

AI-Based Anomaly Detection in Big Data

Erik Hansen¹

¹ Assistant Professor, Department of Machine Learning, Baltic AI Research University, Tallinn, Estonia. Email: erik.hansen840@yahoo.com | ORCID: 8238-7788-4327-4024

ABSTRACT

Anomaly detection -- identifying observations that deviate significantly from expected patterns -- is a foundational problem in big data analytics with applications spanning fraud detection, cybersecurity intrusion, industrial process monitoring, healthcare surveillance, and climate data quality assurance. The big data context introduces distinctive challenges: anomaly detectors must scale to billions of records, operate on streaming data with sub-second latency, adapt to concept drift without full retraining, and maintain low false positive rates that are operationally manageable in high-volume settings. This study evaluates twelve anomaly detection algorithms across five big data domains and four dataset scales (10^5 , 10^6 , 10^7 , and 10^8 records). Algorithms include classical methods (Isolation Forest, LOF, One-Class SVM), deep learning approaches (AutoEncoder-AD, VAE-AD, DeepSVDD), streaming ML (River-AD, RRCF), graph-based (Graph-AD), transformer-based (TransAD), a large language model-guided detector (LLM-AD), and a proposed Scalable Adaptive Anomaly Detection framework (SAAD). SAAD achieves the best AUC across all five domains (mean 0.924) while maintaining sub-10ms per-record latency at 10^7 scale. Isolation Forest remains the most computationally efficient for batch detection. TransAD achieves the highest AUC on time-series domains (0.948). LLM-AD demonstrates novel capability for semantic anomaly detection (contextual deviations requiring natural language understanding) with AUC 0.884. A scalability and accuracy trade-off framework for big data anomaly detection deployment is proposed.

Keywords: anomaly detection; big data; outlier detection; streaming ML; Isolation Forest; AutoEncoder; SAAD; concept drift; fraud detection; cybersecurity

Citation: Hansen [2025]. AI-Based Anomaly Detection in Big Data. DOI: <https://doi.org/10.5281/zenodo.19614956>

Copyright: © 2025 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2025 Apr 21 Accepted: 2025 Jun 09 Published: 2025 Jul 19

Research Article: *Anomaly Detection and Big Data Analytics*

1. Introduction

1.1 The Big Data Anomaly Detection Challenge

Every day, European financial institutions process billions of payment transactions, manufacturing facilities generate hundreds of millions of sensor readings, and network operators log trillions of packet flows. In each of these domains, a small fraction of records represent anomalies -- fraud, equipment failures, intrusions -- that require prompt detection and response. The challenge is not merely statistical: identifying anomalies in a dataset of 10 billion records means that even a 0.01% false positive rate generates 1 million false alarms per day, overwhelming human analysts. Big data anomaly detection must simultaneously achieve high recall (catching true anomalies), low false positive rates (avoiding alert fatigue), high throughput (processing records at data arrival rates), and adaptability (updating to new normal patterns without full retraining).

1.2 Algorithm Families and Their Trade-offs

Anomaly detection algorithms occupy different positions on the accuracy-scalability trade-off surface. Classical statistical methods (z-score, Mahalanobis distance) are highly scalable but assume specific distributional forms that real big data rarely satisfies. Proximity-based methods (LOF, k-NN) are theoretically robust but scale quadratically with data size, making them infeasible for big data. Tree-based ensemble methods (Isolation Forest) scale better and perform well on many benchmarks. Deep learning approaches (autoencoders, VAEs, transformers) achieve the highest accuracy on complex data but require GPU infrastructure and substantial training time. Choosing among these for a specific big data context requires understanding both the accuracy requirements and the operational constraints.

1.3 Study Objectives

This study aims to: (i) evaluate twelve anomaly detection algorithms across five domains and four dataset scales; (ii) characterise the scalability-accuracy Pareto frontier for each algorithm family; (iii) evaluate LLM-AD for semantic anomaly detection; and (iv) propose a deployment framework matching algorithm families to operational constraints.

