We Need to Rethink Benchmarking in Anomaly Detection

Philipp Röchner[*]
University of Mainz
roechner@uni-mainz.de

Simon Klüttermann[*]
TU Dortmund
simon.kluettermann@cs.tu-dortmund.de

Franz Rothlauf
University of Mainz
rothlauf@uni-mainz.de

Daniel Schlör[*]
University of Würzburg
schloer@informatik.uni-wuerzburg.de

Abstract

Despite the continuous proposal of new anomaly detection algorithms and extensive benchmarking efforts, progress seems to stagnate, with only minor performance differences between established baselines and new algorithms. In this position paper, we argue that this stagnation is due to limitations in how we evaluate anomaly detection algorithms. Current benchmarking does not, for example, sufficiently reflect the diversity of anomalies in applications ranging from predictive maintenance to scientific discovery. Consequently, we need to rethink benchmarking in anomaly detection. In our opinion, anomaly detection should be studied using *scenarios* that capture the relevant characteristics of different applications. We identify three key areas for improvement: First, we need to identify anomaly detection scenarios based on a common taxonomy. Second, anomaly detection pipelines should be analyzed end-to-end and by component. Third, evaluating anomaly detection algorithms should be meaningful regarding the scenario's objectives.

1 Introduction

Anomaly detection, sometimes also called outlier or novelty detection [1], has many different applications, such as detecting attacks on networks [2], quality issues in medical data [3], or financial fraud [4]. Although these tasks share the property that anomalies are rare and different from the rest of the data [5], the definition of anomalies varies fundamentally across domains, which makes it difficult to unify anomaly detection research [1].

Because a single benchmark dataset cannot capture all aspects of anomalies, benchmarking is performed on collections of datasets, which capture multiple characteristics of anomalies [6, 7]. This approach distinguishes anomaly detection from other areas of machine learning research, such as image classification, where the problem is more clearly defined and single benchmark datasets drive the field for a period of time [8, 9].

The current evaluation approach for anomaly detection algorithms has limitations: Even though new and highly complex algorithms are frequently proposed, studies often find that established methods such as k-Nearest Neighbors (kNN) [10] and Isolation Forest [11] perform competitively [7, 12, 13]. For example, in the benchmark performed by Han et al. [7] for unsupervised methods, kNN ranked best for detecting global anomalies in terms of ROC-AUC; DeepSVDD [14] and DAGMM [15], the only two methods in the benchmark that use neural networks, ranked last and second last, respectively. Moreover, Bouman et al. [12] identify kNN as the best algorithm for datasets with local anomalies

[*]equal contribution

Preprint. Under review.

and Extended Isolation Forest [16] as the best for global anomalies, both outperforming recently published methods.

To illustrate the limitations of current benchmarking practices that allow simple methods to produce seemingly convincing results, we discuss the following anomaly detection algorithm *Quantiles*:

Observations with extreme values in a feature are more likely to be anomalies. The anomaly score for each observation is the fraction of features with values less than the 5%*-quantile or greater than the* 95%*-quantile of the training data.*

Figure 1 shows a critical difference diagram comparing the performance of our *Quantiles* approach with two state-of-the-art approaches (GoAD [17] and DTE [13]), as well as with two established baselines (DeepSVDD [14] and OCSVM [18]).The training data only contains normal observations, and we use standard hyperparameters. We then rank the models by their ROC-AUC on each dataset of the widely used ADBench benchmark [7] and report their average ranks across all datasets. Algorithms whose performance was not statistically different are connected by horizontal bars. In brackets after each algorithm, we show the number of code lines used to implement it. We follow the PyOD implementations for DeepSVDD and OCSVM and the authors' implementation for GoAD and DTE. Our *Quantiles* approach ranks second, slightly ahead of DeepSVDD, GoAD, and OCSVM; only DTE ranks better. The performance of the *Quantiles* approach was not statistically different from the other methods. Although our simple competitor seems to perform well, it cannot detect certain types of anomalies. Consider, e.g., an anomalous sample in the middle of an annulus of normal points. These results question the common practice of using a large and diverse collection of datasets for benchmarking to demonstrate improvements in anomaly detection [7, 12].

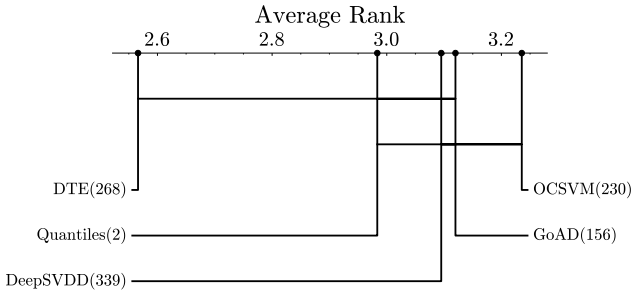

Figure 1: Critical difference diagram. We show the number of code lines used to implement each algorithm in brackets. Our simple *Quantiles* algorithm considers extreme values separately for each variable. It performs similarly to state-of-the-art algorithms and established baselines on the ADBench datasets. But it cannot detect many types of anomalies, e.g., an anomalous sample located within an annulus of normal points. **Takeaway:** These results question the common practice of using a large and diverse collection of datasets for benchmarking to demonstrate improvements in anomaly detection [7, 12].

Motivated by this observation, this position paper argues that **we need to rethink benchmarking in anomaly detection.** We therefore discuss limitations and potential improvements of the current evaluation practice in anomaly detection. In our opinion, there are three main areas where benchmarking needs to evolve: First, anomaly detection should be classified, discussed, and evaluated in scenarios that capture the relevant characteristics of multiple applications, such as anomaly and data type (see Section 2). Second, development decisions and evaluation should consider all individual components of the anomaly detection pipeline, including preprocessing and hyperparameter choices (see Section 3). Third, anomaly detection algorithms should be evaluated in a scenario-specific setting using meaningful evaluation objectives regarding the problem studied, for example, also considering diversity, severity, or fairness (see Section 4).

Our goal is to stimulate discussion within the anomaly detection community about benchmarking and refining evaluation practices. In this position paper, we highlight pressing issues and suggest ways to address them. We invite the research community to discuss and extend our views, aiming for more meaningful evaluations that motivate new approaches.

