Why Out-of-Distribution Detection Experiments Are Not Reliable - Subtle Experimental Details Muddle the OOD Detector Rankings (Supplementary Material)

Kamil Szyc[1] **Tomasz Walkowiak**[1] **Henryk Maciejewski**[1]

[1]Faculty of Information and Communication Technology, Wrocław University of Science and Technology, Wrocław, Poland

A ADDITIONAL SIMULATION RESULTS FOR ARCHITECTURE FEATURES

Additional simulation results (additional metrics: DTACC, TNR at TPR 95%, and AUPR) for Table 2 of the main paper showing the instability of OOD detection metrics for three different variations of CNN architectures (ResNet-101 and ResNet-110). The results are presented in Tables 1,2, and 3. The message from the results presented here is similar to that from Table 2 of the main paper, that changing the details of the architecture leads to different winning OOD methods and can cause large changes in the OOD detection quality metrics. This is especially visible for MaxLogits (ML) and FreeEnergy (FE). Moreover, the winning methods for AUC, DTACC, and AUPR are almost the same (these metrics select the same winning method for a given architecture and task), but the results for TNR are very different.

Table 1: DTACC results for the OOD detection instability caused by the architecture features. Additional results to those presented in Table 2 of the main paper.

Type	ACC	KNN	Mah	ML	MSP	LOF_C	LOF_E	FE
SVHN								
type-0	94.50	**86.05**	83.40	82.64	83.69	79.69	79.80	82.64
type-1	93.64	**87.14**	82.07	74.44	79.73	85.31	85.11	74.41
type-2	92.91	82.84	**83.30**	66.53	72.27	81.97	81.16	66.43
CIFAR-100								
type-0	94.50	79.38	77.98	81.28	81.33	72.31	70.23	**81.31**
type-1	93.64	**81.39**	80.68	80.16	80.12	79.63	78.98	80.17
type-2	92.91	**79.75**	78.19	77.82	79.14	76.21	74.91	77.80

B ADDITIONAL SIMULATION RESULTS FOR INITIAL SEEDS

Additional simulation results (DTACC, TNR, and AUPR metrics) in the experiments reported in the main paper in Table 3 are shown here in Tables 4,5, and 6. We analyze the instability of OOD detection decisions due to the random seeds used during training. We find that the initial seed has a small effect on the close-set accuracy but causes a large variation in all OOD detection metrics used. Moreover, the ranking of the methods is significantly changed regardless of the OOD detection method used.

Accepted for the 39^th^ Conference on Uncertainty in Artificial Intelligence (UAI 2023).

Table 2: TNR at TPR 95% results for the OOD detection instability caused by the architecture features. Additional results to those presented in Table 2 of the main paper.

Type	ACC	KNN	Mah	ML	MSP	LOF_C	LOF_E	FE
SVHN								
type-0	94.50	52.15	**53.67**	40.08	34.27	42.04	43.21	39.73
type-1	93.64	**57.57**	49.91	18.72	27.59	43.70	50.43	17.85
type-2	92.91	36.00	38.87	18.96	17.17	53.12	**49.01**	19.27
CIFAR-100								
type-0	94.50	35.00	33.01	45.89	34.99	21.20	20.44	**45.98**
type-1	93.64	42.14	**42.28**	41.83	32.70	41.09	41.48	42.10
type-2	92.91	36.85	33.22	38.02	29.39	32.81	30.19	**38.60**

Table 3: AUPR results for the OOD detection instability caused by the architecture features. Additional results to those presented in Table 2 of the main paper.

Type	ACC	KNN	Mah	ML	MSP	LOF_C	LOF_E	FE
SVHN								
type-0	94.50	**92.05**	91.12	87.98	88.29	87.13	87.18	87.95
type-1	93.64	**93.43**	89.97	77.71	83.40	90.85	91.49	77.49
type-2	92.91	88.42	88.96	67.26	73.70	**89.92**	88.85	67.23
CIFAR-100								
type-0	94.50	85.47	84.27	87.50	86.70	77.53	75.51	**87.55**
type-1	93.64	**88.12**	87.45	85.82	85.27	86.13	85.24	85.85
type-2	92.91	**86.23**	84.41	82.25	83.00	82.49	81.21	82.29

C ADDITIONAL SIMULATION RESULTS FOR OOD EXAMPLE SELECTION

Additional simulation results (ACC, DTACC, TNR, and AUPR metrics) for Table 4 of the main paper (we also show here the rank of the methods). We analyze the instability of the OOD detection metrics as a result of the random selection of OOD examples (necessary to keep the 1:1 ratio between ID and OOD data). The results are presented in Tables 7,8,9, and 10. It can be seen that the random selection of OOD examples has almost no influence on the method ranks, regardless of the reported metric.

D ADDITIONAL SIMULATION RESULTS FOR TRAIN-TEST SPLIT

Additional simulation results (DTACC, TNR, and AUPR metrics) for Table 5 of the main paper. We analyze the instability of the OOD detection metrics as a result of the train-test split for close-set task. The results are presented in Tables 11, 12, and 13. Similarly to the initial seeds of the nearest-neighbor training, the train-test split has a small effect on the nearest-neighbor accuracy, but causes a large variation in all OOD detection metrics used. Furthermore, the ranking of the methods changes significantly regardless of the OOD detection method used.

