18th International Conference on Space Operations, Montreal, Canada, 26 - 30 May 2025.


SpaceOps-2025, ID # 152

# The Making of the European Space Agency Benchmark for Anomaly Detection in Satellite Telemetry

**Krzysztof Kotowski[a,*], Christoph Haskamp[b,*], Jacek Andrzejewski[a], Bogdan Ruszczak[a,c], Jakub Nalepa[d,a], Daniel Lakey[e], Peter Collins[f], Aybike Kolmas[g], Mauro Bartesaghi[h], Jose Martínez-Heras[i], Gabriele De Canio[f,*]**

[a] *KP Labs, Poland, {kkotowski, jandrzejewski, bruszczak, jnalepa}@kplabs.pl*
[b] *Airbus Defence and Space GmbH, Germany, christoph.haskamp@airbus.com*
[c] *Opole University of Technology, Poland, b.ruszczak@po.edu.pl*
[d] *Silesian University of Technology, Poland, jnalepa@ieee.org*
[e] *CGI Deutschland B.V. & Co. KG, Germany, daniel.lakey@cgi.com*
[f] *European Space Operations Centre, European Space Agency, Germany, {peter.collins, gabriele.decanio}@esa.int*
[g] *Telespazio Germany GmbH, Germany, aybike.kolmas@telespazio.de*
[h] *LSE Space GmbH, Germany, mauro.bartesaghi@lsespace.com*
[i] *Solenix GmbH, Germany, jose.martinez@solenix.ch*
* Corresponding Authors

**Abstract**

The recently published European Space Agency Benchmark for Anomaly Detection in Satellite Telemetry (ESA-ADB, https://doi.org/10.5281/zenodo.12528696) is the first large-scale, real-life, carefully annotated benchmark of its kind. It is a significant milestone in the AI for Automation ($A^2I$) Roadmap of the European Space Operations Centre (ESOC), which allows for training and validating advanced algorithms for multivariate time series anomaly detection and related tasks in satellite telemetry monitoring. It took over a year of close cooperation between spacecraft operations and machine learning (ML) engineers to prepare this curated dataset and evaluation pipeline addressing the needs of both communities. This paper gives insights and lessons learned from the process of creating ESA-ADB. It focuses on key technical and organizational challenges encountered and solved during the project to build a common understanding between ML and space operations perspectives.

**Keywords:** anomaly detection, satellite telemetry, AI, benchmark, dataset

**Acronyms/Abbreviations**
AI – Artificial Intelligence
ESA-ADB – European Space Agency Benchmark for Anomaly Detection in Satellite Telemetry
ESOC – European Space Operations Centre
ML – Machine Learning
MLE – Machine Learning Engineer
SOE – Spacecraft Operations Engineer
TSAD – Time Series Anomaly Detection

## 1. Introduction

With the exponentially growing number and complexity of spacecraft, the scalability of telemetry monitoring is a crucial aspect for spacecraft operations centers worldwide. Timely detection and resolution of anomalies in telemetry ensure the safety and effectiveness of a mission and its objectives. Currently, spacecraft operations engineers (SOEs) use computer-aided systems that alarm for out-of-limit measurements or predefined anomalous patterns [1], [2]. However, more sophisticated and previously unseen anomalies need a long process of manual identification and analysis, reducing available time and spectrum of potential resolution options. To mitigate this, advanced automatic anomaly detection systems are actively developed by all key players in the domain of space operations, including ESA/ESOC [3], [4], NASA [2], CSA [5], CNES [6], DLR [7], and JAXA [8]. Such systems are often based on machine learning (ML) algorithms which need significant amounts of data to learn from. Thus, one of the key challenges considered in the Artificial Intelligence for Automation ($A^2I$) Roadmap [9], [10] of ESOC was to prepare a comprehensive dataset for training and validating ML-based anomaly detection algorithms in satellite telemetry from real missions. We addressed this issue by creating the ESA-ADB benchmark [11] – the first large-scale, AI-ready, carefully annotated dataset of its kind, accompanied with space operations-oriented evaluation pipeline, metrics, and results from popular algorithms. It contains long fragments of spacecraft telemetry from 3 large ESA missions with

very different characteristics and purposes. The results of the benchmark showed that new algorithms are necessary to fulfil requirements and expectations of SOEs (discussed in Section 3), so we recently started a Kaggle competition to engage a wider ML community to propose new solutions. All materials related to the ESA-ADB are publicly available under the links provided in Table 1. These materials are actively used in several on-going ESA projects [10], [12], [13], [14]. They include data, the code of the complete benchmarking pipeline, summary articles, and software tools used in the project.

