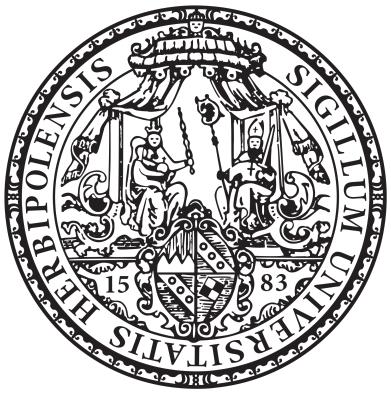
Time Series Anomaly Detection Benchmarking

Philip Spaier

3110375



Seminararbeit

Lehrstuhl für Wirtschaftsinformatik und Business Analytics
Universität Würzburg

Betreuer: Prof. Dr. Gunther Gust
Assistent: Viet Nguyen

W"urzburg, den 05.04.2025

Contents

List of Figures II

List of Tables II

1 Literature Review 1
- 1.1 Time Series Data and Anomaly Detection Definition . . . 1
- 1.2 Relevant Fields . . . 2
- 1.3 Detection Methods . . . 3
  - 1.3.1 Degree of Supervision . . . 3
  - 1.3.2 Architecture . . . 3
  - 1.3.3 Technique . . . 4
- 1.4 Performance Metrics . . . 5
  - 1.4.1 Point-wise or Range-wise . . . 5
  - 1.4.2 Threshold-dependent or Threshold-independent . . . 6
  - 1.4.3 Definition and Classifgication of Metrics . . . 6
- 1.5 State of Benchmarking . . . 7

2 Dataset Analysis 7

3 Replication of TSB-AD Benchmark Results 7

4 Dataset Creation 7

5 Conclusion 7

6 Zitieren und Referenzieren 7

7 Abbildungen 8

8 Tabellen 8

9 Formeln 9

Bibliography 11

A Anhang A 12

List of Figures

1 Siegel der Universität 8

List of Tables

1 Evaluation Measures 7
2 Meine Tabelle 8

Abstract

Eine Kurzzusammenfassung der Vorgehensweise und der wesentlichen Ergebnisse.

Allgemeine Merkmale

- Objektivität: Es soll sich jeder persönlichen Wertung enthalten.
- Kürze: Es soll so kurz wie möglich sein.
- Verständlichkeit: Es weist eine klare, nachvollziehbare Sprache und Struktur auf.
- Vollständigkeit: Alle wesentlichen Sachverhalte sollen enthalten sein.
- Genauigkeit: Es soll genau die Inhalte und die Meinung der Originalarbeit wiedergeben.

1 Literature Review

Time Series Anomaly Detection (TSAD), as a subcategory of the broader field of Anomaly Detection, has seen increased attention since the start of the twenty first century. With the internet having established itself as a persistent and omnipresent force in every imaginable aspect of human life, time series data can be found in abundance. Modern developments in Internet-of-Things (IoT) applications, the digitization of financial data, and a massive rise in the consumption of streaming services have contributed to an exponential growth of time series data [source needed]. This in turn has made the manual search of potential anomalies in many fields completely infeasible, leading to an increased demand for automated anomaly detection methods. While there is a continuously growing repertoire of such automated detection methods, the lack of a generally accepted and reliable benchmark makes not just further developments but also the selection of appropriate models difficult. In the following sections of this literature review, I will provide the reader with a better understanding of context independent Time Series Anomaly Detection, the most commonly applied methods, and the current state of benchmarking.

