Deep Anomaly Detection on Tennessee Eastman Process Data*

Fabian Hartung[1,2], Billy Joe Franks[1], Tobias Michels[1], Dennis Wagner[1], Philipp Liznerski[1], Steffen Reithermann[1], Sophie Fellenz[1], Fabian Jirasek[1], Maja Rudolph[3], Daniel Neider[4], Frank Rhein[5], Heike Leitte[1], Chen Song[6], Benjamin Kloepper[6], Stephan Mandt[7], Michael Bortz[8], Jakob Burger[9], Hans Hasse[1], and Marius Kloft[1]

[1] RPTU, Kaiserslautern, Germany kloft@cs.uni-kl.de
[2] BASF SE, Germany
[3] Bosch AI, USA
[4] Technische Universität Dortmund, Germany
[5] Karlsruher Institut für Technologie, Germany
[6] ABB Corporate Research Center Ladenburg, Germany
[7] University of California Irvine, USA
[8] Fraunhofer ITWM, Germany
[9] Technische Universität München, Germany

Abstract. This work evaluates comprehensively and analyzes modern unsupervised (deep-learning-based) anomaly detection methods operating on chemical process data. We use the Tennessee Eastman process dataset, which has been the benchmark data set for chemical process data for nearly three decades. This extensive study will examine anomaly detection methods in industrial applications to determine their best choice. The benchmark results let us conclude that reconstruction-based methods are superior, followed by variational autoencoders, GAN networks, and forecasting-based methods. We extend our evaluation of Hartung et al. by several shallow baseline methods.

Keywords: Anomaly detection · Chemical Process Data · Benchmark · Tennessee Eastman process · Time series

1 Introduction

Detecting data deviating from normality - Anomaly detection (AD) - is crucial in several application domains, from identifying social media bots and fake online reviews to crucial medical and industrial applications, e.g., tumor and fault detection. AD is highly important, especially in safety-critical applications like chemical plants, where failing to recognize anomalies may lead to serious failures, injuries, or even worse. Consequently, much literature on machine learning and AD in chemical processes has been published [4, 28, 44]. Since its introduction

[*] This work is a shortened version but with added baseline evaluations of a previously published journal article by the same authors [13].

three decades ago, the Tennessee Eastman process (TEP) has been established as a standard litmus test for learning-based AD on chemical process data. Most of the new methods are benchmarked on its dataset [8,37]. However, the majority (excluding [3, 4, 30, 34, 45, 49]) of scientific publications evaluate mostly shallow unsupervised anomaly detection methods but do not include neural networks. Since deep neural networks have enabled most of the progress in artificial intelligence during the last 12 years, we propose that shallow machine learning is inadequate for complex, structured data like chemical processes.

Early papers about deep AD on times series (TS) were mostly based on reconstruction [2, 14, 16, 17, 23, 24, 27, 47, 50, 51] or forecasting objectives [6, 9, 15, 25, 29] only. But there is another class of AD methods based on generative models–variational autoencoders (VAEs) [12, 20, 32, 43, 46, 48] and generative adversarial neural networks (GANs) [5,10,19,21,31,39,53]. To get the best parts from all worlds, some hybrid methods combine the above techniques [40, 52]. Adapting the success of supervised classifiers, "one-class classification" trains a network in a way that normal samples are concentrated to a hypersphere [38] or hyperplane [41]. This has recently been applied for AD on TS [40, 42]. A more direct application of classifiers relies on "auxiliary anomalies" [22, 35] or actively querying anomaly labels [18]. Goyal et al. [11] trained a network to distinguish normal training data from synthetically generated anomalies. One of the newest concepts of TS-AD is using self-supervised learning and designing an auxiliary training objective like predicting which transformation was applied to data [36].

With the present work, we intend to evaluate the above-mentioned deep AD approaches on the TEP. This work is an extension of the evaluation of Hartung et al. with its wide range of 27 unsupervised deep AD methods for TS regarding their detection accuracy on the TEP data. We added three shallow methods as baselines. This analysis represents the first - and by far the most comprehensive - evaluation of modern unsupervised AD methods on chemical process data. The results of this study also provide sound advice on which AD methods might best perform on real chemical process data. With the goal of autonomizing the running of chemical processes, establishing deep AD in these would open the route for new, yet unexplored, ways to control them and increase safety and profit for industrial applications and workers.

