




Anomaly Detectors for Multivariate Time Series: The Proof of the Pudding is in the Eating

Phillip Wenig 
Hasso Plattner Institute
University of Potsdam
Potsdam, Germany
phillip.wenig@hpi.de

Sebastian Schmidl 
Hasso Plattner Institute
University of Potsdam
Potsdam, Germany
sebastian.schmidl@hpi.de

Thorsten Papenbrock 
Philipps University of Marburg
Marburg, Germany
papenbrock@informatik.uni-marburg.de

Abstract—Anomaly detection is a popular task in time series analytics and researchers have, therefore, developed a plethora of algorithms to solve it. While most algorithms focus on univariate time series, one family of anomaly detection algorithms specializes on multivariate data. Because existing studies benchmark on non-meaningful datasets and often only within uni- or multivariate algorithm families, it is unclear whether multivariate solutions are actually superior on multivariate data.

In this study, we compare univariate and multivariate approaches on common multivariate benchmark times series to demonstrate that existing benchmark datasets cannot highlight the strengths of multivariate anomaly detection algorithms. We though demonstrate such strengths with a simple, generated dataset that contains a special type of anomaly, which we call *correlation anomaly*. Our experimental results, therefore, call for novel types of benchmark datasets whose anomalies actually facilitate the multidimensional nature of the data.

I. ANOMALIES IN MULTIVARIATE TIME SERIES

An anomaly in a time series describes a pattern that deviates w.r.t. some measure, model, or embedding from the regular patterns of the time series. Anomalies are often the result of special events, such as heart failures in cardiology [1], structural defects in jet turbine engineering [2], or ecosystem disturbances in earth sciences [3]. Due to their importance in many domains, researchers have developed a multitude of discovery algorithms, which have been surveyed and evaluated in various benchmarks [4]–[16]. One family of anomaly detection algorithms specializes on *multivariate* time series, which are time series with more than one channel. The superiority of this algorithm family over univariate algorithms on multivariate data has, however, never been shown because existing benchmarks either compare algorithms only within these families [5], [6], [17], they compare them on only univariate data [4], [7], [10], they show only the superiority of univariate algorithms over multivariate algorithms [8], or they come to inconclusive results w.r.t. multivariate capabilities [18]. As we will show in this study, the lack of empirical evidence for the strengths of multivariate algorithms is not only due to the absence of experimental publications but also because existing multivariate benchmark datasets contain predominantly simple (maybe even trivial [5], [8], [9], [19]) anomalies that also univariate algorithms can detect.

Our experiments evaluate seven state-of-the-art *univariate* anomaly detection algorithms and six state-of-the-art *multi-*

variate anomaly detection algorithms on 14 popular multivariate datasets and compare their ROC-AUC and PR-AUC scores. Contrary to common assumptions, the measured scores are overall better for univariate approaches than for multivariate ones. With a special type of anomaly, namely *correlation anomalies* (as also recognized by related works [20], [21]), we demonstrate that this observation is not because multivariate algorithms are useless but because existing benchmark datasets do not contain (many) anomalies that multivariate anomaly detection algorithms are specialized for. Hence, with this study, we make the following contributions:

- (1) We integrate *correlation anomalies* into a time series anomaly taxonomy as a new type of anomaly that highlights the special strengths of multivariate anomaly detection algorithms (Section II).
- (2) We demonstrate the shortcomings of multivariate benchmark datasets by showing that univariate algorithms can detect their anomalies more accurately than most existing multivariate anomaly detection algorithms (Section III).
- (3) We investigate two concrete multivariate datasets to differentiate anomalies that do and do not require multivariate detection approaches (Section IV).

II. TYPES AND PROPERTIES OF ANOMALIES

For a comprehensive detection of anomalies in multivariate time series, different types of anomalies need to be considered. In this section, we classify these anomaly types with a novel anomaly taxonomy for multivariate time series (Figure 1). We also define the concept of *anomaly arity* that captures the number of channels an anomaly manifests itself in.

a) Anomaly Taxonomy: Existing literature on time series anomaly detection (e.g., [4], [9], [11], [12], [17], [22]–[24]) relies on the general outlier/anomaly taxonomy proposed by Chandola et al. [14]. This taxonomy considers three different types of anomalies, which are *point*, *collective*, and *contextual* anomalies, and is applicable to all kinds of data formats (temporal, spacial, relational, etc.). We specialize this taxonomy to anomalies in multivariate time series by adding another level for *locality* and adding the new anomaly type *correlation anomaly*, which can be observed only in multivariate time series. Figure 1 shows the extended taxonomy with an example for each anomaly type. The first level distinguishes between

