Systems & Control Transactions

Original Research Article
Peer Reviewed Conference Proceeding
ESCAPE 36 - European Symposium on Computer Aided Process Engineering
Sheffield, UK. 21-24 June 2026

PSE PRESS

Beyond Tennessee Eastman: Benchmarking Deep Anomaly Detection on Real-World Pilot-Scale Continuous Distillation Data

Fabian Hartung[ac†], Aparna Muraleedharan[b†], Marius Kloft[a] and Jakob Burger[b*]

[a] RPTU Kaiserslautern, Department of Machine Learning, Kaiserslautern, Germany
[b] Technical University of Munich, Laboratory for Chemical Process Engineering, Straubing, Germany
[c] BASF, Gas Treatment, Monheim am Rhein, Germany
* Corresponding Author: burger@tum.de.

ABSTRACT

Anomaly detection is essential for ensuring the safe and efficient operation of chemical plants. Although many deep-learning-based methods have been proposed in recent years, their evaluation remains largely limited to synthetic benchmarks such as the Tennessee Eastman Process (TEP) [1]. While these simulators enable controlled and reproducible comparisons, they fail to capture the noise characteristics, operational complexity, and irregular fault dynamics of real industrial plants, leaving the practical generalizability of many methods unclear. In this work, we extend our earlier ESCAPE study [2] beyond water-based systems to industrially relevant chemical processes. We analyze multivariate time-series data from two continuously operated pilot-plant scenarios at the Technical University of Munich, namely *n*-butanol/water heteroazeotropic distillation and poly(oxymethylene) ether purification, whose datasets were recently published at NeurIPS 2025 [3]. Using the open-source TimeSeAD library [4], we benchmark 30 anomaly detection methods, including 26 deep-learning-based and 4 classical approaches, under a unified preprocessing, model-selection, and evaluation pipeline. Performance is assessed using the F1-score and the area under the precision–recall curve (AUPRC). Our results show a substantial performance drop when moving from synthetic to real process data, with average scores far below those commonly reported for TEP. No single method performs consistently best across all datasets, and the ranking depends strongly on the chosen metric and process scenario. These findings highlight the limitations of synthetic benchmarks and underscore the need for more realistic industrial datasets, process-aware methods, and evaluation practices that better reflect real operating conditions.

Keywords: Machine Learning, Anomaly Detection, Continuous Distillation, Pilot Plant Data, Heteroazeotropic distillation, Tennessee Eastman Process Data

1. INTRODUCTION

Continuous distillation is the most widely used separation technique in the chemical industry, enabling steady-state operation through continuous feed input and simultaneous product withdrawal [5, 6]. Anomaly detection (AD) is therefore critical, as accidents in chemical plants—although rare—can result from leaks, fires, or explosions and may harm personnel, equipment, or the environment. Over the past two decades, numerous data-driven AD techniques have been developed, ranging from classical multivariate statistical to principal component analysis, and, more recently, unsupervised machine-learning-based approaches [7, 8].

Most existing AD methods have been developed and evaluated on simulated benchmarks such as the Tennessee Eastman Process (TEP) [1]. Despite their strong reported performance, it remains unclear to what extent models trained and validated on simulated data generalize to real industrial processes operating under

† These authors contributed equally to this work.

non-ideal and evolving conditions. Cheng et al. [7] report that only 12% of the processing methods in their analysis have been applied to real industrial data, primarily due to data confidentiality and limited public availability. This highlights a key challenge: while simulated data represent idealized scenarios with a limited set of fault cases, real industrial data are far more complex, influenced by factors such as minor operational disturbances, equipment aging, and complex chemical mixtures. Industrial process data are typically proprietary, as they contain sensitive information on instrumentation, operating conditions, process variables, and control strategies, and are therefore rarely accessible in the public domain.

To address this gap, we recently published process data from three scenarios of a continuous distillation mini-plant at the TUM Campus in Straubing [3]. This publicly available dataset [9] is used to train and evaluate AD methods. By benchmarking a broad range of anomaly detection approaches from the literature on our data, this work provides a more realistic assessment of model performance under practical industrial operating conditions.