E ADDITIONAL SIMULATION RESULTS FOR AUGMENTATION STRATEGIES

Additional simulation results (DTACC, TNR, and AUPR metrics) for Table 6 of the main paper. We analyze the instability of the OOD detection metrics as a result of the augmentation strategy used. The results are presented in Tables 14, 15, and

Table 4: DTACC results for the OOD detection instability caused by the different random seeds used during training. Additional results to those presented in Table 3 of the main paper.

MobileNet with closed set ACC = 74.75±0.31								
	CIFAR-100 vs SVHN				CIFAR-100 vs CIFAR-10			
	AUPR		Rank		AUPR		Rank	
Method	mean±std	delta	mean±std	range	mean±std	delta	mean±std	range
KNN	64.98±6.95	22.21	5.50±0.92	3-6	60.62±1.54	4.66	4.90±0.54	4-6
Mah	69.67±7.07	19.88	4.50±1.28	2-6	57.45±1.96	7.18	5.80±0.60	4-6
ML	81.59±4.88	18.64	1.50±0.92	1-4	76.53±0.57	2.03	0.00±0.00	0-0
MSP	77.58±4.42	16.10	3.70±1.00	2-6	75.69±0.35	1.05	2.00±0.00	2-2
LOF_C	82.04±3.73	11.27	1.40±1.20	0-3	64.94±1.77	6.02	3.00±0.00	3-3
LOF_E	74.96±4.18	12.79	3.50±1.20	1-6	62.37±1.30	5.06	4.30±0.46	4-5
FE	82.02±5.08	19.46	0.90±1.51	0-5	76.40±0.63	2.30	1.00±0.00	1-1
ResNet with closed set ACC = 92.73±0.27								
	CIFAR-10 vs SVHN				CIFAR-10 vs CIFAR-100			
	AUPR		Rank		AUPR		Rank	
Method	mean±std	delta	mean±std	range	mean±std	delta	mean±std	range
KNN	74.46±13.32	44.52	3.90±1.70	1-6	62.31±7.34	21.98	4.20±0.98	3-6
Mah	77.13±6.45	23.18	4.20±1.08	2-6	62.08±6.30	20.44	3.60±0.49	3-4
ML	82.92±1.53	4.90	1.80±1.40	0-4	87.07±0.49	1.47	1.00±0.00	1-1
MSP	83.83±1.41	4.00	0.90±1.04	0-3	83.78±0.63	1.86	2.00±0.00	2-2
LOF_C	69.58±6.56	23.29	5.70±0.46	5-6	57.04±3.34	8.96	5.70±0.64	4-6
LOF_E	81.83±7.37	29.18	1.70±1.27	0-3	58.86±3.43	11.91	4.50±1.02	3-6
FE	82.77±1.53	4.90	2.80±1.40	1-5	87.11±0.49	1.50	0.00±0.00	0-0

16. It confirms the conclusions presented in the main paper, i.e., a large impact of augmentation techniques on OOD results, and with SVHN as OOD, almost any OOD method can be considered the best by choosing the appropriate augmentation method.

F ADDITIONAL SIMULATION RESULTS FOR TEXT BASED OOD

Additional simulation results (DTACC, TNR, and AUPR metrics) for Table 7 of the main paper. We analyze the instability of OOD detection decisions as an effect of different random seeds used during training for text classification based on BERT (transformer-based) representations. The results are presented in Table 17. They confirm the conclusions presented in the main paper that the rank of the OOD method could be selected in almost any order just by peeking at the seed used during training.

Table 5: TNR at TPR 95% results for the OOD detection instability caused by the different random seeds used during training. Additional results to those presented in Table 3 of the main paper.

MobileNet with closed set ACC = 74.75±0.31								
	CIFAR-100 vs SVHN				CIFAR-100 vs CIFAR-10			
	TNR		Rank		TNR		Rank	
Method	mean±std	delta	mean±std	range	mean±std	delta	mean±std	range
KNN	7.48±4.74	13.48	5.90±0.30	5-6	7.13±0.75	2.82	4.00±0.77	3-5
Mah	14.20±8.10	23.12	4.30±1.42	1-5	5.69±1.32	4.91	5.60±0.92	3-6
ML	27.34±7.64	28.77	1.90±0.83	1-3	18.81±0.72	2.25	0.20±0.40	0-1
MSP	22.39±4.99	18.86	3.10±0.70	2-4	18.07±0.48	1.75	1.50±0.67	0-2
LOF_C	30.52±6.39	22.70	0.90±1.22	0-3	6.83±1.08	3.38	4.30±1.10	3-6
LOF_E	16.95±7.06	24.63	3.70±1.27	1-6	6.87±0.88	2.89	4.10±0.83	3-5
FE	28.64±8.89	33.78	1.20±1.47	0-5	18.35±0.90	3.14	1.30±0.64	0-2