1.1 Time Series Data and Anomaly Detection Definition

Time Series Data, as used in the rest of this thesis, shall be defined as follows: a sequence of data or observations, typically indexed by or associated with specific timestamps, collected in chronological order over a period of time. For the purpose of analysis, continuous signals must be converted into individual data points. Each datapoint can either represent a binary state (1 or 0), be a numerical value measured on a ratio scale (eg. number of occurrences), or a numerical value measured on an interval scale (eg. temperature on a Celsius scale). A time series with a dimensionality of one (only a single feature) will be referred to as "univariate", while a time series with higher dimensionality (multiple features) will be referred to as "multivariate".
An anomaly will be defined as follows: an abnormal, rarely occurring data point or sequence, that has to be be detectable with exclusively context independent methods. Individual anomalous data points will be referred to as "point based" anomalies. Multiple consecutive anomalous points, each of which might be unremarkable on their own, while displaying unusual behavior as a sequence, will be referred to as "sequence based" or "collective" anomalies (Liu and Paparrizos, 2024, p. 3; Chalapathy and Chawla, 2019, p. 8). A separate category of anomalies would be context dependent ones. Those are data points or sequences, possibly indistinguishable from normal ones if analyzed without context, but if combined with additional information about the field or time series, are considered anomalous (Chalapathy and Chawla, 2019, pp. 7-8). Context dependent anomalies will not be topic of the research presented here.
Given those definitions, Time Series Anomaly Detection is therefore the task of correctly and autonomously identifying anomalies within a given time series.

Insert Images of point vs sequence.

1.2 Relevant Fields

The following is an overview of fields relying on Time Series Anomaly Detection. It is a non-exhaustive list, simply highlighting some of the most prominent use cases to provide context.

Illicit Activity and Fraud Detection: With the global financial system relying primarily on digital transactions, it has become crucial to detect fraudulent activities as quickly and accurately as possible. A particularly obvious example is credit card fraud, creating an estimated yearly loss in the billions of dollar (Zhou, Xun et al., 2018, p. 2). Companies like Visa and Mastercard put great emphasis on being able to detect anomalous transactions in real time to then analyses them and prevent potential harm to their customers (*Visa Acceptance Solutions* 2025). While credit card fraud is a prominent application, the scope of financial anomaly detection extends significantly further, playing a critical role in the operations of stock exchanges, brokerage firms, and banks. These institutions leverage anomaly detection techniques to identify various illicit activities, ensure market integrity, manage operational risks, and comply with stringent regulatory requirements (*Deutsche Börse* 2025).

Healthcare: Healthcare critically relies on analyzing physiological signals, such as those captured by the electrocardiogram (ECG), which provides vital time series data reflecting the heart's electrical activity. While historically, ECG analysis has focused on identifying established patterns of known heart diseases, this approach often fails to detect rare or atypical anomalies that do not fit predefined categories, potentially missing critical conditions. To address this issue, Time Series Anomaly Detection has been introduced for the purpose of detecting such rare anomalies that would go unnoticed by conventional pattern classification (Jiang et al., 2024, p. 1-2).

Website Traffic: A common threat faced by web-services are so called Denial or Service (DoS) and Distributed Denial of Service (DDoS) attacks. These include hitting a webserver with so many requests that the systems becomes inoperational and can no longer service legitimate users (*Bundesamt für Sicherheit in der Informationstechnik* 2025). A significant challenge in detecting these attacks is that the malicious traffic can often mimic normal network traffic, making it difficult for traditional packet-based intrusion detection systems or statistical methods reliant on fixed thresholds to accurately identify attacks, especially when they are hidden within legitimate flows. Time series analysis allows systems to observe and distinguish the instant changes in network traffic that indicate an attack, even when individual packets or simple statistics are insufficient. Time series anomaly detection provides a means to autonomously identify and localize potentially harmful deviations within the network traffic and thereby ensure the availability and reliability of services (Fouladi, Ermiş, and Anarim, 2020, pp. 1-2).

The list extends far beyond the fields named above. Time Series Anomaly Detection can be also found in astronomy (Huijse et al., 2014), earth sciences, manufacturing (Zamanzadeh Darban et al., 2024, p. 1), cybersecurity, and law enforcement (Boniol et al., 2024, p. 1).

1.3 Detection Methods

Detection methods, in common descriptions and within the scientific literature, are often grouped or distinguished by a variety of aspects. This categorization can sometimes lack a consistent taxonomy. To provide a clearer framework, I will now systematically explain and categorize these methods through three key perspectives:

- Degree of supervision
- Architecture
- Technique

1.3.1 Degree of Supervision

Unsupervised models operate on data without any explicit labels distinguishing normal from anomalous instances. While they don't require pre-labeled data, they typically do require a training or fitting phase. During this phase, the model learns the inherent structure, patterns, distributions, or densities from the unlabeled dataset.