2. RELATED WORK

The detection of anomalous behavior in industrial systems has a long tradition in process engineering. Early work framed fault diagnosis as a multivariate statistical process control problem, relying on dimension-reduction techniques such as PCA, PLS, or SIMCA to construct control charts that identify deviations from regular operation [10, 11]. Subsequent approaches explored probabilistic models, including Gaussian mixture models [12], and kernel-based one-class methods [13]. While these “shallow” models have proven effective at small to medium scales, their ability to capture the complex, nonlinear dynamics of chemical plants is inherently limited. As a result, recent advances in deep learning have been transferred to the process domain. Ruff et al. [14] introduced deep one-class classification, followed by numerous studies proposing autoencoder-, GAN-, and VAE-based architectures tailored to multivariate time series from simulators such as the TEP [15, 16].

Comprehensive surveys [17, 18] categorize AD techniques into probabilistic, distance-based, domain-based, reconstruction-based, forecasting-based, and classification-based approaches. Our previous work presented the first large-scale comparison of 27 deep AD algorithms on the TEP and demonstrated the strong performance of reconstruction-based models [15]. While these results are promising, it remains unclear whether models that perform well on synthetic benchmarks generalize to real industrial processes. The present paper addresses this gap by evaluating the same methodological spectrum on three newly released real plant datasets [9]. In practice, anomalies in real plants rarely manifest as abrupt, well-defined events; instead, they often evolve gradually, resemble normal process variability, or are partially masked by control actions.

A key challenge in AD benchmarking is choosing evaluation protocols that account for the temporal nature of anomalies. Point-wise metrics such as the F1-score and AUPRC are still most commonly reported, as they summarize precision–recall trade-offs in a single value [19, 20]. However, they ignore event duration and detection latency. Consequently, different metrics can yield substantially different conclusions about model performance. To mitigate this issue, the literature has proposed window-based scores (NAB, NAB-modified) [21], time-to-detect measures [22], and cost-sensitive utility functions [23]. In this work, we report both F1 and AUPRC in time-series-adapted variants.

The continued popularity of the TEP is driven by its realistic flowsheet, stochastic disturbances, and public availability. However, several limitations warrant caution: (a) faults are introduced as persistent step changes over entire test runs, amplifying anomaly signals and biasing results toward high recall; (b) only a single operating point is represented, whereas real plants undergo grade changes and load variations; (c) the simplified control structure lacks advanced layers such as MPC or APC; and (d) hyperparameter optimization on a small set of standard faults increases the risk of overfitting and score inflation. Recent studies have shown that even trivial baselines, such as an untrained LSTM autoencoder, can achieve F1-scores above 0.93 on TEP [24]. These limitations suggest that results obtained on synthetic benchmarks should be interpreted with care when assessing real-world applicability, underscoring the need for more rigorous and realistic benchmarks—an objective addressed by the datasets used in this study.

3. DATASETS

We use recently published steady-state multivariate time-series data [9] generated on a pilot-scale continuous distillation mini-plant with a capacity of 5 t/a located at the TUM Campus Straubing. The setup comprises two distillation columns and a decanter, instrumented with 34 sensors and actuators. Three operating scenarios of increasing industrial relevance are considered: water runs, a heteroazeotropic *n*-butanol/water system, and a reactive poly(oxymethylene) ether (OME) system. Water runs were performed on a single column, yielding approximately 520 h of steady-state operation. Experiments with the *n*-butanol/water system used two columns and a decanter, with campaigns averaging 10 h and 80 h of steady-state data. The OME system was operated in a single-column configuration (30h of data).

All process variables were recorded at a 30s sampling interval, sufficient to capture relevant steady-state

dynamics. Measurements include column and reboiler temperatures, mass flow rates, column head pressure, pump power, and differential pressure signals. A detailed description of the plant setup, process, and instrumentation diagram, sensors and actuators, control strategies, data preparation, and occurring anomalies is provided in the related work [9]. Startup and shutdown phases were removed prior to analysis, leaving only steady-state operation. During routine operation, a range of typical industrial anomalies was observed, including clogging, reboiler level instabilities, overheating, instrument malfunctions, pressure control deviations, and human operating errors. When specific anomaly types did not occur naturally, they were deliberately induced under controlled conditions to ensure a representative yet realistic range of fault behaviours, cf. [9] for details.

Anomalies were labeled following the classification commonly used in anomaly detection studies [15]. For each anomaly, metadata describing its type and cause, affected sensors, and timestamps are provided in the related work [9]. We additionally document the time of fault introduction, its first observable manifestation in the data, and the return to regular operation. Detailed dataset statistics are given in [3]; for the continuous water, butanol, and OME datasets, anomalous samples account for 20%/25%, 24%/41%, and 4%/36% of train/test time steps, respectively. Table 1 summarizes the key characteristics distinguishing our pilot-scale datasets from the TEP benchmark, namely: (1) real sensor noise levels 5–20× higher than in simulation, (2) closed-loop control that actively responds to and masks faults, and (3) gradual fault development rather than step changes.