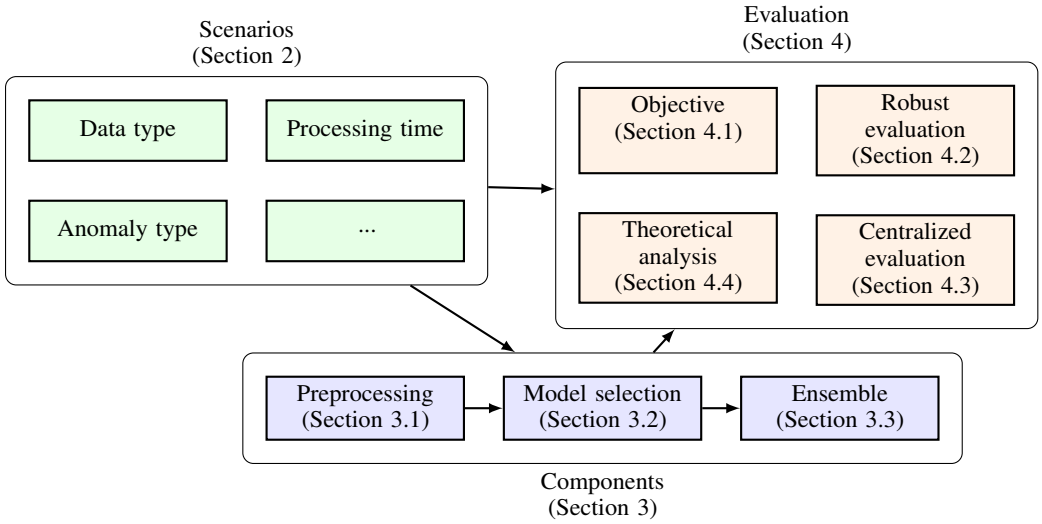

Figure 2: The proposed scenario-specific benchmarking. Each scenario describes relevant characteristics of multiple anomaly detection applications. The chosen scenario determines both the anomaly detection components and the algorithm evaluation.

2 Scenario-specific Benchmarking

2.1 From Applications to a Taxonomy for Anomaly Detection Scenarios

Anomalies are often defined as patterns in data that deviate from normal behavior. These patterns originate from the unusual behavior of the underlying generative process [19, 20]. This perspective requires defining normal behavior and quantifying deviations from it. Without explicit and practical definitions of normality, domain experts often determine the semantic boundary between normal data, noise, and anomalies, tailoring it to their specific anomaly detection task [20]. Thus, either through explicit definitions or post-hoc analysis, the properties that distinguish anomalies from normal data are application dependent.

Therefore, newly proposed anomaly detection approaches and their evaluations must consider the application. Since anomalies differ widely across tasks, application domains, and data types [21], one cannot expect a one-size-fits-all solution for the *best* anomaly detection algorithm [22], which is in line with the no-free-lunch theorem. Instead, we expect specialized algorithms to perform well in their intended applications. This finding is also supported by empirical studies [22, 12].

The current practice of evaluating anomaly detection approaches using collections of benchmark datasets that cover various domains and tasks has been widely adopted. This practice does not, however, account for those differences. As Bouman et al. [12] recently criticized, little is known about the anomalies in commonly used benchmark datasets. Treating results from highly dissimilar benchmark datasets equally can produce paradoxical findings, as our anecdotal example from Figure 1 shows. These effects are theoretically grounded and are known as Simpson's paradox [23]. When applied to anomaly detection benchmarking, new algorithms can outperform previous state-of-the-art algorithms for a particular group of datasets related to one application. This improvement is missed when aggregating across all benchmark datasets.

In our opinion, **treating all benchmark datasets equally when evaluating algorithms and seeking the best overall performance hinders scientific progress in anomaly detection**. We base this position on our observation that an aggregated view overlooks the characteristics of specific applications.

Because there are so many applications, it is impractical to compare anomaly detection algorithms for each one separately. Therefore, we believe that fair comparisons require considering the specific properties of *scenarios*. Scenarios describe relevant aspects of different anomaly detection applications, such as anomaly and data type, or evaluation objective. This type of comparison provides a detailed

view highlighting improvements in specific scenarios. To abstract anomaly detection applications into scenarios, one can define a taxonomy covering the important characteristics of each scenario. Using this taxonomy, we can derive scenario-specific subsets of benchmark datasets.

Various anomaly detection taxonomies have been proposed [19, 24–31, 7]. Most taxonomies focus on learning paradigms (supervised, semi-supervised, or unsupervised), algorithm type [32], or application aspects [33, 34]. The data-centric taxonomy of anomaly characteristics proposed by Foorthuis [21] categorizes anomaly detection by datasets, algorithms, and objectives. However, categorizing anomaly detection methods by application or algorithm family can overlook important aspects such as anomaly characteristics (e.g., local versus global, sequence versus point versus clustered, univariate versus multivariate, distance versus density versus clustering anomalies, or anomaly ratio), (intrinsic) data dimensionality, time criticality (e.g., real-time versus batch processing), and evaluation constraints. In particular, there may be connections between these aspects. For instance, some algorithms may be designed to detect specific types of anomalies, while others may be able to detect diverse types of anomalies.

2.2 Partially Scenario-specific Benchmarks

Several recent studies have recognized the limitations of aggregated benchmarks in anomaly detection and have begun to shift toward more scenario-specific evaluations.

ADBench [7] illustrates how aggregated rankings can obscure meaningful performance differences. While the overall performance is similar across methods, breaking down the results by anomaly type reveals substantial differences. For example, despite its lower overall rank, LOF outperforms other algorithms on local anomalies. The authors note that anomaly type information can be more informative than label information alone, implicitly supporting our argument for scenario-specific evaluations.

In time series anomaly detection, TSB-AD [35] differentiates anomaly types and aligns evaluation with task-specific goals. Although it still aggregates across dimensions such as dataset size and complexity, TSB-AD provides a basis for more nuanced benchmarking.

In medical imaging, BMAD [36] organizes datasets and tasks around clinical contexts and diagnostic goals. This approach naturally allows for evaluation by scenario.

In cybersecurity, datasets such as UNSW-NB15, BoT-IoT, ToN-IoT, and CSE-CIC-IDS2018, as analyzed in [37], report performance per attack category (e.g., DoS, reconnaissance, exploits). This approach allows for more targeted evaluations. However, practical considerations, such as analyst workload and false positive costs, remain underaddressed.