Semi-supervised models are trained exclusively on data that is known or assumed to be 'normal.' They do not require labeled anomalies for training. The model learns a precise representation or boundary of this normal behavior. During deployment, any new data instance that significantly deviates from this learned model of normalcy is flagged as an anomaly.

Supervised models require a dataset where both normal and anomalous instances are explicitly labeled beforehand. The model is then trained to learn the distinguishing features or decision boundaries that separate these classes, effectively treating anomaly detection as a (often highly imbalanced) classification problem. (Boniol et al., 2024, pp. 5-6; Liu and Paparrizos, 2024, p. 3; Schmidl, Wenig, and Papenbrock, 2022, p. 3-4).

1.3.2 Architecture

Statistical models identify anomalies by relying on statistical assumptions to detect deviations from expected data distributions. They often involve fitting a distribution model to the data and measuring abnormality based on probabilities or distances from the calculated distribution. Statistical models often require a threshold to be set beforehand(Liu and Paparrizos, 2024, p. 6-7; Fouladi, Ermiş, and Anarim, 2020, p. 1).

Neural Network based models are a collection of distributed, adaptive, non-linear processing units with adjustable weights (Guresen and Kayakutlu, 2011, p. 427). They rely on a training dataset and are often semi-supervised. Deep neural networks, a subcategory of neural networks, model spacial and temporal dependencies (Liu and Paparrizos, 2024, p. 6-7; Zamanzadeh Darban et al., 2024, p. 6).

Foundational Models utilize transfer learning, using knowledge from a different class of tasks and then applying it on the target task. These models are pre-trained and are then being fine-tuned (Bommasani et al., 2022, p. 4). In the context of TSAD, those models are GPT

models fine tuned on time series data, general purpose time series models, or originally time series classification models now used for anomaly detection (Liu and Paparrizos, 2024, p. 7).

1.3.3 Technique

Distance based models work on the idea that anomalous points or sequences will further away when using a distance measurement. They can be either be compared to their nearest neighbor, all other points/subsequences, or cluster centers (Schmidl, Wenig, and Papenbrock, 2022, p. 6). Such distances are calculated in various ways depending on the model and implementation, with the most common definitions being the Euclidean distance or the Z-normalized Euclidean distance. Distance based models use only the x- and y-axis data, with no labels being required (Boniol et al., 2024, p. 8).

Forecasting models learn the normal patters of a time series and, often using a sliding context window, forecast the next datapoint in the series. The forecasted and actual data points are then compared, with the difference being used for an anomaly score. Given a high enough anomaly score, a point is considered an anomaly. Such models are usually semi-supervised (Schmidl, Wenig, and Papenbrock, 2022, p. 4-5).

Isolation Tree Models use ensembles of random trees, selecting random features and splits, to separate points or sequences from each other. It operates on the idea that anomalies require fewer steps to be separated from the rest of the data than normal points/sequences. For each point/sequence, the distance from the root is calculated. The shorter a distance is, the more likely is a point/sequence to be an anomaly. These models can be both unsupervised and supervised (Schmidl, Wenig, and Papenbrock, 2022, p. 6-7)

Distribution based models estimate a distribution of the time series and then score individual points or sequences as anomalous or normal based on it. Anomalous points are expected to have a low probability. Alternatively to probabilities, the anomaly score can also be calculated using likelihoods or distances. These models are generally unsupervised or occasionally semi-supervised (Schmidl, Wenig, and Papenbrock, 2022, p. 6).

Graph based models methods turn time series data, or parts of it, into a graph structure. This graph represents the different types of patterns (subsequences) found in the data as nodes, and how these patterns follow each other over time as connections (edges) between the nodes. Anomalies are then determined based on usual structures or behaviors found in the graph (Boniol et al., 2024, p. 23-24). Graph based time series models can be further divided in multiple subcategories, including AutoEncoder- and GAN-based methods, as well as predictive graph models (Ho, Karami, and Armanfard, 2025).

