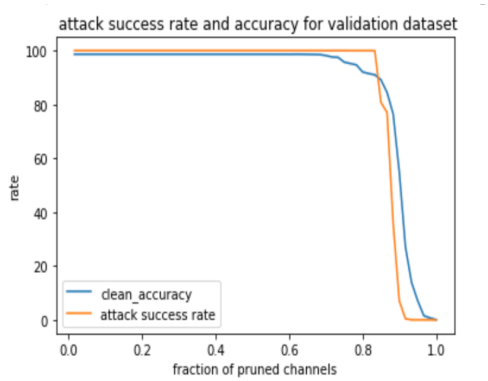
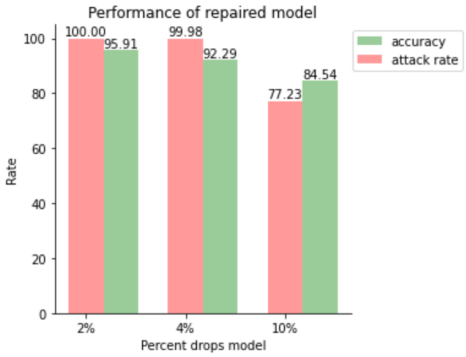
We prune the convolution layer based on the last pooling average activation across the entire validation dataset. We need to prune the conv_3 layer.
The attack success rate when the accuracy drops at least 30%: 6.980168009006668

On pruning the model, we use the clean validation data set and test it on the test data set. The accuracy and attack success rate would look like this for the validation dataset -.



The attack success rate does not significantly decrease in this case; thus, we can see that the prune defense is not very effective. The attack success rate is acceptable but not great because the accuracy is too much. According to my theory, the attack strategy is a prune immune attack, and the poisoned data are kept with the pruned model.

Graph showing Performance of repaired mode:



model	text_acc	attack_rate
repaired_2%	95.908028	100.000000
repaired_4%	92.291504	99.984412
repaired_10%	84.544037	77.233048