While these efforts represent meaningful progress beyond purely aggregated benchmarking, they remain only partially scenario-specific. Many benchmarks combine anomaly types, tasks, and data characteristics or lack alignment with real-world constraints. Therefore, a more systematic framework is needed that considers context, task, and anomaly type. Nevertheless, these contributions demonstrate that curating and expanding existing datasets is feasible to reflect practical considerations better.

2.3 Examples of Anomaly Detection Scenarios

We present potential scenarios for scenario-specific benchmarking (see Table 1). These diverse examples highlight the necessity of such benchmarking and the key characteristics of scenarios. We emphasize, however, that collaboration with domain experts is essential to ensure real-world relevance.

The number of scenarios should be practical to avoid fragmentation in research while addressing diverse domain needs. We believe a low double-digit count is sufficient to capture the complexity of most applications and allow for detailed distinctions of sub-scenarios.

Real-time Detection of Sequential Anomalies Regarding Severity For autonomous driving, we aim to identify unexpected events that deviate from the typical driving environment. Therefore, we examine video and sensor data with a time component. To react in time, we need to process the data in real time. Ideally, the model output is not binary as is typically studied in anomaly detection. Instead,

Table 1: Examples of anomaly detection scenarios. Each scenario describes relevant characteristics of multiple applications. We emphasize, however, that collaboration with domain experts is essential to ensure real-world relevance.

Application		Scenario			
Task	Domain	Anomaly Type	Data Type	Processing time	Objective
Object detection	Autonomous driving	Sequence anomalies	Multidimensional, sequential data	Real-time	Severity
Quality control	Healthcare	Clustered anomalies	Numerical and categorical tabular data	Batches	Diversity
Fraud detection	Finance	Sparse, contextual anomalies	Numerical and categorical tabular data	Near real-time	Risk score, robustness

we are interested in the severity of unexpected events. Depending on the severity, the autonomous driving system will respond differently. There may be no response for low severity. For high severity, the system might try to avoid obstacles or brake. A corresponding scenario generalizing this application is the real-time detection of severe sequence anomalies in multidimensional sequences.

Batch Detection of Diverse Clustered Anomalies The second application is the automated quality control of medical data. In this application, the goal is to detect implausible patient information in databases using anomaly detection. Since this application is not time-critical, batch processing is sufficient. The data is often tabular, with numeric and categorical variables, but it can also include images or genetic data. Regarding data quality, domain experts are often interested in more than just detecting low-quality data; they want to find the root causes of quality problems. To detect these root causes, they must identify quality problems caused by multiple root causes. Detecting many quality problems caused by the same root cause may be less relevant. Therefore, in this application, we are interested in finding diverse anomalies related to different root causes. The relevant scenario for this application is detecting diverse, clustered anomalies in numerical and categorical tabular data in batches.

Near Real-time Detection of Rapidly Changing Contextual Anomalies Another application of anomaly detection is in the financial service industry. High-volume transactions, such as ATM withdrawals or bank transfers, are processed in near real-time. Detecting fraud is a primary concern and operational challenge for financial institutions. Each transaction is characterized by numerical and categorical variables that can be enriched with graph-based features. Underlying data distributions are highly dynamic. For example, prices and spending behavior change when cardholders travel to different countries. Since fraudsters constantly adapt their scams, detection approaches must respond accordingly. Rather than using binary labels, analysts often want to quantify severity and uncertainty using risk scores. Predictions must be made within a short time frame — from milliseconds to seconds — creating latency and throughput constraints for near real-time processing. A benchmark scenario should therefore focus on detection severity, false alarms, near real-time inference, and robustness to changing anomaly patterns.

3 Scenario-specific Anomaly Detection Components

Researchers often propose end-to-end anomaly detection pipelines that include specific data preprocessing, hyperparameters, and evaluation procedures [38–40, 14, 17]. However, this approach complicates understanding improvements, making it unclear whether enhancements originate from new algorithms or different data preprocessing. As a result, insights from these studies are limited. Even when components of the pipeline are improved, their effects on other pipelines remain unclear. Jiang et al. [41] support this idea, arguing that a modular perspective, rather than an end-to-end approach, enhances the performance of deep one-class anomaly detection.

Studying specific anomaly detection components is crucial due to the varying requirements of different scenarios (see Section 2), setting it apart from other machine learning fields where such differentiation is less critical. Therefore, **we advocate for a modular approach to anomaly detection research.** Instead of only focusing on the overall performance of end-to-end systems, researchers should

analyze individual components within anomaly detection pipelines, incorporating ablation studies and insights from existing literature [41]. This shift could lead to improved performance and new research directions.

3.1 Data Preprocessing

Anomaly detection pipelines typically start with data preprocessing, often affecting detection performance. Unlike other fields, this step is critical and depends on the scenario being studied. For example, Kandanaarachchi et al. [42] have shown that for tabular data, the normalization method used can have a significant impact on the performance of the anomaly detection algorithm. Normalization becomes even more critical when considering complex datasets containing categorical variables, text, or images, which are often mapped to low-dimensional embeddings before anomaly detection. For example, Han et al. [7] uses embeddings for image and text data and then applies anomaly detection algorithms to the tabular data.

Data preprocessing steps are rarely explicitly designed for anomaly detection, as pointed out by Kandanaarachchi et al. [42]. For example, when processing numerical tabular data with neural networks, variables are usually min-max scaled to the interval $[0, 1]$ or to zero mean and unit variance. However, both transformations are sensitive to large values, which can negatively affect subsequent anomaly detection algorithms. Robust standardization could be beneficial, such as transforming the data to zero median and unit median absolute deviation to the median.

Understanding how existing preprocessing steps interact with other anomaly detection components and proposing new preprocessing steps could open up unexplored research directions, such as how to design embeddings for text and image datasets that capture the differences between anomalies and normal data well for subsequent tabular anomaly detection algorithms.

3.2 Model Selection

Researchers constantly propose new anomaly detection algorithms [43–53, 15, 54, 14, 13]. Selecting a *good* one from these algorithms is challenging, especially when there are few or no ground-truth labels. Additionally, choosing appropriate hyperparameters can be crucial for effective anomaly detection and often depends on the studied task [55].