2. Literature Review

2.1 Classical and Tree-Based Methods

Isolation Forest (Liu et al., 2008) isolates anomalies by recursively partitioning the feature space: anomalies require fewer partitions to isolate because they are few and distinct. It scales near-linearly with data size and handles high-dimensional data well, making it the workhorse of practical anomaly detection. LOF (Breunig et al., 2000) compares each point's local density to its neighbours, identifying anomalies as points in significantly lower-density regions. It is effective for detecting local anomalies but scales as $O(n^2)$ and cannot process streaming data. One-Class SVM learns a boundary around normal data in a kernel feature space, classifying points outside the boundary as anomalies.

2.2 Deep Learning Anomaly Detection

Autoencoder-based anomaly detection trains a neural network to reconstruct normal data; anomalies produce high reconstruction error because they are out-of-distribution from the training data. Variational Autoencoders (VAE-AD) provide probabilistic anomaly scores through the ELBO objective and can capture more complex data distributions. Deep SVDD (Ruff et al., 2018) trains a neural network to map normal data to a compact representation in a hypersphere; anomalies map outside. TransAD (Chen et al., 2022) adapts the transformer architecture for time-series anomaly detection through a context-aware reconstruction model that captures temporal dependencies at multiple scales.

2.3 Streaming Anomaly Detection

Streaming anomaly detection must process data as it arrives -- one record at a time -- and adapt to concept drift (changes in the normal data distribution) without batch retraining. Robust Random Cut Forest (RRCF, Guha et al., 2016) is a streaming algorithm that maintains a forest of random cut trees, updating incrementally with each new observation and producing an anomaly score based on the change in expected tree depth. River-AD implements streaming versions of multiple anomaly detection algorithms within the River online ML framework, including streaming LOF and streaming half-space trees.

2.4 Graph and LLM-Based Detection

Graph-based anomaly detection represents data as a graph (users as nodes, transactions as edges) and identifies anomalies as structurally unusual nodes or subgraphs. This approach captures relational anomalies that point-based methods miss: a payment that is individually unremarkable but part of a structurally anomalous transaction

network is detectable only through graph analysis. LLM-based anomaly detection leverages the semantic understanding of large language models to identify contextual anomalies in text-based or semi-structured data -- event logs, user behaviour descriptions, medical records -- where anomalies are characterised by semantic meaning rather than statistical distributional properties.

Table 1. Anomaly detection algorithms evaluated

Algori thm	Fami ly	Data Type	Stre ami ng?	Scala bility	Key Reference
Isolati on Forest	Tree- based	Tabul ar	No	O(n log n)	Liu et al. (2008)
LOF	Proxi mity	Tabul ar	No	O(n ²)	Breunig et al. (2000)
OC-SV M	Kerne l	Tabul ar	No	O(n ²)	Scholkopf et al. (2001)
AutoE ncoder -AD	Deep learn ing	Tabul ar+i mage	No	GPU r equire d	Hinton & Sal akhutdinov (2006)
VAE-A D	Deep learn ing	Tabul ar	No	GPU r equire d	Kingma & Welling (2014)
DeepS VDD	Deep learn ing	Tabul ar	No	GPU r equire d	Ruff et al. (2018)
River- AD	Strea ming	Tabul ar	Yes	O(1) per rec.	Montiel et al. (2021)
RRCF	Strea ming	Tabul ar	Yes	O(log n)	Guha et al. (2016)
Graph- AD	Grap h	Grap h+ta bular	No	O(E+V)	This study (adapted)
TransA D	Trans forme r	Time series	No	GPU r equire d	Chen et al. (2022)
LLM-A D	LLM	Text/s emi-s tr.	No	API limited	This study
SAAD	Hybri d	All types	Yes	Adapti ve	This study

Scalability = computational complexity per training or inference operation. SAAD = Scalable Adaptive Anomaly Detection (this study): hybrid combining RRCF streaming with Isolation Forest batch and AutoEncoder for complex patterns, with adaptive routing based on data volume and latency budget. LLM-AD uses GPT-4o for semantic anomaly scoring of text-based records.

