Lab 2

What is a backdoored detector model?

A backdoored detector model is a machine learning model that has been intentionally trained to include a "backdoor" or hidden trigger that can be used to manipulate the model's behavior. This type of model is typically used in adversarial machine learning, where the goal is to create a model that appears to be accurate and reliable, but that can be easily manipulated by an attacker.

Backdoored detector models are often used in situations where the reliability and security of a machine learning model are critical, such as in medical diagnostic systems or security applications. By incorporating a hidden backdoor into the model, an attacker can gain control over the model and use it to make incorrect predictions or to bypass security measures.

What is pruning?

Pruning a model refers to the process of removing unnecessary or redundant parameters from a trained machine learning model to make it more efficient and easier to deploy. This can be done by identifying and removing parameters that have little impact on the model's performance, or by applying regularization techniques that encourage the model to learn more compact representations of the data.

How can we use pruning as a defensive measure?

Pruning can be used as a defensive measure in machine learning by applying pruning techniques to a trained model to remove unnecessary or redundant parameters that could be exploited by an attacker. By pruning the model, we can reduce its size and complexity, making it more difficult for an attacker to manipulate the model's behavior or to uncover any hidden backdoors that may have been introduced during training.

Dataset:

The YouTube Face Dataset (YFCC100M) is a large-scale collection of video and image data that has been widely used in research on facial recognition and related topics. The dataset contains more than 100 million videos and images, which have been annotated with metadata including tags, descriptions, and user-generated tags.

Results of experiments:

	Test Data Accuracy	Attack Success Rate
Original Model	98.62	100
Pruned 0.02 Model	98.55	100
Pruned 0.04 Model	92.29	99.98
Pruned 0.1 Model	84.54	77.20
Repaired 0.02 Model	98.526	100
Repaired 0.04 Model	92.12	99.98
Repaired 0.1 Model	84.33	77.20

Loop Index	Clean Test Data Accuracy	ASR	% of Channels Pruned
30	98.62042089	100.00%	50
31	98.62042089	100.00%	51.66666667
32	98.62042089	100.00%	53.33333333
33	98.62042089	100.00%	55
34	98.61262666	100.00%	56.66666667
35	98.61262666	100.00%	58.33333333
36	98.60483242	100.00%	60
37	98.60483242	100.00%	61.66666667
38	98.58924396	100.00%	63.33333333
39	98.5502728	100.00%	65
40	98.53468433	100.00%	66.66666667
41	98.5268901	100.00%	68.33333333
42	98.26968044	100.00%	70
43	97.88776306	100.00%	71.66666667
44	97.66173032	100.00%	73.33333333
45	95.90023383	100.00%	75
46	95.52611068	99.98%	76.66666667
47	95.05845674	99.98%	78.33333333
48	92.29150429	99.98%	80
49	91.80826189	99.98%	81.66666667
50	91.30943102	99.98%	83.33333333
51	89.84411535	80.65%	85
52	84.54403741	77.21%	86.66666667