ResNet with closed set ACC = 92.73±0.27								
	CIFAR-10 vs SVHN				CIFAR-10 vs CIFAR-100			
	TNR		Rank		TNR		Rank	
Method	mean±std	delta	mean±std	range	mean±std	delta	mean±std	range
KNN	19.29±19.23	46.65	4.00±2.45	1-6	10.74±3.22	11.39	4.50±0.92	3-6
Mah	26.30±10.85	39.80	3.00±1.34	1-5	13.53±3.53	12.28	3.10±0.30	3-4
ML	23.22±3.01	9.68	2.90±1.14	1-4	40.19±1.47	4.43	0.80±0.40	0-1
MSP	25.50±2.19	7.50	2.30±1.10	0-4	28.58±1.52	4.68	2.00±0.00	2-2
LOF_C	17.87±9.01	28.28	4.40±2.06	0-6	8.38±1.50	3.97	5.80±0.40	5-6
LOF_E	37.73±12.18	50.00	0.40±0.92	0-3	9.85±1.47	4.97	4.60±0.49	4-5
FE	22.71±3.16	10.65	4.00±1.18	2-5	40.36±1.54	4.20	0.20±0.40	0-1

Table 6: AUPR results for the OOD detection instability caused by the different random seeds used during training. Additional results to those presented in Table 3 of the main paper.

Method	CIFAR-100 vs SVHN AUPR mean±std	CIFAR-100 vs SVHN AUPR delta	CIFAR-100 vs SVHN Rank mean±std	CIFAR-100 vs SVHN Rank range	CIFAR-100 vs CIFAR-10 AUPR mean±std	CIFAR-100 vs CIFAR-10 AUPR delta	CIFAR-100 vs CIFAR-10 Rank mean±std	CIFAR-100 vs CIFAR-10 Rank range
MobileNet with closed set ACC = 74.75±0.31								
KNN	64.98±6.95	22.21	5.50±0.92	3-6	60.62±1.54	4.66	4.90±0.54	4-6
Mah	69.67±7.07	19.88	4.50±1.28	2-6	57.45±1.96	7.18	5.80±0.60	4-6
ML	81.59±4.88	18.64	1.50±0.92	1-4	76.53±0.57	2.03	0.00±0.00	0-0
MSP	77.58±4.42	16.10	3.70±1.00	2-6	75.69±0.35	1.05	2.00±0.00	2-2
LOF_C	82.04±3.73	11.27	1.40±1.20	0-3	64.94±1.77	6.02	3.00±0.00	3-3
LOF_E	74.96±4.18	12.79	3.50±1.20	1-6	62.37±1.30	5.06	4.30±0.46	4-5
FE	82.02±5.08	19.46	0.90±1.51	0-5	76.40±0.63	2.30	1.00±0.00	1-1

Method	CIFAR-10 vs SVHN AUPR mean±std	CIFAR-10 vs SVHN AUPR delta	CIFAR-10 vs SVHN Rank mean±std	CIFAR-10 vs SVHN Rank range	CIFAR-10 vs CIFAR-100 AUPR mean±std	CIFAR-10 vs CIFAR-100 AUPR delta	CIFAR-10 vs CIFAR-100 Rank mean±std	CIFAR-10 vs CIFAR-100 Rank range
ResNet with closed set ACC = 92.73±0.27								
KNN	74.46±13.32	44.52	3.90±1.70	1-6	62.31±7.34	21.98	4.20±0.98	3-6
Mah	77.13±6.45	23.18	4.20±1.08	2-6	62.08±6.30	20.44	3.60±0.49	3-4
ML	82.92±1.53	4.90	1.80±1.40	0-4	87.07±0.49	1.47	1.00±0.00	1-1
MSP	83.83±1.41	4.00	0.90±1.04	0-3	83.78±0.63	1.86	2.00±0.00	2-2
LOF_C	69.58±6.56	23.29	5.70±0.46	5-6	57.04±3.34	8.96	5.70±0.64	4-6
LOF_E	81.83±7.37	29.18	1.70±1.27	0-3	58.86±3.43	11.91	4.50±1.02	3-6
FE	82.77±1.53	4.90	2.80±1.40	1-5	87.11±0.49	1.50	0.00±0.00	0-0

Table 7: ACC results for the OOD detection instability caused by random selection of OOD images. Additional results (methods rank added) for experiments reported in Table 4 of the main paper.

MobileNet								
	CIFAR-100 vs SVHN				CIFAR-100 vs CIFAR-10			
Method	AUC		Rank		AUC		Rank	
	mean±std	delta	mean±std	range	mean±std	delta	mean±std	range
KNN	75.56±0.18	0.90	5.35±0.48	5-6	60.67±0.22	1.14	5.00±0.00	5-5
Mah	75.52±0.20	0.87	5.65±0.48	5-6	57.55±0.24	1.18	6.00±0.00	6-6
ML	81.72±0.15	0.80	2.00±0.00	2-2	76.88±0.18	0.85	0.00±0.00	0-0
MSP	78.25±0.18	0.87	4.00±0.00	4-4	76.09±0.16	0.96	2.00±0.00	2-2
LOF_C	84.31±0.15	0.74	0.00±0.00	0-0	68.41±0.20	0.94	3.00±0.00	3-3
LOF_E	78.87±0.18	0.97	3.00±0.00	3-3	64.31±0.22	1.12	4.00±0.00	4-4
FE	81.84±0.14	0.80	1.00±0.00	1-1	76.70±0.18	0.89	1.00±0.00	1-1