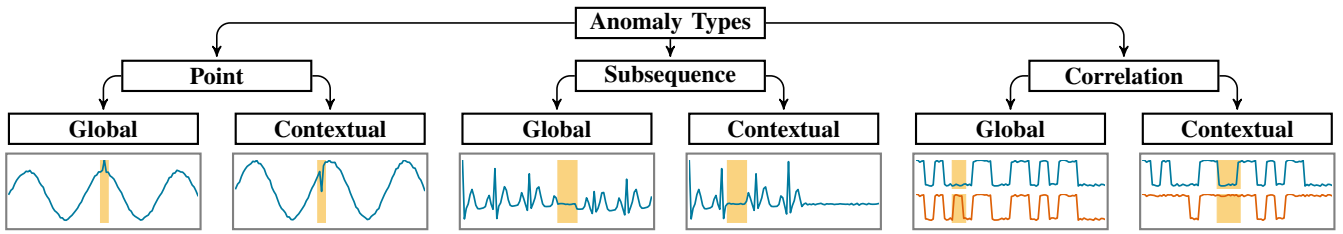


Fig. 1. Taxonomy of anomaly types and localities, with illustrative example visualizations.

point, *subsequence*, and *correlation* anomalies and the second level differentiates *global* and *contextual* anomalies:

Global point anomalies correspond to *outliers* or *point anomalies* in related work [14] and denote individual points of a time series, whose values fall outside the global range of all normal data points.

Contextual point anomalies are also referred to as *contextual anomalies* in related work [14] and denote individual data points with a significantly different value as their surrounding, i. e., contextual data points; the same value might be considered normal in a different context.

Global subsequence anomalies are also known as *collective anomalies* [14] and describe contiguous data points with a value pattern that differs from all other patterns in the entire, i. e., global time series; all individual values might be normal, but their sequence is rare (or even unique).

Contextual subsequence anomalies mark a collection of contiguous data points whose pattern is anomalous not necessarily w. r. t. the entire time series but within their specific environment; all individual values might still be normal in their contexts.

Global correlation anomalies represent series of contiguous multivariate data points whose values contradict a correlation of at least two channels. Given that the values of two (or more) channels are linearly related in the (global) time series, the correlation anomaly violates this linear relationship; the values and value pattern in every single channel might still be normal w. r. t. other values and patterns in that channel.

Contextual correlation anomalies are correlation anomalies that violate a contextually limited linear relationship of two (or more) time series channels. If a correlation is observed for a (statistically significant) subsequence, a contextual correlation anomaly is a local contradiction of this relationship; similar to other contextual anomalies, the correlation might not hold for the entire time series, but its violation is considered anomalous in certain contexts that follow the correlation.

Correlation anomalies are a type of anomaly that exists only in multivariate time series. If they do not also manifest as subsequence or point anomalies in their individual channels, algorithms need to consider the interaction of multiple channels for their detection. We used the concept of correlation to generate a new time series collection named CoMuT (**C**orrelated **M**ultivariate **T**ime Series) with different numbers,

shapes, and lengths of correlation anomalies.

b) *Behavioral and Technical Classifications*: Existing literature proposes to further distinguish subsequence anomalies (both global and contextual) into different behavioral or technical types. Depending on the source of the definition, subsequence anomalies are divided into (i) shapelet, seasonal, and trend anomalies [5], (ii) position, pattern change, and artifact anomalies [25], (iii) pattern-related and frequency-related anomalies [9], or (iv) spike, contextual, flip, speedup, noise, cutoff, scale, wander, and average anomalies [24]. The taxonomy can incorporate these additional types in a third level below global and contextual subsequence anomalies. The proposed types are useful primarily for the generation of representative synthetic benchmark datasets, but neither of them serves to highlight special discovery capabilities for multivariate time series.

c) *Anomaly Arity*: When considering multivariate time series, anomalies might manifest in any number of channels. We, thus, define the *arity* of an anomaly as the number of channels affected by an anomaly. An n -ary anomaly impacts n channels of an m -dimensional time series, where $n \leq m$. To ease the arity reasoning, we denote 1-ary anomalies as *single* anomalies, n -ary anomalies with $1 < n < m$ as *multiple* anomalies, and m -ary anomalies as *all* anomalies. Univariate time series can contain only *single* anomalies, and correlation anomalies are always either *multiple* or *all* anomalies.