Surveys [56, 1] guide practitioners' algorithm choices by reporting the highest average performance on several benchmark datasets. However, these surveys tend to conclude that there is no statistical difference in the overall performance of different algorithms [7, 12]. To better understand the differences between algorithms, Han et al. [7] and Bouman et al. [12] examine the performance of anomaly detection algorithms for different anomaly types. The choice of hyperparameters is well studied in the supervised setting [57]. However, there is a significant gap in the unsupervised or one-class case because algorithms lack feedback to guide hyperparameter selection.

Some recent papers have studied the automatic and unsupervised selection of algorithms and hyperparameters [58–62]. For example, Zhao et al. [58] use meta-learning to predict good models for a given task. However, their algorithm selection is limited by the meta-learning assumption that multiple labeled datasets similar to the one studied are available. Ma et al. [63] found that the proposed automatic model selection approaches provide limited benefit to practitioners.

We believe this is due to the broad and diverse nature of anomaly detection. To effectively guide the selection of algorithms and hyperparameters, surveys should examine not only model performance for different types of anomalies, but also the information provided by relevant anomaly detection scenarios, as discussed in Section 2. Regarding the automated selection of algorithms and hyperparameters, we advocate for approaches that consider scenario-specific characteristics.

3.3 Anomaly Detection Ensembles

Ensemble methods that combine different anomaly detection methods often outperform individual models [64, 65]. However, best practices for ensemble methods are rarely studied because using multiple models exponentially increases the number of choices researchers must make when designing an ensemble, making such analyses complex. This problem is particularly challenging for unsupervised anomaly detection because there is no feedback when combining different ensemble components.

Thus, while ensembles usually improve performance, this is based on implicit knowledge and personal experience [66].

A modular approach can help reduce the exponential number of combinations by individually identifying best practices for each anomaly detection component. We can further simplify this by using an appropriate taxonomy and specifying the studied scenario. For example, if researchers want to detect local *and* global anomalies, they can choose the ensemble components accordingly.

4 Scenario-specific Evaluation

The evaluation of anomaly detection algorithms typically relies on metrics like ROC-AUC and AUC-PR, which assess the ranking of observations based on anomaly scores against binary ground-truth labels. However, the effectiveness of these evaluations can vary depending on the algorithm and the context in which it is applied. For example, a classical approach like kNN [67] may not work well when applied directly to high-dimensional images, but it can work efficiently with proper preprocessing [7].

In our opinion, anomaly detection algorithms need to be evaluated on a scenario-by-scenario basis. This means evaluating not only the standard metrics but also considering the specific anomaly detection scenario, associated objectives, and relevant evaluation protocols (see Section 2). We also need to take into account the various components involved in anomaly detection (see Section 3).

Current rank-based evaluations often oversimplify the complexity of anomaly detection, focusing mainly on the relationship between ranked scores and binary labels. They frequently overlook critical aspects such as the diversity and severity of anomalies, the algorithms' robustness against adversarial attacks, and fairness considerations. Essential scenario-specific factors include robustness to noise in healthcare applications [68], fairness in business screenings [69], and interpretability in engineering for prompt problem resolution [70].

4.1 Evaluation Objectives

We propose studying performance objectives that focus on properties relevant to anomaly detection tasks. For instance, in identifying quality issues in a database, prioritizing fewer but more diverse and severe anomalies is often more valuable than detecting many similar, insignificant ones (see Section 2.3). Diversity is a well-studied concept for recommender systems [71].

Current rank-based evaluations, like ROC-AUC, do not optimally reflect the importance of incorrectly ranked observations, particularly those ranked higher, which can be more critical for practitioners. To address this, top-heavy objectives such as normalized discounted cumulative gain should be considered [72], following the argument of Aggarwal [20].

We encourage the anomaly detection community to suggest additional evaluation objectives. Expanding the rank-based evaluation can lead to new research directions, including proposing new objectives and datasets for existing algorithms or developing new algorithms that enhance these properties. Some approaches incorporate human-guided feedback to refine anomaly types [73, 74]. By exploring scenario-specific objectives beyond rank-based objectives, we can align anomaly detection research with practical needs and improve algorithm performance in specific scenarios.

4.2 Robust Evaluation

Evaluating anomaly detection algorithms involves considering the uncertainty introduced by both the algorithms and the benchmark datasets. Many algorithms, like Isolation Forests [75], include random components, while datasets often represent a single sample from complex generative processes for both anomalies and normal observations. If a dataset has insufficient anomalies, its findings may not generalize well. For example, the ADBench [7] benchmark includes datasets with very few anomalies, leading to potential missing types in data subsets used for certain evaluation setups, such as train-test or k-fold cross-validation.

Unfortunately, many algorithms are only trained and evaluated once on a fixed dataset, which does not capture the variability of different dataset splits or seeds. This limits the generalizability of results and raises questions about their real-world relevance.

Researchers in anomaly detection should consider sampling variability and random factors in their evaluations. Repeating assessments with multiple initializations and reporting performance variability statistically, such as through confidence intervals [76] or bootstrapping [77, 78], is essential. However, small sample sizes can complicate statistical interpretations [76, 79]. Creating twin-datasets from the same generative processes could make evaluations more robust [80].

Overall, a robust evaluation that considers both dataset and model variability ensures reliable performance estimates and meaningful assessments of advancements in anomaly detection research.

4.3 Centralized Evaluation

Benchmark datasets lose relevance over time as they are used to design and optimize algorithms [81]. In other words, there is a risk of overfitting algorithms to these benchmarks, limiting their generalizability to new real-world datasets. Anomaly detection is particularly susceptible to this issue because benchmark datasets contain rare anomalies that researchers must assume they have no prior knowledge of. However, these anomalies often influence algorithm design and hyperparameter choices, which can lead to significant performance improvements. These choices may explain why anomaly detection research papers often omit details on network architecture or hyperparameters [82]. Furthermore, these choices are nearly undetectable to the research community in the resulting papers.

We believe that reliable anomaly detection research requires a way to limit such unwanted optimizations. We want to highlight an approach that we believe solves this problem: Kasieczka et al. [83] published an uncontaminated dataset[2] while keeping anomalous data private. This approach ensures that the evaluation is fair. While the implementation has shortcomings, such as a limited set of possible taxonomies and a small number of anomalies, we believe that anomaly detection should be evaluated similarly. Although benchmarks serve a similar purpose, researchers have little incentive to design algorithms and select hyperparameters without relying on ground-truth labels. Therefore, we argue that evaluation strategies commonly used in online leaderboards and challenges should also be adopted for anomaly detection, complementing traditional benchmarks.

