TU Delft

Delft University of Technology

Unmasking overestimation

a re-evaluation of deep anomaly detection in spacecraft telemetry

Herrmann, Lars; Bieber, Marie; Verhagen, Wim J.C.; Cosson, Fabrice; Santos, Bruno F.

DOI
10.1007/s12567-023-00529-5

Publication date
2024

Document Version
Final published version

Published in
CEAS Space Journal

Citation (APA)
Herrmann, L., Bieber, M., Verhagen, W. J. C., Cosson, F., & Santos, B. F. (2024). Unmasking overestimation: a re-evaluation of deep anomaly detection in spacecraft telemetry. *CEAS Space Journal*, *16*(2), 225-237. https://doi.org/10.1007/s12567-023-00529-5

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

CEAS Space Journal
https://doi.org/10.1007/s12567-023-00529-5

ORIGINAL PAPER

Unmasking overestimation: a re-evaluation of deep anomaly detection in spacecraft telemetry

Lars Herrmann[1] · Marie Bieber[1] · Wim J. C. Verhagen[2] · Fabrice Cosson[3] · Bruno F. Santos[1]

Received: 5 June 2023 / Revised: 28 October 2023 / Accepted: 2 November 2023
© The Author(s) 2024

Abstract

As the volume of telemetry data generated by satellites and other complex systems continues to grow, there is a pressing need for more efficient and accurate anomaly detection methods. Current techniques often rely on human analysis and preset criteria, presenting several challenges including the necessity for expert interpretation and continual updates to match the dynamic mission environment. This paper critically examines the use of deep anomaly detection (DAD) methods in addressing these challenges, evaluating their efficacy on real-world spacecraft telemetry data. It exposes limitations in current DAD research, highlighting the tendency for performance results to be overestimated and suggesting that simpler methods can sometimes outperform more complex DAD algorithms. By comparing established metrics for anomaly detection with newly proposed ones, this paper aims to improve the evaluation of DAD algorithms. It underscores the importance of using less accuracy-inflating metrics and offers a comprehensive comparison of DAD methods on popular benchmark datasets and real-life satellite telemetry data. Among the DAD methods examined, the LSTM algorithm demonstrates considerable promise. However, the paper also reveals the potential limitations of this approach, particularly in complex systems that lack a single, clear predictive failure channel. The paper concludes with a series of recommendations for future research, including the adoption of best practices, the need for high-quality, pre-split datasets, and the investigation of other prediction error methods. Through these insights, this paper contributes to the improved understanding and application of DAD methods, ultimately enhancing the reliability and effectiveness of anomaly detection in real-world scenarios.

Keywords Deep anomaly detection · Real-life satellite telemetry data · Anomaly detection metrics · Time-series anomaly detection

✉ Bruno F. Santos
B.F.Santos@tudelft.nl

Lars Herrmann
peschkelars@googlemail.com

Marie Bieber
M.T.Bieber@tudelft.nl

Wim J. C. Verhagen
wim.verhagen@rmit.edu.au

Fabrice Cosson
Fabrice.Cosson@esa.int

[1] Faculty of Aerospace Engineering, Delft University of Technology, Kluyverweg 1, 2629 HS Delft, Zuid-Holland, The Netherlands

[2] Aerospace Engineering and Aviation, RMIT University, Carlton, VIC 3053, Australia

[3] European Space Research & Technology Centre, ESA, Keplerlaan 1, 2200 AG Noordwijk, Zuid-Holland, The Netherlands

Abbreviations

AE	Autoencoder
AUC	Area under the curve
CNN	Convolutional neural network
DAD	Deep anomaly detection
ESA	European Space Agency
GAN	Generative adversarial networks
GAT	Graph attention network
GMM	Gaussian mixture model
GNN	Graph neural networks
kNN	k-nearest neighbor
LSTM	Long short-term memory
MSL	Mars Science Laboratory
NASA	National Aeronautics and Space Administration
NPT	Nonparametric dynamic thresholding
OCSVM	One-class support vector machine
PA	Point adjust
PCA	Principal component analysis

Published online: 17 February 2024

RNN	Recurrent neural network
SMAP	Soil Moisture Active Passive Satellite
SMD	Server machine dataset
SWAT	Secure water treatment
USAD	UnSupervised anomaly detection
VAE	Variational autoencoder

1 Introduction

Satellites and other complex systems generate increasing amounts of telemetry data that can be analyzed by terrestrial systems. Monitoring this data is crucial for ensuring the success of spacecraft operations and missions. Anomaly detection is a key method to prevent spacecraft loss due to undetected flaws or slow responses to hazards. However, most current anomaly detection methods rely on human evaluation of aggregated data and out-of-limit checks with established criteria. These methods have significant disadvantages, as they require specialized expertise and effort to organize and analyze the data.

In the coming years, these challenges are expected to intensify due to ongoing advancements in computer and storage capacities. As a result, the volume of telemetry data will significantly increase, placing greater demands on technical resources and data aggregation techniques. Deep learning for anomaly detection in high-dimensional time-series data has shown promising results with recent advancements in neural network architecture and increased processing power. Some deep learning algorithms perform better than traditional anomaly detection techniques on real-world time-series challenges [1], with reported F_1 scores greater than 0.9, indicating highly accurate deep anomaly detection (DAD) capabilities. However, the widely used point adjust (PA) method [2] in modern DAD research has faced criticism in recent publications, mainly due to its tendency to overestimate accuracy [3–6]. This paper aims to address the aforementioned shortcomings by applying and evaluating DAD methods on real-world spacecraft telemetry data. In doing so, it contributes to the state of the art in the following ways:

- We compare established metrics for anomaly detection with newly proposed metrics, aiming to improve the measurement and evaluation of anomaly detection algorithms. By considering a range of metrics, we provide a more robust framework for assessing algorithm performance.
- We investigate the performance of anomaly detection methods with different levels of complexity using metrics that are less likely to overestimate accuracy. This allows for a more comprehensive assessment of the methods' effectiveness.
- The comparison and evaluation of the anomaly detection methods are conducted on two popular benchmark datasets as well as on a real-life dataset of satellite telemetry data. This provides a comprehensive and realistic assessment of the methods' performance in different scenarios.

The remainder of this paper is organized as follows. Section 2 provides an introduction to the background and current state of the art in anomaly detection. Section 3 describes the metrics, anomaly detection algorithms, and thresholding methods employed in this study for the comparison and evaluation of the methods. Section 4 provides a comprehensive overview of the datasets utilized in the case studies, offering pertinent information about each dataset. Furthermore, it presents the detailed results obtained from the case studies and engages in an in-depth discussion of the findings. Finally, in Sect. 5, we summarize the key findings and limitations of the study and suggest potential directions for further research and improvement in the field of DAD.