Table 1: Process comparison of TEP data [1] vs data available from the pilot plant [9]

Aspect	TEP	Butanol Data	OME Data
Fault Behavior	Step Changes	Gradual Evolution	Very Gradual
Control Mode	Often open-loop	Closed-loop automatic	Closed-loop automatic
Sensor Noise	Roughly 0.1%	0.5-2%	0.5-2%
Time Constants	Fast	Medium 10-30 min	Slow 30-60 min

The Supporting Information compares faulty and normal runs from TEP and our plant, illustrating that anomalies in our data are more gradual and subtle. Additional examples in [9] further show that behavior appearing anomalous to a neutral observer may in fact correspond to normal and expected plant operation.

4. COMPUTATIONAL EXPERIMENTS

All experiments are conducted using **TimeSeAD**, our open-source benchmark library for time-series anomaly detection [4]. The evaluated methods were benchmarked using their implementations in the open-source TimeSeAD framework and trained by the authors on the considered datasets under a unified preprocessing, hyperparameter-selection, and evaluation protocol. The framework handles data ingestion, stratified splitting, normalization, hyperparameter selection, and metric aggregation, ensuring full reproducibility across datasets and methods. Training data are split into folds for model training and validation across different hyperparameter configurations. To ensure a fair comparison, the total runtime is limited to 24 hours per method. The best-performing configuration is then applied to compute anomaly scores on the test data. For each dataset, sensor channels with constant readings are removed, and the initial transient phase is discarded. The remaining channels are standardized to zero mean and unit variance within each dataset to stabilize optimization [25].

We benchmark 30 state-of-the-art anomaly detection methods, grouped into the following categories:

- **Reconstruction-based (8 methods):** Models are trained to reconstruct normal data accurately; anomalies are identified when reconstruction errors exceed a threshold.
- **Prediction-based (5 methods):** Temporal or probabilistic dynamics are learned to predict future values; large deviations between predictions and observations indicate anomalies.
- **GAN-based generative (3 methods):** A generator and discriminator are trained adversarially; anomalies are poorly reconstructed or classified as unlikely by the discriminator.
- **VAE-based generative (6 methods):** Models learn a latent probabilistic representation while optimizing reconstruction fidelity; anomalies correspond to low likelihood or high reconstruction error.
- **Hybrid (4 methods):** Multiple paradigms are combined to exploit complementary strengths, typically integrating representation learning with explicit anomaly scoring.
- **Shallow baselines (4 methods):** Classical, non-deep approaches based on statistical or simple machine-learning principles, included as interpretable and computationally efficient reference methods.

Evaluation Metrics

Model performance is assessed using the F1-score and the area under the precision–recall curve (AUPRC), both derived from true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN). Precision measures the fraction of correctly detected anomalies among all detections, while recall quantifies the fraction of actual anomalies that are successfully identified.

$$Precision = \frac{TP}{TP+FP} \quad Recall = \frac{TP}{TP+FN}$$

The F1-score is defined as the harmonic mean of precision and recall:

$$F1 = 2.\frac{Precision.Recall}{Precision + Recall}$$

In addition, we report the AUPRC, which balances precision and recall across all detection thresholds and is particularly well suited to the highly imbalanced nature of anomaly detection datasets. However, conventional point-wise definitions of these metrics are ill-suited for time-series data, where anomalies typically occur as contiguous temporal segments rather than isolated points.

To address this, we use the time-series-aware versions of F1 and AUPRC implemented in the TimeSeAD library [26]. These formulations extend the classical TP/FP/FN definitions by evaluating the temporal overlap between predicted and ground-truth anomaly intervals instead of comparing individual time points. Each predicted segment contributes to precision and recall in proportion to its overlap with one or more true anomaly segments. Partial overlaps are weighted by their intersection length and penalized when a predicted segment covers multiple true segments, or vice versa. The resulting time-series-aware precision and recall are obtained by aggregating these weighted overlaps across all predicted and true intervals, and the F1-score and AUPRC are then computed analogously to their classical definitions.

