https://github.com/IlPodd/ML_Sec-Robustness_Evaluation_Across_Classes

Variation of Accuracy and Robustness across classes

PODDIGHE GABRIELE
70/90/00337

Robust Bench

A STANDARDIZED BENCHMARK FOR ADVERSARIAL ROBUSTNESS

- Reasonable computational requirements
- Model Zoo and Leaderboard
- AutoAttack Evaluation
- L∞ and L2 threat models

https://robustbench.github.io

AUTOATTACK

AN ENSEMBLE OF COMPLEMENTARY ATTACKS DESIGNED TO ESTIMATE ADVERSARIAL ROBUSTNESS

- APGD-CE
- APGD-DLR
- FAB
- SQUARE

CROCE ET AL.(2020)

Models chosen

L∞, EPS = 8/255, CIFAR-10

PENG2023ROBUST

- RaWideResNet-70-16
- 267.72M parameters

WANG2023BETTER

- WideResNet-70-16
- 266.79M parameters

WANG2023BETTER

- WideResNet-28-10
- 36.47M parameters

BAI2023IMPROVING_EDM

- ResNet-152 + WideResNet-70-16 + mixing network,
- 566.92M parameters

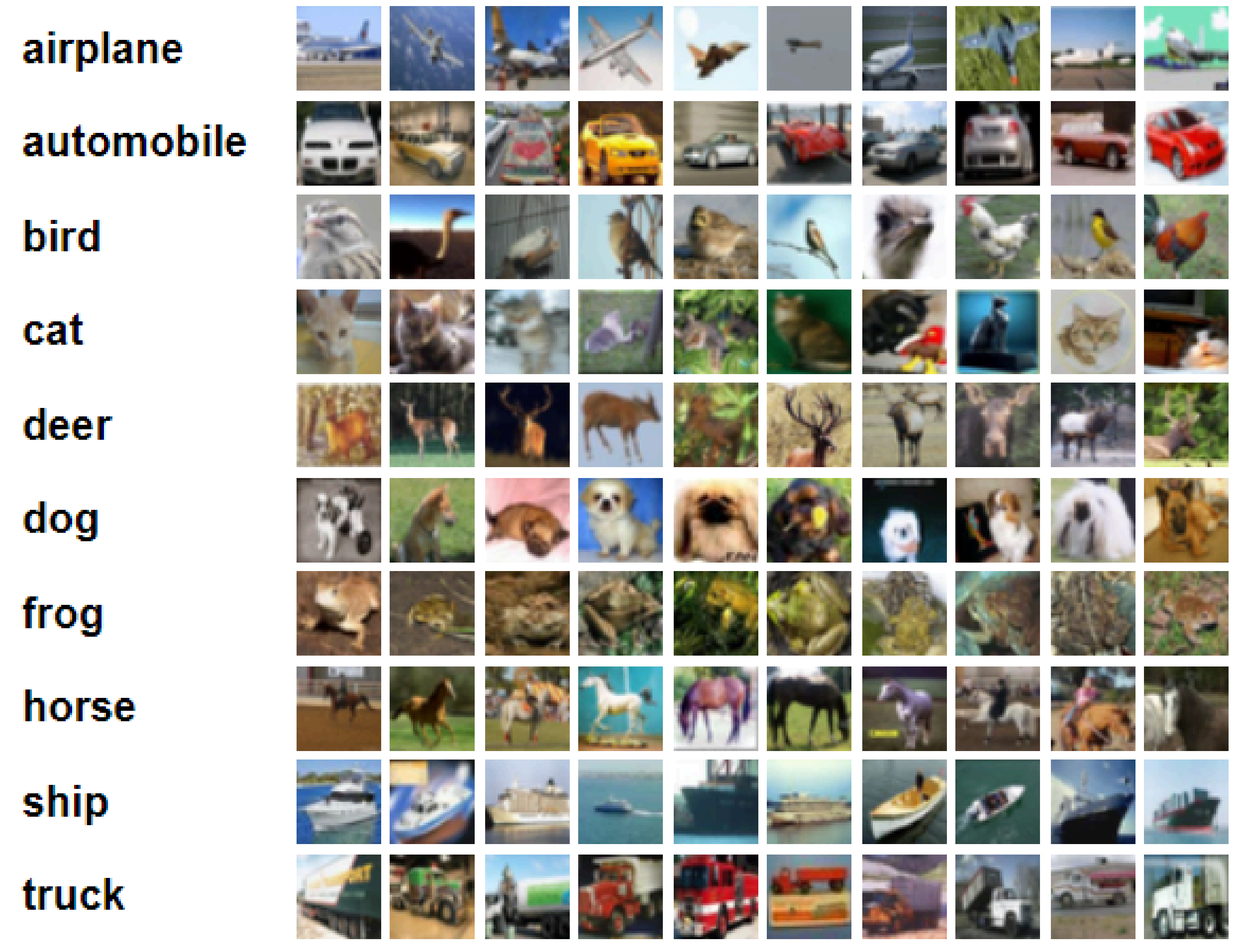
CUI2023DECOUPLED

- WideResNet-28-10
- 36.48M parameters

CIFAR-10

Image classification dataset

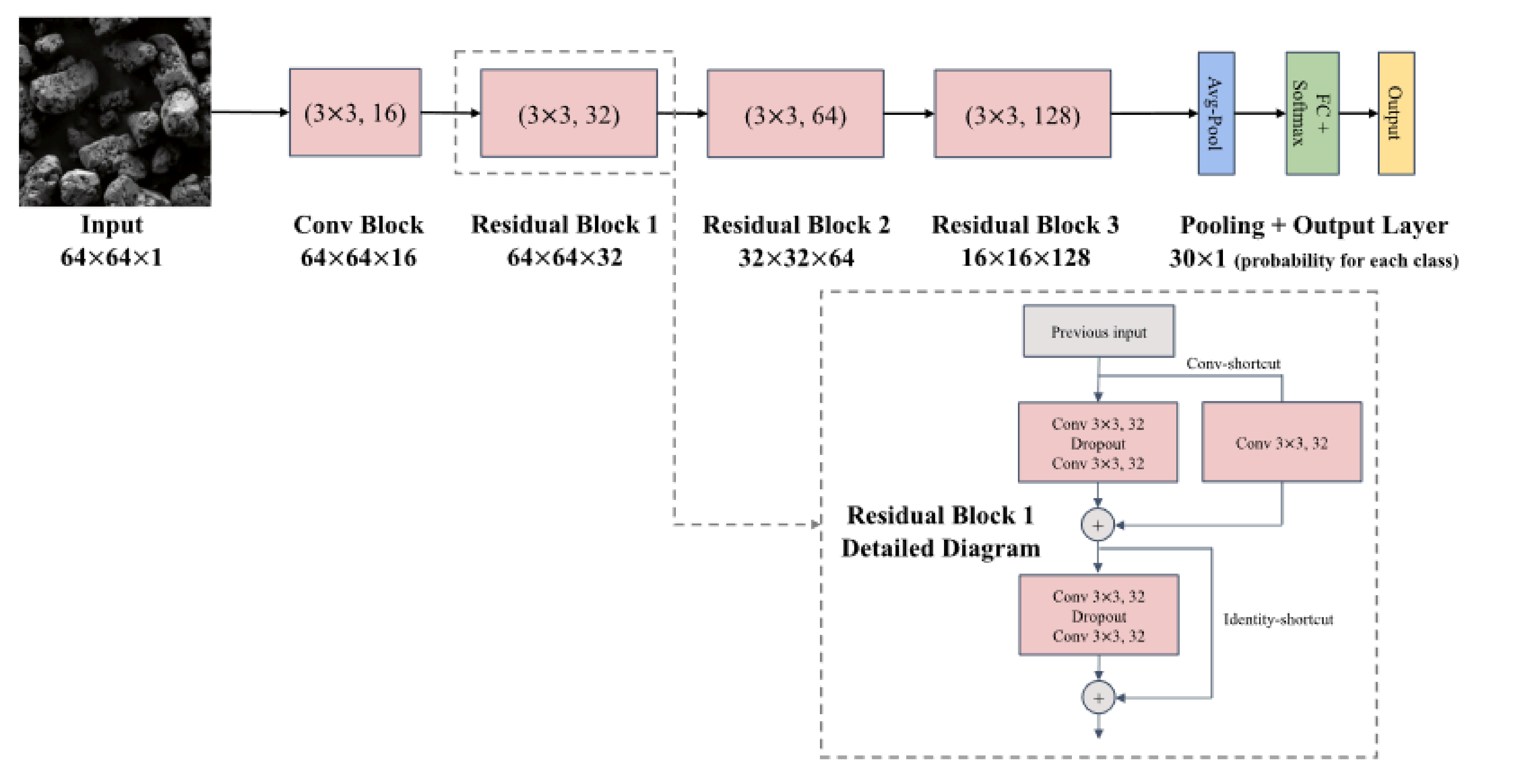
- 60k color images
- 32x32 pixels
- 10 classes