2 Literature review

2.1 State of the art

Numerous studies have investigated various aspects of DAD on satellite data, exploring diverse topics within this field. At the core of anomaly detection algorithms lies the data they analyze. Time-series data, characterized by a sequence of time-dependent variables, represents a distinctive type of input data with its own unique properties and challenges [7, 8]. Various efforts to characterize the nature of anomalous data have been documented in the literature. Hawkins' definition of an anomaly is: "an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism" [9]. According to Barnett and Lewis, an anomaly is "an observation or subset of observations, which appears to be inconsistent with the remainder of that set of data" [10]. Time-series anomalies are typically classified into three types: Point, Contextual, and Collective anomalies [11]. They can also be further divided into more specific subsets that depend on the domain being analyzed. For example, Tang et al. defined six patterns to categorize vibration anomalies [12].

Choi et al. categorized classic anomaly detection approaches into time/frequency domain analysis, statistical models, distance-based models, auto-regressive models, and clustering models [7]. Basora et al. groups distance-based and clustering models together and expands the classification to include ensemble-based, domain-based, and subspace-based methods [13]. Telemetry data from spacecraft is

typically analyzed in the time domain using simple limit checking with upper and lower limits for the observed values [14]. However, fixed thresholds can be limiting for dynamic systems. To address this, adaptive limit checking has been developed [15].

Traditional methods for anomaly detection face limitations in scaling with increasing dimensionality and large data volumes. In contrast, deep learning methods, specifically DAD, have shown superior performance in such scenarios [16–18]. Recent research has focused on DAD to overcome these challenges, and it has been successfully applied to various tasks across different domains.

The advancement in deep learning architecture and increase in data and computational resources have resulted in deep learning models performing some tasks at a human level, even surpassing it in certain cases. This has also fueled extensive research in the field of diagnostics. The most common deep learning architectures that are used in anomaly detection are convolutional neural network (CNN) and recurrent neural network (RNN), specifically the long short-term memory (LSTM) networks [19], autoencoder (AE) and variational autoencoder (VAE) [20], generative adversarial networks (GAN) [21], graph attention network (GAT) [22] and transformers [23].

DAD models aim to minimize an objective loss function during training, which depends on the model architecture and relates to abnormality decision criteria. These models output an anomaly score, which is a numeric value that indicates the probability of a sample being abnormal, and samples are labeled as anomalous when the score exceeds a certain threshold. While domain experts used to set the threshold empirically, it is now determined based on training results, either through performance evaluation on validation data for labeled data or by using extreme value theory for non-labeled data [24]. DAD can be categorized into three types depending on the method used to calculate the anomaly score: reconstruction error, prediction error, and dissimilarity [7]. However, the first two criteria are more commonly used than the third.

Autoencoders, variational autoencoders, generative adversarial networks, and transformers are examples of models that typically use reconstruction errors to obtain an anomaly score. They learn low-dimensional representations of the data and map them to the input space to calculate residuals by comparing the reconstructed values with the original data. Reconstruction-based methods assume that anomalies lose information when mapped to a lower dimensional space and cannot be effectively reconstructed. Therefore, high reconstruction errors suggest a high chance of being anomalous [25]. Prediction error methods use a model to fit the given data and predict future values. The difference between the model output and the actual values is used to identify anomalies. Commonly used models for prediction error anomaly scores include LSTM, CNN, graph neural networks (GNN), and transformers. Dissimilarity-based models measure distance or similarity between data instances. Objects that are distant from a cluster or distribution are considered anomalies. A table of current state-of-the-art algorithms with their corresponding architecture, anomaly criterion and benchmark scores can be seen in Table 1.

2.2 Limitations in the state of art

While many papers claim to have unsupervised algorithms, several of them suffer from data leakage, especially many of those that achieve the best results. Data leakage refers to the utilization of information during the

Table 1 Performance comparison of recent deep anomaly detection algorithms

Name	Year	Network type	Anomaly criterion	F_{1PA} Benchmark results			
				SMD	MSL	SMAP	SWAT
Anomaly transformer [18]	2021	Transformer	Reconstruction error, association discrepancy	92,33	93,59	96,69	94,07
ImDiffusion [28]	2023	Transformer, diffusion model	Prediction error	94,88	87,79	91,75	87,09
BeatGAN [29]	2019	GAN, autoencoder	Reconstruction error	78,1	87,53	69,61	73,92
TadGAN [25]	2020	GAN	Reconstruction error, critique Score		62,3	70,4	
MAD-GAN [30]	2019	LSTM, GAN	Reconstruction error		87,47	81,31	0,77
MTAD-GAT [22]	2020	GAT, Attention	Prediction error, reconstruction error		90,84	90,13	
OmniAnomaly [26]	2019	RNN, VAE	Reconstruction error	88,57	89,89	84,34	
THOC [17]	2020	RNN, One-class network	Dissimilarity		93,67	95,18	88,09
USAD [31]	2020	Autoencoder, adverse training	Reconstruction error	93,82	91,09	81,86	84,6
GTA [32]	2021	Transformer, GNN	Prediction error		91,11	90,04	91
LSTM [14]	2018	LSTM	Prediction error		69	71	

Server Machine Dataset (SMD) [26], Mars Science Laboratory (MSL) and Small Active Passive Satellite (SMAP) [14], Secure Water Treatment (SWAT) [27]

model training process that would not be available at prediction time. Algorithms that really provide unsupervised results often rely on extreme value theory [14, 22, 25, 26, 29] or use a discriminator network [29]. However, many top-performing algorithms create the illusion of being unsupervised but actually utilize the test data to determine the threshold [17, 31, 32].

Classic cross-validation with a standard split can be considered data leakage when applied to most time-series datasets. This is because it relies on the assumptions of independence and identical distribution, which may not hold in many real-world scenarios. The independence assumption implies that the values in the time series are not influenced by previous or future values. However, in many engineering systems, there are dynamics and interdependencies that violate this assumption. For example, in a mechanical system, the current state of the system may depend on its past states or external factors. This violates the independence assumption and can lead to biased and inaccurate performance estimates when using standard cross-validation. Similarly, the identical distribution assumption assumes that each observation in the time series is drawn from the same underlying probability distribution. However, in engineering systems, it is common for the distribution to change over time due to factors such as wear and tear, aging, or external influences. Therefore, the identical distribution assumption may not hold, and using standard cross-validation can introduce bias and inaccuracies in the evaluation of anomaly detection methods. There exist other methods, for example cross-validation on a rolling basis, that could be used. Additionally, some algorithms use the input of the assumed or known fraction of outliers to determine the threshold [18, 33]. This approach has two issues. Firstly, the true outlier fraction is often unknown in real-life datasets. Secondly, it is not appropriate to apply the outlier fraction to the training dataset, as the training data is typically assumed to consist of normal instances.

Recent DAD research has reported high anomaly detection scores, leading to a perceived increase in accuracy. However, many studies use a method called PA [14] which artificially inflates metric scores and significantly improves real positive identification. Some authors have identified problems with PA and proposed new metrics [3–6]. They found out that many DAD algorithms do perform worse than a random signal or an untrained network when using PA to compare the results, demonstrating its inherent flaw. Additionally, the benchmark datasets used by the research community to compare DAD algorithms have been criticized. Wu and Keogh argue that these benchmarks are flawed and create an illusion of progress due to issues such as triviality, unrealistic anomaly density, mislabeled ground truth, and run-to-failure bias [34].