4.4 Theoretical Analysis

Theoretical analysis is rarely utilized to support new algorithms in anomaly detection, which is surprising given its potential to overcome limitations like the shortage of datasets for empirical evaluation. This may originate from the field's broad nature and unclear design goals. For example, while an anomaly score based on density distribution can be effective in one-class scenarios, as low-density samples are more likely to represent anomalies, it may not suit unsupervised cases where clustered anomalies could mask each other and would be in high-density regions [84].

To address this, we can define specific scenarios with meaningful assumptions (see Section 2), allowing for generalization across different approaches within a scenario. A well-defined taxonomy enables the creation of synthetic datasets with specific properties, offering deeper insights than standard benchmark datasets. For instance, Klüttermann and Müller [85] highlight an algorithm's failure to detect rare anomalies, using synthetic experiments to identify the minimum frequency required for successful detection.

5 Conclusions

This paper argues for rethinking how we benchmark anomaly detection algorithms, criticizing the one-size-fits-all approach that overlooks the diversity of anomalies in different applications. We propose three areas for improvement based on *scenarios* that capture the relevant characteristics of different applications: identifying anomaly detection scenarios based on a common taxonomy, examining detection pipelines both end-to-end and in components, and ensuring evaluation methods are robust and tailored to each scenario. We aim to detail these scenarios in future papers and encourage collaboration among researchers in the field.

[2] https://lhco2020.github.io/homepage/

6 Alternative Views

6.1 Number versus Relevance of Benchmarking Datasets

Alternative view: While promoting a scenario-specific perspective, it is crucial to consider whether evaluating on a limited subset of datasets is preferable to using a wide range of them. Large benchmark collections offer a common reference point for comparing anomaly detection methods, enabling researchers to understand how their methods generalize to other contexts. This approach focuses on methodological advancements and faster development within established frameworks.

Recognizing dataset differences and their underlying assumptions facilitates post-hoc analysis, revealing how specific dataset properties influence results. Such analyses can identify state-of-the-art advancements and inconsistencies across domains, helping determine whether a new method truly outperforms others or addresses niche problems. Practitioners can then select the *optimal* algorithm for similar dataset, while researchers can pinpoint algorithmic families with comparable performance.

In summary, a detailed analysis on various datasets and levels is key to uncovering broader patterns and avoiding misleading conclusions drawn from aggregated results.

Our view: While benchmarking suites serve as useful reference points, a one-size-fits-all evaluation fails to capture the diversity of anomaly detection. Combining datasets from different backgrounds can create misalignments due to varying labeling standards. The distinction between anomalies and normal observations often depends on their relevance to domain experts. Therefore, it is important to consider this factor in benchmarking rather than solely relying on the decisions of benchmark curators.

Relying on post hoc analyses to evaluate performance across numerous datasets risks identifying random variations, necessitating a Bonferroni correction that could mask meaningful insights. We propose a more effective method: first, define your research questions and hypotheses, then design models and experiments to address them. This approach fosters research that is more focused and aligned with real-world applications than the pursuit of an unrealistic overall *best* solution.

6.2 Anomaly Detection Pipelines vs. Components

Alternative view: Individual anomaly detection components cannot detect anomalies; only complete pipelines of components can do so. Consequently, researchers typically design anomaly detection pipelines to maximize end-to-end performance. They rarely expect components to perform well in other anomaly detection pipelines. Benchmark studies evaluating component combinations would increase effort exponentially.

Our view: While anomaly detection components are often optimized for specific pipelines, it is reasonable to expect that some will perform well in novel combinations. By exploring these combinations, we can uncover components that enhance performance generalizing beyond the original setup. Proposing individual anomaly detection components rather than whole pipelines allows researchers to clarify their contributions and foster future research. Additionally, designing algorithms based on a single pipeline using ground-truth labels can lead to overfitting, which may hinder generalization to similar scenarios.

6.3 One vs. Multiple Evaluation Objectives

Alternative view: While scenario-specific objectives may be helpful for some tasks, a larger set of objectives makes it easier to choose an objective in which a proposed algorithm performs well. This can make research arbitrary and make it difficult for practitioners to choose a good algorithm. By sticking to one evaluation objective, which may be suboptimal in specific scenarios, we have a reliable evaluation.

Our view: We agree that defining objectives solely to show an algorithm's superiority over baselines is undesirable. Instead, objectives should be determined by the scenario rather than the algorithm itself. The key question is: Which relevant properties should be measured in my scenario? Ideally, the research community agrees on specific evaluation objectives for each scenario, which is preferable to using general objectives that are less relevant.

References

[1] Madalina Olteanu, Fabrice Rossi, and Florian Yger. Meta-survey on outlier and anomaly detection. *Neurocomputing*, 555:126634, 2023. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2023.126634. URL https://www.sciencedirect.com/science/article/pii/S0925231223007579.

[2] Aleksandar Lazarevic, Levent Ertöz, Vipin Kumar, Aysel Ozgur, and Jaideep Srivastava. A comparative study of anomaly detection schemes in network intrusion detection. In *SDM*, pages 25–36. SIAM, 2003.

[3] Philipp Röchner and Franz Rothlauf. Unsupervised anomaly detection of implausible electronic health records: a real-world evaluation in cancer registries. *BMC Medical Research Methodology*, 23(1):125, 2023.

[4] Waleed Hilal, S. Andrew Gadsden, and John Yawney. Financial fraud: A review of anomaly detection techniques and recent advances. *Expert Syst. Appl.*, 193:116429, 2022.

[5] D. M. Hawkins. *Identification of Outliers*. Monographs on Applied Probability and Statistics. Springer, 1980.

[6] Guilherme Oliveira Campos, Arthur Zimek, Jörg Sander, Ricardo J. G. B. Campello, Barbora Micenková, Erich Schubert, Ira Assent, and Michael E. Houle. On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. *Data Min. Knowl. Discov.*, 30(4):891–927, 2016.

[7] Songqiao Han, Xiyang Hu, Hailiang Huang, Mingqi Jiang, and Yue Zhao. Adbench: Anomaly detection benchmark. In *Neural Information Processing Systems (NeurIPS)*, 2022.

[8] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Master's thesis, University of Tront*, 2009.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE Computer Society, 2009.

[10] Sridhar Ramaswamy, Rajeev Rastogi, and Kyuseok Shim. Efficient algorithms for mining outliers from large data sets. In *SIGMOD Conference*, pages 427–438. ACM, 2000.