ResNet								
	CIFAR-10 vs SVHN				CIFAR-10 vs CIFAR-100			
Method	AUC		Rank		AUC		Rank	
	mean±std	delta	mean±std	range	mean±std	delta	mean±std	range
KNN	87.80±0.17	0.88	1.00±0.00	1-1	59.80±0.27	1.50	4.00±0.00	4-4
Mah	85.26±0.18	1.02	3.00±0.00	3-3	58.67±0.27	1.43	5.00±0.00	5-5
ML	83.88±0.12	0.63	4.00±0.00	4-4	87.73±0.13	0.69	1.00±0.00	1-1
MSP	86.69±0.11	0.53	2.00±0.00	2-2	86.20±0.13	0.66	2.00±0.00	2-2
LOF_C	72.08±0.21	1.39	6.00±0.00	6-6	55.56±0.23	1.28	6.00±0.00	6-6
LOF_E	91.40±0.12	0.63	0.00±0.00	0-0	62.73±0.22	1.09	3.00±0.00	3-3
FE	83.75±0.12	0.62	5.00±0.00	5-5	87.76±0.13	0.69	0.00±0.00	0-0

Table 8: DTACC results for the OOD detection instability caused by random selection of OOD images. Additional results for experiments reported in Table 4 of the main paper.

MobileNet								
	CIFAR-100 vs SVHN				CIFAR-100 vs CIFAR-10			
			Rank		DTACC		Rank	
Method	mean±std	delta	mean±std	range	mean±std	delta	mean±std	range
KNN	69.48±0.17	0.82	5.01±0.10	5-6	58.62±0.18	0.94	5.00±0.00	5-5
Mah	69.12±0.17	0.81	5.99±0.10	5-6	56.29±0.20	1.03	6.00±0.00	6-6
ML	74.77±0.17	0.87	2.00±0.00	2-2	70.98±0.17	0.94	0.11±0.31	0-1
MSP	70.86±0.18	0.77	4.00±0.00	4-4	70.10±0.16	0.76	2.00±0.00	2-2
LOF_C	76.69±0.17	0.86	0.00±0.00	0-0	64.54±0.16	0.89	3.00±0.00	3-3
LOF_E	72.44±0.18	0.75	3.00±0.00	3-3	61.30±0.19	0.94	4.00±0.00	4-4
FE	75.08±0.17	0.78	1.00±0.00	1-1	70.94±0.18	1.03	0.89±0.31	0-1

ResNet								
	CIFAR-10 vs SVHN				CIFAR-10 vs CIFAR-100			
			Rank		DTACC		Rank	
Method	mean±std	delta	mean±std	range	mean±std	delta	mean±std	range
KNN	82.23±0.17	0.88	1.00±0.00	1-1	58.78±0.21	1.02	4.04±0.31	3-5
Mah	77.90±0.18	1.14	3.97±0.97	3-5	58.52±0.21	1.10	4.93±0.26	4-5
ML	77.89±0.16	0.72	3.49±0.50	3-4	80.44±0.16	0.81	0.84±0.37	0-1
MSP	80.78±0.18	0.87	2.00±0.00	2-2	80.05±0.15	0.87	2.00±0.00	2-2
LOF_C	66.53±0.21	1.31	6.00±0.00	6-6	54.06±0.20	1.12	6.00±0.00	6-6
LOF_E	83.99±0.16	0.72	0.00±0.00	0-0	59.23±0.21	0.93	3.03±0.17	3-4
FE	77.87±0.16	0.73	4.54±0.50	4-5	80.46±0.16	0.83	0.16±0.37	0-1

Table 9: TNR at TPR 95% results for the OOD detection instability caused by random selection of OOD images. Additional results for experiments reported in Table 4 of the main paper.

MobileNet								
	CIFAR-100 vs SVHN				CIFAR-100 vs CIFAR-10			
Method	TNR		Rank		TNR		Rank	
	mean±std	delta	mean±std	range	mean±std	delta	mean±std	range
KNN	15.61±0.30	1.71	6.00±0.00	6-6	6.76±0.22	1.03	5.00±0.00	5-5
Mah	18.76±0.37	1.86	5.00±0.00	5-5	5.53±0.20	1.02	6.00±0.00	6-6
ML	23.86±0.37	1.92	1.00±0.00	1-1	17.44±0.36	1.93	0.34±0.47	0-1
MSP	22.64±0.32	1.54	2.98±0.14	2-3	17.36±0.40	1.84	0.71±0.53	0-2
LOF_C	35.44±0.46	2.66	0.00±0.00	0-0	9.20±0.27	1.23	3.00±0.00	3-3
LOF_E	21.12±0.36	1.76	4.00±0.00	4-4	7.98±0.24	1.17	4.00±0.00	4-4
FE	23.29±0.39	1.96	2.02±0.14	2-3	16.87±0.34	2.07	1.95±0.26	0-2
ResNet								
	CIFAR-10 vs SVHN				CIFAR-10 vs CIFAR-100			
Method	TNR		Rank		TNR		Rank	
	mean±std	delta	mean±std	range	mean±std	delta	mean±std	range
KNN	46.91±0.45	2.46	1.00±0.00	1-1	10.59±0.26	1.49	4.84±0.37	4-5
Mah	39.91±0.47	2.28	2.00±0.00	2-2	13.63±0.33	1.66	3.00±0.00	3-3
ML	22.44±0.38	1.82	4.00±0.00	4-4	41.05±0.40	2.18	0.95±0.22	0-1
MSP	27.27±0.42	2.43	3.00±0.00	3-3	29.45±0.41	2.07	2.00±0.00	2-2
LOF_C	14.08±0.31	1.52	6.00±0.00	6-6	7.69±0.21	1.13	6.00±0.00	6-6
LOF_E	59.76±0.46	2.28	0.00±0.00	0-0	10.95±0.25	1.57	4.16±0.37	4-5
FE	21.87±0.36	1.70	5.00±0.00	5-5	41.27±0.41	2.31	0.05±0.22	0-1

Table 10: AUPR results for the OOD detection instability caused by random selection of OOD images. Additional results for experiments reported in Table 4 of the main paper.