3 Methodology

3.1 Deep anomaly detection

This study seeks to provide a comprehensive representation of the DAD research field, focusing on the two main anomaly detection criteria: reconstruction error and prediction error. This exploration aims to discern whether increasing algorithmic complexity directly corresponds to more precise anomaly detection. To do so, algorithms of varying complexity were chosen, each demonstrating different strengths and attributes in their performance. The selection process encompassed both traditional and innovative anomaly detection techniques. The LSTM algorithm uses prediction error as its anomaly detection criterion, while the USAD and anomaly transformer operate based on reconstruction error. The anomaly transformer algorithm additionally introduces a novel anomaly detection metric, the 'association discrepancy', showcasing the current trend of inventing new techniques to enhance detection capabilities. Furthermore, LSTM algorithm comes from the same domain as the telemetry data of satellites, which is considered to provide a unique perspective to the research. The selection process also considered performance capabilities, with one algorithm chosen specifically for its impressive results in numerous instances.

After careful consideration, the three algorithms selected were: the anomaly transformer, USAD, and LSTM. The anomaly transformer, the most complex among the three, features a transformer network, a learnable Gaussian kernel, and two-phase learning, and introduces its own 'association discrepancy'. USAD, a moderately complex model, necessitates the use of two autoencoders trained in two phases, one of which involves adversarial training. LSTM, on the other hand, represents the simpler end of the spectrum, involving the training of a basic LSTM network to fit the data. These choices ensure a wide-ranging exploration of the current state of DAD research.

3.1.1 Anomaly transformer: time-series anomaly detection with association discrepancy [18]

Xu et al. have proposed an "anomaly transformer" that utilizes the global representation capability of transformers to handle long sequences effectively. The authors use the self-attention mechanism to calculate "association

discrepancy" between "prior association" and "series association". "Series association" refers to the association of a specific time point with the entire data of the chosen sliding window, whereas "prior association" refers to the association with the adjacent region represented by a learnable Gaussian kernel. The assumption of the dataset is that anomalies are rare, and most of the data is "normal." Therefore, a normal data point would have a high association with the whole data series, while an abnormal point would have a higher association with adjacent points containing more abnormal patterns due to continuity. The difference between the prior association and the series association is called association discrepancy. A low discrepancy indicates an anomaly, while a high discrepancy indicates a normal point. The algorithm employs minimax association learning, which is depicted in Fig. 1. In the "minimize" phase, the prior association is adjusted to approximate the series association and adapt to the temporal patterns to decrease the association discrepancy. In the "maximize" phase, the series association is optimized to increase the association discrepancy and focus more on the non-adjacent horizon.

This paper has demonstrated excellent results on benchmark datasets and is one of the first to utilize transformers. The authors use a new association-based detection criterion, which they pair with reconstruction error to obtain an anomaly score. They use the outlier fraction to determine the threshold.

The authors of the study concatenated individual anomaly sequences from the test datasets into a single dataset file [35], which raises two issues. Firstly, calculating the metric score by summing up the classes of the confusion matrix and calculating the metric using this sum can lead to inflated scores. This is explained further in Subsect. 3.2 where an improved method is suggested and used in our experiments. Secondly, by concatenating the anomaly sequences, the dataset becomes discontinuous. The anomaly sequences are not necessarily related to each other and are from different times. This means that at the time step where one sequence transitions to another, there will be a sudden difference in values. Datasets suffer from run-to-failure bias, where anomalies often occur toward the end of a sequence [34]. The sudden change in input at the transition point could lead to flagging the entire time window around that point as an anomaly. While the algorithm correctly detects the anomaly, it may only detect the beginning of a new sequence rather than an actual anomaly. Furthermore, the anomaly sequences in the dataset are independent of each other involving various subsystems and channel types. Therefore, the training and testing is run per anomaly sequence in our experiments.

3.1.2 LSTM: detecting spacecraft anomalies using LSTMs and nonparametric dynamic thresholding [14]

This algorithm utilizes LSTM and nonparametric dynamic thresholding (NPT) to detect spacecraft anomalies. During training, the LSTMs are fitted to the normal operating data of the spacecraft to predict future telemetry data. However, while the input is multivariate, the algorithm only predicts one channel (feature) of the data stream. Therefore the performance of the algorithm might depend on the channel that was selected. The trained LSTMs are then used to generate anomaly scores by calculating the prediction error in the testing phase. In the testing phase, the NPT algorithm is used to set a threshold for the anomaly scores generated by the LSTMs. The threshold is dynamically adjusted based on the past anomaly scores, which allows for the detection of anomalies with varying degrees of severity. Additionally, error pruning techniques are used to ensure that anomalous sequences are not considered as a result of regular noise within a stream. This helps to filter out false positives and improves the accuracy of anomaly detection by focusing on significant deviations from normal patterns. Overall, this algorithm utilizes

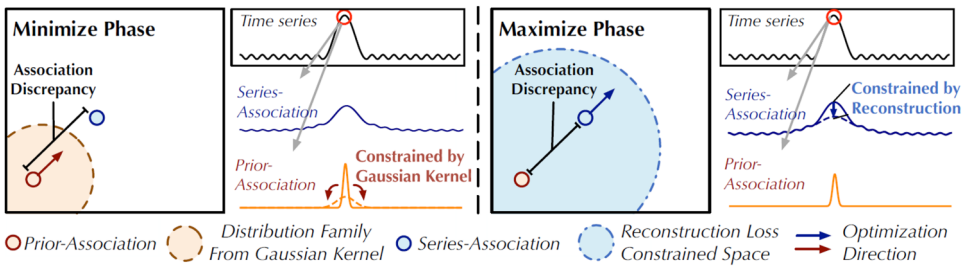

Fig. 1 Minimax association learning as shown by Xu et al. [18]. At the minimize phase, the prior association minimizes the association discrepancy within the distribution family derived by Gaussian kernel. At the maximize phase, the series-association maximizes the association discrepancy under the reconstruction loss

LSTMs to model the spacecraft's normal behavior and dynamic thresholding to detect deviations from the normal behavior, enabling the detection of spacecraft anomalies in real time.

3.1.3 USAD: unsupervised anomaly detection on multivariate time series [31]

Audibert et al. have developed an algorithm that employs two adversarial autoencoder networks, inspired by GAN, to achieve high stability, robustness, and training speed without compromising accuracy. Unsupervised anomaly detection (USAD) consists of an encoder network and two decoder networks, which are combined into an architecture that includes two autoencoders, AE1 and AE2, sharing the same encoder network. The architecture can be seen in Fig. 2. The training of USAD is carried out in two phases. In the first phase, the two autoencoders are trained to reconstruct the normal input windows. In the second phase, the two autoencoders are trained in an adversarial manner, where AE1 attempts to deceive AE2 while AE2 tries to distinguish between real and reconstructed data. The encoder–decoder network is trained on the normal data to learn the temporal patterns and correlations between the variables in the time series. The anomaly scoring mechanism then uses the reconstruction error of the network to generate an anomaly score for each data point. The adversarial training of the encoder–decoder architecture is shown to amplify the reconstruction error and improve stability compared to GAN methods. Additionally, the algorithm introduces a sensitivity threshold that can be adjusted without retraining the model to increase or decrease the detection sensitivity. They use grid search to determine the threshold that gives the best F_1 score.

