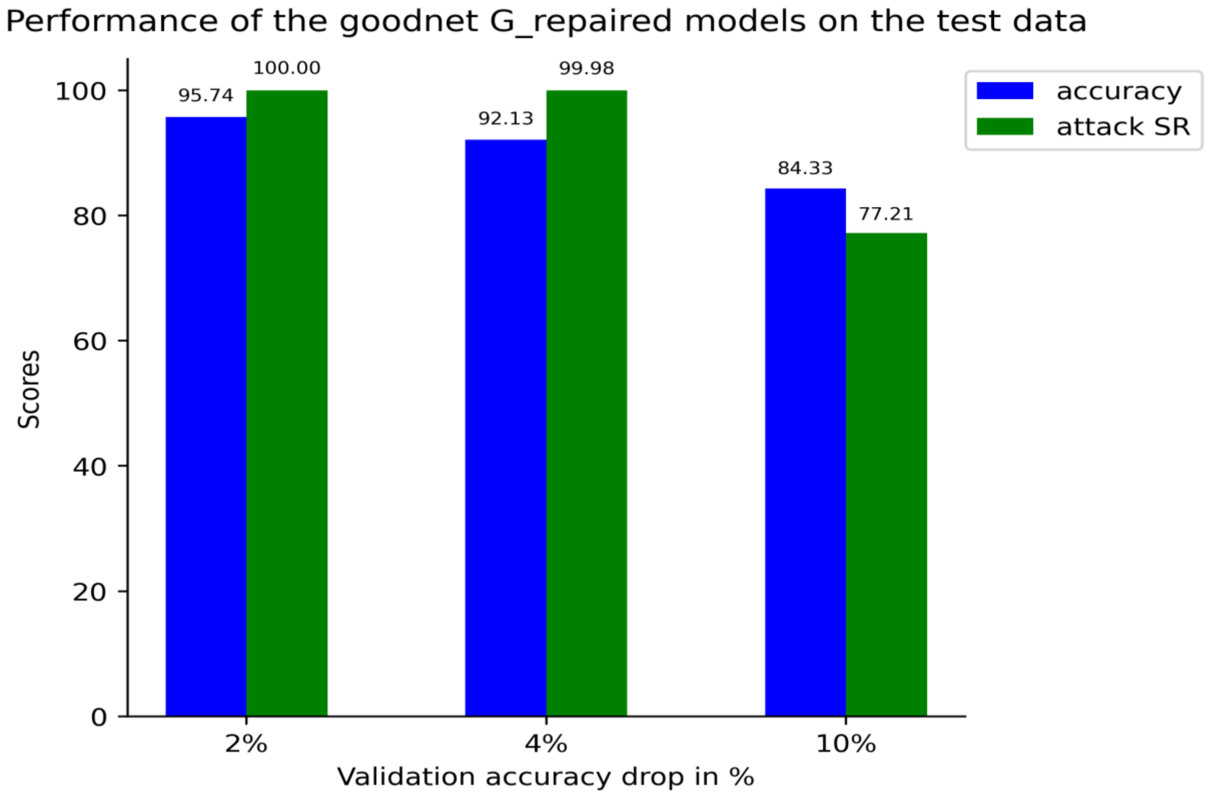
For the construction of goodnet G, each test input is processed through both B and B'. If the classification outputs from B and B' are the same (i.e., class i), then the goodnet outputs class i. Otherwise, it outputs N+1, where N is 1283 in this case (since class numbering starts from 0 to 1282, resulting in a total of 1283 classes).



The plotted results reveal a noticeable decline in the success rate of backdoor attacks as a significant portion of neurons is pruned. Initially, the attack success rate remains consistently around 100%, while the clean classification accuracy remains stable. The process can be elucidated as follows: initially, we prune neurons that are either inactive or poorly activated, irrelevant to both a legitimate network and a compromised one. Subsequently, when the removal of channels exceeds 70% but remains below 83% of their initial quantity, a decrease in clean classification accuracy is observed. This indicates the pruning of neurons responsible for classifying genuine inputs, while those activated by malicious inputs remain intact. Beyond the 83% threshold of neuron removal, both the attack success rate and clean classification accuracy decline. This implies the removal of neurons activated by both clean and malicious inputs.

In this scenario, it becomes possible to maintain a reasonably acceptable clean classification accuracy. For instance, when the validation accuracy drops below 30% of the actual accuracy (approximately 88% of neurons removed), the classification accuracy is around 65%, while the attack success rate is approximately 20%. However, it is noted that completely disabling the backdoor attack leads to a drop in clean classification accuracy, such as reducing the attack success rate to 6% when 90% of neurons are disabled, resulting in a decline in clean classification accuracy to almost 50%.

It is essential to recognize that due to the discrete and finite number of channels, precise pruning to achieve a specific validation accuracy drop below X% of the actual accuracy is not achievable. For instance, with 60 channels, we disable 1.66% of channels at each step. Given that only 32 channels have non-zero activations, actual pruning occurs when these channels are disabled. Thus, when observing a validation accuracy drop below 30% of the actual baseline accuracy, the accuracy effectively drops from 22% in the previous step (when 88.33% of channels were disabled) to 44% in the next step (where 90% of channels were disabled).



The attempts to repair the models appear to be ineffective, as, in most cases, they do not thwart the attack. When the validation accuracy drops by 2% and 4% below the original accuracy, the attack success rate consistently surpasses the prediction accuracy. This outcome persists because the repaired badnets (B') still exhibit a 100% success rate. These findings indicate the presence of a pruning-aware attack, where the attacker encoded the backdoor behavior into the same neurons used for classifying clean data.

A notable shift occurs when the validation accuracy drops by 10% and 30% below the original accuracy. In this scenario, the validation accuracy outperforms the attack success rate. However, due to the pruning-aware nature of the attack, where the attacker utilized the same set of neurons as the original model for classification, removing these neurons decreases the attack success rate and leads to a decline in clean data classification accuracy. This is evident in the bar plots above. Even when using a model with almost 90% of neurons pruned (corresponding to a validation accuracy drop of 30% below the original accuracy), the achieved accuracies are only slightly above the chance level. This makes the pruning defense less effective against this specific type of attack.

Notably, Goodnet (G) accuracy is slightly lower than that of repaired networks (B') since Goodnet removes some labels that were misclassified by Badnet.