[11] Zhiguo Ding and Minrui Fei. An anomaly detection approach based on isolation forest algorithm for streaming data using sliding window. In *ICONS*, pages 12–17. International Federation of Automatic Control, 2013.

[12] Roel Bouman, Zaharah Bukhsh, and Tom Heskes. Unsupervised anomaly detection algorithms on real-world data: How many do we need? *J. Mach. Learn. Res.*, 25:105:1–105:34, 2024.

[13] Victor Livernoche, Vineet Jain, Yashar Hezaveh, and Siamak Ravanbakhsh. On diffusion modeling for anomaly detection. In *ICLR*. OpenReview.net, 2024.

[14] Lukas Ruff, Nico Görnitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Robert A. Vandermeulen, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 4390–4399. PMLR, 2018.

[15] Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Dae ki Cho, and Haifeng Chen. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *International Conference on Learning Representations*, 2018. URL https://api.semanticscholar.org/CorpusID:51805340.

[16] Sahand Hariri, Matias Carrasco Kind, and Robert J. Brunner. Extended isolation forest. *CoRR*, abs/1811.02141, 2018.

[17] Liron Bergman and Yedid Hoshen. Classification-based anomaly detection for general data. In *ICLR*. OpenReview.net, 2020. URL http://dblp.uni-trier.de/db/conf/iclr/iclr2020.html#BergmanH20.

[18] Mr Moya. One-class classifier networks for target recognition applications. In *Proc. World Congress on Neural Networks, 1993*, pages 797–801, 1993.

[19] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3):15:1–15:58, July 2009. ISSN 0360-0300. doi: 10.1145/1541880.1541882. URL https://doi.org/10.1145/1541880.1541882.

[20] Charu C. Aggarwal. *Outlier Analysis*. Springer, 2013. ISBN 978-1-4614-6396-2. URL http://dx.doi.org/10.1007/978-1-4614-6396-2.

[21] Ralph Foorthuis. On the nature and types of anomalies: a review of deviations in data. *International Journal of Data Science and Analytics*, 12(4):297–331, October 2021. ISSN 2364-415X, 2364-4168. doi: 10.1007/s41060-021-00265-1. URL https://link.springer.com/10.1007/s41060-021-00265-1.

[22] Ece Calikus, Sławomir Nowaczyk, Anita Sant'Anna, and Onur Dikmen. No free lunch but a cheaper supper: A general framework for streaming anomaly detection. *Expert Systems with Applications*, 155:113453, October 2020. ISSN 0957-4174. doi: 10.1016/j.eswa.2020.113453. URL https://www.sciencedirect.com/science/article/pii/S0957417420302773.

[23] E. H. Simpson. The Interpretation of Interaction in Contingency Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 13(2):238–241, 1951. ISSN 0035-9246. URL https://www.jstor.org/stable/2984065. Publisher: [Royal Statistical Society, Oxford University Press].

[24] Sherenaz Al-Haj Baddar, Alessio Merlo, and Mauro Migliardi. Anomaly Detection in Computer Networks: A State-of-the-Art Review. *J. Wirel. Mob. Networks Ubiquitous Comput. Dependable Appl.*, 5(4):29–64, 2014.

[25] Ruchi Makani and B. V. R. Reddy. Taxonomy of Machine Leaning Based Anomaly Detection and its suitability. *Procedia Computer Science*, 132:1842–1849, January 2018. ISSN 1877-0509. doi: 10.1016/j.procs.2018.05.133. URL https://www.sciencedirect.com/science/article/pii/S1877050918308652.

[26] Gheorghe Sebestyen, Anca Hangan, Zoltan Czako, and Gyorgy Kovacs. A taxonomy and platform for anomaly detection. In *2018 IEEE International Conference on Automation, Quality and Testing, Robotics (AQTR)*, pages 1–6, May 2018. doi: 10.1109/AQTR.2018.8402710. URL https://ieeexplore.ieee.org/document/8402710.

[27] Gilberto Fernandes, Joel J. P. C. Rodrigues, Luiz Fernando Carvalho, Jalal F. Al-Muhtadi, and Mario Lemes Proença. A comprehensive survey on network anomaly detection. *Telecommunication Systems*, 70(3):447–489, March 2019. ISSN 1572-9451. doi: 10.1007/s11235-018-0475-8. URL https://doi.org/10.1007/s11235-018-0475-8.

[28] Raghavendra Chalapathy and Sanjay Chawla. Deep Learning for Anomaly Detection: A Survey, January 2019. URL http://arxiv.org/abs/1901.03407. arXiv:1901.03407 [cs].

[29] Thorsten Wittkopp, Philipp Wiesner, Dominik Scheinert, and Odej Kao. A Taxonomy of Anomalies in Log Data, November 2021. URL http://arxiv.org/abs/2111.13462. arXiv:2111.13462 [cs].

[30] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton van den Hengel. Deep Learning for Anomaly Detection: A Review. *ACM Computing Surveys*, 54(2):1–38, March 2022. ISSN 0360-0300, 1557-7341. doi: 10.1145/3439950. URL http://arxiv.org/abs/2007.02500. arXiv:2007.02500 [cs].

[31] Amnah Aldayri and Waleed Albattah. Taxonomy of Anomaly Detection Techniques in Crowd Scenes. *Sensors*, 22(16):6080, January 2022. ISSN 1424-8220. doi: 10.3390/s22166080. URL https://www.mdpi.com/1424-8220/22/16/6080. Number: 16 Publisher: Multidisciplinary Digital Publishing Institute.

[32] Sondre Sørbø and Massimiliano Ruocco. Navigating the metric maze: a taxonomy of evaluation metrics for anomaly detection in time series. *Data Mining and Knowledge Discovery*, 38 (3):1027–1068, May 2024. ISSN 1573-756X. doi: 10.1007/s10618-023-00988-8. URL https://doi.org/10.1007/s10618-023-00988-8.

[33] Juan M. Estevez-Tapiador, Pedro Garcia-Teodoro, and Jesus E. Diaz-Verdejo. Anomaly detection methods in wired networks: a survey and taxonomy. *Computer Communications*, 27 (16):1569–1584, October 2004. ISSN 0140-3664. doi: 10.1016/j.comcom.2004.07.002. URL https://www.sciencedirect.com/science/article/pii/S0140366404002385.