Reconstruction models learn a time series' features and patterns by encoding normal data into a low dimensional space. Given a test dataset, they compress test data and reconstruct it using their model based on that low-dimensional space. Should a point or sequence of this reconstructed version deviate substantially from the actual data, then it is labeled as anomalous. These models are often considered semi-supervised because they typically use

normal labeled data for training. However, models that do not rely on a training dataset and instead directly encode and reconstruct the test data also exist, operating in an unsupervised manner. (Schmidl, Wenig, and Papenbrock, 2022, p. 5).

Encoder based models operate similarly to reconstruction models. They compress a given time series into a low-dimensional representation, but instead of reconstructing it, they directly compare this compressed version to their model of normal time series. Anomalous points or sequences might have unusual encoded representations, and their deviations from the normal model are then used to calculate an anomaly score (Schmidl, Wenig, and Papenbrock, 2022, p. 5-6).

1.4 Performance Metrics

For the effective evaluation of a models performance, as defined by Paparrizos et al., 2022, metrics have to fullfil the following criteria:

- *Robustness to Lag:* The evaluation measure should be insensitive to slight temporal shifts or lags in anomaly scores.
- *Robustness to Noise:* The evaluation measure should be stable and unaffected by noise in the anomaly scores.
- *Robustness to Anomaly Cardinality Ratio:* The evaluation measure's score should not be influenced by the proportion of anomalies in the data.
- *High Separability between Accurate and Inaccurate Methods:* The measure must effectively distinguish between accurate and inaccurate detection methods.
- *Consistency:* The measure should produce repeatable scores for similar data and consistently rank different methods.

Commonly applied performance measures for TSAD can generally be classified based on two characteristics: Point-wise or Range-wise, and Threshold-dependent or Threshold-independent.

1.4.1 Point-wise or Range-wise

Point-wise evaluation measures look at each anomalous point independently, determining in a binary fashion whether a model classified them correctly as normal or anomalous (Liu and Paparrizos, 2024, p. 7). These measures suffer from a variety of issues. Most crucially, they can unfairly penalize methods that detect only part of an anomalous range or whose detection peak doesn't perfectly align with the labeled range. Further more, they are sensitive to temporal lag. Should an anomalous data point be detected slightly before or after the actual anomaly occurs, a fully point-wise metric will score it with an unreasonably low score (Paparrizos et al., 2022, p. 2778).
Range-wise measures look at anomalies not just from the perspective of individual points but take sequences into consideration. For anomalous sequences, their evaluation can involve determining how much the detected and the actual sequence overlap. Additionally, such measures may incorporate strategies like adequately handling lag (e.g., by considering

an anomaly detected if it's within a specified range of an actual one, even if not at the exact spot) or including a cardinality factor to penalize models that incorrectly segment anomalies (such as detecting multiple short ones for a single large event, or vice-versa) (Liu and Paparrizos, 2024, p. 7).

1.4.2 Threshold-dependent or Threshold-independent

Threshold-dependent measures require a threshold to be set that determines whether an anomaly score classifies a value as anomalous or normal. This can be done based on statistical assumptions, or using dynamic algorithms that adjust to the data and results (Boniol et al., 2024, p. 38-39). Setting these thresholds automatically, however, is often difficult when working with large and diverse datasets, and the chosen thresholds can drastically change a metric's accuracy. Noise and the normal-to-anomalous ratio in a time series can be particularly problematic (Paparrizos et al., 2022, p. 2777-2778).
Threshold-independent measures evaluate the performance of a time series anomaly detection method without needing a specific score cutoff to decide what constitutes an anomaly. Instead of relying on a fixed threshold, they assess how effectively the method's anomaly scores rank true anomalies higher than normal data points across the entire range of scores (Boniol et al., 2024, p. 39-41).

1.4.3 Definition and Classifgication of Metrics

The following defines the most commonly used metrics and classifies them into the above described categories (Paparrizos et al., 2022, p.2776-2780):

- *Precision / Range Precision:* number of correctly identified anomalies over all anomalies.
- *Recall (TPR) / Range Recall:* number of correctly identified anomalies over all anomalies.
- *F-Score / Range F-Score:* Harmonic Mean of Precision and Recall.
- *False Positive Rate (FPR):* number of points wrongly identified as anomalies over the total number of normal points.
- *AUC-ROC:* area under the curve corresponding to TPR on the y-axis and FPR on the x-axis at all threshold levels.
- *AUC-Precision:* area under the curve corresponding to the Recall on the x-axis and Precision on the y-axis at all threshold levels.
- *VUS-ROC:* generating multiple ROC curves for a range of different buffer lengths. These stacked ROC curves form a 3D surface, and VUS-ROC is the volume beneath this surface.
- *VUS-Precision:* generating multiple Precision-Recall curves for a range of different buffer lengths. These stacked PR curves form a 3D surface, and VUS-Precision (VUS-PR) is the volume beneath this surface.