III. BENCHMARKING MULTIVARIATE ALGORITHMS

To evaluate whether state-of-the-art benchmark collections for multivariate anomaly detection contain anomalies that can highlight the strengths of multivariate detection algorithms, we process these collections with both univariate and multivariate anomaly detection algorithms. Based on the results of a comprehensive evaluation on univariate time series [10], we selected seven top performing univariate anomaly detectors¹ and six top performing multivariate anomaly detectors² for the benchmark. To execute the univariate methods on multivariate time series, we use three strategies: SUM_BEFORE transforms all channels of a multivariate time series via element-wise sum into a univariate time series that univariate detectors can process; MAX_AFTER and MEAN_AFTER execute the univariate detectors on every channel and aggregate the resulting

¹Univariate selection: k-Means [26], STAMP [27], Subsequence-IF [28], Sub-LOF [29], NormA-SJ [30], Series2Graph [31], and DWT-MLEAD [32].

²Multivariate selection: k-Means [26], Torsk [33], RBFforest [34], LSTM-AD [35], DBStream [36], and NF [37]

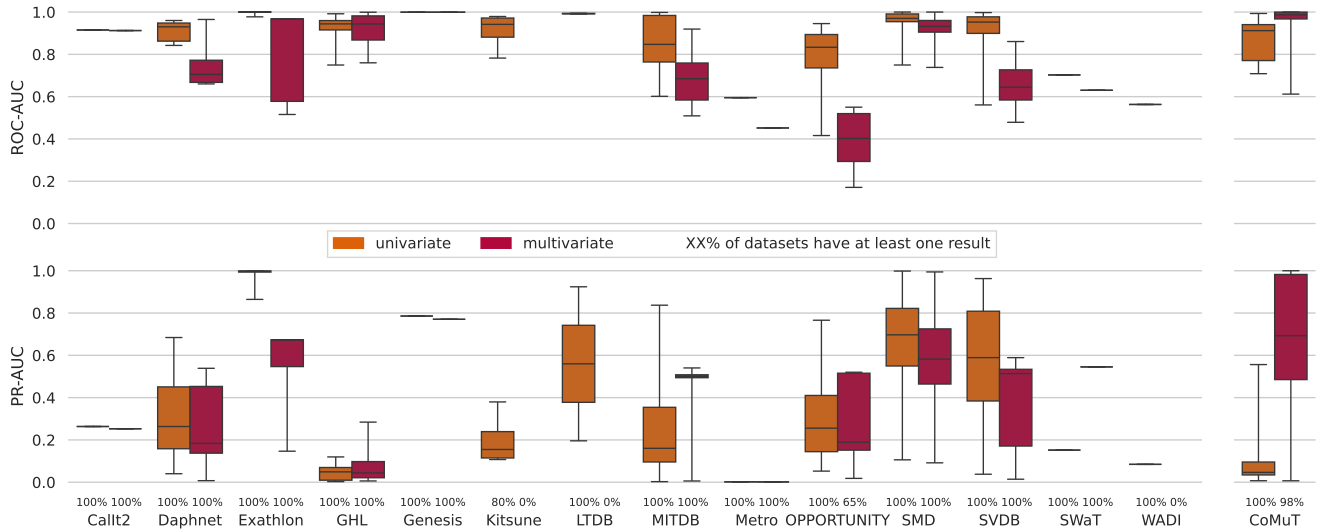


Fig. 2. Max ROC-AUC scores (top) and max PR-AUC scores (bottom) over all algorithms for each time series grouped by collection into box plots.

anomaly scores of each channel into one scoring by taking the element-wise max and mean, respectively. The evaluation uses 14 public multivariate anomaly detection benchmark collections³ and our own, novel CoMuT collection⁴, which consists of 60 generated time series with three, five, seven, and nine dimensions, 10,000 time steps, and different *global correlation anomalies*. We run the experiments with the *TimeEval* [25] evaluation toolkit and provide each execution with 60 GB of RAM and 12 hours of runtime on an Intel Xeon E5-2630 v4 CPU. For all executions, we measure the results with the ROC-AUC [54] and the PR-AUC [55] metrics. Our evaluation, however, considers for every time series in a collection only the maximum ROC-AUC and PR-AUC scores that any univariate algorithm (with any strategy) and any multivariate algorithm could achieve. The results in Figure 2, therefore, show the best results for each of the two anomaly detection families on all time series grouped by collection.

