Why Ranking Anomaly Detection Algorithms Isn't as Reliable as You May Think

No Author Given

No Institute Given

Abstract. Anomaly detection is a safety-critical machine learning problem with applications ranging from fraud detection to network intrusion prevention and industrial monitoring. Despite the large number of proposed anomaly detection algorithms, many novel methods claim state-of-the-art performance. However, many authors do so under benchmark settings that are not aligned with one another. This lack of comparability raises concerns regarding the reproducibility and reliability of anomaly detection benchmarks.

In this work, we systematically study the impact of common benchmarking choices on the stability of algorithm rankings. Using seven representative anomaly detection algorithms and 690 datasets from the OddBench benchmark suite, we analyze how rankings change under varying dataset selections, evaluation metrics, hyperparameter configurations, and random seeds. To quantify this effect, we introduce a rank instability metric measuring the variability of algorithm rankings across benchmark settings.

Our results show that algorithm rankings in anomaly detection are highly unstable. In many cases, almost every competitive algorithm can appear as the best-performing method under some benchmark configuration. Among the studied factors, dataset selection and hyperparameter choice contribute most strongly to ranking uncertainty, while random seeds and evaluation metrics have a comparatively limited impact. Furthermore, we observe that reliable benchmarking requires substantially larger and more diverse dataset collections than the ones commonly used in prior work.

Based on these findings, we recommend that future anomaly detection benchmark studies use at least 200 datasets and explicitly evaluate hyperparameter sensitivity. More broadly, our analysis suggests that minor improvements in benchmark rankings should be interpreted cautiously, as the notion of state-of-the-art performance in anomaly detection is often highly dependent on experimental design choices. Only once major improvements on a large amount of datasets can be observed and are invariant to design choices, it can be assumed that a tangible improvement has been made.

Keywords: Anomaly Detection · Outlier Detection · Benchmarking · Reliability

1 Introduction

Anomaly Detection (AD) is a crucial machine learning problem with many safety-critical applications, ranging from fraud detection [12,16] and failure prevention [5,19] to network intrusion detection [23,24]. Matching this practical relevance, a large variety of anomaly detection algorithms have been proposed over the years, including classical statistical and distance-based approaches [17,21], modern deep learning methods [3,18], and more recently foundation model-based approaches [7,14].

Paradoxically, despite the large methodological diversity, many newly proposed approaches report superior performance over previous methods [7,15,18], which, of course, cannot be true for all of them at once. As a consequence, selecting the "best" anomaly detection algorithm becomes increasingly difficult. While several survey papers and benchmarking studies exist [6,9,13,22], they often fail to identify consistent and practically meaningful differences between algorithms [9]. Furthermore, different studies rely on substantially different sets of algorithms, datasets, evaluation protocols, and hyperparameter settings [6,20], frequently leading to contradictory conclusions.

This raises important questions regarding the reproducibility and reliability of anomaly detection benchmarks. In this work, we therefore investigate the stability of algorithm rankings in anomaly detection and analyze the impact of common benchmarking choices on the resulting conclusions. We will present our methods in the subsequent section and will then evaluate the impact of benchmark choices and benchmark components on rank instability. [1]

2 Methods

To study the reliability of anomaly detection benchmarks, we analyze how algorithm rankings change depending on benchmarking choices. For this purpose, we consider seven anomaly detection algorithms: KNN (k-nearest neighbors) [21], LOF (Local Outlier Factor) [2], IFOR (Isolation Forest) [17], HBOS (Histogram-Based Outlier Score) [8], PCA (Principal Component Analysis) [4], CBLOF (Cluster-Based Local Outlier Factor) [11], and SEAN (Shallow Ensemble ANomaly detection) [15]. We then evaluate how the relative ranking of these algorithms varies across different experimental settings.