[34] Manasi Gyanchandani, J L Rana, and R N Yadav. Taxonomy of Anomaly Based Intrusion Detection System: A Review. *International Journal of Scientific and Research Publications*, 2 (12), 2012.

[35] Qinghua Liu and John Paparrizos. The elephant in the room: Towards a reliable time-series anomaly detection benchmark. *Advances in Neural Information Processing Systems*, 37:108231–108261, 2024.

[36] Jinan Bao, Hanshi Sun, Hanqiu Deng, Yinsheng He, Zhaoxiang Zhang, and Xingyu Li. Bmad: Benchmarks for medical anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4042–4053, 2024.

[37] Mohanad Sarhan, Siamak Layeghy, Nour Moustafa, and Marius Portmann. Netflow datasets for machine learning-based network intrusion detection systems. In *Big data technologies and applications: 10th EAI international conference, BDTA 2020, and 13th EAI international conference on wireless internet, WiCON 2020, virtual event, December 11, 2020, proceedings 10*, pages 117–135. Springer, 2021.

[38] Andrew Brinkerhoff, Chosila Sutantawibul, Robert White, Caio Daumann, Chad Freer, Indara Suarez, Samuel May, Vivan Nguyen, Jonathan Guiang, Bennett Marsh, Darin Acosta, Alex Aubuchon, Emanuela Barberis, Aaron Bundock, Evan Collins, Preston Epps, Johannes Erdmann, Henning Flaecher, Junshen Huang, Ryan Nie, Sudarshan Parameswaran, John Rotter, Kaitlin Salyer, Siddhesh Sawant, Tanvi Sheokand, and Darien Wood. Anomaly detection for automated data quality monitoring in the cms detector, 2025. URL https://arxiv.org/abs/2501.13789.

[39] Ildar N. Idrisov, Divine Okeke, Abdullatif Albaseer, Mohamed Abdallah, and Federico M. Ibanez. Leveraging digital twin and machine learning techniques for anomaly detection in power electronics dominated grid, 2025. URL https://arxiv.org/abs/2501.13474.

[40] Sota Mashiko, Yuji Kawamata, Tomoru Nakayama, Tetsuya Sakurai, and Yukihiko Okada. Anomaly detection in double-entry bookkeeping data by federated learning system with non-model sharing approach, 2025. URL https://arxiv.org/abs/2501.12723.

[41] Minqi Jiang, Chaochuan Hou, Ao Zheng, Songqiao Han, Hailiang Huang, Qingsong Wen, Xiyang Hu, and Yue Zhao. ADGym: Design Choices for Deep Anomaly Detection, October 2023. URL http://arxiv.org/abs/2309.15376. arXiv:2309.15376 [cs].

[42] Sevvandi Kandanaarachchi, Mario A. Muñoz, Rob J. Hyndman, and Kate Smith-Miles. On normalization and algorithm selection for unsupervised outlier detection. *Data Min. Knowl. Discov.*, 34(2):309–354, 2020.

[43] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. LOF: identifying density-based local outliers. In Weidong Chen, Jeffrey F. Naughton, and Philip A. Bernstein, editors, *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, May 16-18, 2000, Dallas, Texas, USA*, pages 93–104. ACM, 2000. doi: 10.1145/342009.335388. URL https://doi.org/10.1145/342009.335388.

[44] Zengyou He, Xiaofei Xu, and Shengchun Deng. Discovering cluster-based local outliers. *Pattern Recognition Letters*, 24(9):1641–1650, 2003. ISSN 0167-8655. doi: https://doi.org/10.1016/S0167-8655(03)00003-5.

[45] Xiaoyi Gu, Leman Akoglu, and Alessandro Rinaldo. Statistical analysis of nearest neighbor methods for anomaly detection. In *NeurIPS*, pages 10921–10931, 2019. URL http://dblp.uni-trier.de/db/conf/nips/nips2019.html#GuAR19.

[46] Zheng Li, Yue Zhao, Xiyang Hu, Nicola Botta, Cezar Ionescu, and George H. Chen. ECOD: unsupervised outlier detection using empirical cumulative distribution functions. *IEEE Trans. Knowl. Data Eng.*, 35(12):12181–12193, 2023.

[47] Markus Goldstein and Andreas Dengel. Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm. In *KI-2012: poster and demo track*, volume 1, pages 59–63. Citeseer, 2012.

[48] Murray Rosenblatt. Remarks on Some Nonparametric Estimates of a Density Function. *The Annals of Mathematical Statistics*, 27(3):832 – 837, 1956. doi: 10.1214/aoms/1177728190.

[49] Mayu Sakurada and Takehisa Yairi. Anomaly detection using autoencoders with nonlinear dimensionality reduction. In *Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis*, MLSDA'14, page 4–11, 2014. ISBN 9781450331593. doi: 10.1145/2689746.2689747.

[50] 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)*, 2020. doi: 10.1109/ICDM50108.2020.00135.

[51] Hans-Peter Kriegel, Peer Kröger, Erich Schubert, and Arthur Zimek. Outlier detection in axis-parallel subspaces of high dimensional data. In Thanaruk Theeramunkong, Boonserm Kijsirikul, Nick Cercone, and Tu-Bao Ho, editors, *Advances in Knowledge Discovery and Data Mining*, pages 831–838, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg. ISBN 978-3-642-01307-2.

[52] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 1530–1538. JMLR.org, 2015.

[53] Adebena Oluwasegun and Jae-Cheon Jung. A multivariate gaussian mixture model for anomaly detection in transient current signature of control element drive mechanism. *Nuclear Engineering and Design*, 402:112098, 2023. ISSN 0029-5493.

[54] Federico Di Mattia, Paolo Galeone, Michele De Simoni, and Emanuele Ghelfi. A survey on gans for anomaly detection, 2021.

[55] Xueying Ding, Lingxiao Zhao, and Leman Akoglu. Hyperparameter sensitivity in deep outlier detection: Analysis and a scalable hyper-ensemble solution. *Advances in Neural Information Processing Systems*, 35:9603–9616, 2022.

[56] Shehroz S. Khan and Michael G. Madden. A survey of recent trends in one class classification. In Lorcan Coyle and Jill Freyne, editors, *Artificial Intelligence and Cognitive Science*, pages 188–197, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. ISBN 978-3-642-17080-5.