MobileNet								
	CIFAR-100 vs SVHN				CIFAR-100 vs CIFAR-10			
	AUPR		Rank		AUPR		Rank	
Method	mean±std	delta	mean±std	range	mean±std	delta	mean±std	range
KNN	74.61±0.18	0.88	5.00±0.00	5-5	59.54±0.20	1.03	5.00±0.00	5-5
Mah	74.29±0.20	0.95	6.00±0.00	6-6	56.80±0.21	1.00	6.00±0.00	6-6
ML	81.01±0.15	0.78	1.32±0.47	1-2	75.60±0.19	0.94	0.00±0.00	0-0
MSP	77.90±0.17	0.92	3.61±0.49	3-4	75.16±0.17	0.90	1.99±0.10	1-2
LOF_C	83.69±0.16	0.84	0.00±0.00	0-0	67.14±0.20	1.00	3.00±0.00	3-3
LOF_E	77.99±0.20	1.03	3.39±0.49	3-4	63.04±0.22	1.25	4.00±0.00	4-4
FE	81.00±0.15	0.80	1.68±0.47	1-2	75.34±0.19	0.93	1.01±0.10	1-2

ResNet								
	CIFAR-10 vs SVHN				CIFAR-10 vs CIFAR-100			
	AUPR		Rank		AUPR		Rank	
Method	mean±std	delta	mean±std	range	mean±std	delta	mean±std	range
KNN	83.84±0.24	1.18	3.00±0.00	3-3	56.83±0.22	1.25	4.99±0.10	4-5
Mah	84.61±0.21	1.10	1.99±0.10	1-2	57.25±0.23	1.12	4.01±0.10	4-5
ML	82.38±0.13	0.72	4.00±0.00	4-4	86.98±0.15	0.71	1.00±0.00	1-1
MSP	85.16±0.15	0.77	1.01±0.10	1-2	84.45±0.18	0.89	2.00±0.00	2-2
LOF_C	70.68±0.21	1.34	6.00±0.00	6-6	54.82±0.21	1.10	6.00±0.00	6-6
LOF_E	90.94±0.13	0.71	0.00±0.00	0-0	60.86±0.22	1.06	3.00±0.00	3-3
FE	82.18±0.13	0.73	5.00±0.00	5-5	87.02±0.15	0.70	0.00±0.00	0-0

Table 11: DTACC results for the OOD detection instability caused by close-set train-test splits. Additional results for experiments reported in Table 5 of the main paper.

MobileNet with closed set ACC = 74.98±0.50								
	CIFAR-100 vs SVHN				CIFAR-100 vs CIFAR-10			
Method			Rank		DTACC		Rank	
	mean±std	delta	mean±std	range	mean±std	delta	mean±std	range
KNN	67.32±2.49	7.25	5.40±1.02	3-6	59.99±1.18	3.99	4.40±0.49	4-5
Mah	70.69±4.78	16.17	4.20±1.47	2-6	55.32±1.61	5.77	6.00±0.00	6-6
ML	78.77±7.22	23.14	1.80±1.47	0-4	76.02±6.91	16.36	0.50±0.50	0-1
MSP	76.31±8.34	23.58	3.10±2.07	0-6	75.65±8.32	19.07	1.40±0.92	0-2
LOF_C	76.61±2.72	8.41	2.10±1.30	0-4	62.56±2.77	8.18	3.00±0.00	3-3
LOF_E	73.28±4.43	14.76	2.80±1.60	0-5	60.18±2.42	6.80	4.60±0.49	4-5
FE	78.96±7.14	23.09	1.60±1.62	0-5	75.92±6.79	16.15	1.10±0.70	0-2

ResNet with closed set ACC = 94.41±0.24								
	CIFAR-10 vs SVHN				CIFAR-10 vs CIFAR-100			
Method			Rank		DTACC		Rank	
	mean±std	delta	mean±std	range	mean±std	delta	mean±std	range
KNN	80.03±2.88	10.99	2.10±1.51	0-5	75.50±3.59	11.70	3.80±0.40	3-4
Mah	79.33±1.95	5.31	3.20±1.66	0-5	76.32±2.52	7.80	3.20±0.40	3-4
ML	79.69±2.77	9.31	2.20±0.75	1-3	81.61±0.36	1.00	1.10±0.54	0-2
MSP	80.80±2.13	8.06	1.00±1.10	0-3	81.66±0.28	0.96	1.00±1.00	0-2
LOF_C	75.32±4.03	14.74	4.60±1.28	1-6	70.79±4.34	13.53	5.00±0.00	5-5
LOF_E	69.69±3.24	12.47	5.90±0.30	5-6	62.76±4.39	14.16	6.00±0.00	6-6
FE	79.68±2.78	9.33	2.00±1.41	0-4	81.61±0.36	1.04	0.90±0.83	0-2

Table 12: TNR at TPR 95% results for the OOD detection instability caused by close-set train-test splits. Additional results for experiments reported in Table 5 of the main paper.

