1 AnoShift - Out-Of-Distribution Anomaly Detection Benchmark - Table 1

- Train data: files $[year]\_subset.parquet$ with $year \in \{2006, 2007, 2008, 2009, 2010\}$
- IID data: files $[year]\_subset\_valid.parquet$ with $year \in \{2006, 2007, 2008, 2009, 2010\}$
- NEAR data: files $[year]\_subset.parquet$ with $year \in \{2011, 2012, 2013\}$
- FAR data: files $[year]\_subset.parquet$ with $year \in \{2014, 2015\}$
- Scripts for reproducing the results are available in '/baselines_OOD_setup'

Table 1: Performance evolution over time for unsupervised methods: `IID` vs `NEAR` vs `FAR`. We report beside the ROC-AUC metric, also the PR-AUC for inliers and PR-AUC for outliers. With bold are the best results per split.

Type	Unsupervised Baselines	`IID`	`NEAR`	`FAR`
		ROC-AUC (%) ↑		
Classical	**OC-SVM** [10] (train 5%)	76.86 ± 0.06	71.43 ± 0.29	49.57 ± 0.09
Classical	**IsoForest** [7]	86.09 ± 0.54	75.26 ± 4.66	27.16 ± 1.69
Classical	**ECOD** [6]	84.76	44.87	49.19
Classical	**COPOD** [5]	85.62	54.24	**50.42**
Classical	**LOF** [2]	91.50 ± 0.88	79.29 ± 3.33	34.96 ± 0.14
Deep	**SO-GAAL** [8]	50.48 ± 1.13	54.55 ± 3.92	49.35 ± 0.51
Deep	**deepSVDD** [9]	**92.67** ± 0.44	**87.00** ± 1.80	34.53 ± 1.62
Deep	**AE** [1] **for anomalies**	81.00 ± 0.22	44.06 ± 0.57	19.96 ± 0.21
Deep	**LUNAR** [4] (train 5%)	85.75 ± 1.95	49.03 ± 2.57	28.19 ± 0.90
Deep	**InternalContrastiveLearning** [11]	84.86 ± 2.14	52.26 ± 1.18	22.45 ± 0.52
Deep	**BERT [3] for anomalies**	84.54 ± 0.07	86.05 ± 0.25	28.15 ± 0.06
		PR-AUC inliers (%) ↑		
Classical	**OC-SVM** [10] (train 5%)	70.84 ± 0.13	41.38 ± 0.29	**15.12** ± 0.04
Classical	**IsoForest** [7]	83.68 ± 3.47	57.06 ± 10.27	9.16 ± 0.18
Classical	**ECOD** [6]	84.47	22.98	13.78
Classical	**COPOD** [5]	**87.86**	29.25	14.55
Classical	**LOF** [2]	84.11 ± 0.96	52.48 ± 4.56	10.15 ± 0.10
Deep	**SO-GAAL** [8]	58.65 ± 5.36	43.52 ± 11.62	10.68 ± 2.42
Deep	**deepSVDD** [9]	82.62 ± 0.52	**71.71** ± 4.85	10.02 ± 0.22
Deep	**AE** [1] **for anomalies**	73.76 ± 0.09	26.16 ± 0.15	8.51 ± 0.01
Deep	**LUNAR** [4] (train 5%)	78.91 ± 1.69	29.36 ± 2.58	9.33 ± 0.11
Deep	**InternalContrastiveLearning** [11]	76.96 ± 2.12	27.28 ± 0.59	8.81 ± 0.05
Deep	**BERT [3] for anomalies**	74.61 ± 0.13	58.94 ± 0.69	8.22 ± 0.02
		PR-AUC outliers (%) ↑		
Classical	**OC-SVM** [10] (train 5%)	67.94 ± 0.21	85.70 ± 0.16	87.27 ± 0.02
Classical	**IsoForest** [7]	81.46 ± 2.52	87.13 ± 2.08	78.33 ± 1.41
Classical	**ECOD** [6]	78.37	74.48	85.90
Classical	**COPOD** [5]	78.19	77.99	85.98
Classical	**LOF** [2]	83.86 ± 0.98	92.34 ± 1.26	81.99 ± 0.05
Deep	**SO-GAAL** [8]	70.38 ± 0.28	87.71 ± 0.74	**92.67** ± 0.13
Deep	**deepSVDD** [9]	**92.65** ± 0.64	94.15 ± 1.05	82.25 ± 0.48
Deep	**AE** [1] **for anomalies**	78.99 ± 0.28	72.97 ± 0.38	75.71 ± 0.05
Deep	**LUNAR** [4] (train 5%)	88.01 ± 1.03	80.91 ± 0.62	79.45 ± 0.30
Deep	**InternalContrastiveLearning** [11]	89.08 ± 0.87	81.93 ± 0.39	77.55 ± 0.50
Deep	**BERT [3] for anomalies**	89.83 ± 0.07	**95.96** ± 0.06	78.38 ± 0.02