	Threshold-dependent	Threshold-independent
Point-wise	Precision Recall False Positive Rate F-Score	AUC-ROC AUC-Precision
Range-wise	Range Precision Range Recall Range F-Score	VUS-ROC VUS-Precision

Table 1: Evaluation Measures

1.5 State of Benchmarking

2 Dataset Analysis

Das ist fett gedruckter Text.

Das ist kursiver Text.

Auflistungen sind oft hilfreich für die Strukturierung:

- Erster Eintrag
- Zweiter Eintrag

Nummerierte Aufzählungen sind oft hilfreich für Reihenfolgen:

1. Erster Eintrag
2. Zweiter Eintrag

3 Replication of TSB-AD Benchmark Results

4 Dataset Creation

5 Conclusion

6 Zitieren und Referenzieren

Beiträge in Fachzeitschriften wie **clemen1989combining** oder Konferenzartikel wie **he2017mask** werden auf diese Weise im Text zitiert. In anderen Fällen möchte man aber in Klammern zitieren (**clemen1989combining**), auch mit mehreren Autoren (**clemen1989combining**; **baumol1958warehous**
he2017mask).

Bei Monographien muss eine Seitenzahl mit angegeben werden (**chollet2018deep**).

So wird eine Webquelle zitiert: **shiny1**. Es kann bei kurzen Informationen im Internet aber

auch reichen die Adresse[1] als Fußnote einzubetten.

So werden andere Teile der Arbeit referenziert: Kapitel 1, Gleichung 1 zeigt...

So verweisen wir auf eine Fußnote [2].

7 Abbildungen

Abbildungen erfordern das package *graphicx*. Idealerweise verwendet man Vektorgrafiken oder hochaufgelöste Bitmaps. Eine gute Variante ist das Verwenden von PDFs.

Figure 1: Siegel der Universität

8 Tabellen

Die Tabular-Umgebung gibt die Anzahl Spalten an, deren Orientierung, Breite und evtl. Zwischenlinien.

Table 2: Meine Tabelle

col1	col2	col3
Multiple row	cell2	cell3
	cell5	cell6
	cell8	cell9

[1] `https://shiny.rstudio.com/tutorial/written-tutorial/lesson1/`
[2] dies ist eine Fußnote

9 Formeln

$$\sum_{i=1}^{N} x_i \tag{1}$$

References

Bommasani, Rishi et al. (2022). *On the Opportunities and Risks of Foundation Models*. arXiv: 2108.07258 [cs.LG]. URL: https://arxiv.org/abs/2108.07258.

Boniol, Paul, Qinghua Liu, Mingyi Huang, Themis Palpanas, and John Paparrizos (2024). *Dive into Time-Series Anomaly Detection: A Decade Review*. arXiv: 2412.20512 [cs.LG]. URL: https://arxiv.org/abs/2412.20512.

Bundesamt für Sicherheit in der Informationstechnik (2025). https://www.bsi.bund.de/EN/Themen/Verbraucherinnen-und-Verbraucher/Cyber-Sicherheitslage/Methoden-der-Cyber-Kriminalitaet/DoS-Denial-of-Service/dos-denial-of-service_node.html. Accessed: 2025-05-10.

Chalapathy, Raghavendra and Sanjay Chawla (2019). *Deep Learning for Anomaly Detection: A Survey*. arXiv: 1901.03407 [cs.LG]. URL: https://arxiv.org/abs/1901.03407.

Deutsche Börse (2025). https://www.deutsche-boerse-cash-market.com/dbcm-en/about-us/organisation-of-the-fwb/market-surveillance-in-germany/market-surveillance-21856?frag=249060. Accessed: 2025-05-10.

