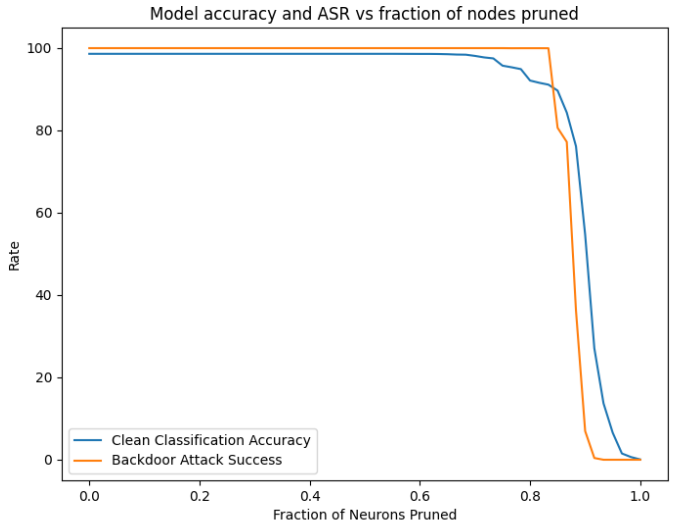
The repaired net model accuracy on clean test and attack success rate (ASR) on poisoned test data is shown below. The repaired net is a ensemble of the original backdoored network and a network that has been pruned to remove inactive neurons by increasing order of their average activation values.

Neurons Pruned	Fraction of Neurons Pruned	Clean Test Accuracy	Poisoned Test ASR
0	0.00000	98.62042%	100.00000%
1	0.01667	98.62042%	100.00000%
2	0.03333	98.62042%	100.00000%
3	0.05000	98.62042%	100.00000%
4	0.06667	98.62042%	100.00000%
5	0.08333	98.62042%	100.00000%
6	0.10000	98.62042%	100.00000%
7	0.11667	98.62042%	100.00000%
8	0.13333	98.62042%	100.00000%
9	0.15000	98.62042%	100.00000%
10	0.16667	98.62042%	100.00000%
11	0.18333	98.62042%	100.00000%
12	0.20000	98.62042%	100.00000%
13	0.21667	98.62042%	100.00000%
14	0.23333	98.62042%	100.00000%
15	0.25000	98.62042%	100.00000%
16	0.26667	98.62042%	100.00000%
17	0.28333	98.62042%	100.00000%
18	0.30000	98.62042%	100.00000%
19	0.31667	98.62042%	100.00000%
20	0.33333	98.62042%	100.00000%
21	0.35000	98.62042%	100.00000%
22	0.36667	98.62042%	100.00000%
23	0.38333	98.62042%	100.00000%
24	0.40000	98.62042%	100.00000%
25	0.41667	98.62042%	100.00000%
26	0.43333	98.62042%	100.00000%
27	0.45000	98.62042%	100.00000%
28	0.46667	98.62042%	100.00000%
29	0.48333	98.62042%	100.00000%
30	0.50000	98.62042%	100.00000%
31	0.51667	98.62042%	100.00000%
32	0.53333	98.62042%	100.00000%
33	0.55000	98.62042%	100.00000%
34	0.56667	98.61263%	100.00000%
35	0.58333	98.60483%	100.00000%
36	0.60000	98.59704%	100.00000%
37	0.61667	98.59704%	100.00000%
38	0.63333	98.57366%	100.00000%
39	0.65000	98.52689%	100.00000%
40	0.66667	98.44115%	100.00000%
41	0.68333	98.40998%	100.00000%
42	0.70000	98.11380%	100.00000%
43	0.71667	97.74747%	100.00000%
44	0.73333	97.50585%	100.00000%
45	0.75000	95.74435%	100.00000%
46	0.76667	95.34684%	99.97662%
47	0.78333	94.90257%	99.98441%
48	0.80000	92.12783%	99.98441%
49	0.81667	91.58223%	99.98441%
50	0.83333	91.13016%	99.97662%
51	0.85000	89.68044%	80.64692%
52	0.86667	84.33359%	77.20966%
53	0.88333	76.16524%	36.26656%
54	0.90000	54.67654%	6.96025%
55	0.91667	27.06937%	0.42089%
56	0.93333	13.70226%	0.00000%

57	0.95000	6.56274%	0.00000%
58	0.96667	1.51988%	0.00000%
59	0.98333	0.64692%	0.00000%
60	1.00000	0.07015%	0.00000%



According to Liu et al. the pruning defense happens in 3 phases. In the first phase, neurons that neither effect the activation on clean or poisoned inputs are removed. These neurons are dormant so removing them neither effects the model accuracy nor the backdoor ASR. This correlates to the removal of the first 33 neurons where model accuracy and ASR are unchanged shown above.

The second phase removes neurons that are activated by the poisoned data but not the clean input. This is the area where the model accuracy remains high but the backdoor ASR starts to dip. However, unlike the paper, we see the model accuracy dipping before the backdoor ASR. Model ASR remains generally high until the 51st neuron is removed, when we see a sharp decline in ASR. The sweet spot for accuracy and ASR seems to be when 53 neurons are removed (0.833 neurons removed as a fraction of total neurons) where the model accuracy is 76% while ASR is 36%.

The last phase removes neurons that correspond to clean inputs leading to a sharp decline in accuracy. This happens right after the sweet spot discussed above. After removing the final neuron the accuracy drops to 0.07% which is what you would expect from a completely random network (1/1283 = 0.078%).

In general, the pruning defense was not as successful as expected since model accuracy had to be sacrificed significantly without seeing a complete elimination of ASR.

References:

1. Liu, K., Dolan-Gavitt, B., & Garg, S. (2018). Fine-pruning: Defending against backdooring attacks on Deep Neural Networks. Research in Attacks, Intrusions, and Defenses, 273–294. https://doi.org/10.1007/978-3-030-00470-5_13