We intentionally restrict our study to comparatively lightweight anomaly detection methods with low computational overhead. This choice is motivated by two considerations. First, our goal is not to identify the overall best-performing algorithm, but rather to investigate the sensitivity of benchmark conclusions to methodological choices. Second, limiting the computational complexity enables us to perform a large number of repeated benchmark evaluations across many datasets and settings. As benchmark data, we use the 690 datasets provided by the OddBench benchmark suite [6]. To reduce computational cost, especially for

[1] To increase the reproducibility of our research, our code is freely available at anonymous.4open.science/r/reproduce-34FF

distance-based methods such as KNN and LOF, each dataset is limited to at most 5000 randomly sampled training and test instances. Unless stated otherwise, we further select a random subset of 100 datasets for each benchmark configuration.

To quantify the stability of algorithm rankings, we introduce our novel *rank instability* metric. Let $\mathrm{Rank}(\mathrm{algo}, \mathrm{setting})$ denote the rank achieved by an algorithm under a specific benchmark setting. We define the overall instability as the mean standard deviation of ranks across settings:

$$\sigma_{\mathrm{rank}} = \mathrm{mean}_{\mathrm{algo}} \left(\mathrm{std}_{\mathrm{setting}} \big(\mathrm{Rank}(\mathrm{algo}, \mathrm{setting}) \big) \right)$$

A low σ_{rank} indicates stable rankings across benchmark settings, whereas higher values indicate that rankings strongly depend on experimental design.

Finally, benchmarking anomaly detection algorithms requires several experimental design choices. In this work, we focus on four major sources of variability:

(1) Dataset selection,
(2) Evaluation metric (ROC-AUC [10] or AUCPR [1]),
(3) Hyperparameter selection,
(4) Random seed selection.

To measure the impact of each factor individually, we simulate 100 random choices of the respective parameter while keeping all remaining components fixed.[2] For each source of variability, we compute the resulting rank instability σ_{rank}.

3 Benchmark Choices and Benchmark Reliability

As we evaluate a large number of benchmark configurations, we observe that the ranking between algorithms changes substantially depending on the specific experimental design. In particular, varying the selected datasets, evaluation metrics, hyperparameter configurations, and random seeds can lead to entirely different conclusions regarding which algorithm performs best. Figure 1 illustrates the distribution of the best-performing algorithm across all evaluated benchmark settings. Remarkably, five of the seven studied algorithms achieve the top rank in more than 10% of all settings. Only HBOS and CBLOF rarely appear as the best-performing methods. HBOS achieves the top rank in only one out of 1016 evaluated settings, while CBLOF was never observed as the best-performing algorithm in our studies.

These observations have two important implications. First, they demonstrate that it is comparatively easy to make a reasonably strong anomaly detection algorithm appear state-of-the-art by selecting favorable benchmark settings. In our experiments, at most eight randomly sampled settings are sufficient to find

[2] For random seeds and hyperparameter configurations, we sample 10 random values per algorithm and combine them randomly across algorithms to obtain 100 distinct benchmark settings.

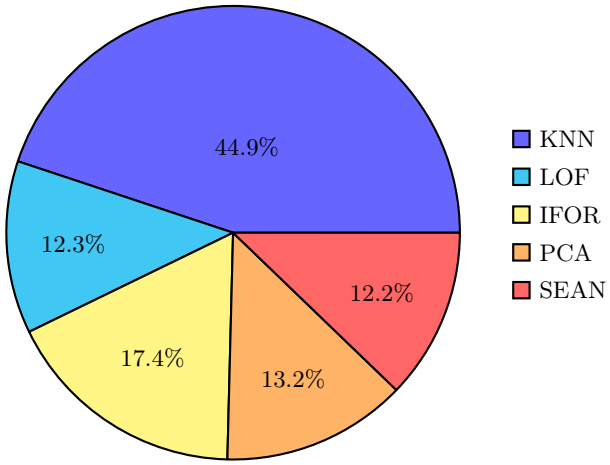

Fig. 1. Distribution of the best-performing algorithm across different benchmark settings. Most algorithms become the top-ranked method under at least some experimental configurations.

a configuration in which a competitive algorithm achieves the best overall performance. Second, the results also indicate that this effect has practical limits: consistently underperforming algorithms cannot easily be made competitive solely through benchmark selection. To better understand the origin of these ranking instabilities, we systematically investigate the impact of individual benchmarking choices on reproducibility and reliability in the following section.