Fouladi, Ramin Fadaei, Orhan Ermiş, and Emin Anarim (2020). "A DDoS attack detection and defense scheme using time-series analysis for SDN". In: *Journal of Information Security and Applications* 54, p. 102587. ISSN: 2214-2126. DOI: https://doi.org/10.1016/j.jisa.2020.102587. URL: https://www.sciencedirect.com/science/article/pii/S2214212620307560.

Guresen, Erkam and Gulgun Kayakutlu (2011). "Definition of artificial neural networks with comparison to other networks". In: *Procedia Computer Science* 3. World Conference on Information Technology, pp. 426–433. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2010.12.071. URL: https://www.sciencedirect.com/science/article/pii/S1877050910004461.

Ho, Thi Kieu Khanh, Ali Karami, and Narges Armanfard (2025). *Graph Anomaly Detection in Time Series: A Survey*. arXiv: 2302.00058 [cs.LG]. URL: https://arxiv.org/abs/2302.00058.

Huijse, Pablo, Pablo A. Estevez, Pavlos Protopapas, Jose C. Principe, and Pablo Zegers (2014). "Computational Intelligence Challenges and Applications on Large-Scale Astronomical Time Series Databases". In: *IEEE Computational Intelligence Magazine* 9.3, pp. 27–39. DOI: 10.1109/MCI.2014.2326100.

Jiang, Aofan, Chaoqin Huang, Qing Cao, Yuchen Xu, Zi Zeng, Kang Chen, Ya Zhang, and Yanfeng Wang (2024). *Anomaly Detection in Electrocardiograms: Advancing Clinical Diagnosis Through Self-Supervised Learning*. arXiv: 2404.04935 [cs.CV]. URL: https://arxiv.org/abs/2404.04935.

Liu, Qinghua and John Paparrizos (2024). "The elephant in the room: Towards a reliable time-series anomaly detection benchmark". In: *Advances in Neural Information Processing Systems* 37, pp. 108231–108261.

Paparrizos, John, Paul Boniol, Themis Palpanas, Ruey S. Tsay, Aaron Elmore, and Michael J. Franklin (July 2022). “Volume under the surface: a new accuracy evaluation measure for time-series anomaly detection”. In: *Proc. VLDB Endow.* 15.11, pp. 2774–2787. ISSN: 2150-8097. DOI: 10.14778/3551793.3551830. URL: https://doi.org/10.14778/3551793.3551830.

Schmidl, Sebastian, Phillip Wenig, and Thorsten Papenbrock (2022). “Anomaly Detection in Time Series: A Comprehensive Evaluation”. In: *Proceedings of the VLDB Endowment (PVLDB)* 15.9, pp. 1779–1797. DOI: 10.14778/3538598.3538602.

Visa Acceptance Solutions (2025). https://www.visaacceptance.com/en-us/solutions/ai-driven-fraud-management.html. Accessed: 2025-05-10.

Zamanzadeh Darban, Zahra, Geoffrey I. Webb, Shirui Pan, Charu Aggarwal, and Mahsa Salehi (Oct. 2024). “Deep Learning for Time Series Anomaly Detection: A Survey”. In: *ACM Computing Surveys* 57.1, pp. 1–42. ISSN: 1557-7341. DOI: 10.1145/3691338. URL: http://dx.doi.org/10.1145/3691338.

Zhou, Xun, Cheng, Sicong, Zhu, Meng, Guo, Chengkun, Zhou, Sida, Xu, Peng, Xue, Zhenghua, and Zhang, Weishi (2018). “A state of the art survey of data mining-based fraud detection and credit scoring”. In: *MATEC Web Conf.* 189, p. 03002. DOI: 10.1051/matecconf/201818903002. URL: https://doi.org/10.1051/matecconf/201818903002.

A Anhang A

Hiermit versichere ich, die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie die Zitate deutlich kenntlich gemacht zu haben.
Ich erkläre weiterhin, dass die vorliegende Arbeit in gleicher oder ähnlicher Form noch nicht im Rahmen eines anderen Prüfungsverfahrens eingereicht wurde.

Würzburg, den 15. Mai 2025 VORNAME NACHNAME