Table 1. Links to all materials related to the ESA-ADB benchmark

| | |
|---|---|
| Dataset | https://zenodo.org/records/12528696 |
| Code | https://github.com/kplabs-pl/ESA-ADB |
| Article [11] | https://arxiv.org/abs/2406.17826 |
| Annotation procedures [15] | https://publications.jrc.ec.europa.eu/repository/handle/JRC135493 |
| Annotation tool [16] | https://oxi.kplabs.pl/ |
| Kaggle competition [17] | https://www.kaggle.com/competitions/esa-adb-challenge |

ESA-ADB is a result of close cooperation between SOEs and ML engineers (MLEs). Conclusions and lessons learned from this cooperation are the main focus of the current paper. The covered topics include the making of the common nomenclature (Section 2), requirements for anomaly detection algorithms in space operations (Section 3), a discussion on the proper evaluation procedures and metrics (Section 4), and case studies of particularly interesting fragments from the dataset (Section 5).

## 2. Common language between machine learning and space operations

Developing a shared language and consistent terminology between MLEs and SOEs was a critical outcome of extensive bilateral meetings in the making of the ESA-ADB. In this section, we discuss 3 key agreements about the naming of event categories and telemetry channels. However, there were numerous other nomenclature settlements, i.e., on using the taxonomy of anomaly types by Blázquez-García et al. [18] and calling non-anomalous samples *nominal*.

### *2.1. Anomalies, rare nominal events, communication gaps, and invalid segments*

The annotation process of ESA-ADB was run in close cooperation between SOEs and MLEs as described in [15]. After marking all the anomalies reported by SOEs in the Anomaly Report Tracking System (ARTS) (artsops.esa.int) used at ESOC, MLEs quickly identified several outliers in the telemetry data that were not mentioned in the reports but detected by ML algorithms. Most of them turned out to be planned or commanded maneuvers, operations, and mission events, so they were not reported as anomalies in ARTS and should not be alarmed to SOEs as such. However, many of these cases cannot be distinguished from actual anomalies without expert knowledge and access to data sources not available in ESA-ADB. Therefore, data-driven ML approaches inevitably detect them as anomalies, but it would be unfair to count them as false alarms and penalize algorithms for detecting them. A similar problem occurs for most ML algorithms in case of communication gaps (see Section 5.3). This problem would hinder any reliable benchmarking of ML solutions.

To address this problem, 4 categories of events were annotated in the dataset: *anomalies*, *rare nominal events*, *communication gaps, and invalid segments*. They are defined in Table 2. This division helps to properly assess the results of different ML-based anomaly detectors, e.g., by ignoring detections of *rare nominal events*, *communication gaps*, and *invalid segments*. Out of the 4 categories, only *anomalies* should be always reported to SOEs. The ideal algorithm should not alarm for *rare nominal events*, but it is practically impossible to distinguish between novel *rare nominal events* and *anomalies* without additional context, so as agreed with SOEs, it is acceptable if an anomaly detection system shows a false alarm for the first occurrence of the specific *rare nominal event* type, but it should not alarm for any subsequent occurrences of similar *rare nominal events*.

Table 2. Definitions of 4 categories of events in ESA-ADB

| Event category | Definition | Typical examples | Alarming |
|---|---|---|---|
| **Anomaly** | Atypical, rare, unplanned, and unwanted change in the telemetry. | Micrometeorite impacts, solar flares, hardware or software failures, latch-ups, unexpected attitude disturbances, unexpected responses to telecommands | Every occurrence should be alarmed. |
| **Rare nominal event** | Atypical and rare but expected or planned change in the telemetry. It can be triggered by known telecommands (commanded rare event) or by any other non-commanded special event in the mission timeline. | *Commanded*: maneuvers, resets, calibrations, switching devices on/off<br><br>*Non-commanded*: planned autonomous operations, eclipses, lunar transitions | Only the first occurrence of a rare nominal event from each class may be alarmed. Subsequent occurrences should not be alarmed. |
| **Communication gap** | Unusually long gap in the telemetry (missing data in some or all channels) not directly related to known anomalies. | Problems with the ground infrastructure, effects of resets | It should not be alarmed unless explicitly stated to do so. |
| **Invalid segment** | Fragment of telemetry data containing invalid or forbidden values not directly related to known anomalies. It is neither nominal nor anomalous. | Telemetry does not meet clearly defined validity rules of the mission | It should not be alarmed unless explicitly stated to do so. |

### *2.2. Telemetry parameter vs. channel*

From the perspective of an SOE, the terms *parameters* and *channels* in spacecraft telemetry have distinct meanings and roles in monitoring and controlling a spacecraft. A *parameter* represents a specific measured or derived value related to the spacecraft's state, health, or performance (e.g., battery voltage, solar array current, or reaction wheel speed); and a *channel* is the specific data stream or location from which the telemetry data for a *parameter* is acquired. Hence, SOEs say that they usually look for anomalies in *parameters*, not in *channels*. This fact turned out to be very problematic in communication with MLEs, because it collides with the fundamental ML nomenclature in which the *parameter* already has a couple of different important meanings:

- a *parameter* of the model that is updated during the training, i.e., a single weight of the neural network;
- a *parameter* (or hyperparameter) of the algorithm which controls its behavior;
- a *parameter* of a statistical test, e.g., mean or variance of the expected distribution.