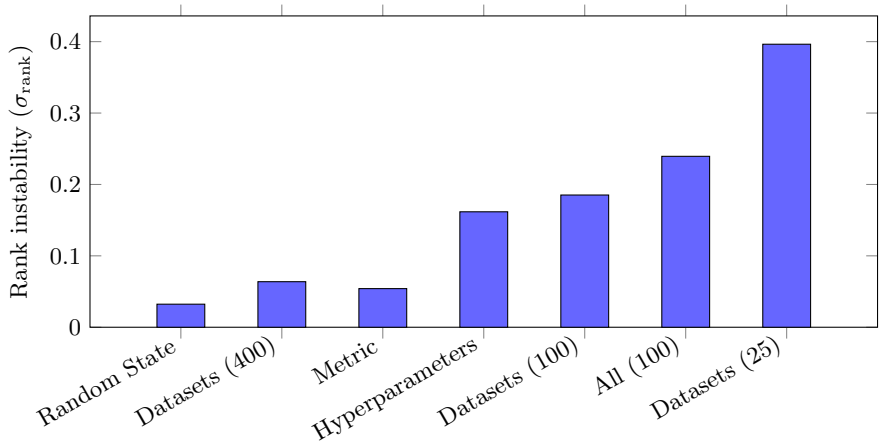

Fig. 2. Impact of different benchmarking choices on the rank instability σ_{rank}. Hyperparameter selection and dataset amount (25, 100, 400) contribute substantially more to ranking uncertainty than random seeds or evaluation metrics.

Figure 2 shows the benchmarking results using our four sources of variability, namely dataset selection, evaluation metrics, hyperparameter selection, and random seed selection. Our analysis demonstrates that benchmark rankings in anomaly detection are substantially affected by experimental design choices. At first glance, the observed instability of approximately $\sigma_{\text{rank}} \approx 0.34$ may appear moderate. However, under a Gaussian interpretation, this implies that only roughly 68% of observed rankings lie within ± 0.34 ranks of the expected ranking. Given the comparatively small ranking differences between algorithms, this uncertainty becomes highly significant in practice. Indeed, across the seven studied algorithms, the average ranking varies by only 1.51 positions overall, ranging from 3.37 for KNN to 4.88 for CBLOF. Consequently, an uncertainty of 0.34 ranks represents a substantial fraction of the total observed performance gap between methods.

Importantly, not all benchmark components contribute equally to rank instability. The influence of random seeds is comparatively negligible, while the choice of evaluation metric has a smaller but still noticeable effect. In contrast, hyperparameter selection introduces substantial ranking uncertainty. Since our study focuses primarily on classical anomaly detection methods with comparatively limited hyperparameter spaces, this effect is likely still underestimated. Modern deep learning-based methods typically involve substantially more hyperparameters and are more sensitive to optimization choices, potentially leading to even larger reproducibility issues.

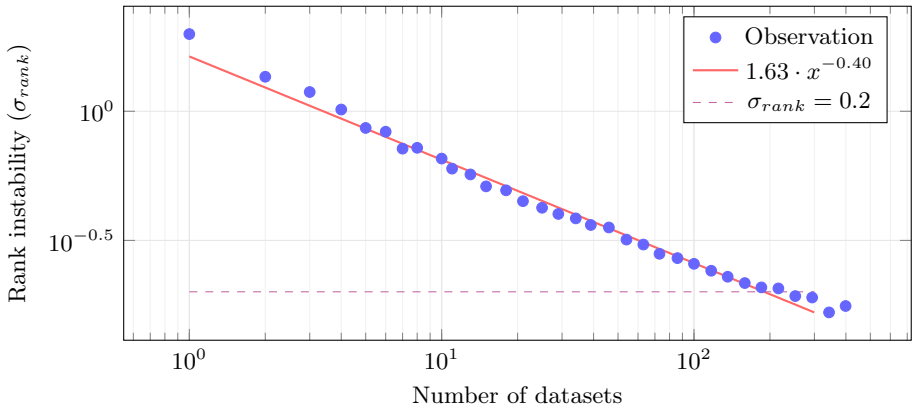

Fig. 3. Relationship between the number of datasets and rank instability σ_{rank}. Increasing the number of datasets improves ranking stability, but the effect saturates due to remaining sources of uncertainty.