3.2 Metrics

The F_1 score is a commonly used metric for evaluating time-series anomaly detection algorithms. It is important to note that most studies use the PA technique before scoring the performance of an algorithm, as it was shown that using the F_1 score without PA does underestimate the the detection capability [3]. For example, the authors of the anomaly transformer say: "We adopt the widely-used adjustment strategy... This strategy is justified from the observation that an abnormal time point will cause an alert and further make the whole segment noticed in real-world applications".

The authors of the USAD algorithm rightfully criticise the overestimation of the PA method when another algorithm performs better on the dataset by stating: "..the advantage obtained with the point-adjust which validates whole segments of good prediction despite having potentially missed several abnormalities". However, they continue to use the technique on the other datasets not stating the results without the PA method.

However, recent studies in deep anomaly detection report high F_1 scores, leading to a perceived increase in accuracy. This is because they use the PA technique which was first used by Xu et al. [2]. The principle of PA can be seen in Fig. 3, where all instances in an anomalous sequence are considered true positives when at least one anomaly is detected within the sequence. This technique greatly amplifies the detection of true positives and artificially inflates the F_1 score [3–6]. Hence, they propose new metrics.

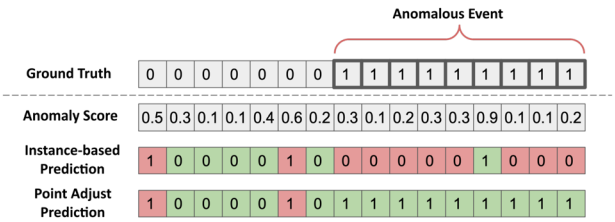

Fig. 3 This figure taken from Doshi et al. [4] demonstrates the comparison between the commonly used PA evaluation method and the traditional instance-based evaluation

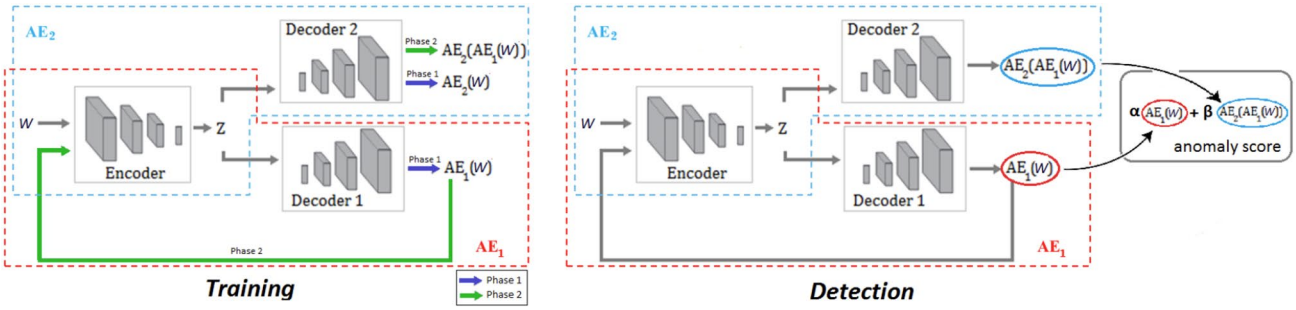

Fig. 2 Architecture of the USAD algorithm as proposed by Audibert et al. [31] illustrating the information flow at training (left) and detection stage (right)

- F_{1C} score: According to Garg et al., a perfect anomaly detection algorithm should be able to detect at least one anomalous data point per anomaly event without any false positives[5]. Therefore, they proposed a new metric called Composite F score (F_{1C}). The F_{1C} score is calculated similar to the F_1 score by taking the harmonic mean of precision and recall. However, the recall is calculated event-based instead of instance-based: $Recall_{event} = \frac{TP_e}{TP_e+FN_e}$
- $F_{1PA\%k}$: Kim et al. argue that PA overestimates detection accuracy, while using F_1 score without PA underestimates accuracy due to incomplete test set labeling [3]. The authors suggest a new metric, called $F_{1PA\%k}$, which can address the problems of over- and underestimation. This metric is similar to PA, but it only considers an event as detected when the proportion of correctly identified instances in the event exceeds a threshold value k. If a user wants to remove the dependency on a specific threshold value k, it is recommended to measure the area under the curve (AUC) of $F_{1PA\%k}$ by gradually increasing k from 0 to 100. This approach allows for a comprehensive evaluation of the model's performance across various threshold values, providing a more robust assessment of anomaly detection capability and is referred to as $F_{1PA\%k\text{-}AUC}$.

Although these new metrics are generally considered an improvement over the F_{1PA}, they have their advantages and disadvantages. To evaluate which metrics perform well in different scenarios, several example signals were created and can be seen in Fig. 4. The first simulated anomaly detector is a random signal with a 0.01 probability of flagging a time step as an anomaly. Y1 to Y4 are constructed signals that correctly detect portions of the anomaly, and some of them also have false positive sequences during normal operation. Y5 is a special case where the alarm is raised the whole time except for one instance where a false negative occurs in the actual anomaly. The example signals can be seen in Fig. 4 and their results for the four different metrics can be found in Table 2. As previously stated, the F_1 score underestimates the capabilities of a detection algorithm. All scores are relatively low except for Y4, where most of the anomalous event was captured. When detecting an anomalous event, it is not essential whether the whole anomalous segment was detected or only a fraction, since an operator would investigate the system as soon as an alarm is raised. Therefore, one could argue that detector Y1 is better than Y4 as no false positives were raised. However, the F_1 score does not capture this. Therefore, PA was created, as it was reasoned that an anomaly detector should have minimal false positives and detect an anomalous event. Looking at the F_{1PA} scores, it can be observed that the metric overestimates the detection capabilities. All detectors achieve high scores, and the random detector almost reaches a perfect score. The F_{1C} score improves on this, as the deviation between a good detector (Y1) and a worse one (Y3) becomes more significant, and the random signal scores fewer points. The $F_{1PA\%k}$ metric rightly scores the random signal even lower. However, it is sensitive to the parameter k. In this case, a k of 20 was selected, and it can be seen that signal Y1 scored low because the fraction of detected anomalous instances in the anomaly event was less than 20%. The signal Y5 can be regarded as a poor detector, since it raises an alarm for almost all the time, yet none of the metrics appear to identify this. The $F_{1PA\%k\text{-}AUC}$ method is not sensitive to threshold selection; however, signals Y1, Y2 and Y3 perform worse than the bad signal Y5. In conclusion, the F_{1C} score is better than the other metrics while it still has problems with clearly marking the signal Y5 as a bad detector. Therefore, the comparison of the results in Sect. 4 will be measured with the F_{1C} score.