[57] Matthias Feurer, Katharina Eggensperger, Stefan Falkner, Marius Lindauer, and Frank Hutter. Auto-sklearn 2.0: Hands-free automl via meta-learning. *arXiv:2007.04074 [cs.LG]*, 2020.

[58] Yue Zhao, Ryan Rossi, and Leman Akoglu. Automatic unsupervised outlier model selection. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 4489–4502. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/23c894276a2c5a16470e6a31f4618d73-Paper.pdf.

[59] Yue Zhao, Ryan A. Rossi, and Leman Akoglu. Automating outlier detection via meta-learning, 2021. URL https://arxiv.org/abs/2009.10606.

[60] Sebastian Schmidl, Phillip Wenig, and Thorsten Papenbrock. Hypex: Hyperparameter optimization in time series anomaly detection. In *Datenbanksysteme für Business, Technologie und Web*, 2023. URL https://api.semanticscholar.org/CorpusID:257104910.

[61] Chi Wang, Qingyun Wu, Markus Weimer, and Erkang (Eric) Zhu. Flaml: A fast and lightweight automl library. In *Fourth Conference on Machine Learning and Systems (MLSys 2021)*, April 2021. URL https://www.microsoft.com/en-us/research/publication/flaml-a-fast-and-lightweight-automl-library/.

[62] Henrique O Marques, Ricardo JGB Campello, Jörg Sander, and Arthur Zimek. Internal evaluation of unsupervised outlier detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 14(4):1–42, 2020.

[63] Martin Q. Ma, Yue Zhao, Xiaorong Zhang, and Leman Akoglu. The need for unsupervised outlier model selection: A review and evaluation of internal evaluation strategies. *SIGKDD Explor. Newsl.*, 25(1):19–35, July 2023. ISSN 1931-0145. doi: 10.1145/3606274.3606277. URL https://doi.org/10.1145/3606274.3606277.

[64] Alvin Chiang, Esther David, Yuh-Jye Lee, Guy Leshem, and Yi-Ren Yeh. A study on anomaly detection ensembles. *Journal of Applied Logic*, 21, 12 2016. doi: 10.1016/j.jal.2016.12.002.

[65] Simon Klüttermann and Emmanuel Müller. Evaluating and comparing heterogeneous ensemble methods for unsupervised anomaly detection. In *IJCNN*, 2023.

[66] Charu Aggarwal and Saket Sathe. *Outlier Ensembles*. Springer, 04 2017. ISBN 978-3-319-54764-0. doi: 10.1007/978-3-319-54765-7.

[67] Jian Tang, Zhixiang Chen, Ada Wai-Chee Fu, and David Wai-Lok Cheung. Enhancing effectiveness of outlier detections for low density patterns. In *PAKDD*, volume 2336 of *Lecture Notes in Computer Science*, pages 535–548. Springer, 2002.

[68] S. Subha and J. G. R. Sathiaseelan. Effective anomaly detection approach to classify noisy data using robust noise detection and removal technique in iot healthcare data. *SN Computer Science*, 4(5):522, Jul 2023. ISSN 2661-8907. doi: 10.1007/s42979-023-01890-2. URL https://doi.org/10.1007/s42979-023-01890-2.

[69] Y. Dai, M. Jayaratne, and B. Jayatilleke. Question-aware outlier answer detection for fairer ai scoring of interviews. In *38th Annual Conference of the Society for Industrial & Organizational Psychology (SIOP)*, apr 2023.

[70] Guoxiang Zhong, Fagui Liu, Jun Jiang, and C. L. Philip Chen. Causeformer: Interpretable anomaly detection with stepwise attention for cloud service. *IEEE Transactions on Network and Service Management*, 21(1):637–652, 2024. doi: 10.1109/TNSM.2023.3299846.

[71] Matevz Kunaver and Tomaz Pozrl. Diversity in recommender systems - A survey. *Knowl. Based Syst.*, 123:154–162, 2017.

[72] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, 2002.

[73] M. Lochner and B.A. Bassett. Astronomaly: Personalised active anomaly detection in astronomical data. *Astronomy and Computing*, 36:100481, 2021. ISSN 2213-1337. doi: https://doi.org/10.1016/j.ascom.2021.100481. URL https://www.sciencedirect.com/science/article/pii/S2213133721000354.

[74] Michelle Lochner and Lawrence Rudnick. Astronomaly protege: Discovery through human-machine collaboration, 2024. URL https://arxiv.org/abs/2411.04188.

[75] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation-based anomaly detection. *ACM Trans. Knowl. Discov. Data*, 6(1):3:1–3:39, 2012.

[76] Kendrick Boyd, Kevin H Eng, and C David Page. Area under the precision-recall curve: point estimates and confidence intervals. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III 13*, pages 451–466. Springer, 2013.

[77] Diego Didona and Paolo Romano. On bootstrapping machine learning performance predictors via analytical models. *CoRR*, abs/1410.5102, 2014.

[78] Michael A Kouritzin, Stephen Styles, and Beatrice-Helen Vritsiou. A bootstrap algorithm for fast supervised learning, 2023. URL https://arxiv.org/abs/2305.03099.

[79] B Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

[80] Olawale Salaudeen and Moritz Hardt. Imagenot: A contrast with imagenet preserves model rankings. *arXiv preprint arXiv:2404.02112*, 2024.

[81] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet?, 2019. URL https://arxiv.org/abs/1902.10811.

[82] Haoyi Fan, Fengbin Zhang, and Zuoyong Li. Anomalydae: Dual autoencoder for anomaly detection on attributed networks. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5685–5689, 2020. URL https://api.semanticscholar.org/CorpusID:211069030.

[83] Gregor Kasieczka, Benjamin Nachman, David Shih, Oz Amram, Anders Andreassen, Kees Benkendorfer, Blaz Bortolato, Gustaaf Brooijmans, Florencia Canelli, Jack H Collins, et al. The lhc olympics 2020 a community challenge for anomaly detection in high energy physics. *Reports on progress in physics*, 84(12):124201, 2021.

[84] Albin Iivari. Anomaly detection techniques for unsupervised machine learning, 2022. URL https://www.diva-portal.org/smash/get/diva2:1674371/FULLTEXT01.pdf. 15 credits.

[85] Simon Klüttermann and Emmanuel Müller. About test-time training for outlier detection, 2024. URL https://arxiv.org/abs/2404.03495.