5. RESULTS

Detailed results for all 30 evaluated methods on the three pilot-plant datasets, including F1-score and AUPRC, are provided in the supplementary material (Tables S8–S10). Table S11 reports the mean performance across all three datasets, while Table S12 lists the corresponding results on the Tennessee Eastman Process (TEP) for comparison. To give a compact overview of score distributions, Table 2 summarizes the average, minimum, and maximum F1-scores achieved per dataset.

A clear performance gap emerges when comparing simulated TEP data with real pilot-plant data. While TEP achieves high average scores (F1 = 0.859, AUPRC = 0.906), performance on the pilot-plant datasets is substantially lower across all cases (F1 = 0.532, AUPRC = 0.361). No method exceeds an F1-score of 0.718 or an AUPRC of 0.597, indicating that none of the evaluated approaches approaches the benchmark performance observed on synthetic data.

Despite this overall performance degradation, deep-learning-based methods consistently outperform shallow baselines on the real-world datasets. Multiple deep learning approaches achieve substantially higher F1-scores than the best-performing shallow baseline across all settings, reaffirming the advantage of deep representation learning even under realistic industrial conditions (Table S11 – results for AUPRC in supporting information). Among the deep-learning-based approaches, the strongest performers vary by dataset: LSTM-2S2-P achieves the best F1-score on Butanol and Water, whereas generative VAE methods, particularly LSTM-VAE and LSTM-DVAE, perform best on OME. At the method-type level, hybrid approaches achieve the highest mean F1-score across the three pilot datasets, whereas, for deep methods, reconstruction-based methods rank first in mean AUPRC, highlighting strong dataset- and metric-dependence.

Table 2: Dataset-level F1-score statistics across evaluated AD methods showing average, minimal, and maximal F1-Score, and the span (AUPRC results in appendix)

F1	average	min	max	Δ
Butanol	0.538	0.291	0.624	0.333
OME	0.629	0.215	0.718	0.503
Water	0.428	0.104	0.531	0.427
Average	0.532	0.267	0.590	0.323
TEP	0.859	0.497	0.911	0.414

Table 3: F1-scores of the best-performing deep-learning and shallow methods, and the mean across all three datasets

	n-butanol	OME	water	mean
Best Deep	0.624	**0.718**	0.531	**0.624**
Best Shallow	0.588	0.666	0.433	0.562

6. DISCUSSION

Our earlier benchmark study on the TEP dataset [11–15] showed that most deep anomaly-detection methods achieve near-perfect performance when anomalies appear as persistent fault states. In contrast, our evaluation

of real pilot-plant distillation data reveals a markedly different picture. When anomalies are transient, gradual, or intermittent, as is typical in industrial practice, the same methods achieve at best an F1-score of 0.718. This pronounced performance degradation highlights a key limitation of current approaches: their strong reliance on temporally persistent and clearly separable fault signatures, which are common in synthetic benchmarks but rare in real plants. Importantly, this performance gap cannot be attributed to label uncertainty. In our datasets, the onset and resolution times of anomalies are known, allowing us to rule out annotation noise as a confounding factor. Instead, the observed differences directly reflect model behavior under realistic operating conditions.

Across the three datasets and both evaluation metrics, no single method or method family consistently dominates. However, across all datasets, several deep-learning-based methods outperform all shallow baselines in terms of F1-score, and in two out of three datasets also with respect to AUPRC. While GenAD [27] and AnomalyTransfer [28] achieve comparatively low average F1-scores, they exhibit stronger AUPRC performance on selected datasets. Hybrid approaches show a slight average advantage but also pronounced dataset-specific weaknesses, underscoring the lack of robust generalization across fault types and evaluation metrics.

The root cause of this discrepancy lies in the development of industrial faults. Real process anomalies typically evolve gradually due to fouling, clogging, drift, or degradation rather than abrupt step changes. The fault scenarios in our datasets [3, 9] were explicitly designed to reflect this behavior. Gradual faults are inherently more difficult to detect: early deviations often remain within normal variability, fault signatures evolve slowly, and closed-loop control systems may partially compensate disturbances, thereby masking their effects. Controllers suppress deviations in controlled variables while redistributing the effects of anomalies into manipulated variables, such as flow rates or reboiler duties. Consequently, faults may be physically present but only indirectly observable through control actions, in stark contrast to TEP, where faults are introduced as step disturbances with clear onsets and pronounced responses.