2 Benchmarking Deep TS-AD on TEP

The TEP is a process simulation of a chemical plant [26]. We use a version of its data available online [1] and referenced in [37]. It provides 20 different types of anomalies and corresponding simulations of 53 parameters - generated every three minutes for 25 hours for training data and 48 hours for test data.

To evaluate the examined algorithms on the TEP, we compare the F1-score and area under the precision-recall curve (AUPRC). Both are the most commonly used metrics in AD. Anomaly detectors generate an anomaly score for each point of a TS. If it exceeds a learned threshold score, the point in time is considered anomalous. The proportion of correctly detected anomalies is called precision. Meanwhile, recall means the proportion of correctly detected anomalies among all true anomalies. Combining precision and recall in one metric yields the F1-score, which can be calculated at every single point of the TS. The total F1-score averages all single F1-scores over the whole TS. Given a dataset and the ordering of all data points regarding a binary decision value, in our case derived from the anomaly score, the precision-recall curve plots for every possible threshold the respective precision against the recall. The AUPRC is a general measure of a model's performance.

We implemented all methods in the same Python environment for an equal and fair comparison and used PyTorch [33] for training and evaluation. After separating a quarter of the training dataset for methods requiring an unlabeled validation set, the test dataset was divided into five equal-sized folds. The remaining folds were then used for the evaluation, excluding neighboring folds to avoid time dependencies. Finally, the performance in F1-Score and AUPRC over these folds was averaged. For a fair comparison and hyperparameter tuning, the size of each method's parameter grid was chosen to ensure a training and evaluation time of 24 hours for every method. We use two thresholding methods as shallow baselines based on the interquartile range (IQR) and the min and max in the training data (OOS), respectively. We also include a baseline using the distance to the mean weighted by feature variance (WMD). Table 1 shows the implemented methods with reference to their publications, performance results, and rankings.

3 Discussion and Conclusion

The results mark reconstruction-based methods as the best performers on average, although one GAN method (BeatGAN) ranks best. On average, the generative methods rank in midfield - VAE performing better than GAN - and the forecasting-based and hybrid methods performing worse than the rest. Both metrics show similar results except for GMM-GRU-VAE, LSTM-AE-OC-SVM, and TCN-S2S-P. Since all deep methods achieve scores higher than the three shallow methods, we conclude that more complex multivariate TS - especially in chemical processes - need deep methods to correctly detect all kinds of anomalies.

Considering future work, the TEP data is synthetic, and real data is preferable. However, no widely accepted benchmark of real-world data is available yet. All methods achieved high scores of 0.9 and above. This could be caused by the synthetic data with defined faults placed in a fault-free run. It will be interesting to compare this evaluation with real-world data in the future. The challenge here will be uncovering the data and correctly labeling its anomalies. Even though the F1-score and AUPRC are state of the art for comparison, they lack assessing more extended periods and some typical characteristics of TS [7, 17].

This benchmark can guide further research and practitioners in choosing a method for AD on chemical TS.

Table 1. This table shows the performance of all evaluated methods. The table lists each method's reference, the best F1-score, and the best AUPRC for each method. The table lists the ranking according to the F1-score, AUPRC, and mean. The methods are sorted according to the best mean of F1-score and AUPRC.