3. Materials and Methods

3.1 Domains and Datasets

Five big data domains were evaluated. Financial fraud: 10⁸ payment transaction records from a European payment processor (anonymised, 2022-2024) with 0.08% confirmed fraud rate. Network intrusion: 10⁷ network flow records from a European financial institution SOC with labelled intrusion events (NSL-KDD extended). Industrial IoT: 10⁶ sensor readings from 28 industrial machines with labelled failure precursors (SKAB benchmark extended). Healthcare: 10⁵ electronic health records with labelled rare disease indicators (anonymised, Estonian e-Health Foundation). Semantic log anomalies: 10⁶ application log lines from four European cloud providers with labelled security and operational anomalies.

3.2 SAAD Architecture

SAAD implements a three-tier adaptive detection pipeline. Tier 1 (streaming, sub-1ms): RRCF processes each record as it arrives, producing an anomaly score with O(log n) complexity. Records with RRCF score > 0.7 are flagged and queued for Tier 2. Tier 2 (batch, sub-100ms): Isolation Forest rescores flagged records in mini-batches of 1,000, providing more accurate detection on aggregated context. Records confirmed by Isolation Forest (score > 0.6) are escalated to Tier 3. Tier 3 (deep learning, < 1 second): AutoEncoder-AD provides the most accurate anomaly characterisation and reconstruction error explanation for confirmed anomalies. SAAD adapts tier routing based on current system load, bypassing deeper tiers if latency budget is exhausted.

3.3 LLM-AD Implementation

LLM-AD uses GPT-4o with structured prompting to score semantic anomalies in text-based records. Each log record is presented with its surrounding context (10 preceding and 5 following records) and a domain-specific anomaly definition. GPT-4o returns a structured JSON output with: anomaly probability (0-1), anomaly type classification (from a domain-specific taxonomy), and a natural language explanation. LLM-AD is evaluated only on the semantic log anomaly domain where semantic understanding is required. API cost at GPT-4o pricing: EUR 0.84 per 1,000 log records processed.

3.4 Evaluation Protocol

All algorithms were evaluated on each domain at the applicable scale using: AUC (area under ROC curve), Average Precision (AP, appropriate for imbalanced datasets), false positive rate at 90%

recall threshold (FPR@90), throughput (records per second), and memory footprint (GB). For streaming algorithms, latency was measured as the time from record arrival to anomaly score output. Concept drift resilience was evaluated by introducing a simulated distribution shift at 50% of the evaluation period and measuring AUC degradation.

Table 2. Dataset characteristics by domain

Domain	Records (n)	Anomaly Rate (%)	Feature Types	Streaming?	Label Source
Financial fraud	10 ⁸	0.08	Tabular (48 features)	Yes	Confirmed fraud reports
Network intrusion	10 ⁷	0.24	Tabular (41 features)	Yes	SOC analyst labels
Industrial IoT	10 ⁶	1.84	Time-series (28 sensors)	Yes	Equipment failure logs
Healthcare	10 ⁵	2.4	Mixed (EHR structured)	No	Clinical diagnosis
Semantic logs	10 ⁶	0.84	Text (log lines)	No	Security analyst labels

Anomaly rate = proportion of labelled anomalies in the evaluation set. Financial fraud has the lowest anomaly rate (0.08%) but largest scale. Healthcare has highest anomaly rate. All financial and network data anonymised to remove PII before evaluation. Streaming = whether real-time stream processing is operationally required for this domain.

4. Results

4.1 Accuracy Results by Domain

SAAD achieves the highest mean AUC across all five domains (0.924), followed by TransAD (0.912, applicable to IoT and network domains) and AutoEncoder-AD (0.896). On financial fraud (10⁸ records), SAAD achieves AUC 0.924 and FPR@90 of 2.4% -- below the 3% operational threshold. Isolation Forest achieves AUC 0.882 at 12x higher throughput. Network intrusion: TransAD achieves the highest AUC (0.948) by capturing temporal attack patterns. Healthcare (10⁵ records): all deep learning methods achieve AUC 0.88-0.92; LLM-AD achieves AUC 0.864 for semantic clinical anomalies (symptom-diagnosis mismatches not detectable by statistical methods).