Table 2 Comparison of evaluation methods for anomaly detection

Name	F_1	F_{1PA}	F_{1C}	$F_{1PA\%k=20}$	$F_{1PA\%k\text{-}AUC}$
Random	0.02	0.99	0.63	0.02	0.03
Y1	0.33	1	1	0.33	0.47
Y2	0.4	0.91	0.75	0.91	0.55
Y3	0.35	0.83	0.6	0.83	0.50
Y4	0.91	0.91	0.91	0.91	0.91
Y5	0.66	0.67	0.67	0.67	0.67

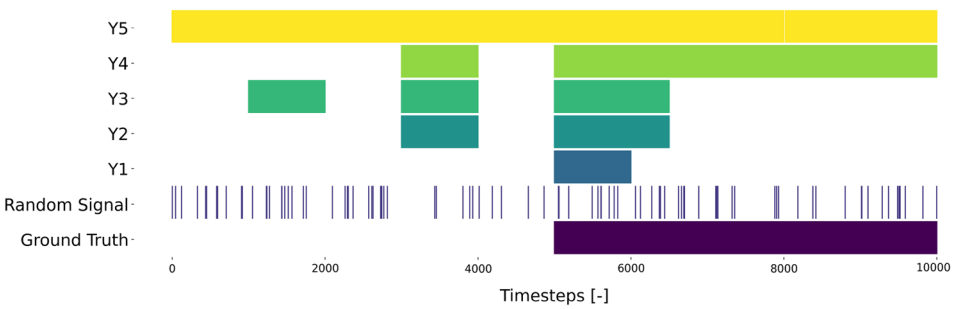

Fig. 4 Comparison of example detectors in anomaly detection using selected evaluation methods for anomaly detection. The corresponding metric scores can be seen in Table 2

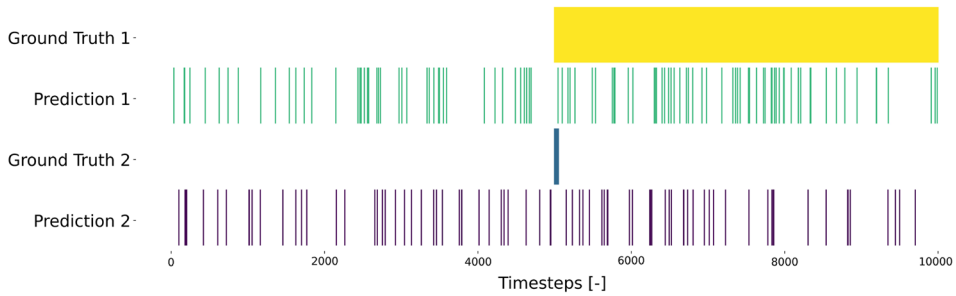

Fig. 5 Impact of combining metric scores on anomaly detection performance

To obtain a final metric score for a dataset with multiple anomaly sequences, there are two options for combining the individual results: taking the average of the metric results for all anomaly sequences or summing up the classes of the confusion matrix and calculating the metric using this sum. However, the latter approach tends to give better results when using the PA method, contributing to the perceived increase in accuracy. The following example illustrates this:

Figure 5 depicts a dataset comprising two distinct anomaly sequences. The first anomaly, accounting for approximately 50% of the anomaly sequence, is more readily detectable compared to the relatively brief second anomaly. A random signal, having a 0.01 probability, is applied as a detector, yielding a resultant F_{1PA} score.

Upon computation of the average F_{1PA} score, the outcome is determined to be 0.497. The first anomaly garners a score of 0.995, while the second anomaly, due to a lack of true positives, scores 0. Conversely, when the total metric is computed via the summation of the confusion matrix elements, the score is substantially higher at 0.979. This elevation in score is attributed to the PA mechanism, which assigns the entirety of the first anomaly sequence as true positives. Given the longer duration of the first anomaly, it significantly contributes to the true positive and false negative classifications, thereby influencing the computed metric. This calculation methodology inherently undermines the significance of shorter and more challenging-to-detect anomalies. Consequently, the study opts for the utilization of the average score methodology.

3.3 Comparative evaluation approach

In this paper, the three selected DAD methods are compared to several popular classic anomaly detection methods and three baseline methods. The four classic methods used for comparison are Gaussian mixture model (GMM), k-nearest neighbor (kNN), one-class support vector machine (OCSVM), and principal component analysis (PCA). These methods have been widely used in anomaly detection research and serve as established benchmarks for comparison [36–39] and are implemented with the Python library PyOD [33]. Along with the selected algorithms, the importance of baseline testing in the evaluation of anomaly detection methods is highlighted, as underscored by the findings of Kim et al. [3]. They asserted that despite attaining high test scores, some detection methods might not exhibit improvement when compared to simple baseline methodologies. This underlines the significance of maintaining a fair and comprehensive evaluation process by including basic benchmark methods.

Therefore, this study incorporates three baseline methods to ensure a robust comparative analysis of the performance of the selected algorithms.

- The raw input method: this simple technique calculates the norm of the input vector, serving as a primary benchmark against more sophisticated methods.
- The untrained autoencoder: in this method, the weights of the autoencoder are randomly initialized from a standard normal distribution.
- The random signal method: here, each instance has a probability of $p = 0.01$ of being flagged as an anomaly. This method, despite its inherent randomness, offers a basic statistical baseline against which the detection capabilities of other methods can be assessed.

The incorporation of these baseline methods in the analysis enables a comprehensive performance evaluation of the selected algorithms, setting them against both conventional techniques and simpler benchmark approaches. This facilitates a holistic understanding of their effectiveness in the deep anomaly detection research field.

As mentioned in Subsect. 2, most of the state-of-the-art algorithms utilize thresholding methods that suffer from data leakage, giving them an advantage. To ensure a fair comparison, the same thresholding approach was applied to all algorithms. A grid search was conducted, exploring all possible thresholds, to find the threshold that maximizes the F_1 score, irrespective of the metric that is evaluated on. For the LSTM algorithm, this study reports the results for both thresholding methods: the original nonparametric method proposed by the authors and the

best F_1 method. This approach allows for a comprehensive comparison and evaluation of all algorithms using the same thresholding approach. It also provides an opportunity to assess the performance of the unsupervised NPT method and determine its effectiveness in anomaly detection.

4 Experiments

4.1 Case study description

This section introduces the datasets used to evaluate various DAD algorithms. We use a dual-approach featuring real-world satellite telemetry data from the Sentinel-1 mission and recognized validation datasets such as the Mars Science Laboratory rover and Soil Moisture Active Passive satellite dataset provided by NASA. These diverse datasets facilitate a comprehensive assessment of the DAD methods, shedding light on their performance in both real-world and standardized scenarios. The subsequent subsections detail each dataset and its unique contribution to our study. An overview of the key characteristics of the datasets can be seen in Table 3.