The results in Figure 2 show that multivariate algorithms are, in general, more complex and resource demanding: While the univariate algorithms delivered results on 92% of the time series, multivariate algorithms delivered a result on only 64% of the time series with 17% errors (*NaN*, *inf*, etc.), 4% timeouts, and 15% out of memory. For Kitsune, LTDB, and WADI, all multivariate detectors failed to process even a single time series. Because multivariate algorithms analyze all channels of a time series simultaneously while the univariate algorithms analyze them sequentially, these additional costs are expected. But they do not pay off for the 14 existing datasets in our evaluation.

³Multivariate time series collections: Callt2 [38], [39], Daphnet [38], [40], Exathlon [6], GHL [41], Genesis [42], Kitsune [43], LTDB [44], [45], MITDB [44], [46], Metro [38], [47], OPPORTUNITY [38], [48], SMD [49], SVDB [44], [50], [51], SWAT [52], and WADI [53]

⁴<https://hpi.de/naumann/s/comut>

Both evaluation metrics in Figure 2 show that univariate algorithms reach better results on most of the collections: They are consistently superior w.r.t. ROC-AUC and on average better w.r.t. PR-AUC. This observation indicates that our selection of univariate algorithms might be a bit more aggressive in scoring anomalies compared to the multivariate selection, which pays off in ROC-AUC but not always in PR-AUC (more details about the metrics in related work [56], [57]), but it does not demonstrate the systematic strengths of the multivariate family. The measurements on our new CoMuT time series, however, show that the multivariate algorithms can consistently outperform the univariate algorithms on both metrics. Because CoMuT contains only *global correlation anomalies*, this advantage can clearly be attributed to the algorithms' ability to consider all channels simultaneously. We can also deduce from the results that the 14 existing benchmark datasets do not include many correlation anomalies or other types of anomalies that highlight the strengths of multivariate algorithms. So because existing benchmark datasets lack the necessity to use multivariate algorithms, more datasets with multivariate anomalies, such as *correlation anomalies*, are needed to benchmark multivariate solutions properly for their specific strengths.

To substantiate our claim, we now compare the results of the univariate k-Means (with `MAX_AFTER` strategy) and the multivariate k-Means in Table I. The measurements show that the PR-AUC scores are a little higher for the multivariate version, but still very close, i.e., within standard deviation. Because the performance difference is significantly higher on CoMuT, the benchmark collections probably contain only few or no *correlation anomalies*. The CoMuT scores also show that the univariate version of k-Means cannot find such anomalies at all, although its multivariate version can detect them.

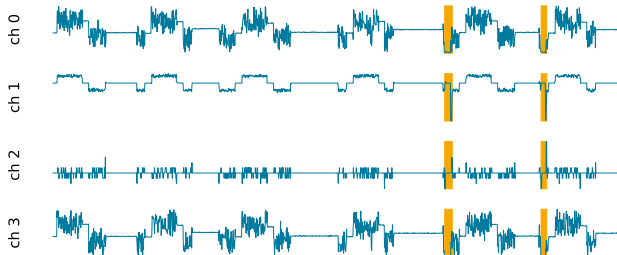


Fig. 3. Excerpt of a Genesis time series with marked anomalies.

IV. DRILL DOWN ON MULTIVARIATE ANOMALIES

As elaborated in the previous sections, univariate anomaly detectors can detect anomalies in certain multivariate time series with simple combination methods. Especially time series that contain *point* or *subsequence anomalies* of any *arity* are no challenge because looking at one anomalous channel alone is often sufficient to find such anomalies. As an example, we take a closer look at the Genesis collection. It consists of only one time series ($\approx 16\,000$ time steps; 18 channels) with two anomalies. Figure 3 shows an excerpt of four channels around these anomalies. The anomalies manifest in the channels 0 and 3 as *global subsequence anomalies* (platforms at a low value), and in the channels 1 and 2 as a combination of *contextual* and *global point anomalies*. For this reason, univariate algorithms can easily detect both *quaternary* anomalies by considering only one of the four channels. The dataset is still somewhat interesting because the other 16 channels of the Genesis time series do not exhibit the anomaly.