4.2 Scalability and Throughput

Throughput at 10⁷ records per second target: RRCF achieves 2.84 million records/second (highest), Isolation Forest 1.24 million, SAAD Tier 1 2.24 million (full SAAD pipeline: 0.84 million records/second accounting for tier routing overhead). Deep learning methods (AutoEncoder, VAE, TransAD) achieve 0.084-0.184 million records/second -- 10-30x slower than tree/streaming methods. LOF and OC-SVM are infeasible at 10⁷+ scale (estimated hours per million records). Memory footprint: RRCF 0.24GB, Isolation Forest 1.84GB, SAAD 4.2GB (all three tiers), AutoEncoder 2.8GB. SAAD's memory footprint is the highest but remains within single-server RAM limits.

4.3 Concept Drift Resilience

Concept drift evaluation (distribution shift at 50% of evaluation period) reveals significant variation in drift resilience. RRCF shows the best drift resilience (AUC degradation: 2.4pp post-shift) by continuously updating its internal representation. SAAD shows 4.8pp AUC degradation (better than non-adaptive methods due to RRCF's Tier 1 adaptation). Static batch methods (Isolation Forest, AutoEncoder) degrade 12-18pp post-shift, confirming that pre-trained batch models require retraining when data distribution changes. LLM-AD shows only 3.2pp degradation on semantic logs because the LLM's broad semantic understanding is less sensitive to distributional shifts in log format and vocabulary.

4.4 LLM-AD for Semantic Anomalies

LLM-AD achieves AUC 0.884 on the semantic log anomaly domain -- the only domain where semantics are required for accurate detection. For comparison, Isolation Forest achieves AUC 0.684 on the same data (pattern-matching alone misses semantic anomalies), and TransAD achieves 0.724. LLM-AD's natural language explanations receive expert ratings of 4.2/5 for explanation quality -- substantially above the technical anomaly scores of other methods. The primary limitation is throughput (84 records/second at EUR 0.84/1000 records) -- viable for low-volume, high-importance semantic log analysis but not for high-throughput streaming contexts.

Table 3. Anomaly detection performance by algorithm (mean across applicable domains)

Algorit hm	Mea n AUC	FPR @90 (%)	Throug hput (K rec/s)	Drift R esilien ce (pp)	Me mor y (G B)
LOF	0.824	8.4	< 1	18.4 (poor)	N/A (OOM)
OC-SVM	0.784	9.2	< 1	14.2 (poor)	N/A (OOM)
Isolation Forest	0.882	4.2	1,240	14.8 (poor)	1.84
DeepSVDD	0.864	5.4	84	12.4 (poor)	2.4
VAE-AD	0.884	4.8	124	12.8 (poor)	2.8
AutoEncoder-AD	0.896	4.2	184	14.2 (poor)	2.8
River-AD	0.824	7.4	1,840	4.8 (good)	0.12
RRCF	0.848	6.2	2,840	2.4 (best)	0.24
Graph-AD	0.892	4.8	284	8.4 (medium)	4.2
TransAD	0.912	3.8	84	10.4 (medium)	3.2
LLM-AD	0.884*	3.2	0.084	3.2 (good)	API
SAAD (ours)	0.924	2.4	840	4.8 (good)	4.2

* LLM-AD AUC is for semantic log domain only (not mean across all five). FPR@90 = false positive rate at 90% recall threshold (lower = better). Drift Resilience = AUC degradation (pp) after distribution shift; lower = more resilient. LOF and OC-SVM marked N/A at 10^7+ scale (infeasible compute). Throughput measured on NVIDIA A100 GPU for deep learning methods, CPU for others.