4.1.1 Real-life satellite telemetry data of ESA satellites

Satellites carry four reaction wheels, a component that presents a compelling use case for applying machine learning in anomaly detection. The accumulated data from these identical reaction wheels—despite variations in satellite sizes—offers an opportunity for holistic prognostics. This data, relatively accessible, encapsulates a mechanical component's degradation over time. Key sensor readings, such as temperature, current, and speed, reflect changes in friction—usually the culprit behind reaction wheel anomalies or failures [40].

The Sentinel-1 mission, comprising two polar-orbiting satellites, Sentinel-1A and Sentinel-1B, launched in April 2014 and April 2015, respectively, was selected as the data source for this study. Its appeal lies in the existence of multiple sequences of anomalous reaction wheels, enabling label generation for testing algorithm efficacy. Seven anomalies, identified by ESA operational personnel, form the basis of our test sets. These are drawn from the data of 7 days before and after each anomaly's onset, while the training sets comprise the preceding 7-day data.

The telemetry data, comprising various features like current, temperature, speed, and torque, is captured at irregular intervals, but usually multiple times per minute. To address the irregularity, the data was averaged within 1-min intervals to provide a consistent time scale for analysis. There were two instances when no data was recorded—18 and 30 min long, respectively—wherein interpolation was used to fill the data gaps. Ultimately, the data was normalized based on the training set values.

Table 3 Key characteristics of the datasets used for this study

	ESA	SMAP	MSL
Total anomalies	7	69	36
Unique telemetry channels	10	55	27
Telemetry values evaluated	146,887	429,735	66,709
Contamination rate	0.36	0.13	0.1
Average anomaly length	1026	826	616

4.1.2 Validation datasets

To ascertain the efficacy of the proposed anomaly detection methodology, it is essential to benchmark its performance against datasets that are widely accepted within the research community. This step becomes particularly crucial considering the limited availability of labeled data and infrequency of anomalies in the use-case dataset. The Mars Science Laboratory (MSL) rover and Soil Moisture Active Passive (SMAP) satellite dataset, released by National Aeronautics and Space Administration (NASA), meets these requirements and serves as an ideal point of reference [14]. This dataset encompasses real-world spacecraft data, with input channels anonymized for security and privacy. It consists of a real-valued telemetry stream and binary commands, which are either sent or received by the corresponding subsystem. This data, labeled by domain experts, provides a robust foundation for evaluating the proposed anomaly detection approach.

4.1.3 Setup

The comparison of the selected DAD methods is conducted against four classic anomaly detection methods and three baselines using three multivariate datasets. In these datasets, each anomaly comprises a single training sequence on which the algorithms are trained. Subsequently, they are tested on the testing sequence, which is labeled to indicate one or more anomalous events.

For the ESA dataset introduced in this paper, the decision was made to target the current channel specifically for prediction when employing LSTM. This choice was informed by the frequent association of reaction wheel failures with friction, and it was anticipated that current measurements would effectively capture this influence. Although a similar influence was expected on temperature, the results were not as promising, potentially due to the delayed impact of friction on temperature.

After training and testing each anomaly for each dataset using different metrics, the average metric score is computed, as explained in Subsect. 3.2. Since anomalies in spacecraft are typically rare, and the number of failures is insufficient to create a validation dataset, hyper-parameter tuning was not performed. The three DAD algorithms use the parameters specified in their respective papers, and the classic anomaly detection methods implemented with the Python library pyod are trained with their default settings. An exception is made for the USAD algorithm, for which a grid search was conducted for the ESA dataset, as autoencoders are highly sensitive to the choice of the latent space size. The parameters used for USAD with the ESA dataset are as follows: window size K=50, sensitivity threshold α=1, and dimension of the latent space m=3.

4.2 Results

In this section, the results of the three DAD algorithms, the four classical algorithms, and the three benchmark methods are first compared based on their metric performance on the three selected datasets in Subsect. 4.2.1. Then a qualitative comparison is conducted on an example anomaly in Subsect. 4.2.2.

4.2.1 Metric results

Table 4 presents the results for the F_{1C} score. It can be observed that the LSTM algorithm achieves the best results in almost all cases, only beaten by the USAD algorithm on the MSL dataset by a fraction. Interestingly, when the LSTM algorithm utilizes nonparametric thresholding instead of the best F_1 score, the F_{1C} score on the ESA dataset improves due to the filtering of false positives. However on the MSL and SMAP dataset, the F_{1C} score drops when using NPT.

Table 4 F_{1C} score for various methods

Dataset	ESA	MSL	SMAP
LSTM (NPT)	**1.000** (↑)	0.413 (↑)	0.579 (↑)
LSTM	0.914 (↑)	**0.567** (↑)	**0.705**(↑)
USAD	0.772 (↓)	**0.574** (↑)	0.393 (↓)
Anomaly transformer	0.336 (↓)	0.201 (↓)	0.263 (↓)
GMM	0.767 (↓)	0.473 (↑)	0.409 (↑)
KNN	0.778 (↓)	0.486 (↑)	0.382 (↓)
OCSVM	0.807 (↓)	0.549 (↑)	0.371 (↓)
PCA	0.820 (↑)	0.382 (↓)	0.355 (↓)
Baseline: raw	0.820	0.387	0.404
Baseline: AE	0.558	0.346	0.344
Baseline: random	0.468	0.191	0.183

Bold indicates the best result or in 2% from the best result

↑ is marked when the score is higher than Baseline: Raw

This is explained by the error pruning which helps out by filtering false positives. In the case of the ESA dataset, this seems to work well and improve the metric score, but on the other two datasets the score drops. This might be due to the error pruning minimum decrease parameter being set too low and wrongfully pruning true alarms. The best F_1 method optimizes specifically for the F_1 score and not the F_{1C} score and therefore scores higher in that metric as can be seen in Table 5. Additionally, it can be observed that the NPT method scores lower on the F_1 metric. This is because while it effectively reduces false positives, it does so at the expense of true positives.

Regarding the baseline methods, the results indicate that the raw signal method achieves results that are difficult for other methods to surpass, particularly on the European Space Agency (ESA) and SMAP dataset where the results are quite high. Comparing the other two deep learning methods to the raw signal baseline, it can be seen that the USAD algorithm outperforms the baseline on the MSL dataset. However, the anomaly transformer does not surpass the baseline results on any of the datasets.

Furthermore, it is observed that all of the classical methods outperform the anomaly transformer on all three datasets. This observation is intriguing as it suggests that the more complex methods, such as USAD and the anomaly transformer, do not perform well in comparison. Conversely, the simple deep learning algorithm, the LSTM, demonstrates excellent performance. This highlights the importance of considering the effectiveness of the algorithm design and complexity, rather than solely relying on the sophistication of the method.

It is evident that, in general, the scores obtained using the ESA dataset surpass those achieved with the other two datasets. This observation suggests that anomalies are more easily detected in the ESA dataset. This phenomenon can be attributed to the contamination rate and the average anomaly length, both of which are presented in Table 3. Sehili and Zhang have previously demonstrated that datasets with a higher contamination rate tend to yield better F_1 scores. Moreover, they express greater confidence in achieving higher F_{1PA} scores when the average length of the anomaly segment increases. Given that the F_{1C} score, like the F_{1PA}, is event based, the same principle holds true.