Architectures

WANG2023-28-10
WANG2023-70-16
CUI2023-DECOUPLED-28-10

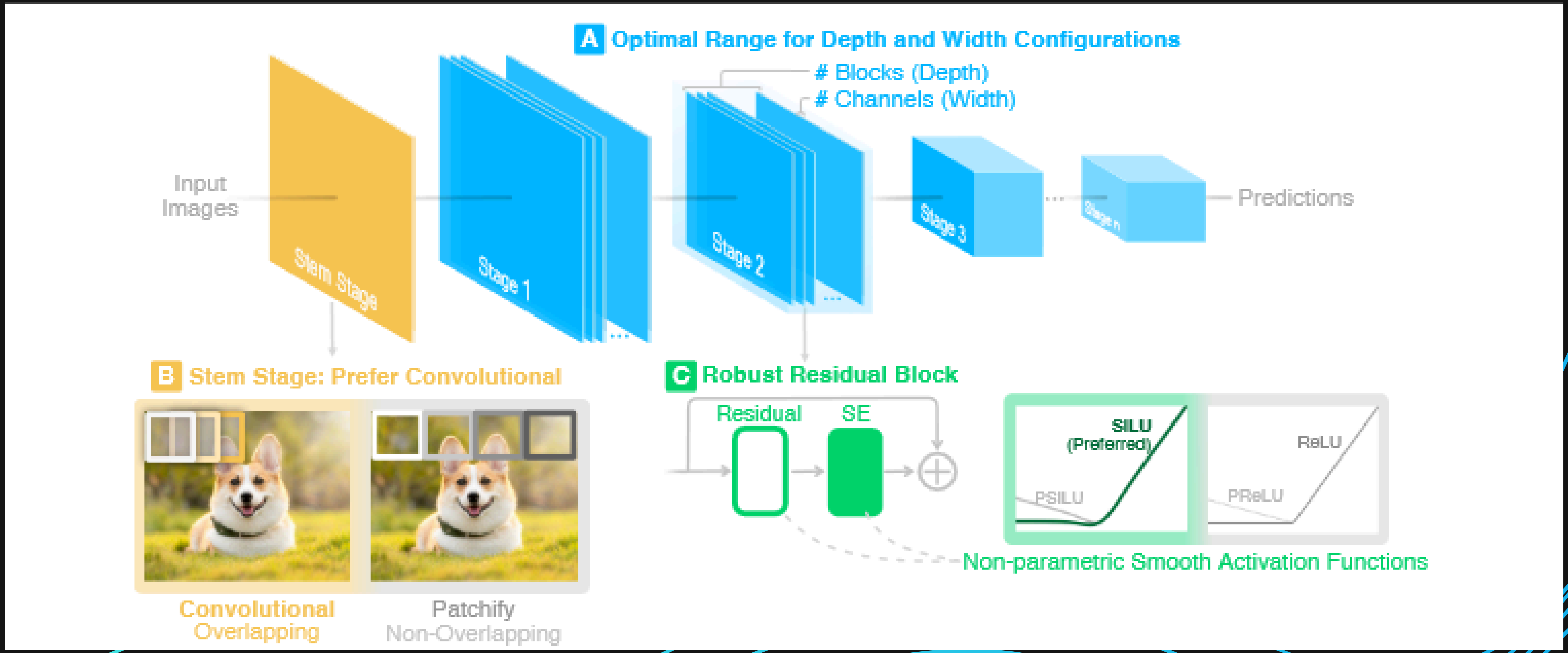
WideResNet-28-10 and 70-16



Architectures

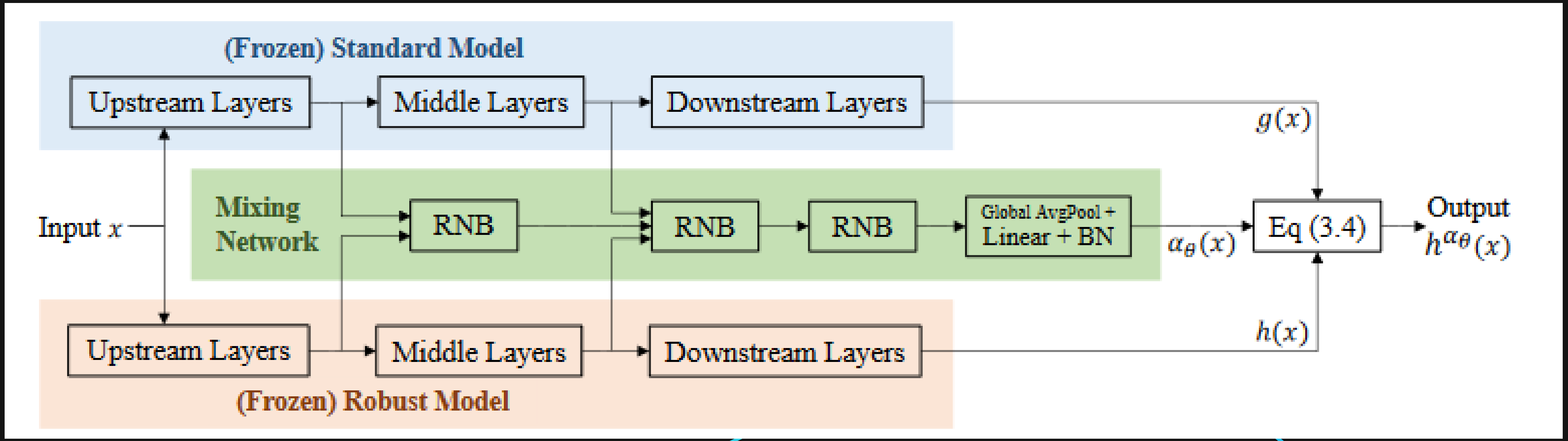
PENG2023ROBUST-70-16

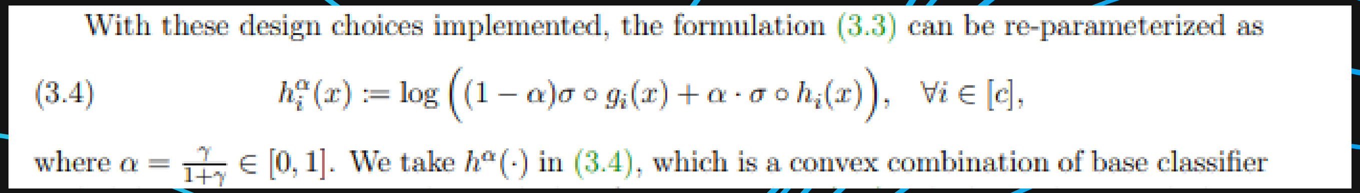
RaWideResNet-70-16



Architectures

RESNET-152 + WIDERESNET-70-16 + MIXING NETWORK,



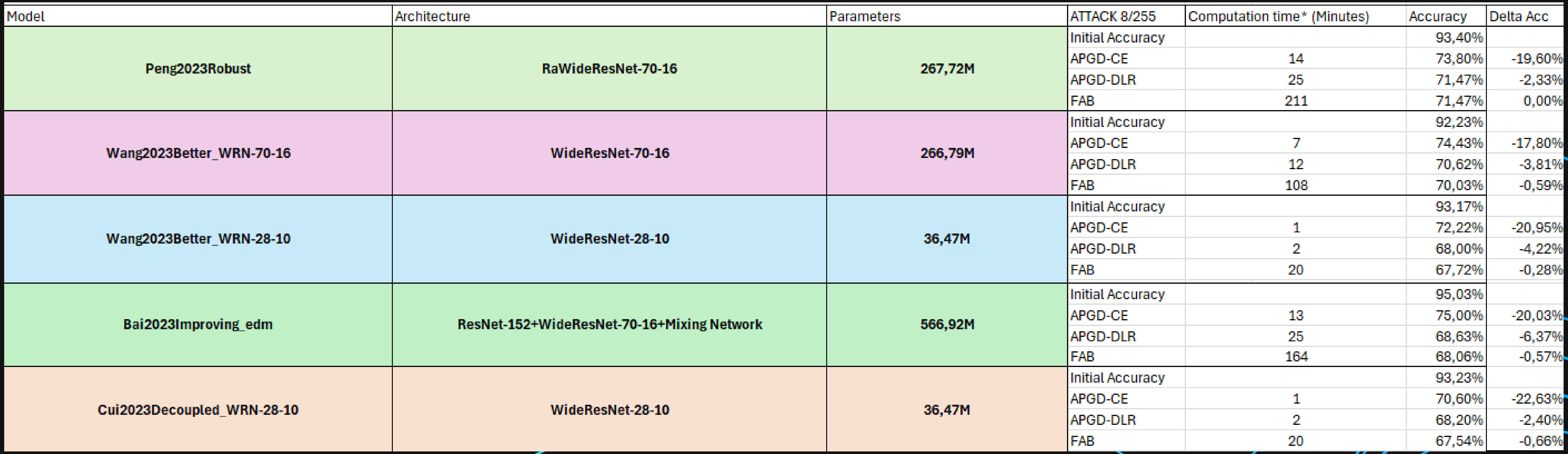

With these design choices implemented, the formulation (3.3) can be re-parameterized as

$$h_i^\alpha(x) := \log\Big((1-\alpha)\sigma \circ g_i(x) + \alpha \cdot \sigma \circ h_i(x)\Big), \quad \forall i \in [c], \tag{3.4}$$

where $\alpha = \frac{\gamma}{1+\gamma} \in [0,1]$. We take $h^\alpha(\cdot)$ in (3.4), which is a convex combination of base classifier