MobileNet with closed set ACC = 74.98±0.50								
	CIFAR-100 vs SVHN				CIFAR-100 vs CIFAR-10			
Method	TNR		Rank		TNR		Rank	
	mean±std	delta	mean±std	range	mean±std	delta	mean±std	range
KNN	11.23±5.80	20.39	5.60±0.92	3-6	6.88±0.68	2.70	3.70±0.90	3-5
Mah	22.90±12.50	41.89	3.70±1.68	0-5	4.72±0.80	2.67	5.90±0.30	5-6
ML	40.71±22.76	61.53	1.80±1.54	0-4	34.28±23.12	53.71	0.00±0.00	0-0
MSP	35.93±21.27	54.64	3.50±1.50	2-6	32.91±22.71	53.42	1.80±0.40	1-2
LOF_C	35.74±9.92	29.41	2.00±1.41	0-4	7.04±1.79	5.24	4.10±0.94	3-6
LOF_E	25.73±11.94	36.14	2.90±1.45	1-5	6.39±1.14	3.96	4.30±0.64	3-5
FE	41.30±22.76	61.75	1.50±1.86	0-6	33.68±22.95	53.34	1.20±0.40	1-2
ResNet with closed set ACC = 94.41±0.24								
	CIFAR-10 vs SVHN				CIFAR-10 vs CIFAR-100			
Method	TNR		Rank		TNR		Rank	
	mean±std	delta	mean±std	range	mean±std	delta	mean±std	range
KNN	31.91±7.03	24.53	2.60±1.80	0-5	28.32±6.16	20.02	3.60±0.66	2-4
Mah	36.49±5.37	18.39	1.80±1.47	0-4	29.78±4.56	14.73	3.20±0.40	3-4
ML	34.03±5.03	15.67	1.60±0.80	0-3	45.84±0.98	3.17	0.90±0.30	0-1
MSP	30.91±2.51	8.81	3.80±1.17	2-6	35.62±1.17	3.83	2.20±0.60	2-4
LOF_C	28.52±10.85	30.70	3.60±2.20	0-6	22.45±6.38	18.43	5.00±0.00	5-5
LOF_E	19.78±7.09	20.81	5.80±0.40	5-6	14.77±3.85	11.76	6.00±0.00	6-6
FE	34.01±5.30	16.92	1.80±1.33	0-4	46.42±0.82	2.51	0.10±0.30	0-1

Table 13: AUPR results for the OOD detection instability caused by close-set train-test splits. Additional results for experiments reported in Table 5 of the main paper.

MobileNet with closed set ACC = 74.98±0.50								
	CIFAR-100 vs SVHN				CIFAR-100 vs CIFAR-10			
Method	AUPR		Rank		AUPR		Rank	
	mean±std	delta	mean±std	range	mean±std	delta	mean±std	range
KNN	70.41±4.32	13.14	5.30±1.19	3-6	60.64±1.21	4.22	4.40±0.49	4-5
Mah	76.12±6.71	23.14	4.00±1.84	0-6	55.18±2.15	7.25	6.00±0.00	6-6
ML	84.32±7.30	23.05	1.90±1.45	0-4	81.07±6.98	16.55	0.40±0.49	0-1
MSP	82.00±7.70	22.49	3.40±1.62	2-6	80.61±7.79	18.30	1.40±0.92	0-2
LOF_C	84.01±3.74	10.84	2.00±1.41	0-4	64.33±3.37	9.98	3.00±0.00	3-3
LOF_E	79.07±6.06	19.76	2.90±1.30	1-4	61.37±2.88	7.97	4.60±0.49	4-5
FE	84.52±7.35	23.10	1.50±1.96	0-5	80.94±6.96	16.40	1.20±0.60	0-2
ResNet with closed set ACC = 94.41±0.24								
	CIFAR-10 vs SVHN				CIFAR-10 vs CIFAR-100			
Method	AUPR		Rank		AUPR		Rank	
	mean±std	delta	mean±std	range	mean±std	delta	mean±std	range
KNN	85.64±3.30	12.28	1.60±1.85	0-5	80.76±4.47	13.86	3.80±0.40	3-4
Mah	86.26±2.14	6.20	1.70±1.49	0-4	81.74±3.54	11.27	3.20±0.40	3-4
ML	85.15±2.80	8.44	2.40±0.80	1-4	87.73±0.61	1.78	1.00±0.00	1-1
MSP	85.74±2.44	9.23	2.20±1.66	0-5	86.80±0.47	1.84	2.00±0.00	2-2
LOF_C	81.40±4.77	16.08	4.20±1.89	0-6	76.07±5.16	15.95	5.00±0.00	5-5
LOF_E	74.88±4.04	14.09	5.90±0.30	5-6	66.05±5.96	18.92	6.00±0.00	6-6
FE	85.10±2.81	8.50	3.00±0.89	1-4	87.78±0.60	1.77	0.00±0.00	0-0

Table 14: DTACC results for the OOD detection instability caused by different augmentation methods used for close model training. Additional results for experiments reported in Table 6 of the main paper.