2 AnoShift - In-Distribution Anomaly Detection Benchmark - Table 2

- Train data: files $[year]\_subset.parquet$
  with $year \in \{2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015\}$
- Test data: files $[year]\_subset\_valid.parquet$
  with $year \in \{2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015\}$
- Scripts for reproducing the results are available in '/baselines_ID_setup' (set full_set=1)

Table 2: Performance in the ID setup. We report beside the ROC-AUC metric, also the PR-AUC for inliers and PR-AUC for outliers. With bold are the best results per split.

Type	Unsupervised Baselines	ROC-AUC (%) ↑
Classical	**OC-SVM** [10] (train 5%)	68.73 ± 6.09
	IsoForest [7] (train 5%)	81.27 ± 0.90
	ECOD [6]	79.41
	COPOD [5]	80.89
	LOF [2]	87.61 ± 1.50
Deep	**SO-GAAL** [8]	49.90 ± 0.58
	deepSVDD [9]	**88.24** ± 0.58
	AE [1] **for anomalies**	64.08 ± 0.23
	LUNAR [4] (train 5%)	78.53 ± 0.21
	InternalContrastiveLearning [11]	66.99 ± 3.59
	BERT [3] for anomalies	79.62 ± 0.77
		PR-AUC inliers (%) ↑
Classical	**OC-SVM** [10] (train 5%)	50.17 ± 4.67
	IsoForest [7] (train 5%)	68.46 ± 0.77
	ECOD [6]	69.31
	COPOD [5]	73.17
	LOF [2]	70.79 ± 1.86
Deep	**SO-GAAL** [8]	39.41 ± 9.75
	deepSVDD [9]	**76.59** ± 2.13
	AE [1] **for anomalies**	47.56 ± 0.17
	LUNAR [4] (train 5%)	63.31 ± 0.51
	InternalContrastiveLearning [11]	50.82 ± 3.54
	BERT [3] for anomalies	63.26 ± 0.59
		PR-AUC outliers (%) ↑
Classical	**OC-SVM** [10] (train 5%)	74.67 ± 3.62
	IsoForest [7] (train 5%)	85.47 ± 0.70
	ECOD [6]	80.73
	COPOD [5]	81.47
	LOF [2]	89.97 ± 1.09
Deep	**SO-GAAL** [8]	79.89 ± 0.18
	deepSVDD [9]	**90.56** ± 0.77
	AE [1] **for anomalies**	78.07 ± 0.05
	LUNAR [4] (train 5%)	88.42 ± 0.14
	InternalContrastiveLearning [11]	80.1 ± 9.63
	BERT [3] for anomalies	84.77 ± 1.70

3 AnoShift - In-Distribution Anomaly Detection Benchmark (years 2006-2010) - Table 3

- Train data: files $[year]\_subset.parquet$ with $year \in \{2006, 2007, 2008, 2009, 2010\}$
- Test data: files $[year]\_subset\_valid.parquet$ with $year \in \{2006, 2007, 2008, 2009, 2010\}$
- Scripts for reproducing the results are available in '/baselines_ID_setup' (set full_set=0)

Table 3: Performance in the ID setup (years corresponding to our original IID split). We report beside the ROC-AUC metric, also the PR-AUC for inliers and PR-AUC for outliers. With bold are the best results per split.