Method	**Method Type**	**F1-Score**	**F1-Score Rank**	**AUPRC**	**AUPRC Rank**	**Total Rank**
BeatGAN [53]	GAN	0.9699	1	0.9896	2	1
TCN-S2S-AE [47]	Reconstr.	0.9632	3	0.9914	1	2
Dense-AE [2]	Reconstr.	0.9631	4	0.9880	3	3
LSTM-AE [24]	Reconstr.	0.9506	5	0.9861	4	4
LSTM-P [25]	Forecasting	0.9693	2	0.9824	8	5
MSCRED [51]	Reconstr.	0.9353	7	0.9842	5	6
Donut [48]	VAE	0.9450	6	0.9829	7	7
LSTM-VAE [43]	VAE	0.9334	11	0.9831	6	8
OmniAnomaly [46]	VAE	0.9336	9	0.9808	12	9
SIS-VAE [20]	VAE	0.9335	10	0.9790	14	10
Untrained-LSTM-AE [17]	Reconstr.	0.9333	13	0.9792	13	11
LSTM-DVAE [32]	VAE	0.9333	16	0.9811	11	12
USAD [2]	Reconstr.	0.9333	12	0.9779	16	13
GMM-GRU-VAE [12]	VAE	0.9291	21	0.9815	10	14
TCN-S2S-P [15]	Forecasting	0.9172	23	0.9821	9	15
LSTM-MAX-AE [27]	Reconstr.	0.9333	18	0.9786	15	16
LSTM-AE-OC-SVM [40]	Hybrid	0.9337	8	0.9511	26	17
LSTM-VAE-GAN [31]	GAN	0.9333	14	0.9735	20	17
GenAD [16]	Reconstr.	0.9333	19	0.9755	19	19
TadGAN [10]	GAN	0.9333	15	0.9690	23	19
STGAT-MAD [50]	Reconstr.	0.9267	22	0.9767	17	21
Mad-GAN [19]	GAN	0.9333	17	0.9621	24	22
MTAD-GAT [52]	Hybrid	0.9097	25	0.9758	18	23
DeepANT/TCN-P [29]	Forecasting	0.9114	24	0.9712	22	24
GDN [6]	Forecasting	0.9078	26	0.9722	21	25
LSTM-2S2-P [9]	Forecasting	0.9327	20	0.9171	27	25
THOC [42]	Hybrid	0.9074	27	0.9618	25	27
WMD	Baseline	0.8956	29	0.8563	28	28
OOS	Baseline	0.8956	28	0.8203	29	29
IQR	Baseline	0.8956	29	0.8125	30	30

Acknowledgments. Part of this work was conducted within the DFG research unit FOR 5359 on Deep Learning on Sparse Chemical Process Data and the BMWK project KEEN (01MK20014U,01MK20014L).

Disclosure of Interests. The authors have no competing interests to declare relevant to this article's content.

References

1. TEP-DATA additional tennessee eastman process simulation data for anomaly detection evaluation. https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/6C3JR1, accessed: 2022-11-04
2. Audibert, J., Michiardi, P., Guyard, F., Marti, S., Zuluaga, M.A.: Usad: Unsupervised anomaly detection on multivariate time series. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 3395–3404 (2020)
3. Chadha, G.S., Islam, I., Schwung, A., Ding, S.X.: Deep convolutional clustering-based time series anomaly detection. Sensors **21**(16), 5488 (2021)
4. Chadha, G.S., Rabbani, A., Schwung, A.: Comparison of semi-supervised deep neural networks for anomaly detection in industrial processes. In: 2019 IEEE 17th international conference on industrial informatics (INDIN). vol. 1, pp. 214–219. IEEE (2019)
5. Deecke, L., Vandermeulen, R., Ruff, L., Mandt, S., Kloft, M.: Image anomaly detection with generative adversarial networks. In: Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I 18. pp. 3–17. Springer (2019)
6. Deng, A., Hooi, B.: Graph neural network-based anomaly detection in multivariate time series. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 4027–4035 (2021)
7. Doshi, K., Abudalou, S., Yilmaz, Y.: Tisat: time series anomaly transformer. arXiv preprint arXiv:2203.05167 (2022)
8. Downs, J.J., Vogel, E.F.: A plant-wide industrial process control problem. Computers & chemical engineering **17**(3), 245–255 (1993)
9. Filonov, P., Lavrentyev, A., Vorontsov, A.: Multivariate industrial time series with cyber-attack simulation: Fault detection using an lstm-based predictive data model. arXiv preprint arXiv:1612.06676 (2016)
10. Geiger, A., Liu, D., Alnegheimish, S., Cuesta-Infante, A., Veeramachaneni, K.: Tadgan: Time series anomaly detection using generative adversarial networks. In: 2020 IEEE International Conference on Big Data (Big Data). pp. 33–43. IEEE (2020)
11. Goyal, S., Raghunathan, A., Jain, M., Simhadri, H.V., Jain, P.: Drocc: Deep robust one-class classification. In: International conference on machine learning. pp. 3711–3721. PMLR (2020)
12. Guo, Y., Liao, W., Wang, Q., Yu, L., Ji, T., Li, P.: Multidimensional time series anomaly detection: A gru-based gaussian mixture variational autoencoder approach. In: Asian Conference on Machine Learning. pp. 97–112. PMLR (2018)
13. Hartung, F., Franks, B.J., 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**(7), 1077–1082 (2023). https://doi.org/https://doi.org/10.1002/cite.202200238, https://onlinelibrary.wiley.com/doi/abs/10.1002/cite.202200238

14. Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A.K., Davis, L.S.: Learning temporal regularity in video sequences. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 733–742 (2016)
15. He, Y., Zhao, J.: Temporal convolutional networks for anomaly detection in time series. In: Journal of Physics: Conference Series. vol. 1213, p. 042050. IOP Publishing (2019)
16. 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**(1), e12683 (2023). https://doi.org/https://doi.org/10.1049/ell2.12683, https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/ell2.12683
17. Kim, S., Choi, K., Choi, H.S., Lee, B., Yoon, S.: Towards a rigorous evaluation of time-series anomaly detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 7194–7201 (2022)
18. Li, A., Qiu, C., Kloft, M., Smyth, P., Mandt, S., Rudolph, M.: Deep anomaly detection under labeling budget constraints. In: International Conference on Machine Learning. pp. 19882–19910. PMLR (2023)
19. Li, D., Chen, D., Jin, B., Shi, L., Goh, J., Ng, S.K.: Mad-gan: Multivariate anomaly detection for time series data with generative adversarial networks. In: Artificial Neural Networks and Machine Learning–ICANN 2019: Text and Time Series: 28th International Conference on Artificial Neural Networks, Munich, Germany, September 17–19, 2019, Proceedings, Part IV. pp. 703–716. Springer (2019)
20. Li, L., Yan, J., Wang, H., Jin, Y.: Anomaly detection of time series with smoothness-inducing sequential variational auto-encoder. IEEE transactions on neural networks and learning systems **32**(3), 1177–1191 (2020)
21. Liu, W., Luo, W., Lian, D., Gao, S.: Future frame prediction for anomaly detection–a new baseline. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6536–6545 (2018)
22. Liznerski, P., Ruff, L., Vandermeulen, R.A., Franks, B.J., Müller, K.R., Kloft, M.: Exposing outlier exposure: What can be learned from few, one, and zero outlier images. arXiv preprint arXiv:2205.11474 (2022)
23. Luo, W., Liu, W., Gao, S.: Remembering history with convolutional lstm for anomaly detection. In: 2017 IEEE International Conference on Multimedia and Expo (ICME). pp. 439–444. IEEE (2017)
24. Malhotra, P., Ramakrishnan, A., Anand, G., Vig, L., Agarwal, P., Shroff, G.: Lstm-based encoder-decoder for multi-sensor anomaly detection. arXiv preprint arXiv:1607.00148 (2016)
25. Malhotra, P., Vig, L., Shroff, G., Agarwal, P., et al.: Long short term memory networks for anomaly detection in time series. In: ESANN. vol. 2015, p. 89 (2015)
26. Manca, G.: 'tennessee-eastman-process' alarm management dataset. IEEE Data-Port (2020)
27. Mirza, A.H., Cosan, S.: Computer network intrusion detection using sequential lstm neural networks autoencoders. In: 2018 26th signal processing and communications applications conference (SIU). pp. 1–4. IEEE (2018)
28. Monroy, I., Escudero, G., Graells, M.: Anomaly detection in batch chemical processes. In: Computer Aided Chemical Engineering, vol. 26, pp. 255–260. Elsevier (2009)