These differences lead to a systematic performance gap, with F1-scores of approximately 0.80–0.90 on TEP compared to 0.55–0.72 on real plant data. Crucially, this gap does not primarily indicate insufficient model complexity or flawed algorithmic design. Instead, it reflects fundamental process-engineering challenges: (i) gradual fault evolution, (ii) active feedback control that masks faults in controlled variables, and (iii) sensor noise levels that are 5–20 times higher than in synthetic benchmarks.

From a deployment perspective, reduced detection performance on real industrial data should therefore be expected rather than interpreted as model failure. Anomalies often become observable first through changes in control actions rather than deviations in process outputs. Effective anomaly-detection systems must explicitly incorporate manipulated variables and control behavior and align detection window sizes with process time scales. For continuous distillation, fault dynamics on the order of 10–60 minutes imply an unavoidable trade-off between detection reliability and alarm latency.

Robust methods must therefore demonstrate stable performance across diverse datasets rather than excelling only on highly structured synthetic benchmarks. Deep-learning-based methods not only outperform shallow approaches under realistic conditions but also offer greater potential for further methodological advances, making them a promising path toward closing the current performance gap between synthetic and real-world benchmarks.

Finally, this work provides the first publicly available pilot-scale distillation benchmark featuring realistic faults, industrial noise characteristics, and closed-loop control interactions. We hope it fosters the development of process-aware anomaly-detection methods that explicitly account for plant physics and control behavior and are validated under realistic industrial conditions, rather than treating chemical plants as generic time-series systems.

7. LIMITATIONS AND FUTURE WORK

While this benchmark focuses on distillation-related process data, several limitations should be acknowledged. Although the datasets and benchmarks presented in this work provide a more realistic assessment of anomaly detection methods than commonly used synthetic benchmarks, they cannot exhaustively represent all anomalies encountered in industrial distillation processes. Rare or safety-critical events, such as severe equipment failures or cascading faults, are inherently difficult or unsafe to reproduce in a pilot-scale plant. Future data collection campaigns could therefore expand the diversity of fault types, include additional separation processes, and capture longer-term degradation phenomena as well as more subtle control-related anomalies. In addition to collecting further real-world data, data augmentation and generation offer substantial potential to mitigate current data limitations [58].

Second, the benchmarking framework relies on the current capabilities of the TimeSeAD library. While TimeSeAD enables reproducible preprocessing, training, and evaluation across a wide range of anomaly detection methods, it does not yet cover all approaches or evaluation protocols proposed in the literature. Future work will extend TimeSeAD with additional models, improved hyperparameter optimization strategies, and alternative evaluation metrics that more explicitly account for

temporal aspects such as detection delay and fault duration. Given the still considerable development potential of deep-learning-based methods, compared to the largely saturated performance of shallow approaches, the integration of transformer-based models, in particular the recently proposed NeuTraL AD method [56], into the library and our benchmark datasets is of high interest.

Third, while F1-score and AUPRC provide complementary perspectives on model performance, they can lead to different rankings and interpretations. To achieve a more comprehensive and task-specific evaluation, future work should investigate additional metrics tailored to anomaly detection, such as the recently proposed ALARM metric [57], which may improve comparability across datasets and operating conditions.

Finally, although the presented pilot-scale and industrial datasets capture realistic process dynamics, noise characteristics, and operational variability, they cannot be assumed to be universally representative of all large-scale industrial systems. Differences in plant design, control strategies, sensor quality, and operating regimes may substantially influence anomaly characteristics and model performance. Consequently, future studies should incorporate data from additional plants and process types to further assess the robustness and generalizability of anomaly detection methods across industrial settings.

8. SUMMARY OF CONTRIBUTIONS

The main contributions of this work are:

- We benchmark 30 anomaly detection methods, including deep-learning-based and classical approaches, on steady-state multivariate time-series data from pilot-scale and industrial continuous distillation processes.
- We show that methods achieving strong performance on synthetic benchmarks, such as the Tennessee Eastman Process, experience a pronounced drop in performance when evaluated on real process data.
- We demonstrate that no single method consistently performs well across datasets and operating conditions, while several deep-learning-based methods outperform shallow baselines under realistic process conditions.
- We highlight the strong dependence of benchmarking outcomes on the choice of evaluation metrics and emphasize the need for time-series–aware evaluation in industrial anomaly detection.

DIGITAL SUPPLEMENTARY MATERIAL

Comparison of TEP data vs data from our pilot plant. Details on the performance of the different AD methods on the Butanol, OME, Water, and TEP datasets. Mean performance of the AD Methods. Supporting Information available at:
https://psecommunity.org/LAPSE:2026.0038

ACKNOWLEDGEMENTS

This work was funded by the German Research Foundation (DFG) and conducted in collaboration with the FOR5359 Research Unit on Deep Learning on Sparse Chemical Process Data.