4.2.2 Qualitative results

In Sect. 3.2, it was established that the F_{1C} score is an improvement over other popular metrics, but still has some limitations. Therefore, it is valuable to qualitatively examine the results of the predictions. Figure 6 illustrates an example anomaly from the ESA dataset along with the corresponding predictions made by the algorithms. This qualitative analysis

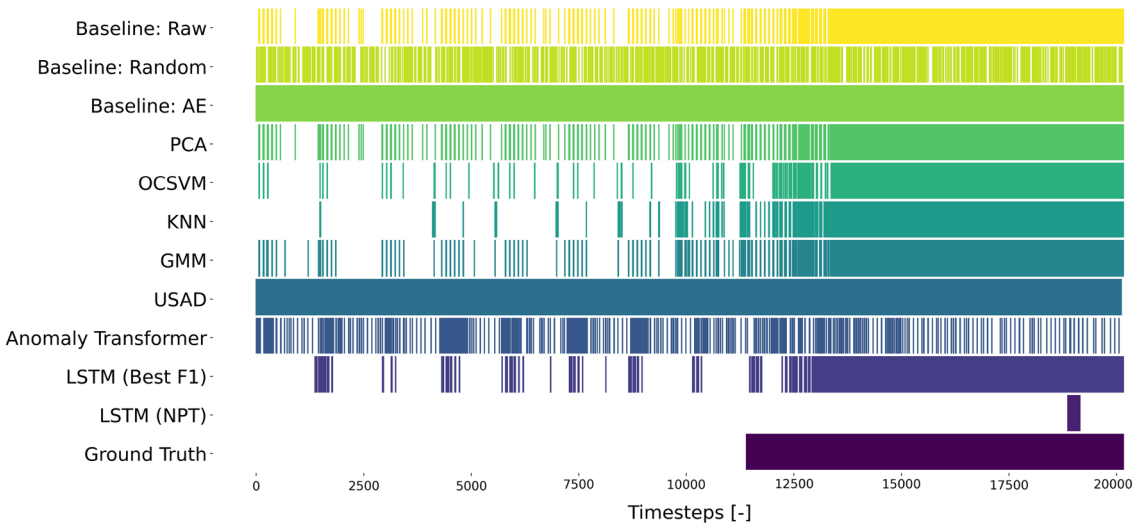

Fig. 6 Qualitative results of one anomaly of the ESA dataset

Table 5 F_1 Score for various methods

Dataset	ESA	MSL	SMAP
LSTM (NPT)	**0.715** (↑)	**0.478** (↑)	**0.544** (↑)
USAD	0.617 (↓)	0.410 (↑)	0.286 (↓)
Anomaly transformer	0.067 (↓)	0.089 (↓)	0.113 (↓)
GMM	0.695 (↑)	0.366 (↑)	0.308 (↑)
KNN	**0.713**(↑)	0.365 (↑)	0.260 (↓)
OCSVM	0.677 (↓)	0.391 (↑)	0.300 (↓)
PCA	0.664 (↑)	0.322 (↓)	0.270 (↓)
Baseline: raw	0.664	0.324	0.313
Baseline: AE	0.494	0.292	0.262
Baseline: random	0.082	0.056	0.013

Bold indicates the best result or in 2% from the best result

↑ is marked when the score is higher than Baseline: raw

Table 6 F_{1PA} score for various methods

Dataset	ESA	MSL	SMAP
LSTM	**1.000** (↑)	0.451 (↓)	0.612 (↓)
LSTM (NPT)	0.928 (↑)	0.643 (↓)	**0.787**(↑)
USAD	0.773 (↓)	**0.684** (↓)	0.517 (↓)
Anomaly transformer	0.907 (↓)	0.568 (↓)	0.595 (↓)
GMM	0.835 (↓)	0.572 (↓)	0.519 (↓)
KNN	0.849 (↓)	0.630 (↓)	0.519 (↓)
OCSVM	0.882 (↓)	0.641 (↓)	0.468 (↓)
PCA	0.889 (↓)	0.471 (↓)	0.451 (↓)
Baseline: raw	0.889	0.482	0.501
Baseline: AE	0.673	0.489	0.477
Baseline: random	0.912	**0.686**	0.642

Bold indicates the best result or in 2% from the best result

↑ is marked when the score is higher than Baseline: Random

provides further insight into the performance and behavior of the algorithms.

In the bottom of the figure, it can be observed that the anomaly starts at approximately 60% of the time series and continues until the end. The baseline autoencoder and the USAD algorithm raise an alarm for almost the entire sequence, thus failing to provide a useful signal. Surprisingly, both methods still achieve a F_{1C} score and F_1 score of 0.61, highlighting the limitations of these metrics as discussed in Subsect. 3.2.

The baseline random method, as expected, randomly raises alarms with a relatively consistent density throughout the sequence. On the other hand, the anomaly transformer raises flags throughout the entire sequence, with a seemingly higher density of alarms before the anomaly.

The raw signal, the four classical methods, and the LSTM algorithm using the best F_1 threshold exhibit a more meaningful signal. They indicate fewer alarms before the anomaly while increasing the density after its onset. One could argue that the LSTM algorithm with the original unsupervised nonparametric thresholding demonstrates the best signal. The anomaly is identified relatively late in the sequence; however, it is important to note that timeliness is not a focus of this research. It avoids false positives and successfully raises an alarm during the anomaly.

Furthermore, as evident in Table 4 and Table 5, it can be observed that the raw signal surpasses some of the anomaly detection methods in terms of performance. Moreover, Table 6 confirms that the F_{1PA} metric is not reliable, as the

random signal outperforms all methods except the LSTM algorithm.

5 Conclusions

This paper illuminates several flaws in current DAD research, revealing that the performance results are often overestimated. Despite the escalating complexity of anomaly detection methods, simpler approaches and even baseline methods outperform some of the more intricate DAD algorithms. While the F_{1C} score represents an improvement over other metrics, the qualitative results indicate that it can still lead to inflated performance for poor detectors. Among the DAD methods examined, the LSTM algorithm demonstrates highly promising results across all datasets and even achieves a perfect score on the real-life dataset. However, it is important to note that this dataset involves a relatively simple subsystem with a clear channel to predict failure, and the real-life dataset had relatively few anomalies due to the reliability of satellites in actual scenarios. The performance of the LSTM algorithm may not be as robust in complex systems lacking a single channel that reliably predicts failure, and a larger dataset is needed to bolster confidence in these results.

For the field to progress, several steps should be taken. Research should adopt best practices such as avoiding data leak, applying truly unsupervised algorithms throughout, abstaining from using measuring or averaging methods that inflate performance results, and applying proper data split methods.