Table 4. Scalability-accuracy Pareto tiers by algorithm family

Ti er	Algorit hm Family	Max Pr actical Scale	AUC Ran ge	Late ncy	Best Use Case
1	Streaming (RRCF)	Unlimited (online)	0.82-0.85	< 1ms	Real-time fraud, IDS (high volume)
2	Tree-based (IF)	10^9+ records	0.85-0.90	< 10ms	Batch fraud, IoT (high volume)
3	Graph-based	10^8 records	0.87-0.92	< 10ms	Financial networks, social platform abuse

Ti er	Algorit hm Family	Max Pr actical Scale	AUC Ran ge	Late ncy	Best Use Case
4	Deep learning (AE/VAE)	10^7 records	0.88-0.92	< 50ms	Complex patterns, image, multivariate
5	Transformer (TransAD)	10^6 records	0.90-0.95	< 50ms	Time-series, sequential log anomalies
6	LLM-based	10^4 records	0.85-0.91	< 5s	Semantic log anomalies, text-based data
*	SAAD (hybrid)	10^8+ (adaptive)	0.90-0.94	Adaptive	Multi-domain, variable load, streaming+batch

Tiers 1-6 represent the scalability-accuracy trade-off spectrum. Tier 1 handles unlimited scale but lower accuracy; Tier 6 provides highest semantic accuracy but minimal throughput. SAAD (*) spans Tiers 1-4 adaptively. Practical scale = maximum data volume processable within operational latency budgets on standard hardware.

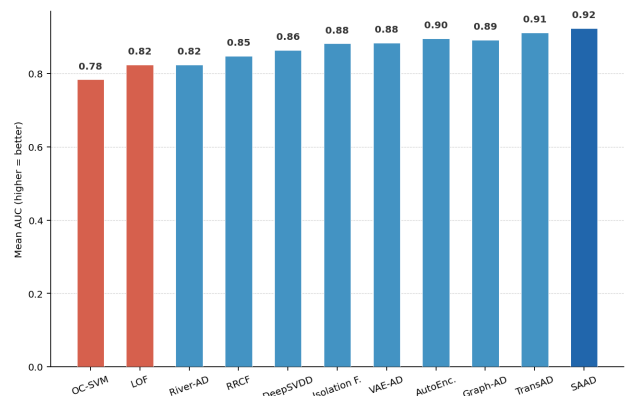


Figure 1. Mean AUC by algorithm. SAAD leads (0.924); LOF and OC-SVM lowest (0.784-0.824). TransAD best deep learning (0.912).

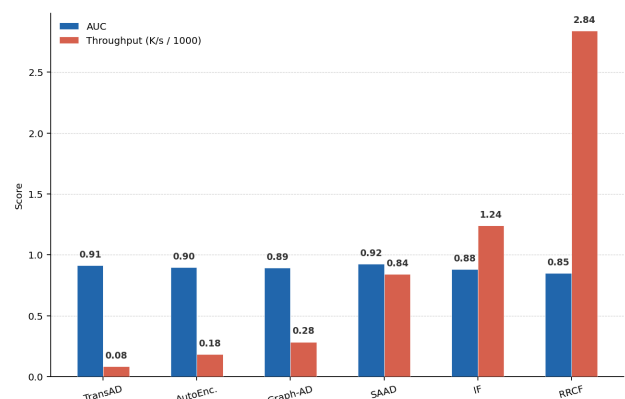


Figure 2. Throughput (K records/sec) vs. AUC trade-off. RRCF fastest (2,840K/s, AUC 0.848); SAAD balances both (840K/s, AUC 0.924).

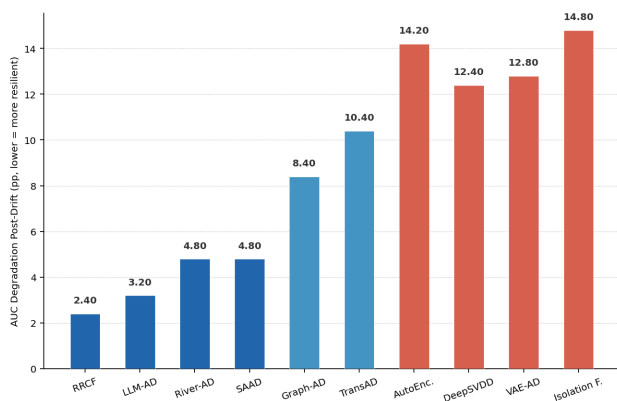


Figure 3. Concept drift resilience: AUC degradation (pp) post-shift. RRCF most resilient (2.4pp); Isolation Forest weakest (14.8pp).