Results

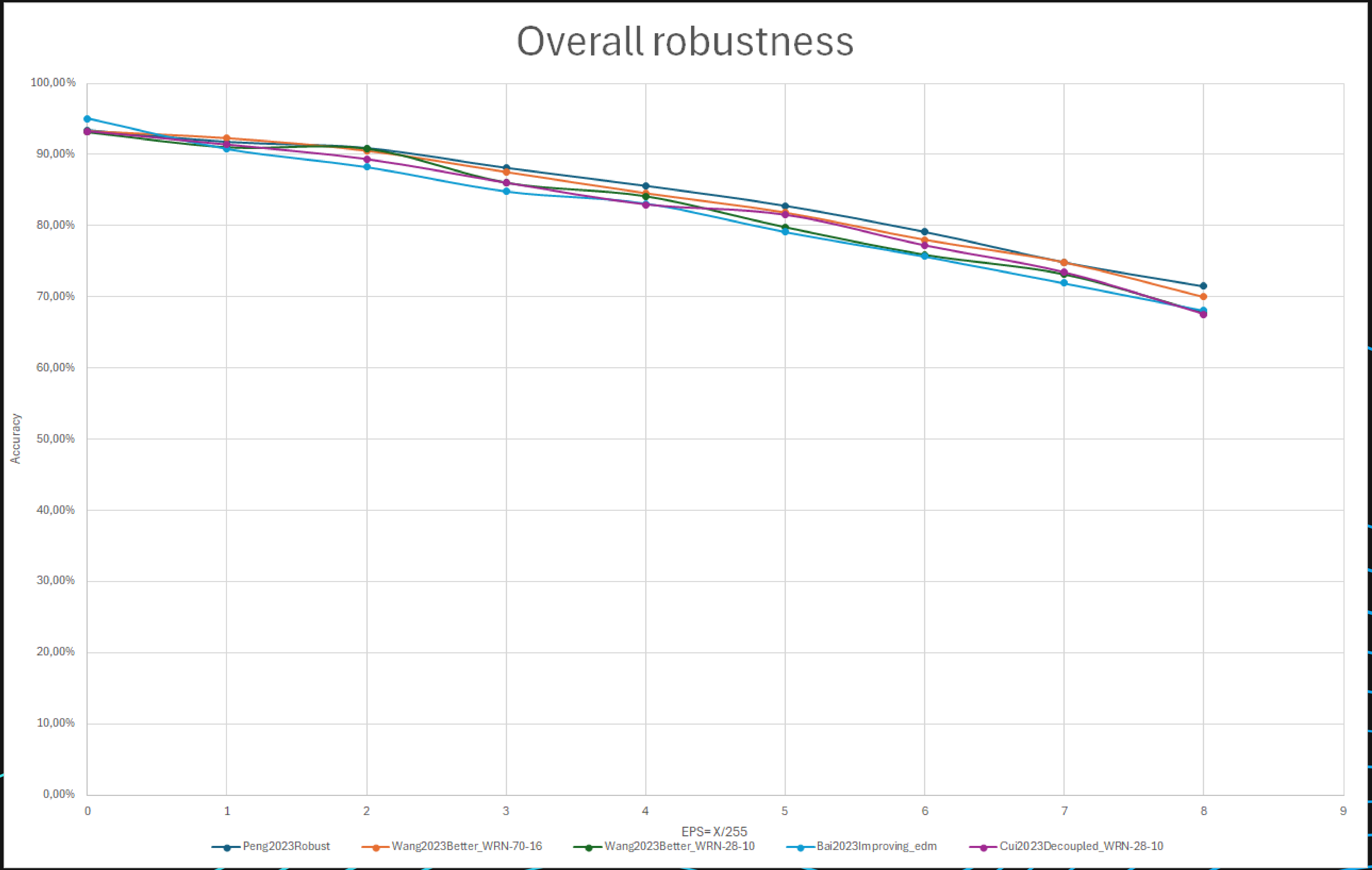
eps = 8/255 on 500 samples



Model	Architecture	Parameters	ATTACK 8/255	Computation time* (Minutes)	Accuracy	Delta Acc
Peng2023Robust	RaWideResNet-70-16	267,72M	Initial Accuracy		93,40%	
			APGD-CE	14	73,80%	-19,60%
			APGD-DLR	25	71,47%	-2,33%
			FAB	211	71,47%	0,00%
Wang2023Better_WRN-70-16	WideResNet-70-16	266,79M	Initial Accuracy		92,23%	
			APGD-CE	7	74,43%	-17,80%
			APGD-DLR	12	70,62%	-3,81%
			FAB	108	70,03%	-0,59%
Wang2023Better_WRN-28-10	WideResNet-28-10	36,47M	Initial Accuracy		93,17%	
			APGD-CE	1	72,22%	-20,95%
			APGD-DLR	2	68,00%	-4,22%
			FAB	20	67,72%	-0,28%
Bai2023Improving_edm	ResNet-152+WideResNet-70-16+Mixing Network	566,92M	Initial Accuracy		95,03%	
			APGD-CE	13	75,00%	-20,03%
			APGD-DLR	25	68,63%	-6,37%
			FAB	164	68,06%	-0,57%
Cui2023Decoupled_WRN-28-10	WideResNet-28-10	36,47M	Initial Accuracy		93,23%	
			APGD-CE	1	70,60%	-22,63%
			APGD-DLR	2	68,20%	-2,40%
			FAB	20	67,54%	-0,66%

Results

eps = x/255



Results

Classes (0:3)

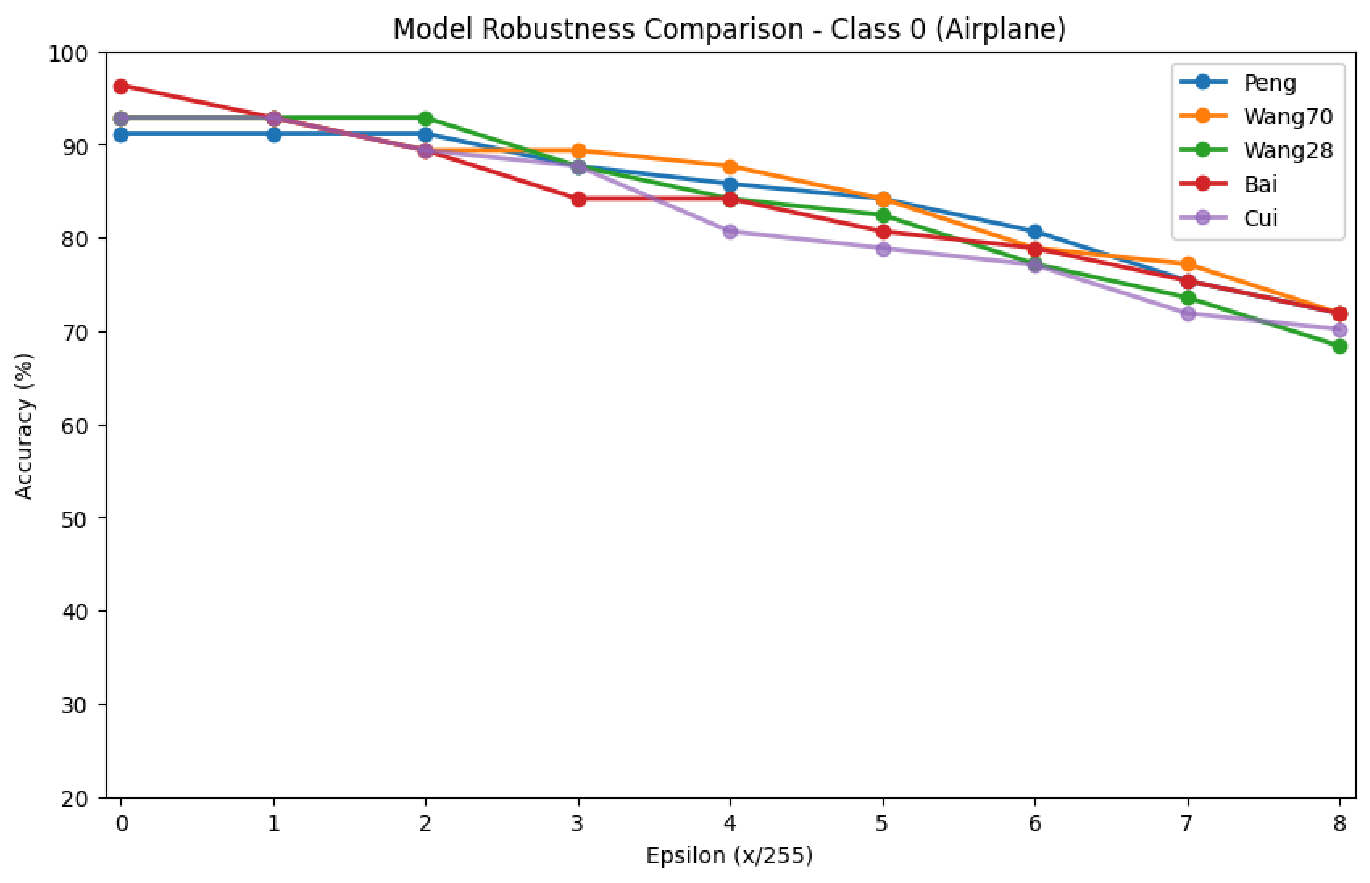
AVG. Drop:

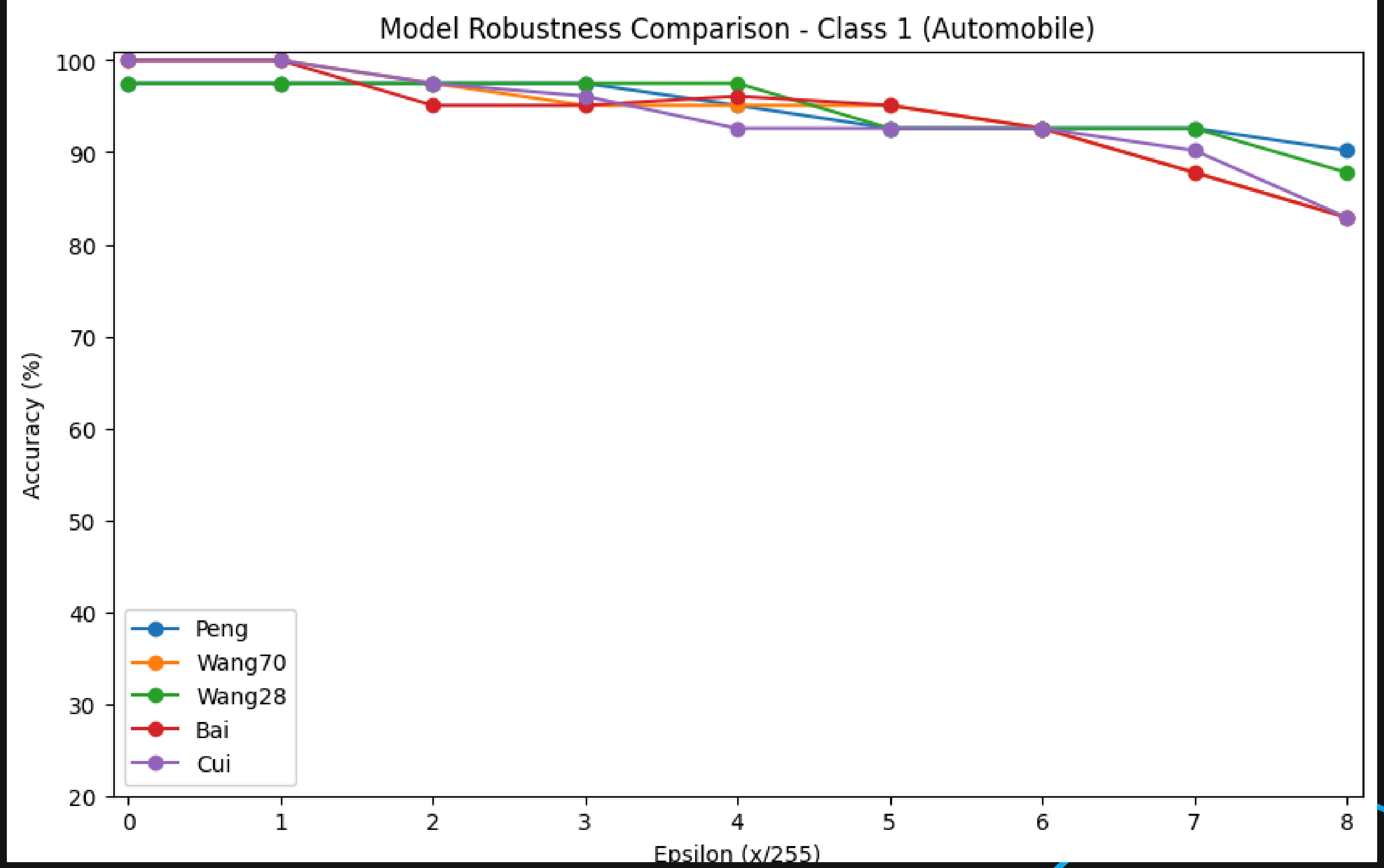
0 - 21.4%

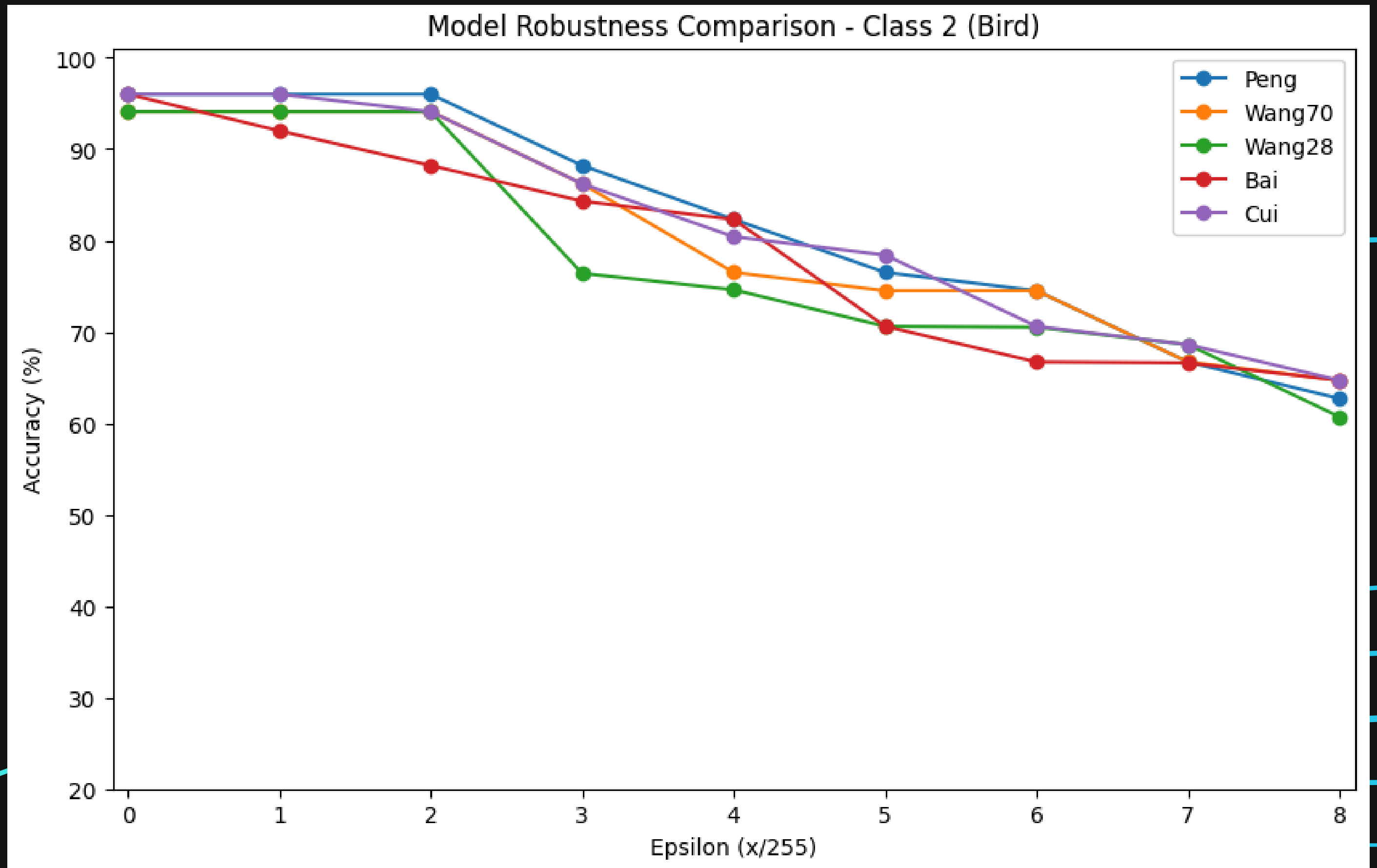
1 - 12.7%

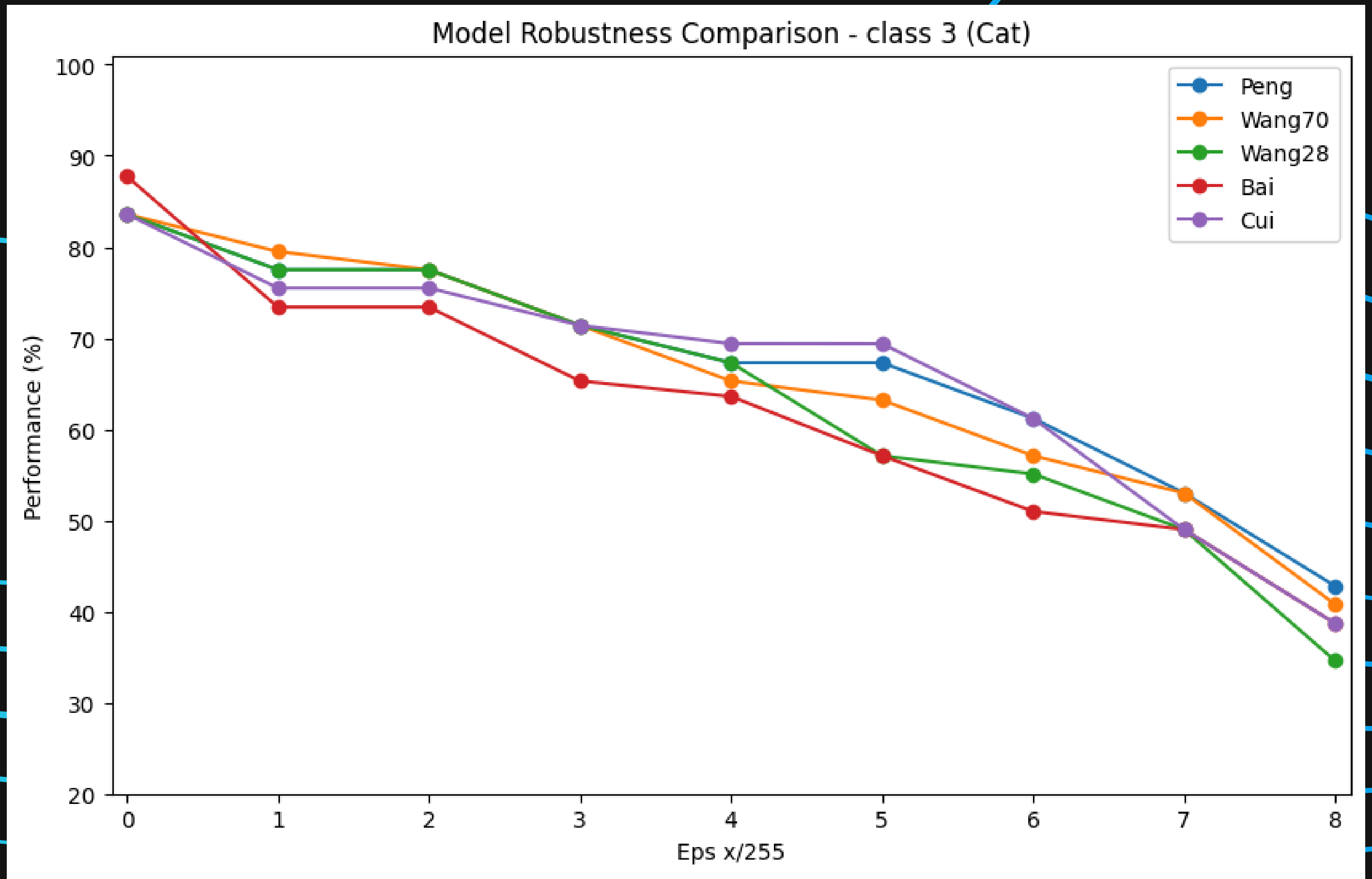
2 - 30.1%

3 - 42.4%


Model Robustness Comparison - Class 0 (Airplane)


Model Robustness Comparison - Class 1 (Automobile)


Model Robustness Comparison - Class 2 (Bird)


Model Robustness Comparison - class 3 (Cat)

Results

Classes (4:7)