In Figure 4, we plotted an excerpt of a CoMuT time series with three channels (in blue), a marked *ternary global correlation anomaly*, and a `SUM_BEFORE` aggregation (in orange). The time series consists of three channels with random mode jumps (and some noise) that highly correlate with each other, i.e., if one channel jumps up or down, all other channels

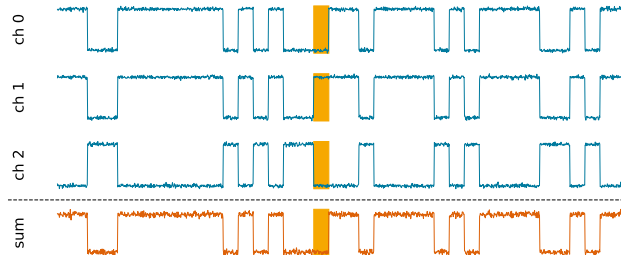


Fig. 4. Excerpt of a CoMuT time series with a visualization of the linear sum of all channels and a marked correlation anomaly.

perform a jump as well. Here, Channel 0 follows the same pattern as Channel 1, and Channel 2 is negatively correlated to Channel 1 and Channel 0. In the marked anomalous sequence, the correlation is violated because Channel 0 jumps 40 time steps too late. While many multivariate algorithms including *k*-Means (see Figure 2 and Table I) can detect this anomaly, hardly any univariate algorithm found it. When inspecting only one mode jump channel with the strategies `MEAN_AFTER` or `MAX_AFTER`, the anomaly is not detectable. Moreover, summing up the channels beforehand with `SUM_BEFORE` also hides the anomalous pattern because the aggregate of the mode jumps happens to be another random mode jump subsequence as illustrated by the orange series in Figure 4 – the negatively correlated channels 1 and 2 simply cancel each other out. Note that if a univariate algorithm could automatically figure out a suitable aggregation that does not hide *correlation anomalies*, this algorithm would be a multivariate solution. In summary, no simple strategy, such as `SUM_BEFORE`, `MAX_AFTER`, or `MEAN_AFTER` enables univariate algorithms to detect *correlation anomalies* effectively. They can, therefore, become a suitable tool to benchmark multivariate algorithms, if we create more benchmark datasets with them.

V. WE NEED BETTER BENCHMARK COLLECTIONS

In this study, we demonstrated that existing multivariate benchmark datasets cannot highlight the strengths of multivariate anomaly detection algorithms. We proposed *correlation anomalies* as a type of anomaly that indicates these strengths, extended existing anomaly typing taxonomies, and generated the CoMuT dataset as a first example for a suitable multivariate benchmark datasets. Our experimental findings call for the creation of further multivariate benchmark datasets and provide direction for experimental designs. They also demonstrate that existing univariate solutions tend to be superior for *point* and *sequence anomalies*; effective anomaly detection projects, therefore, need to run both families of algorithms to effectively detect all types of anomalies.

ACKNOWLEDGEMENTS

We thank Felix Naumann for his support and input to the formulation of the extended anomaly taxonomy.

TABLE I
MEAN PR-AUC (\pm STANDARD DEVIATION) OF UNIVARIATE (`MAX` STRATEGY) AND MULTIVARIATE *k*-MEANS PER COLLECTION.

Collection	Univ. <i>k</i> -Means	Multiv. <i>k</i> -Means
Callt2	0.246 (± 0.000)	0.253 (± 0.000)
Daphnet	0.094 (± 0.057)	0.098 (± 0.061)
Exathlon	0.337 (± 0.184)	0.464 (± 0.274)
GHL	0.028 (± 0.032)	0.066 (± 0.078)
Genesis	0.748 (± 0.000)	0.770 (± 0.000)
Kitsune	0.107 (± 0.000)	-
LTDB	-	-
MITDB	0.077 (± 0.078)	0.085 (± 0.077)
Metro	0.001 (± 0.000)	0.001 (± 0.000)
OPPORTUNITY	-	-
SMD	0.381 (± 0.227)	0.506 (± 0.196)
SVDB	0.204 (± 0.158)	0.217 (± 0.171)
SWAT	-	-
WADI	-	-
CoMuT	0.016 (± 0.009)	0.622 (± 0.244)