Augmentation	ACC	KNN	Mah	ML	MSP	LOF_C	LOF_E	FE
MobileNet CIFAR-100 vs SVHN								
None	53.73	66.25	72.72	69.83	65.33	**74.52**	72.76	70.27
Affine	74.41	71.52	75.14	70.45	70.27	75.70	**78.31**	70.39
CoarseDropout	67.18	60.19	63.25	77.18	71.20	74.05	67.96	**77.75**
ColorJitter	65.98	60.59	62.54	75.07	68.76	**77.47**	73.00	75.77
CropAndPad	72.59	63.02	70.89	77.12	74.16	**78.31**	74.45	77.29
MixUp	68.65	63.30	**78.87**	71.24	70.38	78.70	78.15	69.80
MobileNet CIFAR-100 vs CIFAR-10								
None	53.73	56.07	53.81	**62.74**	62.19	56.76	51.29	62.56
Affine	74.41	60.15	56.33	**72.52**	70.91	64.25	60.15	72.51
CoarseDropout	67.18	59.42	56.02	**67.16**	66.87	59.96	55.65	67.15
ColorJitter	65.98	58.22	56.76	**67.40**	66.50	62.33	57.38	67.31
CropAndPad	72.59	60.57	56.23	**70.25**	69.37	62.40	58.84	70.06
MixUp	68.65	59.47	56.20	**68.70**	68.16	55.34	57.07	67.23
ResNet CIFAR-10 vs SVHN								
None	83.64	71.41	67.70	**74.43**	73.87	74.15	69.59	74.41
Affine	94.76	**86.45**	83.97	85.24	84.40	84.83	81.62	85.23
CoarseDropout	89.27	62.52	59.39	81.27	80.31	64.89	72.69	**81.28**
ColorJitter	87.98	80.35	70.02	77.99	78.00	**82.09**	80.31	77.93
CropAndPad	94.01	84.79	80.48	87.02	85.14	83.11	79.55	**87.06**
MixUp	89.41	77.28	81.31	57.10	**83.12**	72.72	78.72	50.43
ResNet CIFAR-10 vs CIFAR-100								
None	83.64	63.20	62.25	74.41	72.67	57.95	54.94	**74.44**
Affine	94.76	81.31	80.25	**83.07**	82.62	79.84	77.84	83.05
CoarseDropout	89.27	58.27	55.08	78.65	77.41	54.14	54.10	**78.69**
ColorJitter	87.98	74.48	71.70	77.50	76.17	74.64	71.78	**77.50**
CropAndPad	94.01	79.84	77.35	81.99	81.72	76.09	72.06	**82.01**
MixUp	89.41	62.93	66.61	72.72	**76.77**	60.33	57.22	67.73

Table 15: TNR at TPR 95% results for the OOD detection instability caused by different augmentation methods used for close model training. Additional results for experiments reported in Table 6 of the main paper.

Augmentation	ACC	KNN	Mah	ML	MSP	LOF_C	LOF_E	FE
MobileNet CIFAR-100 vs SVHN								
None	53.73	12.47	**31.45**	13.11	11.28	24.10	28.72	12.87
Affine	74.41	28.78	33.28	11.84	13.83	30.81	**41.14**	9.93
CoarseDropout	67.18	5.10	10.09	26.34	20.82	25.26	12.80	**28.50**
ColorJitter	65.98	4.16	10.52	22.22	18.21	**39.95**	25.39	23.19
CropAndPad	72.59	8.77	21.19	23.35	20.32	**35.89**	29.84	23.09
MixUp	68.65	8.72	**40.35**	10.10	13.73	35.25	38.45	7.78
MobileNet CIFAR-100 vs CIFAR-10								
None	53.73	4.22	4.77	**11.15**	10.32	4.26	3.29	10.73
Affine	74.41	6.06	4.59	**19.65**	19.35	7.59	5.09	18.96
CoarseDropout	67.18	5.19	5.12	**14.84**	14.45	6.48	5.15	14.36
ColorJitter	65.98	5.66	6.17	**13.73**	13.14	9.91	5.82	13.50
CropAndPad	72.59	6.58	4.58	**17.92**	16.32	6.98	5.90	17.43
MixUp	68.65	6.68	5.26	14.62	**15.83**	4.94	5.65	13.19
ResNet CIFAR-10 vs SVHN								
None	83.64	25.14	26.36	16.21	15.69	**39.94**	26.22	15.74
Affine	94.76	51.16	47.07	54.64	38.62	51.35	34.75	**55.50**
CoarseDropout	89.27	0.10	5.92	**32.08**	23.34	10.31	26.32	31.95
ColorJitter	87.98	41.32	15.55	23.76	22.63	**54.33**	46.30	23.03
CropAndPad	94.01	47.77	43.02	61.05	41.01	53.79	35.88	**62.23**
MixUp	89.41	32.63	**55.88**	7.73	35.98	19.50	43.04	3.70
ResNet CIFAR-10 vs CIFAR-100								
None	83.64	10.77	11.80	22.91	17.63	11.39	7.54	**23.23**
Affine	94.76	39.08	39.33	47.28	35.74	39.87	36.48	**48.26**
CoarseDropout	89.27	6.09	9.00	33.23	23.39	7.53	6.90	**33.50**
ColorJitter	87.98	22.85	19.93	29.16	21.77	27.42	22.02	**30.12**
CropAndPad	94.01	38.73	35.15	46.50	34.21	33.88	26.85	**46.87**
MixUp	89.41	8.67	13.36	**31.88**	29.22	9.73	7.68	31.14

Table 16: AUPR results for the OOD detection instability caused by different augmentation methods used for close model training. Additional results for experiments reported in Table 6 of the main paper.

