We Need to Rethink Benchmarking in Anomaly Detection

Position Paper Track

Philipp Röchner[*,1,2] Simon Klüttermann[*,3,4] Kevin Kammler[1] Franz Rothlauf[1] Emmanuel Müller[3] Daniel Schlör[*,5]

NeurIPS 2026

[1]Johannes Gutenberg University Mainz [2]University of Southern Denmark [3]TU Dortmund
[4]Carnegie Mellon University [5]University of Würzburg [*]Equal contribution

Status Quo: Benchmarking in Anomaly Detection

Applications

- Network attacks, fraud
- Shared: anomalies differ from normal data
- Differing: anomaly definition, labels, latency

Established Practice

- One dataset is not enough $\Rightarrow$ collections
- Performance averaged over all datasets
- Application context ignored

Symptoms of Aggregated Rankings over Heterogeneous Datasets

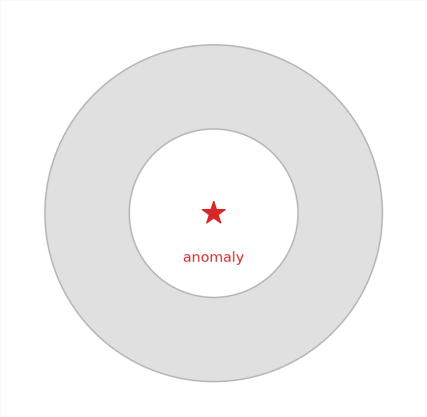
- New detectors keep appearing, yet established baselines (kNN, Isolation Forest) stay competitive Bouman et al., 2024
- Across ADBench's 57 datasets, no method is statistically superior Han et al., 2022

Our position: Benchmark results should be reported per *scenario*, not aggregated over heterogeneous datasets.

A Two-Line Baseline Ranks Second

- *Quantiles*: Score each point by the fraction of its features outside the 5%–95% quantile range
- On ADBench (57 datasets): average rank 2nd, statistically indistinguishable from state-of-the-art detectors Han et al., 2022

But fails on simple examples:



A good aggregate rank over heterogeneous datasets does not indicate a detector's performance for specific applications.

Diagnosis: Aggregation Without Stated Requirements

- Averaging ranks mixes applications with incompatible requirements: within-group trends can vanish or even reverse (Simpson's paradox)
- ADBench's actionable findings come from experiments controlled per anomaly type: LOF wins for local, kNN for global anomalies Han et al., 2022

Consequence: Evaluate algorithms within scenarios that make their *structural properties* explicit.

From Applications to Scenarios

Contextual properties P_{ctx}

- Domain and goal
- e.g. patient monitoring in intensive care

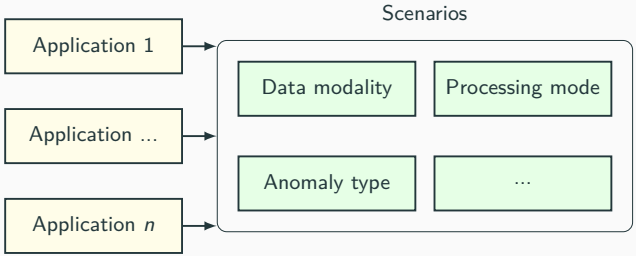
$\xrightarrow{\text{experts}}$

Structural properties P_{str}

- Anomaly type, modality, supervision, ...
- e.g. real-time processing, high recall

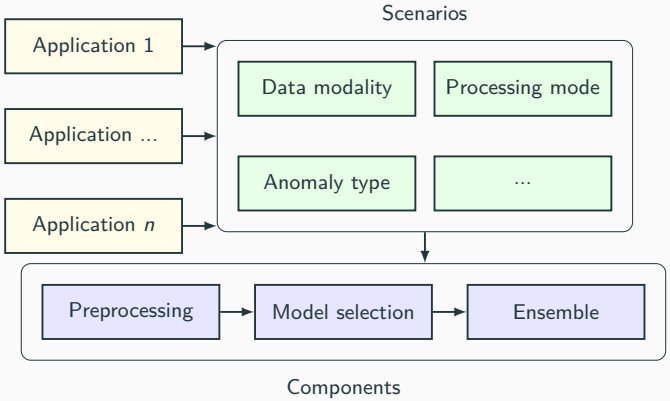
Scenario: Applications sharing the same structural properties regardless of their context.

From Scenarios to Components and Evaluation



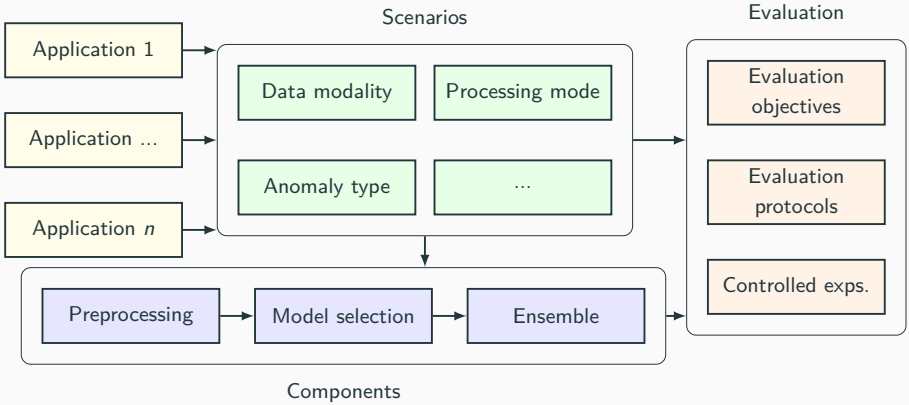
Applications sharing structural properties form a scenario.

From Scenarios to Components and Evaluation



The scenario determines the components of the detection pipeline

From Scenarios to Components and Evaluation



... and the evaluation: objectives, protocols, theory.

Call to Action

1. **Define**: develop the taxonomy and derive scenarios
2. **Curate**: scenario-specific datasets with standardized metadata

Report benchmark results per scenario,
not aggregated over heterogeneous datasets.

Contribute scenarios, applications, and datasets:
go.c1a.eu/benchad