REFERENCES

- [1] S. Ansari, N. Farzaneh, M. Duda, K. Horan, H. B. Andersson, Z. D. Goldberger, B. K. Nallamothu, and K. Najarian, "A review of automated methods for detection of myocardial ischemia and infarction using electrocardiogram and electronic health records," *IEEE Reviews in Biomedical Engineering*, vol. 10, pp. 264–298, 2017.
- [2] M. Woike, A. Abdul-Aziz, and M. Clem, "Structural health monitoring on turbine engines using microwave blade tip clearance sensors," in *Smart Sensor Phenomena, Technology, Networks, and Systems Integration 2014*, vol. 9062, 2014, pp. 167–180.
- [3] H. Cheng, P.-N. Tan, C. Potter, and S. Klooster, "Detection and characterization of anomalies in multivariate time series," in *Proceedings of the SIAM International Conference on Data Mining (SDM)*, 2009, pp. 413–424.
- [4] F. Rewicki, J. Denzler, and J. Niebling, "Is It Worth It? Comparing Six Deep and Classical Methods for Unsupervised Anomaly Detection in Time Series," *Applied Sciences*, vol. 13, no. 3, p. 1778, 2023.
- [5] K.-H. Lai, D. Zha, Y. Zhao, G. Wang, J. Xu, and X. Hu, "Revisiting Time Series Outlier Detection: Definitions and Benchmarks," in *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, 2021.
- [6] V. Jacob, F. Song, A. Stiegler, B. Rad, Y. Diao, and N. Tatbul, "Exathlon: A Benchmark for Explainable Anomaly Detection over Time Series," *Proceedings of the VLDB Endowment (PVLDB)*, vol. 14, pp. 2613–2626, 2021.
- [7] C. Freeman, J. Merriman, I. Beaver, and A. Mueen, "Experimental Comparison and Survey of Twelve Time Series Anomaly Detection Algorithms," *Journal of Artificial Intelligence Research*, vol. 72, pp. 849–899, 2021.
- [8] A. Garg, W. Zhang, J. Samarán, R. Savitha, and C.-S. Foo, "An Evaluation of Anomaly Detection and Diagnosis in Multivariate Time Series," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 6, pp. 2508–2517, 2022.
- [9] J. Paparrizos, Y. Kang, R. S. Tsay, T. Palpanas, and M. J. Franklin, "TSB-AUD: An end-to-end anomaly detection benchmark suite for univariate time-series data," 2021. [Online]. Available: <https://github.com/johnpaparrizos/TSB-UAD>
- [10] S. Schmidl, P. Wenig, and T. Papenbrock, "Anomaly detection in time series: A comprehensive evaluation," in *Proceedings of the VLDB Endowment*, 2022.
- [11] A. Blázquez-García, A. Conde, U. Mori, and J. A. Lozano, "A review on outlier/anomaly detection in time series data," *arXiv preprint arXiv:2002.04236*, 2020.
- [12] M. Braei and S. Wagner, "Anomaly Detection in Univariate Time-series: A Survey on the State-of-the-Art," *arXiv preprint arXiv:2004.00433*, 2020.
- [13] R. Chalapathy and S. Chawla, "Deep Learning for Anomaly Detection: A Survey," *arXiv preprint arXiv:1901.03407*, 2019.
- [14] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys*, vol. 41, no. 3, pp. 1–58, 2009.
- [15] D. Choudhary, A. Kejariwal, and F. Orsini, "On the Runtime-Efficacy Trade-off of Anomaly Detection Techniques for Real-Time Streaming Data," *arXiv preprint arXiv:1710.04735*, 2017.
- [16] A. A. Cook, G. Misirli, and Z. Fan, "Anomaly Detection for IoT Time-Series Data: A Survey," *IEEE Internet of Things Journal*, vol. 7, no. 7, pp. 6481–6494, 2020.
- [17] J. Audibert, P. Michiardi, F. Guyard, S. Marti, and M. A. Zuluaga, "Do deep neural networks contribute to multivariate time series anomaly detection?" *Pattern Recognition*, vol. 132, 2022.
- [18] A. Zhang, S. Deng, D. Cui, Y. Yuan, and G. Wang, "An Experimental Evaluation of Anomaly Detection in Time Series," *Proceedings of the VLDB Endowment (PVLDB)*, vol. 17, no. 3, pp. 483–496, 2024.
- [19] R. Wu and E. J. Keogh, "Current Time Series Anomaly Detection Benchmarks are Flawed and are Creating the Illusion of Progress," *arXiv preprint arXiv:2009.13807*, 2020.
- [20] X. Wang, J. Lin, N. Patel, and M. Braun, "Exact variable-length anomaly detection algorithm for univariate and multivariate time series," *Data Mining and Knowledge Discovery*, 2018.
- [21] J. He, C.-C. M. Yeh, Y. Wu, L. Wang, and W. Zhang, "Mining anomalies in subspaces of high-dimensional time series for financial transactional data," in *Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track*, 2021.