Augmentation	ACC	KNN	Mah	ML	MSP	LOF_C	LOF_E	FE
MobileNet CIFAR-100 vs SVHN								
None	53.73	70.10	**79.54**	73.90	69.96	79.46	78.17	74.31
Affine	74.41	78.77	82.64	74.01	74.72	83.09	**86.26**	73.33
CoarseDropout	67.18	61.15	66.26	83.01	77.49	80.46	71.51	**83.81**
ColorJitter	65.98	60.91	65.98	80.69	75.03	**84.37**	78.94	81.41
CropAndPad	72.59	65.89	76.56	82.31	79.87	**85.33**	81.15	82.36
MixUp	68.65	66.55	**86.10**	74.14	74.85	85.40	85.45	71.32
MobileNet CIFAR-100 vs CIFAR-10								
None	53.73	56.72	54.17	65.31	**65.77**	56.58	49.59	65.07
Affine	74.41	60.38	56.31	**77.79**	76.70	66.11	60.79	77.63
CoarseDropout	67.18	60.59	56.35	71.03	**71.29**	61.06	55.98	70.83
ColorJitter	65.98	58.05	57.53	**71.15**	70.87	65.25	58.27	70.99
CropAndPad	72.59	61.61	56.33	**75.11**	74.58	64.20	59.46	74.98
MixUp	68.65	61.07	56.90	72.59	**72.81**	55.52	57.94	70.77
ResNet CIFAR-10 vs SVHN								
None	83.64	78.25	73.42	78.52	78.04	**81.72**	74.36	78.39
Affine	94.76	**92.07**	90.60	91.47	89.71	91.23	86.77	91.59
CoarseDropout	89.27	58.34	59.88	**86.43**	84.49	66.94	76.27	86.40
ColorJitter	87.98	88.18	75.01	82.39	82.52	**90.64**	88.00	82.15
CropAndPad	94.01	91.22	88.14	92.88	90.21	90.77	86.24	**93.04**
MixUp	89.41	81.64	**87.96**	51.87	87.24	76.64	85.32	38.94
ResNet CIFAR-10 vs CIFAR-100								
None	83.64	66.50	65.28	79.92	77.49	61.49	56.09	**80.03**
Affine	94.76	87.56	87.04	89.06	88.06	85.98	83.93	**89.12**
CoarseDropout	89.27	57.26	53.67	84.87	82.39	55.65	53.88	**84.92**
ColorJitter	87.98	80.17	77.21	83.50	81.39	80.44	76.04	**83.61**
CropAndPad	94.01	86.64	84.53	87.90	86.90	83.12	78.72	**87.98**
MixUp	89.41	65.40	69.89	73.66	**79.85**	62.45	59.04	68.80

Table 17: DTACC, TNR, and AUPR results for the OOD detection instability caused by the different random seeds used during training for text classification by BERT model. Additional results to those presented in Table 7 of the main paper.

BERT with closed set ACC = 97.49±0.11				
Method	DTACC		Rank	
	mean±std	delta	mean±std	range
KNN	72.10±4.16	11.34	2.25±1.85	0-5
Mah	70.75±2.21	7.46	2.75±2.17	0-6
ML	69.08±4.90	16.39	3.38±2.12	0-6
MSP	67.60±7.66	21.69	3.62±2.34	0-6
LOF_C	70.98±3.14	10.93	2.12±1.45	0-5
LOF_E	70.74±3.08	10.75	3.00±1.22	1-5
FE	69.06±4.90	16.39	3.88±1.90	1-6
	TNR at TPR 95%		Rank	
	mean±std	delta	mean±std	range
KNN	28.26±4.28	13.16	1.50±1.22	0-3
Mah	28.07±2.87	8.84	2.00±0.87	1-3
ML	17.97±3.23	9.68	4.62±0.48	4-5
MSP	16.85±2.05	5.26	5.25±0.97	4-6
LOF_C	29.68±3.27	9.26	1.00±0.87	0-2
LOF_E	28.61±3.60	9.95	1.50±1.22	0-3
FE	17.88±3.43	9.68	5.12±0.78	4-6
	AUPR		Rank	
	mean±std	delta	mean±std	range
KNN	73.85±4.15	11.79	1.00±1.32	0-3
Mah	73.21±2.13	6.48	2.00±1.50	0-5
ML	67.03±5.07	16.64	4.88±0.60	4-6
MSP	65.62±7.33	22.60	5.25±1.39	2-6
LOF_C	73.36±2.86	9.04	1.25±0.97	0-3
LOF_E	72.82±2.79	8.72	2.25±0.97	1-4
FE	67.07±5.08	16.80	4.38±0.86	3-6