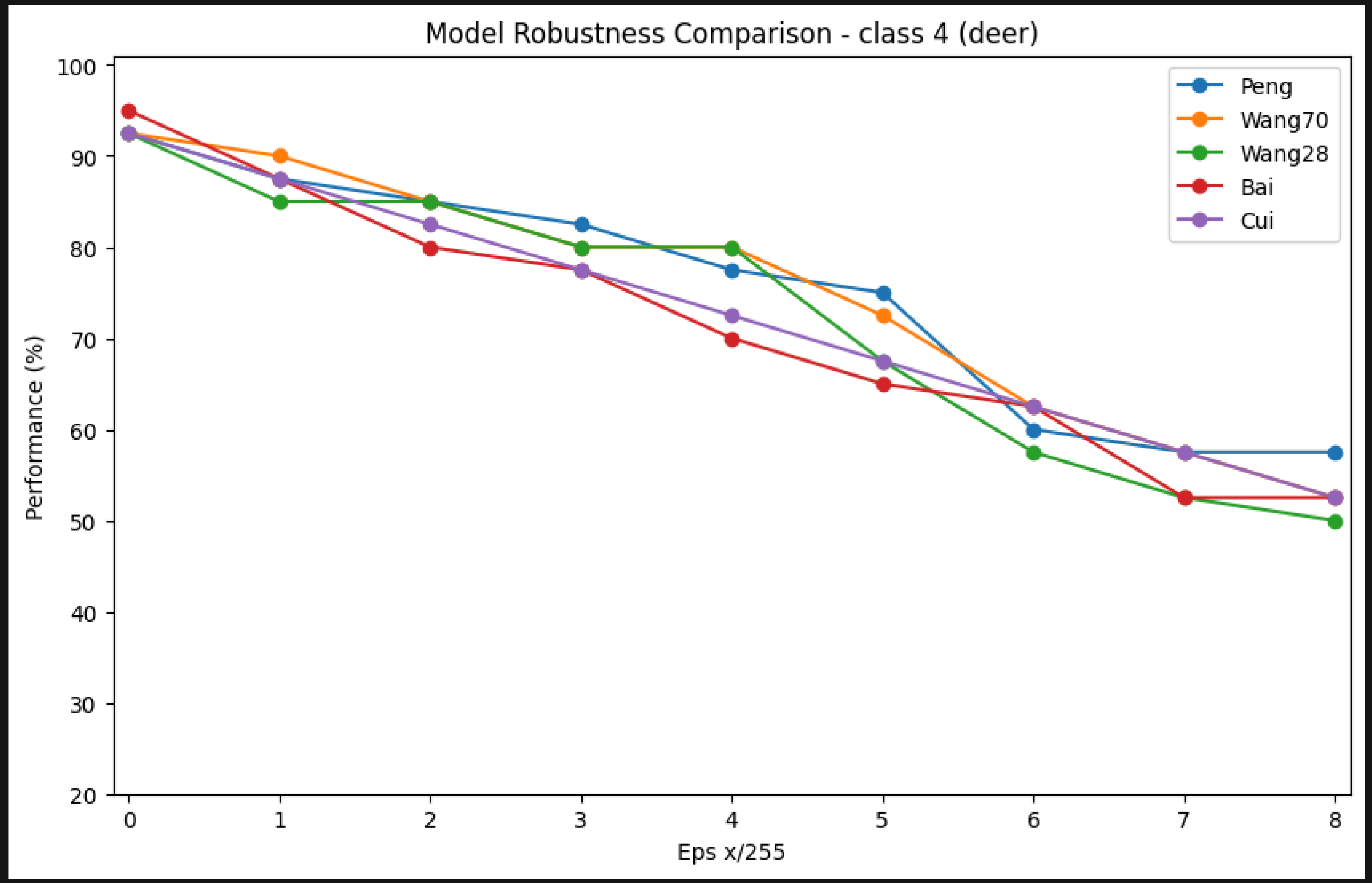
AVG. Drop:

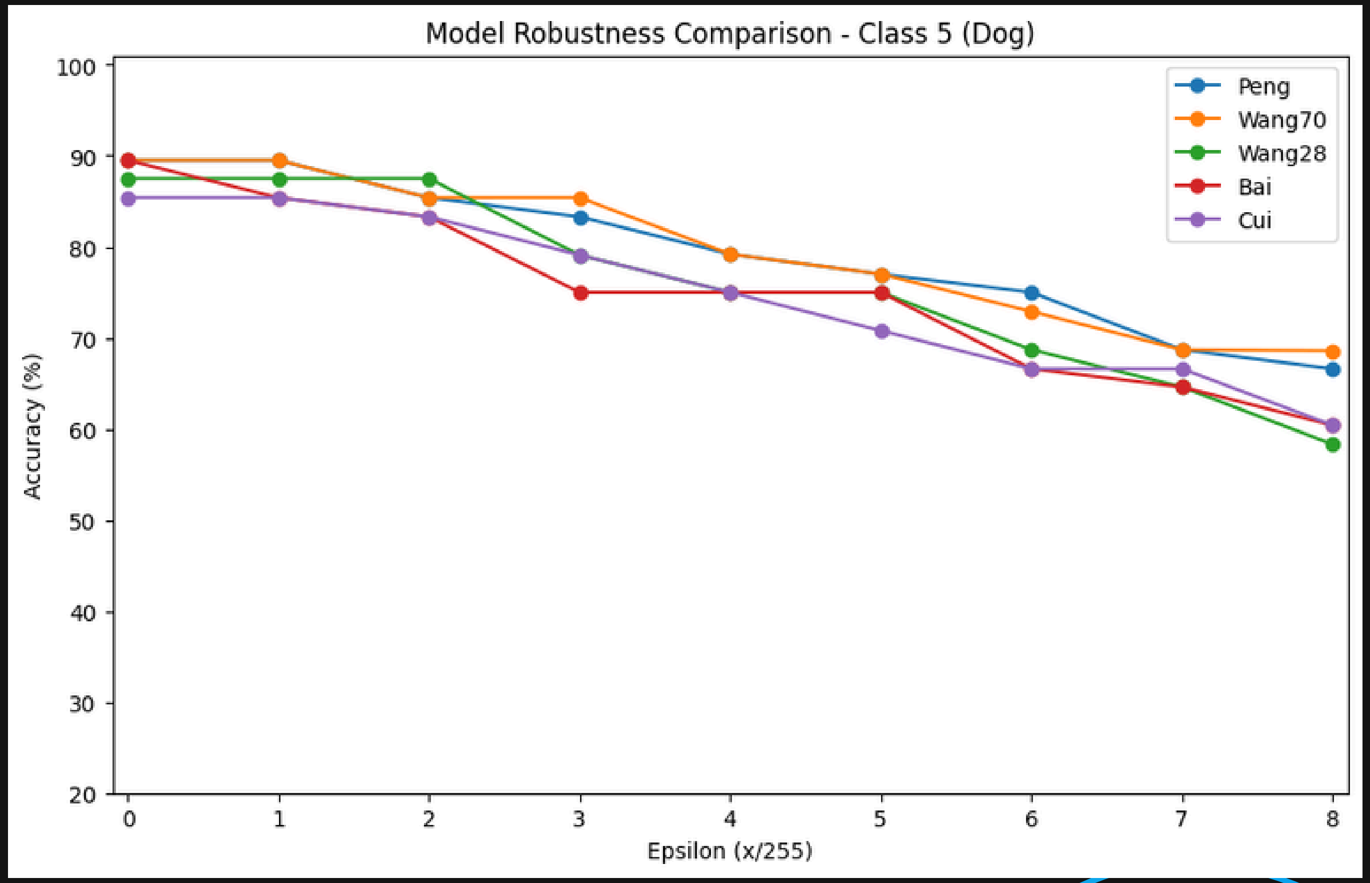
4 - 39.5%

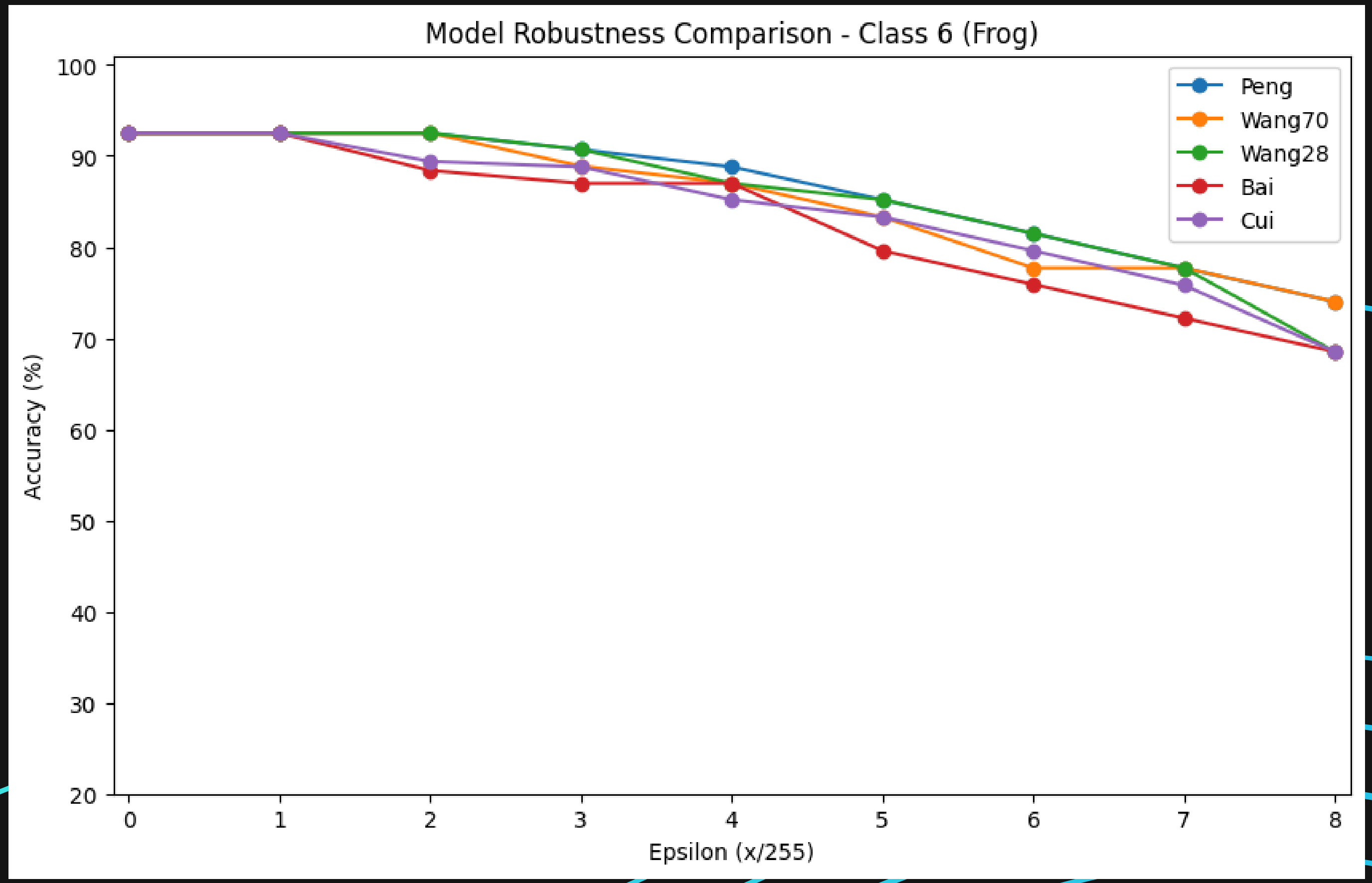
5 - 24.2%

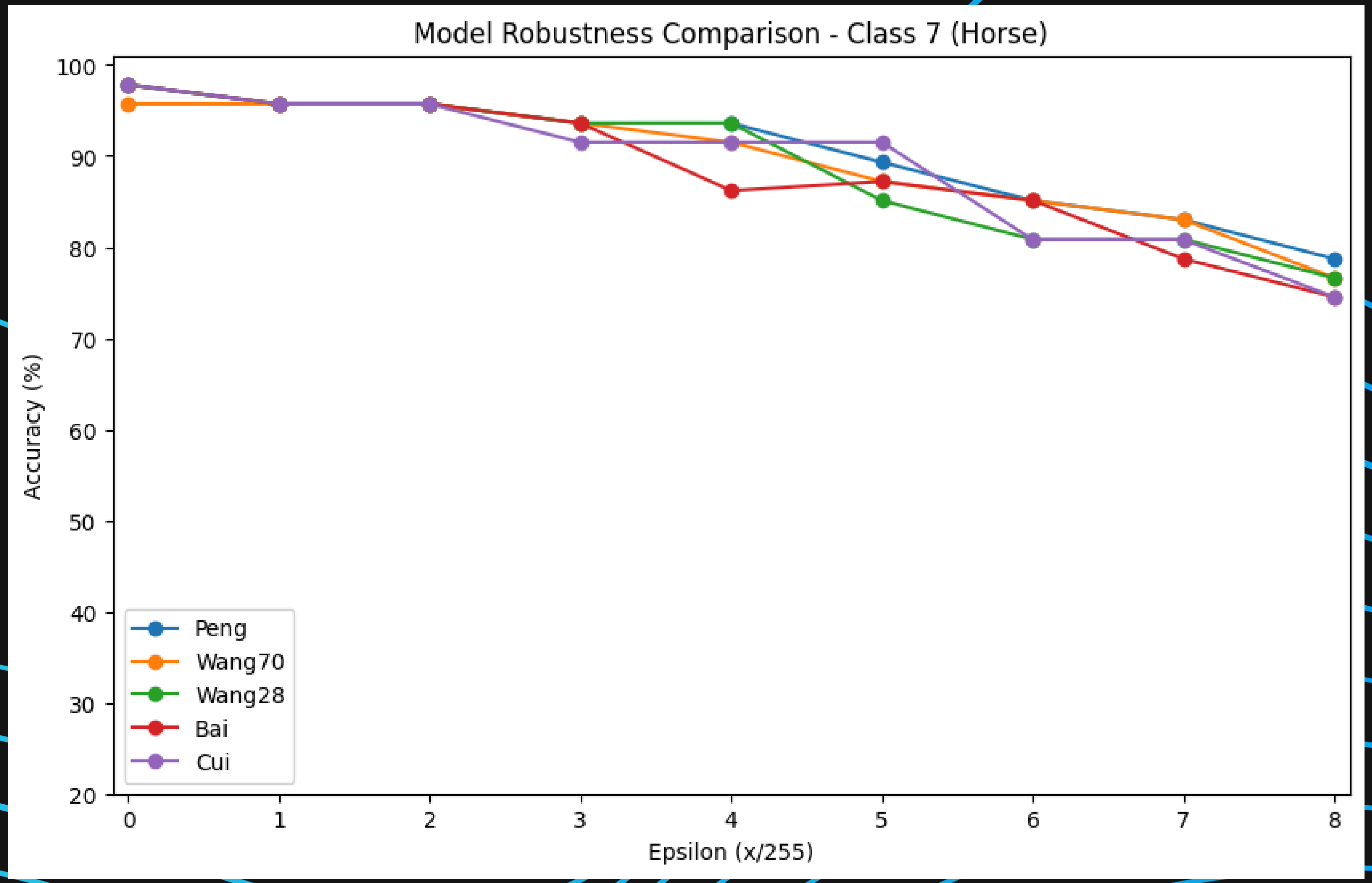
6 - 20.0%

7 - 20.4%


Model Robustness Comparison - class 4 (deer)