- [22] E. Sylligardos, P. Boniol, J. Paparrizos, P. Trahanias, and T. Palpanas, "Choose Wisely: An Extensive Evaluation of Model Selection for Anomaly Detection in Time Series," *Proceedings of the VLDB Endowment*, vol. 16, no. 11, pp. 3418–3432, 2023.
- [23] S. Kim, K. Choi, H.-S. Choi, B. Lee, and S. Yoon, "Towards a Rigorous Evaluation of Time-series Anomaly Detection," *arXiv preprint arXiv:2109.05257*, 2022-01-04.
- [24] M. Goswami, C. Challu, L. Callot, L. Minorics, and A. Kan, "Unsupervised Model Selection for Time-series Anomaly Detection," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.
- [25] P. Wenig, S. Schmidl, and T. Papenbrock, "TimeEval: A benchmarking toolkit for time series anomaly detection algorithms," in *Proceedings of the VLDB Endowment*, 2022.
- [26] T. Yairi, Y. Kato, and K. Hori, "Fault detection by mining association rules from house-keeping data," in *Proceedings of the International Symposium on Artificial Intelligence, Robotics and Automation in Space (SAIRAS)*, vol. 6, 2001.
- [27] C.-C. M. Yeh, Y. Zhu, L. Ulanova, N. Begum, Y. Ding, H. A. Dau, D. F. Silva, A. Mueen, and E. Keogh, "Matrix Profile I: All Pairs Similarity Joins for Time Series: A Unifying View That Includes Motifs, Discords and Shapelets," in *Proceedings of the International Conference on Data Mining (ICDM)*, 2016, pp. 1317–1322.
- [28] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation Forest," in *Proceedings of the International Conference on Data Mining (ICDM)*, 2008, pp. 413–422.
- [29] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," in *Proceedings of the International Conference on Management of Data (SIGMOD)*, 2000, pp. 93–104.
- [30] P. Boniol, M. Linardi, F. Roncallo, T. Palpanas, M. Meftah, and E. Remy, "Unsupervised and Scalable Subsequence Anomaly Detection in Large Data Series," *The VLDB Journal*, 2021.
- [31] P. Boniol and T. Palpanas, "Series2Graph: Graph-based subsequence anomaly detection for time series," *Proceedings of the VLDB Endowment (PVLDB)*, vol. 13, no. 11, p. 14, 2020.
- [32] M. Thill, W. Konen, and T. Bäck, "Time Series Anomaly Detection with Discrete Wavelet Transforms and Maximum Likelihood Estimation," in *Proceedings of the International Conference on Time Series (ITISE)*, 2017.
- [33] N. Heim and J. E. Avery, "Adaptive Anomaly Detection in Chaotic Time Series with a Spatially Aware Echo State Network," *arXiv preprint arXiv:1909.01709*, 2019-09-02.
- [34] J. Ziegelmeier, "Development and comparison of self-learning modules for automated bench test data analysis of transient flight engine development tests," 2019.
- [35] P. Malhotra, L. Vig, G. Shroff, and P. Agarwal, "Long Short Term Memory Networks for Anomaly Detection in Time Series," in *Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, vol. 23, 2015.
- [36] M. Hahsler and M. Bolaos, "Clustering Data Streams Based on Shared Density between Micro-Clusters," *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, vol. 28, no. 6, pp. 1449–1461, 2016.
- [37] A. Ryzhikov, M. Borisyak, A. Ustyuzhanin, and D. Derkach, "Normalizing flows for deep anomaly detection," *arXiv preprint arXiv:1912.09323*, 2019.
- [38] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: <http://archive.ics.uci.edu/ml>
- [39] A. Ihler, J. Hutchins, and P. Smyth, "Adaptive event detection with time-varying poisson processes," in *Proceedings of the International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 2006, pp. 207–216.
- [40] M. Bachlin, M. Plotnik, D. Roggen, I. Maidan, J. Hausdorff, N. Giladi, and G. Troster, "Wearable Assistant for Parkinson's Disease Patients With the Freezing of Gait Symptom," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 2, pp. 436–446, 2010.
- [41] P. Filonov, A. Lavrentyev, and A. Vorontsov, "Multivariate Industrial Time Series with Cyber-Attack Simulation: Fault Detection Using an LSTM-based Predictive Data Model," *arXiv preprint arXiv:1612.06676*, 2016.
- [42] A. von Birgelen and O. Niggemann, "Anomaly Detection and Localization for Cyber-Physical Production Systems with Self-Organizing Maps," in *IMPROVE - Innovative Modelling Approaches for Production Systems to Raise Validatable Efficiency*, 2018, vol. 8, pp. 55–71.