AUTHOR IDENTIFIERS

Author ORCIDs:
Hartung: 0009-0004-5093-3287
Muraleedharan: 0009-0006-7232-1550
Kloft: 0000-0001-6829-3725
Burger: 0000-0002-9376-9438

REFERENCES

1. Downs JJ, Vogel EF. A plant-wide industrial process control problem. Computers & Chemical Engineering 17:245-255 (1993). https://doi.org/10.1016/0098-1354(93)80018-i
2. Muraleedharan A, et al. Benchmarking deep anomaly detection on real process data of a continuous distillation process. ESCAPE-34 / PSE Conf (2024).
3. Wagner D, et al. NoBOOM: Chemical process datasets for industrial anomaly detection. Proc NeurIPS Datasets Benchmarks Track (2025). https://openreview.net/forum?id=qiLboR0ocm
4. Wagner D, et al. TimeSeAD: Benchmarking deep multivariate time-series anomaly detection. Trans Mach Learn Res (2023). https://openreview.net/forum?id=iMmsCI0JsS
5. Liu J, Ren J, Yang Y, Liu X, Sun L. Effective semicontinuous distillation design for separating normal alkanes via multi-objective optimization and control. Chemical Engineering Research and Design 168:340-356 (2021). https://doi.org/10.1016/j.cherd.2021.02.018
6. Safrit BT, Westerberg AW, Diwekar U, Wahnschafft OM. Extending continuous conventional and extractive distillation feasibility insights to batch distillation. Ind. Eng. Chem. Res. 34:3257-3264 (2002). https://doi.org/10.1021/ie00037a012
7. Ji C, Sun W. A review on data-driven process monitoring methods: characterization and mining of industrial data. Processes 10:335 (2022). https://doi.org/10.3390/pr10020335

