

	Model	τ Lm	τ Fm	AUROC Lm	AUROC Fm	AUPR Lm	AUPR Fm	FPR95 Lm	FPR95 Fm	Acc	Rob. Acc	Architecture
CIFAR10	DI0	0.28	0.32	67.49	70.75	70.91	74.28	82.56	80.57	84.36	41.44	WideResNet-28-4
	XU80	0.43	0.45	83.30	84.30	80.50	82.25	83.42	82.44	93.69	63.89	WideResNet-28-10
	MR0	0.68	0.70	92.95	93.73	94.92	95.54	29.76	27.43	79.69	39.12	ResNet-18
	AL0	0.72	0.74	94.67	95.12	95.98	96.28	24.93	22.21	80.38	40.21	ResNet-18
	CU80	0.73	0.75	96.87	97.52	94.42	95.47	17.90	14.77	92.16	67.73	WideResNet-28-10
	WA80	0.74	0.76	96.82	97.50	94.33	95.43	17.60	14.42	92.44	67.31	WideResNet-28-10
	SE10	0.74	0.74	96.03	96.59	94.66	95.52	19.13	16.85	84.59	55.54	ResNet-18
	EN0	0.74	0.76	95.16	95.90	95.07	95.85	24.10	21.22	87.03	49.25	ResNet-50
	TR0	0.74	0.77	94.63	95.52	96.13	96.74	30.93	25.73	80.72	42.23	ResNet-18
	DS0	0.75	0.75	95.80	95.93	95.08	95.23	24.65	23.34	90.06	56.14	XCiT-S12
	MD0	0.75	0.75	95.36	95.29	97.00	97.00	23.23	23.75	81.85	36.91	ResNet-18
	ZH0	0.75	0.77	95.86	96.35	95.65	96.14	24.91	21.74	84.92	53.08	WideResNet-34-10
	AD10	0.82	0.84	97.86	98.26	97.68	98.09	13.26	11.50	80.24	51.06	ResNet-18
CIFAR100	HE1	0.74	0.74	94.43	94.41	97.39	97.39	30.40	30.40	59.23	28.42	WideResNet-28-10
	WU1	0.78	0.79	95.81	95.83	98.00	98.01	23.34	22.55	60.38	28.86	WideResNet-34-10
	RE812	0.80	0.80	96.87	96.91	98.30	98.32	18.06	17.49	62.41	32.06	WideResNet-28-10
	DS1	0.81	0.81	96.78	96.78	98.30	98.30	19.18	18.46	67.34	32.19	XCiT-S12
	CU41	0.82	0.82	97.07	97.09	98.48	98.49	17.21	17.35	64.08	31.65	WideResNet-34-10
	CU81	0.83	0.83	97.41	97.43	98.24	98.25	15.62	15.67	73.85	39.18	WideResNet-28-10
	RI1	0.83	0.83	96.61	96.62	99.05	99.06	18.14	17.70	53.83	18.95	PreActResNet-18
	PA21	0.83	0.83	97.66	97.70	98.82	98.84	13.83	13.80	63.66	31.08	WideResNet-28-10
	WA81	0.83	0.83	97.51	97.50	98.28	98.27	14.96	14.86	72.58	38.83	WideResNet-28-10
	AD21	0.84	0.84	97.46	97.48	98.92	98.93	16.00	15.36	65.45	27.67	ResNet-18
	AD1	0.84	0.84	97.65	97.67	98.99	98.99	13.88	13.34	62.02	27.14	PreActResNet-18
	RE11	0.85	0.84	97.97	97.88	99.05	99.01	13.21	13.36	56.87	28.50	PreActResNet-18
	RA11	0.85	0.85	98.01	98.01	99.08	99.08	12.36	12.20	61.50	28.88	PreActResNet-18

Table 1: Correlations between the input margin and the logit margin (Lm) or the exact Feature margin (Fm) and detection scores on CIFAR10 (ℓ_∞ , $\epsilon = 8/255$).