Model Robustness Comparison - Class 5 (Dog)


Model Robustness Comparison - Class 6 (Frog)


Model Robustness Comparison - Class 7 (Horse)

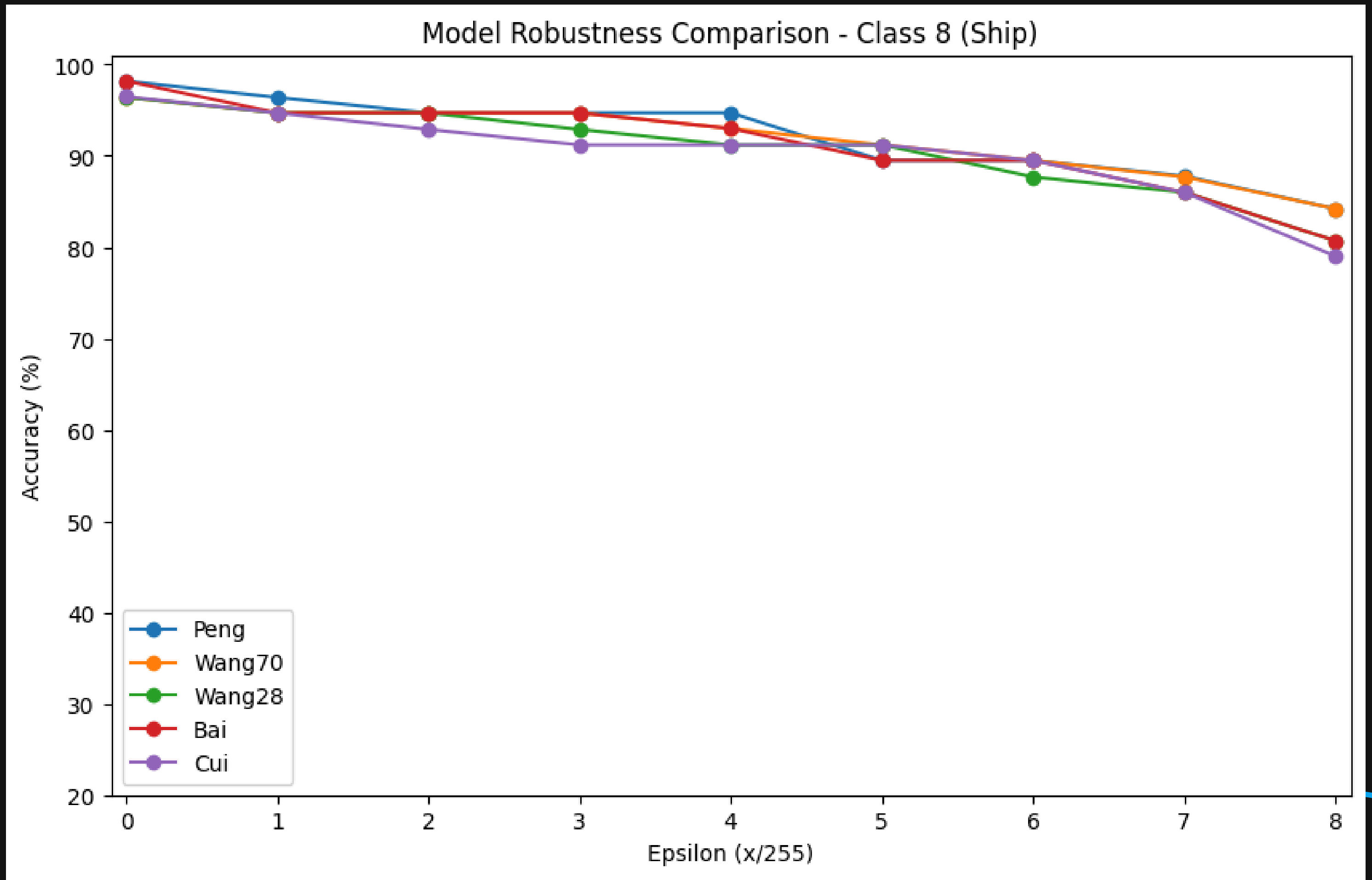
Results

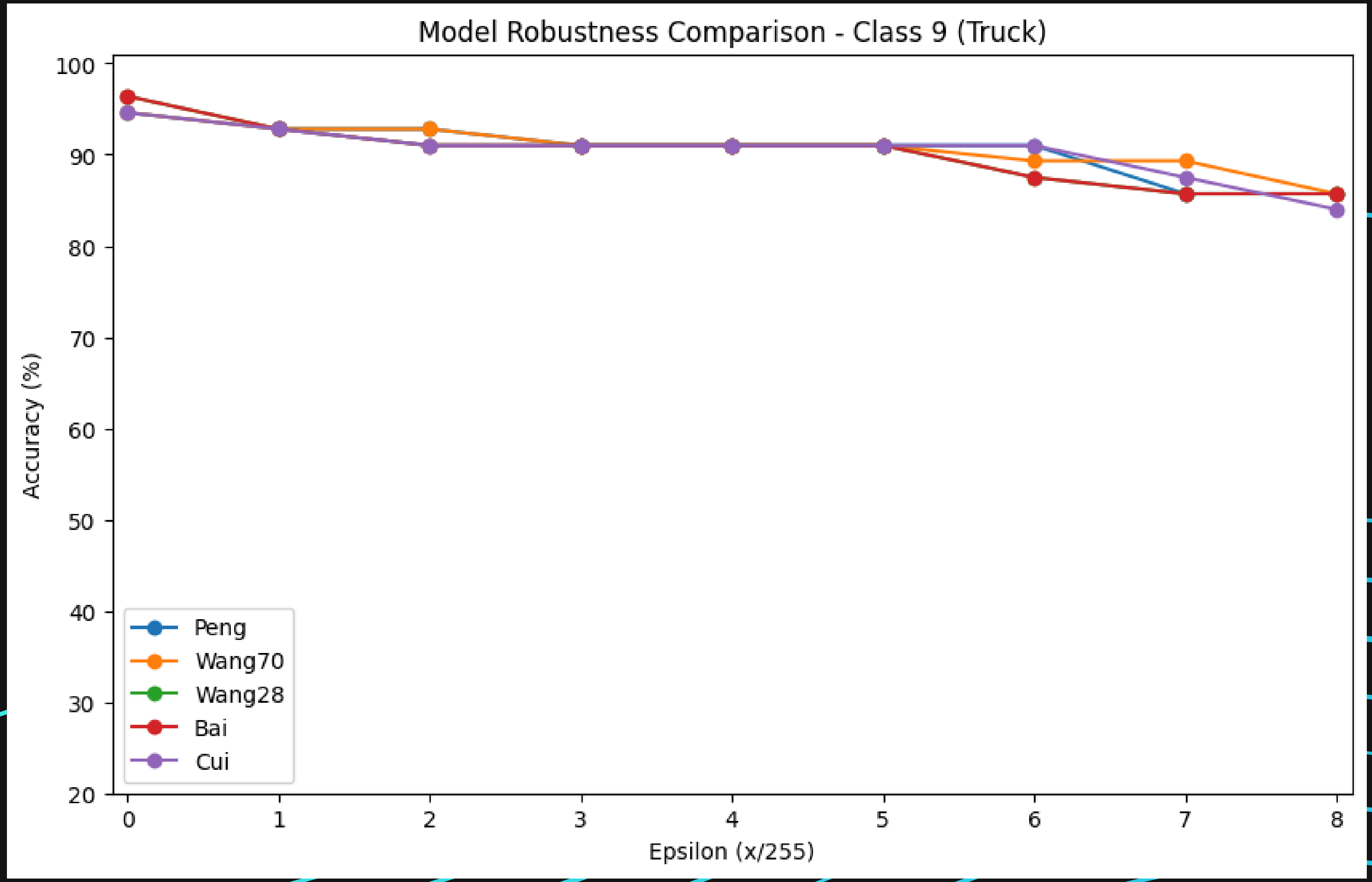
Classes (8:9)