8. Siegel B. Industrial anomaly detection: a comparison of unsupervised neural network architectures. IEEE Sens. Lett. 4:1-4 (2020). https://doi.org/10.1109/lsens.2020.3007880
9. Muraleedharan A, et al. Experimental time series data with and without anomalies from a continuous distillation mini-plant for development of machine learning anomaly detection methods. engrXiv (2025). https://doi.org/10.31224/5631
10. Ku W, Storer RH, Georgakis C. Disturbance detection and isolation by dynamic principal component analysis. Chemometrics and Intelligent Laboratory Systems 30:179-196 (1995). https://doi.org/10.1016/0169-7439(95)00076-3
11. Kresta JV, Macgregor JF, Marlin TE. Multivariate statistical monitoring of process operating performance. Can J Chem Eng 69:35-47 (2009). https://doi.org/10.1002/cjce.5450690105
12. Yin S, Ding SX, Xie X, Luo H. A review on basic data-driven approaches for industrial process monitoring. IEEE Trans. Ind. Electron. 61:6418-6428 (2014). https://doi.org/10.1109/tie.2014.2301773
13. Chiang LH, Russell EL, Braatz RD. Fault diagnosis in chemical processes using fisher discriminant analysis, discriminant partial least squares, and principal component analysis. Chemometrics and Intelligent Laboratory Systems 50:243-252 (2000). https://doi.org/10.1016/s0169-7439(99)00061-1
14. Ruff L, et al. Deep one-class classification. Proc Int Conf Mach Learn 80:4393–4402 (2018).
15. Hartung F, Franks BJ, Michels T, Wagner D, Liznerski P, Reithermann S, Fellenz S, Jirasek F, Rudolph M, Neider D, Leitte H, Song C, Kloepper B, Mandt S, Bortz M, Burger J, Hasse H, Kloft M. Deep anomaly detection on tennessee eastman process data. Chemie Ingenieur Technik 95:1077-1082 (2023). https://doi.org/10.1002/cite.202200238
16. Li X, Wang J, Qin SJ. Temporal convolutional autoencoder for fault detection: A case study on the Tennessee Eastman process. Comput Chem Eng 165:107998 (2022).
17. Chandola V, Banerjee A, Kumar V. Anomaly detection. ACM Comput. Surv. 41:1-58 (2009). https://doi.org/10.1145/1541880.1541882
18. Ruff L, Kauffmann JR, Vandermeulen RA, Montavon G, Samek W, Kloft M, Dietterich TG, Muller KR. A unifying review of deep and shallow anomaly detection. Proc. IEEE 109:756-795 (2021). https://doi.org/10.1109/jproc.2021.3052449
19. Tatbul N, et al. Precision and recall for time series. Adv Neural Inf Process Syst 31:1920–1930 (2018).
20. Hanselmann M, Strauss T, Dormann K, Ulmer H. Canet: an unsupervised intrusion detection system for high dimensional CAN bus data. IEEE Access 8:58194-58205 (2020). https://doi.org/10.1109/access.2020.2982544
21. Lavin A, Ahmad S. Evaluating real-time anomaly detection algorithms -- the numenta anomaly benchmark. 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA) :38-44 (2015). https://doi.org/10.1109/icmla.2015.141
22. Görnitz N, et al. Toward supervised anomaly detection. J Artif Intell Res 46:235–262 (2013). https://doi.org/10.1613/jair.3965
23. Siffer A, Fouque PA, Termier A, Largouet C. Anomaly detection in streams with extreme value theory. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining :1067-1075 (2017). https://doi.org/10.1145/3097983.3098144
24. Fährmann D, Damer N, Kirchbuchner F, Kuijper A. Lightweight long short-term memory variational auto-encoder for multivariate time series anomaly detection in industrial control systems. Sensors 22:2886 (2022). https://doi.org/10.3390/s22082886
25. Heaton J. Ian goodfellow, yoshua bengio, and aaron courville: deep learning. Genet Program Evolvable Mach 19:305-307 (2017). https://doi.org/10.1007/s10710-017-9314-z
26. Hua X, Zhu L, Zhang S, Li Z, Wang S, Deng C, Feng J, Zhang Z, Wu W. Genad: general unsupervised anomaly detection using multivariate time series for large?scale wireless base stations. Electronics Letters 59: (2022). https://doi.org/10.1049/ell2.12683
27. Xu J, et al. Anomaly transformer: Time series anomaly detection with association discrepancy. arXiv 2110.02642 (2022).
28. Abdi H, Williams LJ. Principal component analysis. WIREs Computational Stats 2:433-459 (2010). https://doi.org/10.1002/wics.101
29. Kumar A, Tripathi AR, Satapathy SC, Zhang YD. Sars-net: COVID-19 detection from chest x-rays by combining graph convolutional network and convolutional neural network. Pattern Recognition 122:108255 (2022). https://doi.org/10.1016/j.patcog.2021.108255
30. Tatbul N, et al. Precision and recall for time series. Adv Neural Inf Process Syst 31:1920–1930 (2018).
31. Hanselmann M, Strauss T, Dormann K, Ulmer H. Canet: an unsupervised intrusion detection system for high dimensional CAN bus data. IEEE Access 8:58194-58205 (2020). https://doi.org/10.1109/access.2020.2982544
32. Lavin A, Ahmad S. Evaluating real-time anomaly detection algorithms -- the numenta anomaly benchmark. 2015 IEEE 14th International Conference on Machine Learning and Applications