	Model ID	Kendall tau	AUROC	AUPR	FPR95	Acc	Rob. Acc	Architecture
CIFAR10 (ℓ_2)	DI02	0.66	95.91	93.05	21.94	88.02	66.09	WideResNet-28-4
	AU02	0.69	95.59	87.41	19.30	91.08	72.91	ResNet-50
	SE102	0.72	97.07	90.99	13.25	89.76	74.41	ResNet-18
	WA802	0.74	98.55	93.59	8.00	95.16	83.68	WideResNet-28-10
	RI02	0.74	96.82	92.86	13.98	88.67	67.68	PreActResNet-18
	EN02	0.75	96.92	92.78	13.25	90.83	69.24	ResNet-50
	RO02	0.76	98.19	96.51	10.15	89.05	66.44	WideResNet-28-10
	RE802	0.77	98.72	95.02	6.66	91.79	78.80	WideResNet-28-10
	RE102	0.79	98.58	95.32	7.04	90.33	75.86	PreActResNet-18
	RA102	0.80	99.00	96.78	5.53	90.57	76.15	PreActResNet-18
ImageNet (ℓ_∞)	SS10	0.63	90.95	90.06	36.47	72.56	48.08	ViT-S + ConvStem
	EN10	0.73	94.95	97.34	25.07	62.56	29.22	ResNet-50
	SA11	0.74	95.00	96.50	24.88	64.02	34.96	ResNet-50
	SA10	0.77	95.16	97.99	29.31	52.92	25.32	ResNet-18
	WO10	0.78	96.28	98.34	25.65	55.62	26.24	ResNet-50

Table 2: Correlations between the input margin and the logit margin and detection scores on CIFAR10 (ℓ_2 , $\epsilon = 0.5$) and on ImageNet (ℓ_∞ , $\epsilon = 4/255$).

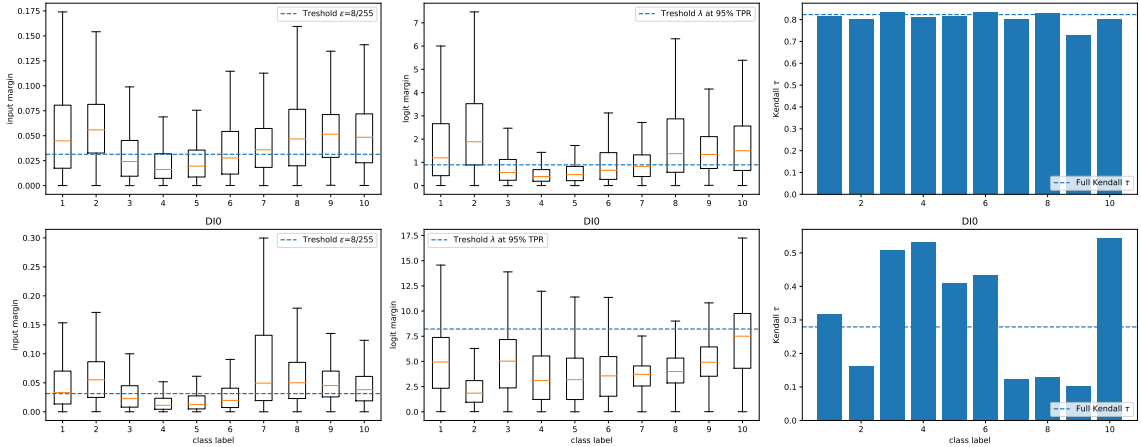


Figure 1: Robustness bias analysis with the input margin (first column) vs the logit margin (second column) and correlation scores per class (third column). On the top row, we have a margin-consistent model, and on the bottom row, we have a weakly margin-consistent model. On the first column (resp. second column), each boxplot represents the distribution of the input margin (resp. logit margin) for the corresponding CIFAR10 class. The dotted blue line indicates the threshold on the first two columns and indicates the correlation score computed over all classes. The threshold λ in the second column is the logit margin threshold at 95% TPR for detection at $\epsilon = 8/255$.