Finally, the number of datasets used in a benchmark has a major influence on ranking stability. We analyze this relationship further in Figure 3. As expected, increasing the number of datasets reduces ranking uncertainty. However, this improvement eventually saturates, as the remaining sources of variability dominate

the instability. Overall, our experiments suggest that approximately 200 datasets are required to reduce the total rank instability to a comparatively reliable level of around $\sigma_{\text{rank}} \approx 0.2$.

4 Conclusion and Outlook

In this work, we presented the first systematic study of how different benchmarking choices affect the reliability of anomaly detection benchmark studies. Our analysis demonstrates that benchmark conclusions in anomaly detection are highly sensitive to experimental design decisions and that algorithm rankings change substantially depending on the selected evaluation setup.

In particular, we showed that it is comparatively feasible to construct benchmark settings for most algorithms in which a newly proposed algorithm outperforms existing methods. At the same time, our experiments reveal that the dominant sources of ranking instability are the number of datasets used and the choice of hyperparameters. In contrast, factors such as random seed selection or the specific choice between common evaluation metrics contribute substantially less to the overall uncertainty.

Based on these findings, we recommend that future anomaly detection benchmark studies should include at least 200 datasets whenever computationally feasible. Furthermore, benchmarks should explicitly evaluate algorithms under varying hyperparameter configurations instead of relying on a single manually selected setting and should include a value such as the herein proposed instability score. If computational resources are limited, increasing dataset diversity appears substantially more beneficial than evaluating additional random seeds or multiple closely related evaluation metrics.

More broadly, our results raise questions regarding the current emphasis on state-of-the-art performance in anomaly detection research. The observed ranking instability suggests that the notion of a universally "best" algorithm is often not robust and may depend strongly on benchmark construction choices. Consequently, we argue that future research should place greater emphasis on robustness, reproducibility, and understanding the conditions under which algorithms perform well, rather than focusing primarily on marginal improvements in benchmark rankings.

References

1. Boyd, K., Eng, K.H., Page, C.D.: Area under the Precision-Recall Curve: Point Estimates and Confidence Intervals. In: Blockeel, H., Kersting, K., Nijssen, S., Železný, F. (eds.) Machine Learning and Knowledge Discovery in Databases. pp. 451–466. Springer Berlin Heidelberg, Berlin, Heidelberg (2013)
2. Breunig, M., Kröger, P., Ng, R., Sander, J.: LOF: Identifying Density-Based Local Outliers. ACM Sigmod Record **29**, 93–104 (06 2000). https://doi.org/10.1145/342009.335388