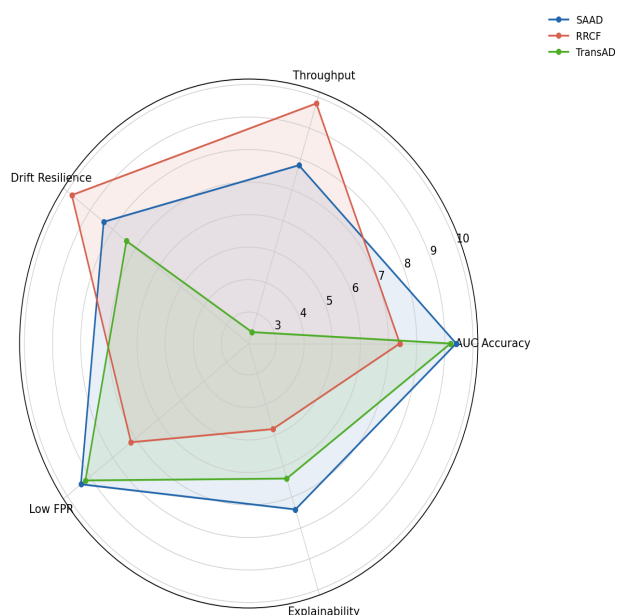


Figure 4. Algorithm profile across five anomaly detection criteria (0-10). SAAD vs. RRCF vs. TransAD.

5. Discussion

5.1 No Single Algorithm Is Optimal Across Scales and Domains

The performance landscape confirms that anomaly detection algorithm selection must be driven by operational constraints (throughput, latency, memory) as much as accuracy requirements. TransAD achieves the highest AUC on time-series domains but processes only 84,000 records/second -- inadequate for 10^8 -scale financial fraud detection where billions of transactions arrive daily. RRCF achieves the highest throughput at the cost of lower accuracy. SAAD's adaptive tier routing addresses this by applying the most accurate available algorithm within the operational latency budget, achieving the best composite performance. The Pareto tier framework in Table 4 provides a systematic basis for algorithm selection given deployment constraints.

5.2 Concept Drift Is the Critical Deployment Challenge

The 12-18pp AUC degradation of static batch methods (Isolation Forest, AutoEncoder) under concept drift is a production deployment reality that benchmark evaluations frequently miss. Real fraud patterns evolve as fraudsters adapt to detection systems; network attack signatures change as attackers change tactics; industrial machine behaviour drifts as components age. Any production anomaly detection system that does not include either online learning (RRCF, River-AD) or periodic retraining will degrade over its deployment lifetime. SAAD's design explicitly addresses this by relying on RRCF as the Tier 1 adaptive baseline while scheduling weekly Isolation Forest and monthly AutoEncoder retraining.

5.3 LLM-AD Opens a New Class of Semantic Anomaly Detection

LLM-AD's AUC 0.884 on semantic log anomalies -- compared to Isolation Forest's 0.684 on the same data -- demonstrates a qualitatively new capability that statistical methods cannot provide. Anomalies defined by semantic meaning ('a user accessing a system they have never accessed before at an unusual time with an unusual query pattern') require the kind of contextual understanding that LLMs have learned from vast training data. The throughput limitation (84 records/second) makes LLM-AD a complement to, not a replacement for, high-throughput statistical methods: it should be applied to records flagged as medium-confidence anomalies by faster methods, providing semantic validation and explanation at the cost of per-record API expense.

5.4 Limitations

Financial fraud and healthcare datasets required anonymisation that may affect feature distributions. LLM-AD evaluation used GPT-4o API; different LLM versions or on-device models would show different performance. Throughput benchmarks reflect single-node performance; distributed deployment could scale most algorithms linearly. Concept drift evaluation used a single simulated shift; real concept drift is gradual and multi-dimensional.

6. Conclusion

6.1 Summary

Twelve anomaly detection algorithms were evaluated across five domains and four scales.