- [43] Y. Mirsky, T. Doitshman, Y. Elovici, and A. Shabtai, "Kitsune: An Ensemble of Autoencoders for Online Network Intrusion Detection," in *Proceedings of the Network and Distributed System Security Symposium (NDSS)*, 2018.
- [44] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals," *Circulation*, vol. 101, no. 23, 2000-06-13.
- [45] A. Goldberger, L. Amaral, L. Glass, J. Hausdorff, P. Ivanov, R. Mark, J. Mietus, G. Moody, C.-K. Peng, and H. Stanley, "The MIT-BIH Long Term Database," 1992.
- [46] G. B. Moody and R. G. Mark, "MIT-BIH Arrhythmia Database," 1992.
- [47] N. Helwig, E. Pignatelli, and A. Schutze, "Condition monitoring of a complex hydraulic system using multivariate statistics," in *Proceedings of the International Instrumentation and Measurement Technology Conference (I2MTC)*, 2015, pp. 210–215.
- [48] D. Roggen, A. Calatroni, M. Rossi, T. Holleczeck, K. Forster, G. Troster, P. Lukowicz, D. Bannach, G. Pirkl, A. Ferscha, J. Doppler, C. Holzmann, M. Kurz, G. Holl, R. Chavarriaga, H. Sagha, H. Bayati, M. Creatura, and J. d. R. Millan, "Collecting complex activity datasets in highly rich networked sensor environments," in *Proceedings of the International Conference on Networked Sensing Systems (INSS)*, 2010, pp. 233–240.
- [49] Y. Su, Y. Zhao, C. Niu, R. Liu, W. Sun, and D. Pei, "Robust Anomaly Detection for Multivariate Time Series through Stochastic Recurrent Neural Network," in *Proceedings of the International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 2019, pp. 2828–2837.
- [50] S. Greenwald, "The MIT-BIH Supraventricular Arrhythmia Database," 1992.
- [51] S. D. Greenwald, "Improved detection and classification of arrhythmias in noise-corrupted electrocardiograms using contextual information," 1990. [Online]. Available: <https://dspace.mit.edu/handle/1721.1/29206>
- [52] J. Goh, S. Adepu, K. N. Junejo, and A. Mathur, "A Dataset to Support Research in the Design of Secure Water Treatment Systems," in *Proceedings of the International Conference on Critical Information Infrastructures Security (CRITIS)*, vol. 10242, 2017, pp. 88–99.
- [53] C. M. Ahmed, V. R. Palleti, and A. P. Mathur, "WADI: A water distribution testbed for research in the design of secure cyber physical systems," in *Proceedings of the International Workshop on Cyber-Physical Systems for Smart Water Networks (CySWATER)*, 2017, pp. 25–28.
- [54] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (roc) curve." *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.
- [55] V. Raghavan, P. Bollmann, and G. S. Jung, "A critical investigation of recall and precision as measures of retrieval system performance," *ACM Transactions on Information Systems (TOIS)*, vol. 7, no. 3, pp. 205–229, 1989.
- [56] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in *Proceedings of the 23rd International Conference on Machine Learning*, 2006, pp. 233–240.
- [57] S. Sørnbø and M. Ruocco, "Navigating the Metric Maze: A Taxonomy of Evaluation Metrics for Anomaly Detection in Time Series," *arXiv preprint arXiv:2303.01272*, 2023.