(ICMLA) :38-44 (2015). https://doi.org/10.1109/icmla.2015.141
33. Görnitz N, et al. Toward supervised anomaly detection. J Artif Intell Res 46:235–262 (2013). https://doi.org/10.1613/jair.3965
34. Siffer A, Fouque PA, Termier A, Largouet C. Anomaly detection in streams with extreme value theory. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining :1067-1075 (2017). https://doi.org/10.1145/3097983.3098144
35. Fährmann D, Damer N, Kirchbuchner F, Kuijper A. Lightweight long short-term memory variational auto-encoder for multivariate time series anomaly detection in industrial control systems. Sensors 22:2886 (2022). https://doi.org/10.3390/s22082886
36. Thill M, Konen W, Bäck T. Time series encodings with temporal convolutional networks. Lecture Notes in Computer Science :161-173 (2020). https://doi.org/10.1007/978-3-030-63710-1_13
37. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science 313:504-507 (2006). https://doi.org/10.1126/science.1127647
38. Malhotra P, et al. LSTM-based encoder-decoder for multi-sensor anomaly detection. arXiv 1607.00148 (2016). https://doi.org/10.48550/arXiv.1607.00148
39. Malhotra P, et al. Long short-term memory networks for anomaly detection in time series. Proc ESANN :89 (2015).
40. Xu H, Feng Y, Chen J, Wang Z, Qiao H, Chen W, Zhao N, Li Z, Bu J, Li Z, Liu Y, Zhao Y, Pei D. Unsupervised anomaly detection via variational auto-encoder for seasonal kpis in web applications. Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18 :187-196 (2018). https://doi.org/10.1145/3178876.3185996
41. Sölch M, et al. Variational inference for on-line anomaly detection in high-dimensional time series. Stat 1050:23 (2016). https://doi.org/10.48550/arXiv.1602.07109
42. Su Y, Zhao Y, Niu C, Liu R, Sun W, Pei D. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining :2828-2837 (2019). https://doi.org/10.1145/3292500.3330672
43. Li L, Yan J, Wang H, Jin Y. Anomaly detection of time series with smoothness-inducing sequential variational auto-encoder. IEEE Trans. Neural Netw. Learning Syst. 32:1177-1191 (2021). https://doi.org/10.1109/tnnls.2020.2980749
44. Park D, Hoshi Y, Kemp CC. A multimodal anomaly detector for robot-assisted feeding using an lstm-based variational autoencoder. IEEE Robot. Autom. Lett. 3:1544-1551 (2018). https://doi.org/10.1109/lra.2018.2801475
45. Audibert J, Michiardi P, Guyard F, Marti S, Zuluaga MA. USAD. Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining :3395-3404 (2020). https://doi.org/10.1145/3394486.3403392
46. Guo Y, et al. Multidimensional time series anomaly detection: A GRU-based Gaussian mixture variational autoencoder approach. Proc Asian Conf Mach Learn :97–112 (2018).
47. He Y, Zhao J. Temporal convolutional networks for anomaly detection in time series. J. Phys.: Conf. Ser. 1213:042050 (2019). https://doi.org/10.1088/1742-6596/1213/4/042050
48. Mirza AH, Cosan S. Computer network intrusion detection using sequential LSTM neural networks autoencoders. 2018 26th Signal Processing and Communications Applications Conference (SIU) :1-4 (2018). https://doi.org/10.1109/siu.2018.8404689
49. Said Elsayed M, Le-Khac NA, Dev S, Jurcut AD. Network anomaly detection using LSTM based autoencoder. Proceedings of the 16th ACM Symposium on QoS and Security for Wireless and Mobile Networks :37-45 (2020). https://doi.org/10.1145/3416013.3426457
50. Geiger A, Liu D, Alnegheimish S, Cuesta-Infante A, Veeramachaneni K. Tadgan: time series anomaly detection using generative adversarial networks. 2020 IEEE International Conference on Big Data (Big Data) :33-43 (2020). https://doi.org/10.1109/bigdata50022.2020.9378139
51. Zhan J, Wang S, Ma X, Wu C, Yang C, Zeng D, Wang S. Stgat-mad : spatial-temporal graph attention network for multivariate time series anomaly detection. ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) :3568-3572 (2022). https://doi.org/10.1109/icassp43922.2022.9747274
52. Li D, Chen D, Jin B, Shi L, Goh J, Ng SK. MAD-GAN: multivariate anomaly detection for time series data with generative adversarial networks. Lecture Notes in Computer Science :703-716 (2019). https://doi.org/10.1007/978-3-030-30490-4_56
53. Zhao H, Wang Y, Duan J, Huang C, Cao D, Tong Y, Xu B, Bai J, Tong J, Zhang Q. Multivariate time-series anomaly detection via graph attention network. 2020 IEEE International Conference on Data Mining (ICDM) :841-850 (2020). https://doi.org/10.1109/icdm50108.2020.00093

54. Munir M, Siddiqui SA, Dengel A, Ahmed S. Deepant: a deep learning approach for unsupervised anomaly detection in time series. IEEE Access 7:1991-2005 (2019). https://doi.org/10.1109/access.2018.2886457
55. Deng A, Hooi B. Graph neural network-based anomaly detection in multivariate time series. AAAI 35:4027-4035 (2021). https://doi.org/10.1609/aaai.v35i5.16523
56. Qiu C, et al. Self-supervised anomaly detection with neural transformation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021). https://doi.org/10.1109/TPAMI.2024.3519543
57. Wagner, D. et al. Formally Exploring Time-Series Anomaly Detection Evaluation Metrics. In Proceedings of the AISTATS (to appear in 2026).
58. Manduchi, L., et al. (2024). On the Challenges and Opportunities in Generative AI. arXiv. https://doi.org/10.48550/arXiv.2403.00025