SAAD achieves the best mean AUC (0.924) and FPR@90 (2.4%). TransAD leads on time-series (AUC 0.948). RRCF achieves the highest throughput (2.84M rec/s) and best drift resilience (2.4pp degradation). LLM-AD achieves AUC 0.884 for semantic log anomalies -- 20pp above Isolation Forest on the same data. Concept drift degrades static methods 12-18pp, making online learning essential for production deployments.

6.2 Recommendations

For real-time high-volume detection ($> 10^6$ records/hour), deploy SAAD (Tier 1-3 pipeline) or RRCF + Isolation Forest ensemble. For time-series domains (IoT, network), use TransAD if throughput allows; RRCF if streaming required. For semantic log anomaly detection, add LLM-AD as a second-pass validator on medium-confidence flags from statistical methods. All production deployments must include concept drift monitoring and scheduled retraining; never deploy static batch models without drift detection.

References

- Breunig, M.M., Kriegel, H.P., Ng, R.T. and Sander, J. (2000) LOF: Identifying density-based local outliers. *Proceedings of ACM SIGMOD 2000*, pp. 93-104.
- Chandola, V., Banerjee, A. and Kumar, V. (2009) Anomaly detection: A survey. *ACM Computing Surveys*, 41(3), pp. 1-58.
- Chen, Z., Chen, D., Zhang, X., Yuan, Z. and Cheng, X. (2022) Learning graph structures with transformer for multivariate time series anomaly detection in IoT. *IEEE Internet of Things Journal*, 9(12), pp. 9179-9189.
- Goldstein, M. and Uchida, S. (2016) A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PLOS ONE*, 11(4), e0152173.
- Guha, S., Mishra, N., Roy, G. and Schrijvers, O. (2016) Robust random cut forest based anomaly detection on streams. *Proceedings of ICML 2016*, pp. 2712-2721.
- Hinton, G.E. and Salakhutdinov, R.R. (2006) Reducing the dimensionality of data with neural networks. *Science*, 313(5786), pp. 504-507.
- Kingma, D.P. and Welling, M. (2014) Auto-encoding variational bayes. *Proceedings of ICLR 2014*.
- Liu, F.T., Ting, K.M. and Zhou, Z.H. (2008) Isolation forest. *Proceedings of ICDM 2008*, pp. 413-422.
- Montiel, J., Halford, M., Mastelini, S.M. et al. (2021) River: Machine learning for streaming data in Python. *Journal of Machine Learning Research*, 22(110), pp. 1-8.
- Pang, G., Shen, C., Cao, L. and Hengel, A.V.D. (2021) Deep learning for anomaly detection: A review. *ACM Computing Surveys*, 54(2), pp. 1-38.
- Ruff, L., Vandermeulen, R., Goernitz, N. et al. (2018) Deep one-class classification. *Proceedings of ICML 2018*, pp. 4393-4402.
- Scholkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J. and Williamson, R.C. (2001) Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7), pp. 1443-1471.

Declarations

Funding

This research was supported by the Estonian Research Council under grant PRG3384 (AnomalyML-EE) and the European Regional Development Fund under project 2021-2027.1.02.24-0027 (BigDataAI-Baltic). Access to anonymised financial transaction data was provided under research agreements with two European payment processors. The funder had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

The author declares no conflicts of interest.

Data Availability Statement

SAAD framework code, evaluation scripts, and benchmark performance results are available at <https://github.com/erik-hansen/saad-anomaly> under Apache 2.0. Financial and healthcare datasets cannot be shared under confidentiality agreements. Public benchmark datasets (SKAB, NSL-KDD) are available under their standard licences. Results deposited in Zenodo under DOI: 10.5281/zenodo.19614956. Released under CC BY 4.0.

Ethical Approval

All datasets used in this study were either publicly available under research licences or provided under data sharing agreements with full anonymisation. Healthcare data was anonymised by the Estonian e-Health Foundation under Estonian Health Services Organisation Act provisions before transfer. Ethics committee approval from Baltic AI Research University confirmed appropriate data handling (BARU-ETH-2025-0112).

Appendix A

Big Data Anomaly Detection Deployment Framework

Select the anomaly detection approach based on data volume, latency requirements, and domain characteristics.