29. Munir, M., Siddiqui, S.A., Dengel, A., Ahmed, S.: Deepant: A deep learning approach for unsupervised anomaly detection in time series. Ieee Access **7**, 1991–2005 (2018)
30. Neubürger, F., Saeid, Y., Kopinski, T.: Variational-autoencoder architectures for anomaly detection in industrial processes
31. Niu, Z., Yu, K., Wu, X.: Lstm-based vae-gan for time-series anomaly detection. Sensors **20**(13), 3738 (2020)
32. Park, D., Hoshi, Y., Kemp, C.C.: A multimodal anomaly detector for robot-assisted feeding using an lstm-based variational autoencoder. IEEE Robotics and Automation Letters **3**(3), 1544–1551 (2018)
33. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
34. Plakias, S., Boutalis, Y.S.: A novel information processing method based on an ensemble of auto-encoders for unsupervised fault detection. Computers in Industry **142**, 103743 (2022)
35. Qiu, C., Li, A., Kloft, M., Rudolph, M., Mandt, S.: Latent outlier exposure for anomaly detection with contaminated data. In: International conference on machine learning. pp. 18153–18167. PMLR (2022)
36. Qiu, C., Pfrommer, T., Kloft, M., Mandt, S., Rudolph, M.: Neural transformation learning for deep anomaly detection beyond images. In: International Conference on Machine Learning. pp. 8703–8714. PMLR (2021)
37. Rieth, C.A., Amsel, B.D., Tran, R., Cook, M.B.: Issues and advances in anomaly detection evaluation for joint human-automated systems. In: Advances in Human Factors in Robots and Unmanned Systems: Proceedings of the AHFE 2017 International Conference on Human Factors in Robots and Unmanned Systems, July 17- 21, 2017, The Westin Bonaventure Hotel, Los Angeles, California, USA 8. pp. 52–63. Springer (2018)
38. Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S.A., Binder, A., Müller, E., Kloft, M.: Deep one-class classification. In: International conference on machine learning. pp. 4393–4402. PMLR (2018)
39. Sabokrou, M., Khalooei, M., Fathy, M., Adeli, E.: Adversarially learned one-class classifier for novelty detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3379–3388 (2018)
40. Said Elsayed, M., Le-Khac, N.A., Dev, S., Jurcut, A.D.: Network anomaly detection using lstm based autoencoder. In: Proceedings of the 16th ACM Symposium on QoS and Security for Wireless and Mobile Networks. pp. 37–45 (2020)
41. Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., Williamson, R.C.: Estimating the support of a high-dimensional distribution. Neural computation **13**(7), 1443–1471 (2001)
42. Shen, L., Li, Z., Kwok, J.: Timeseries anomaly detection using temporal hierarchical one-class network. Advances in Neural Information Processing Systems **33**, 13016–13026 (2020)
43. Sölch, M., Bayer, J., Ludersdorfer, M., van der Smagt, P.: Variational inference for on-line anomaly detection in high-dimensional time series. stat **1050**, 23 (2016)
44. Song, B., Suh, Y.: Narrative texts-based anomaly detection using accident report documents: The case of chemical process safety. Journal of Loss Prevention in the Process Industries **57**, 47–54 (2019)
45. Spyridon, P., Boutalis, Y.S.: Generative adversarial networks for unsupervised fault detection. In: 2018 European Control Conference (ECC). pp. 691–696. IEEE (2018)

46. Su, Y., Zhao, Y., Niu, C., Liu, R., Sun, W., Pei, D.: Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. pp. 2828–2837 (2019)
47. Thill, M., Konen, W., Bäck, T.: Time series encodings with temporal convolutional networks. In: Bioinspired Optimization Methods and Their Applications: 9th International Conference, BIOMA 2020, Brussels, Belgium, November 19–20, 2020, Proceedings 9. pp. 161–173. Springer (2020)
48. Xu, H., Chen, W., Zhao, N., Li, Z., Bu, J., Li, Z., Liu, Y., Zhao, Y., Pei, D., Feng, Y., et al.: Unsupervised anomaly detection via variational auto-encoder for seasonal kpis in web applications. In: Proceedings of the 2018 world wide web conference. pp. 187–196 (2018)
49. Yang, X., Feng, D.: Generative adversarial network based anomaly detection on the benchmark tennessee eastman process. In: 2019 5th International conference on control, automation and robotics (ICCAR). pp. 644–648. IEEE (2019)
50. 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. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3568–3572. IEEE (2022)
51. Zhang, C., Song, D., Chen, Y., Feng, X., Lumezanu, C., Cheng, W., Ni, J., Zong, B., Chen, H., Chawla, N.V.: A deep neural network for unsupervised anomaly detection and diagnosis in multivariate time series data. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 1409–1416 (2019)
52. 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. In: 2020 IEEE International Conference on Data Mining (ICDM). pp. 841–850. IEEE (2020)
53. Zhou, B., Liu, S., Hooi, B., Cheng, X., Ye, J.: Beatgan: Anomalous rhythm detection using adversarially generated time series. In: IJCAI. vol. 2019, pp. 4433–4439 (2019)