For effective comparisons, the community needs high-quality, pre-split datasets and metrics that correlate with good performance. This research compared algorithms using the same thresholding, but future work should also compare different unsupervised thresholding methods.

For satellite operators, further research could explore whether other prediction error methods such as convolutional neural networks and transformers also perform well on telemetry data. Modifying the LSTM to predict more than one channel could enhance explainability and potentially improve performance. Moreover, it would be beneficial to determine which subsystems are best suited to these methods.

In conclusion, while our findings highlight the potential of deep learning in anomaly detection, they also underscore the need for a more critical and nuanced approach to evaluating its effectiveness in real-world applications. The ongoing advancement in deep learning and the ever-increasing volume of satellite telemetry data open up exciting avenues for further research and innovation in this field.

Author contributions Conceptualization LH, MB, BS and WV. Methodology LH. Software LH. Validation LH. Writing—original draft preparation LH. Writing—review and editing MB, WV. Visualization LH. Supervision MB. Funding acquisition WMB. All authors have read and agreed to the published version of the manuscript.

Funding This research has received funding from the European Space Agency's co-sponsored PhD program under contract number 4000131846/20/NL/MH/hm.

Declarations

Conflict of interest The authors have no conflict of interests to declare that are relevant to the content of this article.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Pang, G., Shen, C., Cao, L., Hengel, A.V.D.: Deep learning for anomaly detection: a review. ACM Comput. Surv. (CSUR) **54**(2), 1–38 (2021)
2. 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)
3. 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)
4. Doshi, K., Abudalou, S., Yilmaz, Y.: Tisat: time series anomaly transformer. arXiv preprint arXiv:2203.05167 (2022)
5. Garg, A., Zhang, W., Samaran, J., Savitha, R., Foo, C.-S.: An evaluation of anomaly detection and diagnosis in multivariate time series. IEEE Trans Neural Netw Learn Syst **33**(6), 2508–2517 (2021)
6. El Amine Sehili, M., Zhang, Z.: Multivariate time series anomaly detection: fancy algorithms and flawed evaluation methodology. arXiv e-prints, 2308 (2023)
7. Choi, K., Yi, J., Park, C., Yoon, S.: Deep learning for anomaly detection in time-series data: review, analysis, and guidelines. IEEE Access **9**, 120043–120065 (2021)
8. Hamilton, J.D.: Time series analysis. Princeton University Press, Princeton (2020)
9. Hawkins, D.M.: Identification of outliers, vol. 11. Springer, London (1980)
10. Barnett, V., Lewis, T.: Outliers in statistical data. Wiley Series in probability and mathematical statistics, applied probability and statistics (1984)

11. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: a survey. ACM Comput. Surv. (CSUR) **41**(3), 1–58 (2009)
12. Tang, Z., Chen, Z., Bao, Y., Li, H.: Convolutional neural network-based data anomaly detection method using multiple information for structural health monitoring. Struct. Control Health Monitor. **26**(1), 2296 (2019)
13. Basora, L., Olive, X., Dubot, T.: Recent advances in anomaly detection methods applied to aviation. Aerospace **6**(11), 117 (2019)
14. Hundman, K., Constantinou, V., Laporte, C., Colwell, I., Soderstrom, T.: Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 387–395 (2018)
15. Yairi, T., Nakatsugawa, M., Hori, K., Nakasuka, S., Machida, K., Ishihama, N.: Adaptive limit checking for spacecraft telemetry data using regression tree learning. In: 2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No. 04CH37583), vol. 6. IEEE, pp. 5130–5135 (2004)
16. Chalapathy, R., Chawla, S.: Deep learning for anomaly detection: a survey. arXiv preprint arXiv:1901.03407 (2019)
17. Shen, L., Li, Z., Kwok, J.: Timeseries anomaly detection using temporal hierarchical one-class network. Adv. Neural Inform. Process. Syst. **33**, 13016–13026 (2020)
18. Xu, J., Wu, H., Wang, J., Long, M.: Anomaly transformer: time series anomaly detection with association discrepancy. arXiv preprint arXiv:2110.02642 (2021)
19. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015)
20. Bank, D., Koenigstein, N., Giryes, R.: Autoencoders. arXiv preprint arXiv:2003.05991 (2020)
21. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Commun ACM **63**(11), 139–144 (2020)
22. 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). IEEE, pp. 841–850 (2020)
23. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Adv. Neural Inform. Process. Syst. **30** (2017)
24. Broadwater, J.B., Chellappa, R.: Adaptive threshold estimation via extreme value theory. IEEE Trans. Signal Process. **58**(2), 490–500 (2009)
25. 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). IEEE, pp. 33–43 (2020)
26. 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)
27. Goh, J., Adepu, S., Junejo, K.N., Mathur, A.: A dataset to support research in the design of secure water treatment systems. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 10242 LNCS, 88–99 (2017) https://doi.org/10.1007/978-3-319-71368-7_8
28. Chen, Y., Zhang, C., Ma, M., Liu, Y., Ding, R., Li, B., He, S., Rajmohan, S., Lin, Q., Zhang, D.: Imdiffusion: imputed diffusion models for multivariate time series anomaly detection. arXiv preprint (2023). arXiv:2307.00754
29. 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)
30. 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. Springer, pp. 703–716 (2019)
31. 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)
32. Chen, Z., Chen, D., Zhang, X., Yuan, Z., Cheng, X.: Learning graph structures with transformer for multivariate time-series anomaly detection in iot. IEEE Internet of Things J. **9**(12), 9179–9189 (2021)
33. Zhao, Y., Nasrullah, Z., Li, Z.: Pyod: a python toolbox for scalable outlier detection. J. Mach. Learn. Res. **20**(96), 1–7 (2019)
34. Wu, R., Keogh, E.: Current time series anomaly detection benchmarks are flawed and are creating the illusion of progress. IEEE Trans. Knowl. Data Eng. (2021). https://doi.org/10.1109/TKDE.2021.3112126
35. Xu, J., Wu, H.: Code repository for anomaly-transformer (2022). https://github.com/thuml/Anomaly-Transformer
36. Reynolds, D.A., et al.: Gaussian mixture models. Encycl. Biom. **741**, 659–663 (2009)
37. Mucherino, A., Papajorgji, P.J., Pardalos, P.M.: k-nearest neighbor classification, pp. 83–106. Springer, New York (2009). https://doi.org/10.1007/978-0-387-88615-2_4
38. Schölkopf, B., Williamson, R.C., Smola, A., Shawe-Taylor, J., Platt, J.: Support vector method for novelty detection. Adv. Neural Inform. Process. Syst. **12** (1999)
39. Shyu, M.-L., Chen, S.-C., Sarinnapakorn, K., Chang, L.: A novel anomaly detection scheme based on principal component classifier. Technical report, Miami Univ Coral Gables Fl Dept of Electrical and Computer Engineering (2003)
40. Bialke, W., Hansell, E.: A newly discovered branch of the fault tree explaining systemic reaction wheel failures and anomalies. In: Proceedings of the European Space Mechanisms and Tribology Symposium, pp. 20–22 (2017)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.