Type	Unsupervised Baselines	ROC-AUC (%) ↑
Classical	**OC-SVM** [10] (train 5%)	76.86 ± 0.06
	IsoForest [7]	86.09 ± 0.54
	ECOD [6]	84.76
	COPOD [5]	85.62
	LOF [2]	91.50 ± 0.88
Deep	**SO-GAAL** [8]	50.48 ± 1.13
	deepSVDD [9]	**92.67** ± 0.44
	AE [1] **for anomalies**	81.00 ± 0.22
	LUNAR [4] (train 5%)	85.75 ± 1.95
	InternalContrastiveLearning [11]	84.86 ± 2.14
	BERT [3] for anomalies	84.54 ± 0.07
		PR-AUC inliers (%) ↑
Classical	**OC-SVM** [10] (train 5%)	70.84 ± 0.13
	IsoForest [7]	83.68 ± 3.47
	ECOD [6]	84.47
	COPOD [5]	**87.86**
	LOF [2]	84.11 ± 0.96
Deep	**SO-GAAL** [8]	58.65 ± 5.36
	deepSVDD [9]	82.62 ± 0.52
	AE [1] **for anomalies**	73.76 ± 0.09
	LUNAR [4] (train 5%)	78.91 ± 1.69
	InternalContrastiveLearning [11]	76.96 ± 2.12
	BERT [3] for anomalies	74.61 ± 0.13
		PR-AUC outliers (%) ↑
Classical	**OC-SVM** [10] (train 5%)	67.94 ± 0.21
	IsoForest [7]	81.46 ± 2.52
	ECOD [6]	78.37
	COPOD [5]	78.19
	LOF [2]	83.86 ± 0.98
Deep	**SO-GAAL** [8]	70.38 ± 0.28
	deepSVDD [9]	**92.65** ± 0.64
	AE [1] **for anomalies**	78.99 ± 0.28
	LUNAR [4] (train 5%)	88.01 ± 1.03
	InternalContrastiveLearning [11]	89.08 ± 0.87
	BERT [3] for anomalies	89.83 ± 0.07

References

[1] Charu C Aggarwal. An introduction to outlier analysis. In *Outlier analysis*, pages 1–34. Springer, 2017.

[2] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. LOF: identifying density-based local outliers. In *SIGMOD International Conference on Management of Data*, 2000.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *NAACL*, 2019.

[4] Adam Goodge, Bryan Hooi, See-Kiong Ng, and Wee Siong Ng. Lunar: Unifying local outlier detection methods via graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6737–6745, 2022.

[5] Zheng Li, Yue Zhao, Nicola Botta, Cezar Ionescu, and Xiyang Hu. Copod: copula-based outlier detection. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 1118–1123. IEEE, 2020.

[6] Zheng Li, Yue Zhao, Xiyang Hu, Nicola Botta, Cezar Ionescu, and George Chen. Ecod: Unsupervised outlier detection using empirical cumulative distribution functions. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[7] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation-based anomaly detection. *ACM Trans. Knowl. Discov. Data*, 2012.

[8] Yezheng Liu, Zhe Li, Chong Zhou, Yuanchun Jiang, Jianshan Sun, Meng Wang, and Xiangnan He. Generative adversarial active learning for unsupervised outlier detection. *IEEE Transactions on Knowledge and Data Engineering*, 32(8):1517–1528, 2019.

[9] Lukas Ruff, Robert A. Vandermeulen, Nico Görnitz, Lucas Deecke, Shoaib A. Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International Conference on Machine Learning ICML*, 2018.

[10] Bernhard Schölkopf, Robert C. Williamson, Alexander J. Smola, John Shawe-Taylor, and John C. Platt. Support vector method for novelty detection. In *Advances in Neural Information Processing Systems, NIPS*, 1999.

[11] Tom Shenkar and Lior Wolf. Anomaly detection for tabular data with internal contrastive learning. In *International Conference on Learning Representations*, 2021.