AVG. Drop:

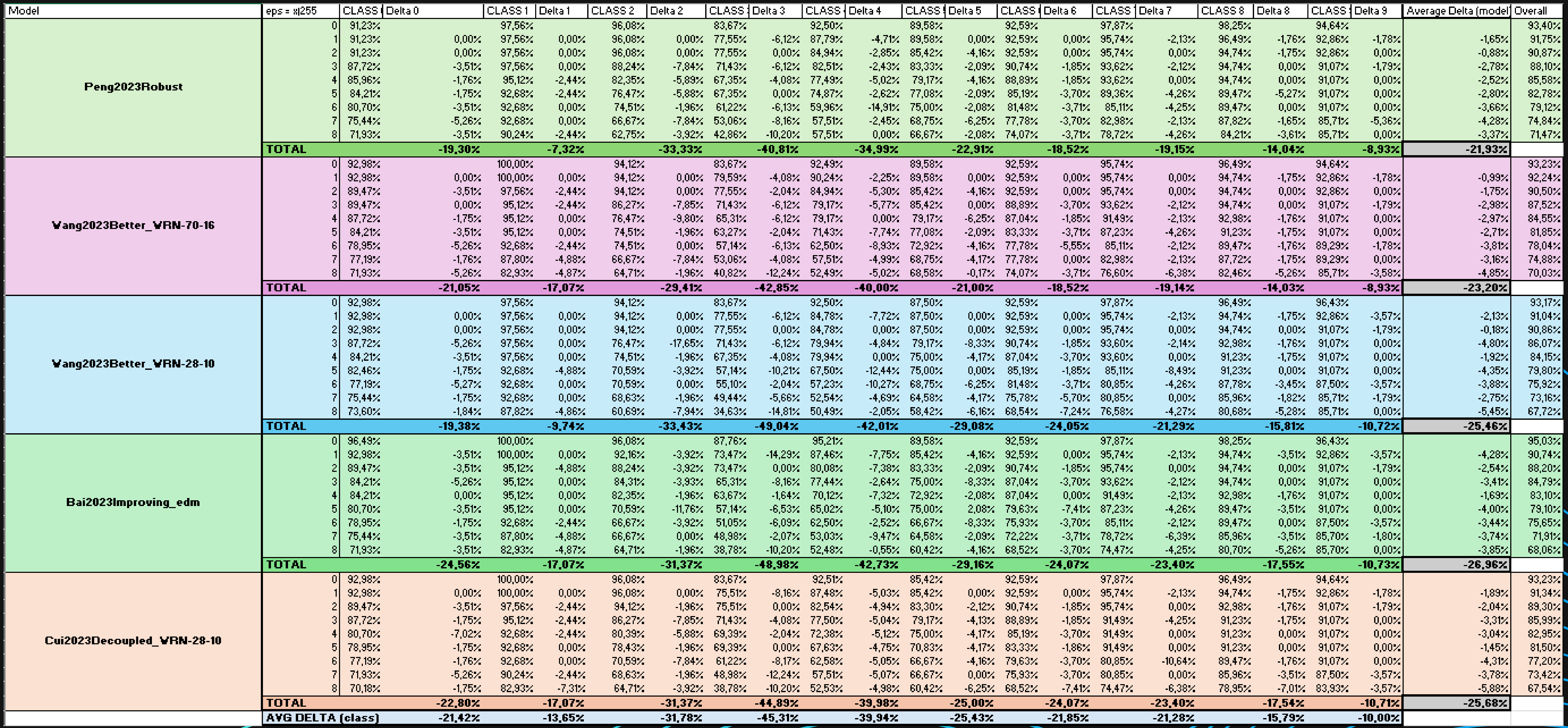
8 - 14.7%

9 - 10.0%





Overall Results



Model	eps = x/255	CLASS	Delta 0	CLASS 1	Delta 1	CLASS 2	Delta 2	CLASS	Delta 3	CLASS	Delta 4	CLASS	Delta 5	CLASS	Delta 6	CLASS	Delta 7	CLASS 8	Delta 8	CLASS	Delta 9	Average Delta (model)	Overall
Peng2023Robust	0	91,23%		97,56%		96,08%		83,67%		92,50%		89,58%		92,59%		97,87%		98,25%		94,64%			93,40%
	1	91,23%	0,00%	97,56%	0,00%	96,08%	0,00%	77,55%	-6,12%	87,79%	-4,71%	89,58%	0,00%	92,59%	0,00%	95,74%	-2,13%	96,49%	-1,76%	92,86%	-1,78%	-1,65%	91,75%
	2	91,23%	0,00%	97,56%	0,00%	96,08%	0,00%	77,55%	0,00%	84,94%	-2,85%	85,42%	-4,16%	92,59%	0,00%	95,74%	0,00%	94,74%	-1,75%	92,86%	0,00%	-0,88%	90,87%
	3	87,72%	-3,51%	97,56%	0,00%	88,24%	-7,84%	71,43%	-6,12%	82,51%	-2,43%	83,33%	-2,09%	90,74%	-1,85%	93,62%	-2,12%	94,74%	0,00%	91,07%	-1,79%	-2,78%	88,10%
	4	85,96%	-1,76%	95,12%	-2,44%	82,35%	-5,89%	67,35%	-4,08%	77,49%	-5,02%	79,17%	-4,16%	88,89%	-1,85%	93,62%	0,00%	94,74%	0,00%	91,07%	0,00%	-2,52%	85,58%
	5	84,21%	-1,75%	92,68%	-2,44%	76,47%	-5,88%	67,35%	0,00%	74,87%	-2,62%	77,08%	-2,09%	85,19%	-3,70%	89,36%	-4,26%	89,47%	-5,27%	91,07%	0,00%	-2,80%	82,78%
	6	80,70%	-3,51%	92,68%	0,00%	74,51%	-1,96%	61,22%	-6,13%	59,96%	-14,91%	75,00%	-2,08%	81,48%	-3,71%	85,11%	-4,25%	89,47%	0,00%	91,07%	0,00%	-3,66%	79,12%
	7	75,44%	-5,26%	92,68%	0,00%	66,67%	-7,84%	53,06%	-8,16%	57,51%	-2,45%	68,75%	-6,25%	77,78%	-3,70%	82,98%	-2,13%	87,82%	-1,65%	85,71%	-5,36%	-4,28%	74,84%
	8	71,93%	-3,51%	90,24%	-2,44%	62,75%	-3,92%	42,86%	-10,20%	57,51%	0,00%	66,67%	-2,08%	74,07%	-3,71%	78,72%	-4,26%	84,21%	-3,61%	85,71%	0,00%	-3,37%	71,47%
	TOTAL		**-19,30%**		**-7,32%**		**-33,33%**		**-40,81%**		**-34,99%**		**-22,91%**		**-18,52%**		**-19,15%**		**-14,04%**		**-8,93%**	**-21,93%**	
Wang2023Better_WRN-70-16	0	92,98%		100,00%		94,12%		83,67%		92,49%		89,58%		92,59%		95,74%		96,49%		94,64%			93,23%
	1	92,98%	0,00%	100,00%	0,00%	94,12%	0,00%	79,59%	-4,08%	90,24%	-2,25%	89,58%	0,00%	92,59%	0,00%	95,74%	0,00%	94,74%	-1,75%	92,86%	-1,78%	-0,99%	92,24%
	2	89,47%	-3,51%	97,56%	-2,44%	94,12%	0,00%	77,55%	-2,04%	84,94%	-5,30%	85,42%	-4,16%	92,59%	0,00%	95,74%	0,00%	94,74%	0,00%	92,86%	0,00%	-1,75%	90,50%
	3	89,47%	0,00%	95,12%	-2,44%	86,27%	-7,85%	71,43%	-6,12%	79,17%	-5,77%	85,42%	0,00%	88,89%	-3,70%	93,62%	-2,12%	94,74%	0,00%	91,07%	-1,79%	-2,98%	87,52%
	4	87,72%	-1,75%	95,12%	0,00%	76,47%	-9,80%	65,31%	-6,12%	79,17%	0,00%	79,17%	-6,25%	87,04%	-1,85%	91,49%	-2,13%	92,98%	-1,76%	91,07%	0,00%	-2,97%	84,55%
	5	84,21%	-3,51%	95,12%	0,00%	74,51%	-1,96%	63,27%	-2,04%	71,43%	-7,74%	77,08%	-2,09%	83,33%	-3,71%	87,23%	-4,26%	91,23%	-1,75%	91,07%	0,00%	-2,71%	81,85%
	6	78,95%	-5,26%	92,68%	-2,44%	74,51%	0,00%	57,14%	-6,13%	62,50%	-8,93%	72,92%	-4,16%	77,78%	-5,55%	85,11%	-2,12%	89,47%	-1,76%	89,29%	-1,78%	-3,81%	78,04%
	7	77,19%	-1,76%	87,80%	-4,88%	66,67%	-7,84%	53,06%	-4,08%	57,51%	-4,99%	68,75%	-4,17%	77,78%	0,00%	82,98%	-2,13%	87,72%	-1,75%	89,29%	0,00%	-3,16%	74,88%
	8	71,93%	-5,26%	82,93%	-4,87%	64,71%	-1,96%	40,82%	-12,24%	52,49%	-5,02%	68,58%	-0,17%	74,07%	-3,71%	76,60%	-6,38%	82,46%	-5,26%	85,71%	-3,58%	-4,85%	70,03%
	TOTAL		**-21,05%**		**-17,07%**		**-29,41%**		**-42,85%**		**-40,00%**		**-21,00%**		**-18,52%**		**-19,14%**		**-14,03%**		**-8,93%**	**-23,20%**	
Wang2023Better_WRN-28-10	0	92,98%		97,56%		94,12%		83,67%		92,50%		87,50%		92,59%		97,87%		96,49%		96,43%			93,17%
	1	92,98%	0,00%	97,56%	0,00%	94,12%	0,00%	77,55%	-6,12%	84,78%	-7,72%	87,50%	0,00%	92,59%	0,00%	95,74%	-2,13%	94,74%	-1,75%	92,86%	-3,57%	-2,13%	91,04%
	2	92,98%	0,00%	97,56%	0,00%	94,12%	0,00%	77,55%	0,00%	84,78%	0,00%	87,50%	0,00%	92,59%	0,00%	95,74%	0,00%	94,74%	0,00%	91,07%	-1,79%	-0,18%	90,86%
	3	87,72%	-5,26%	97,56%	0,00%	76,47%	-17,65%	71,43%	-6,12%	79,94%	-4,84%	79,17%	-8,33%	90,74%	-1,85%	93,60%	-2,14%	92,98%	-1,76%	91,07%	0,00%	-4,80%	86,07%
	4	84,21%	-3,51%	97,56%	0,00%	74,51%	-1,96%	67,35%	-4,08%	79,94%	0,00%	75,00%	-4,17%	87,04%	-3,70%	93,60%	0,00%	91,23%	-1,75%	91,07%	0,00%	-1,92%	84,15%
	5	82,46%	-1,75%	92,68%	-4,88%	70,59%	-3,92%	57,14%	-10,21%	67,50%	-12,44%	75,00%	0,00%	85,19%	-1,85%	85,11%	-8,49%	91,23%	0,00%	91,07%	0,00%	-4,35%	79,80%
	6	77,19%	-5,27%	92,68%	0,00%	70,59%	0,00%	55,10%	-2,04%	57,23%	-10,27%	68,75%	-6,25%	81,48%	-3,71%	80,85%	-4,26%	87,78%	-3,45%	87,50%	-3,57%	-3,88%	75,92%
	7	75,44%	-1,75%	92,68%	0,00%	68,63%	-1,96%	49,44%	-5,66%	52,54%	-4,69%	64,58%	-4,17%	75,78%	-5,70%	80,85%	0,00%	85,96%	-1,82%	85,71%	-1,79%	-2,75%	73,16%
	8	73,60%	-1,84%	87,82%	-4,86%	60,69%	-7,94%	34,63%	-14,81%	50,49%	-2,05%	58,42%	-6,16%	68,54%	-7,24%	76,58%	-4,27%	80,68%	-5,28%	85,71%	0,00%	-5,45%	67,72%
	TOTAL		**-19,38%**		**-9,74%**		**-33,43%**		**-49,04%**		**-42,01%**		**-29,08%**		**-24,05%**		**-21,29%**		**-15,81%**		**-10,72%**	**-25,46%**	
Bai2023Improving_edm	0	96,49%		100,00%		96,08%		87,76%		95,21%		89,58%		92,59%		97,87%		98,25%		96,43%			95,03%
	1	92,98%	-3,51%	100,00%	0,00%	92,16%	-3,92%	73,47%	-14,29%	87,46%	-7,75%	85,42%	-4,16%	92,59%	0,00%	95,74%	-2,13%	94,74%	-3,51%	92,86%	-3,57%	-4,28%	90,74%
	2	89,47%	-3,51%	95,12%	-4,88%	88,24%	-3,92%	73,47%	0,00%	80,08%	-7,38%	83,33%	-2,09%	90,74%	-1,85%	95,74%	0,00%	94,74%	0,00%	91,07%	-1,79%	-2,54%	88,20%
	3	84,21%	-5,26%	95,12%	0,00%	84,31%	-3,93%	65,31%	-8,16%	77,44%	-2,64%	75,00%	-8,33%	87,04%	-3,70%	93,62%	-2,12%	94,74%	0,00%	91,07%	0,00%	-3,41%	84,79%
	4	84,21%	0,00%	95,12%	0,00%	82,35%	-1,96%	63,67%	-1,64%	70,12%	-7,32%	72,92%	-2,08%	87,04%	0,00%	91,49%	-2,13%	92,98%	-1,76%	91,07%	0,00%	-1,69%	83,10%
	5	80,70%	-3,51%	95,12%	0,00%	70,59%	-11,76%	57,14%	-6,53%	65,02%	-5,10%	75,00%	2,08%	79,63%	-7,41%	87,23%	-4,26%	89,47%	-3,51%	91,07%	0,00%	-4,00%	79,10%
	6	78,95%	-1,75%	92,68%	-2,44%	66,67%	-3,92%	51,05%	-6,09%	62,50%	-2,52%	66,67%	-8,33%	75,93%	-3,70%	85,11%	-2,12%	89,47%	0,00%	87,50%	-3,57%	-3,44%	75,65%
	7	75,44%	-3,51%	87,80%	-4,88%	66,67%	0,00%	48,98%	-2,07%	53,03%	-9,47%	64,58%	-2,09%	72,22%	-3,71%	78,72%	-6,39%	85,96%	-3,51%	85,70%	-1,80%	-3,74%	71,91%
	8	71,93%	-3,51%	82,93%	-4,87%	64,71%	-1,96%	38,78%	-10,20%	52,48%	-0,55%	60,42%	-4,16%	68,52%	-3,70%	74,47%	-4,25%	80,70%	-5,26%	85,70%	0,00%	-3,85%	68,06%
	TOTAL		**-24,56%**		**-17,07%**		**-31,37%**		**-48,98%**		**-42,73%**		**-29,16%**		**-24,07%**		**-23,40%**		**-17,55%**		**-10,73%**	**-26,96%**	
Cui2023Decoupled_WRN-28-10	0	92,98%		100,00%		96,08%		83,67%		92,51%		85,42%		92,59%		97,87%		96,49%		94,64%			93,23%
	1	92,98%	0,00%	100,00%	0,00%	96,08%	0,00%	75,51%	-8,16%	87,48%	-5,03%	85,42%	0,00%	92,59%	0,00%	95,74%	-2,13%	94,74%	-1,75%	92,86%	-1,78%	-1,89%	91,34%
	2	89,47%	-3,51%	97,56%	-2,44%	94,12%	-1,96%	75,51%	0,00%	82,54%	-4,94%	83,30%	-2,12%	90,74%	-1,85%	95,74%	0,00%	92,98%	-1,76%	91,07%	-1,79%	-2,04%	89,30%
	3	87,72%	-1,75%	95,12%	-2,44%	86,27%	-7,85%	71,43%	-4,08%	77,50%	-5,04%	79,17%	-4,13%	88,89%	-1,85%	91,49%	-4,25%	91,23%	-1,75%	91,07%	0,00%	-3,31%	85,99%
	4	80,70%	-7,02%	92,68%	-2,44%	80,39%	-5,88%	69,39%	-2,04%	72,38%	-5,12%	75,00%	-4,17%	85,19%	-3,70%	91,49%	0,00%	91,23%	0,00%	91,07%	0,00%	-3,04%	82,95%
	5	78,95%	-1,75%	92,68%	0,00%	78,43%	-1,96%	69,39%	0,00%	67,63%	-4,75%	70,83%	-4,17%	83,33%	-1,86%	91,49%	0,00%	91,23%	0,00%	91,07%	0,00%	-1,45%	81,50%
	6	77,19%	-1,76%	92,68%	0,00%	70,59%	-7,84%	61,22%	-8,17%	62,58%	-5,05%	66,67%	-4,16%	79,63%	-3,70%	80,85%	-10,64%	89,47%	-1,76%	91,07%	0,00%	-4,31%	77,20%
	7	71,93%	-5,26%	90,24%	-2,44%	68,63%	-1,96%	48,98%	-12,24%	57,51%	-5,07%	66,67%	0,00%	75,93%	-3,70%	80,85%	0,00%	85,96%	-3,51%	87,50%	-3,57%	-3,78%	73,42%
	8	70,18%	-1,75%	82,93%	-7,31%	64,71%	-3,92%	38,78%	-10,20%	52,53%	-4,98%	60,42%	-6,25%	68,52%	-7,41%	74,47%	-6,38%	78,95%	-7,01%	83,93%	-3,57%	-5,88%	67,54%
	TOTAL		**-22,80%**		**-17,07%**		**-31,37%**		**-44,89%**		**-39,98%**		**-25,00%**		**-24,07%**		**-23,40%**		**-17,54%**		**-10,71%**	**-25,68%**	
	AVG DELTA (class)		**-21,42%**		**-13,65%**		**-31,78%**		**-45,31%**		**-39,94%**		**-25,43%**		**-21,85%**		**-21,28%**		**-15,79%**		**-10,00%**		

Observations

CLASSES

A single class can directly affect the overall robustness of the model, this leads to an issue on explainability: Are our results related to the model or by the dataset's properties and how it is used?

PARAMETERS

The number of parameters doesn't looks to affect the robustness of the model, but it seems to affect the time needed to compute the adversarial sample by AutoAttack, this together with the model architecture complexity

MODEL ARCHITECTURE

The reasoning behind the model architecture seems to be the an important factor in robustness. A deeper analysis with different models and datasets in training is required.