3. Böing, B., Klüttermann, S., Müller, E.: Post-robustifying deep anomaly detection ensembles by model selection. In: 2022 IEEE International Conference on Data Mining (ICDM) (2022)
4. Callegari, C., Gazzarrini, L., Giordano, S., Pagano, M., Pepe, T.: A Novel PCA-Based Network Anomaly Detection. IEEE International Conference on Communications (ICC) pp. 1 – 5 (07 2011). https://doi.org/10.1109/icc.2011.5962595
5. Carvalho, T.P., Soares, F.A., Vita, R., Francisco, R.d.P., Basto, J.P., Alcalá, S.G.: A Systematic Literature Review of Machine Learning Methods Applied to Predictive Maintenance. Computers & Industrial Engineering **137**, 106024 (2019)
6. Ding, X., Klüttermann, S., Wen, H., Chen, Y., Akoglu, L.: MacrOData: New Benchmarks of Thousands of Datasets for Tabular Outlier Detection (2026), https://arxiv.org/abs/2602.09329
7. Ding, X., Wen, H., Klütterman, S., Akoglu, L.: From Zero to Hero: Advancing Zero-Shot Foundation Models for Tabular Outlier Detection (2026), https://arxiv.org/abs/2602.03018
8. Goldstein, M., Dengel, A.R.: Histogram-based Outlier Score (HBOS): A fast Unsupervised Anomaly Detection Algorithm. KI-2012: poster and demo track (2012)
9. Han, S., Hu, X., Huang, H., Jiang, M., Zhao, Y.: Adbench: Anomaly detection benchmark. In: Advances in neural Information Processing Systems (2022)
10. Hanley, J., Mcneil, B.: The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve. Radiology **143**, 29–36 (05 1982). https://doi.org/10.1148/radiology.143.1.7063747
11. He, Z., Xu, X., Deng, S.: Discovering cluster-based local outliers. Pattern Recognition Letters **24**(9-10), 1641–1650 (2003)
12. Hilal, W., Gadsden, S.A., Yawney, J.: Financial Fraud: A Review of Anomaly Detection Techniques and Recent Advances. Expert Systems with Applications (2022)
13. Klüttermann, S., Gupta, S., Müller, E.: Evaluating Anomaly Detection Algorithms: The Role of Hyperparameters and Standardized Benchmarks. In: Proceedings of the 2025 IEEE International Conference on Data Science and Advanced Analytics (DSAA) (2025)
14. Klüttermann, S., Katzke, T., Nguyen, P.H., Müller, E.: FoMo X: Modular Explainability Signals for Outlier Detection Foundation Models (2026), https://arxiv.org/abs/2603.17570
15. Klüttermann, S., Peka, V., Doebler, P., Müller, E.: Towards Highly Efficient Anomaly Detection for Predictive Maintenance. In: International Conference on Machine Learning and Applications (ICMLA) (2024). https://doi.org/10.1109/ICMLA61862.2024.00261
16. Levin, M.Z.R.A.I.: Election Forensics: Using Machine Learning and Synthetic Data for Possible Election Anomaly Detection. PLoS ONE **14** (2019). https://doi.org/10.1371/journal.pone.0223950
17. Liu, F.T., Ting, K.M., Zhou, Z.H.: Isolation forest. In: IEEE International Conference on Data Mining (ICDM) (2008)
18. Livernoche, V., Jain, V., Hezaveh, Y., Ravanbakhsh, S.: On Diffusion Modeling for Anomaly Detection. In: International Conference on Learning Representations (ICLR) (2024)
19. Miljković, D.: Fault Detection Methods: A Literature Survey. In: 2011 Proceedings of the 34th International Convention MIPRO. pp. 750–755 (2011)
20. Olteanu, M., Rossi, F., Yger, F.: Meta-survey on Outlier and Anomaly Detection. Neurocomputing **555**, 126634 (2023). https://doi.org/https://doi.org/10.1016/j.neucom.2023.126634

21. Ramaswamy, S., Rastogi, R., Shim, K.: Efficient Algorithms for Mining Outliers from Large Data Sets. In: SIGMOD 2000. p. 427–438. Association for Computing Machinery (2000). https://doi.org/10.1145/342009.335437
22. Ruff, L., Kauffmann, J., Vandermeulen, R., Montavon, G., Samek, W., Kloft, M., Dietterich, T., Müller, K.R.: A Unifying Review of Deep and Shallow Anomaly Detection. Proceedings of the IEEE **PP**, 1–40 (02 2021)
23. Salo, F., Nassif, A.B., Essex, A.: Dimensionality Reduction with IG-PCA and Ensemble Classifier for Network Intrusion Detection. Computer Networks **148**, 164–175 (2019). https://doi.org/https://doi.org/10.1016/j.comnet.2018.11.010
24. Xu, W., Jang-Jaccard, J., Liu, T., Sabrina, F., Kwak, J.: Improved Bidirectional GAN-Based Approach for Network Intrusion Detection Using One-Class Classifier. MDPI Computers **11**(6) (2022). https://doi.org/10.3390/computers11060085