Furthermore, from the ML point of view, satellite telemetry is a specific type of multivariate (or multichannel) time series data, where the "variates" are sometimes called *channels*, variables, or features, but almost never *parameters*. Thus, after several repeated nomenclature collisions in the project, it was collectively agreed that the term *channel* should be used instead of a *parameter* for purposes of the ESA-ADB and similar ML-oriented benchmarks.

### *2.3. Target and non-target channels*

Spacecraft telemetry in ESA-ADB includes many different types of channels and telecommands that may be relevant to deciding if an event is an anomaly or not. However, not every single information source should be actually monitored for anomalies in practice. The clearest example is telecommands, which are essential for detecting certain anomalies (such as incorrect responses to commands) but are not expected to contain anomalies themselves. The same is true for some channels representing auxiliary mission information or external forces (e.g., status flags, counters, payload activity, or space weather data). These *non-target* channels may contain outliers that are, however, irrelevant (or nominal) for SOEs unless there is any atypical reaction in the *target* channels. This proposed division into *non-target* (auxiliary or external) and *target* (annotated and monitored for anomalies) channels is rarely considered in anomaly detection datasets and benchmarks (with the exception of the CATS dataset [19]) but is crucial to concentrate the attention of the benchmark on the most relevant channels. The metrics in ESA-ADB are calculated only for *target* channels. *Non-target* channels may and should be used as input features for algorithms.

The main drawback of this approach is that the selection of *target* and *non-target* channels is somewhat subjective, dependent on the available subset of channels, and it may turn out that some algorithms could be able to properly handle some *non-target* channels by discovering some unknown relationships in the data.

## 3. Functional requirements for anomaly detection algorithms in space operations

Real-life satellite telemetry poses a significant challenge to the majority of typical ML-based anomaly detection algorithms. It is an especially challenging example of a multivariate time series with many specific problems and complexities related to its:

- high dimensionality and volume (years of recordings from up to thousands of channels per satellite [20]);
- complex characteristics (i.e., varying sampling frequencies across time and channels; data gaps caused by idle states and communication problems; trends connected with the degradation of spacecraft components; concept drifts related to different operational modes and mission phases);
- complex web of dependencies between channels;
- a large variety of channel types (i.e., different ranges of physical measurements, categorical status flags, counters, and binary telecommands);
- inherent noise and measurement errors due to the influence of the space environment.
- the difficulty of encoding operational context (i.e., mission plans, commanded activities, and associated expected deviations from the norm)

Additionally, SOEs have a specific set of stringent requirements for anomaly detection systems, such as a clear decision boundary between nominal and anomalous samples, a minimal number of false positives, the ability of the model to learn from the feedback (active learning), and built-in explainability features to provide a list of affected channels.

When working on ESA-ADB, SOEs and MLEs compiled a common list of the nine most important requirements for anomaly detection algorithms in space operations. They are summarized in Table 3. To the best of our knowledge, no existing algorithm fulfills all those requirements. ESA-ADB was created to open up possibilities to change this situation.

Table 3. Nine most important requirements for anomaly detection algorithms in space operations. The levels of *shall* and *should* are used according to the Technical requirements specification ECSS standard *[21]*.

| ID | Level | Requirement | Posed by | Justification |
|---|---|---|---|---|
| R1 | shall | provide binary responses | SOEs | A clear boundary is needed to decide if something should be alarmed to SOEs or not (i.e., 0 – nominal, 1 – anomaly). SOEs cannot rely on abstract continuous anomaly scores from algorithms. |
| R2 | shall | be able to model dependencies between multiple channels | MLEs, SOEs | Satellite telemetry contains hundreds of interconnected channels and there are many examples of anomalies that can be detected only when using information from multiple channels at once. |
| R3 | should | allow for online streaming anomaly detection | MLEs | Even if satellite telemetry at the ground segment is not currently processed in the actual online manner (because data packets are collected during infrequent communication windows), algorithms should not require any future samples to generate predictions in practical applications. |
| R4 | should | provide a list of channels affected by a detected anomaly | SOEs | There are thousands of channels in real missions, so the information about affected channels would save a lot of SOEs' work. Also, it would increase the trustworthiness of the algorithm. |
| R5 | should | learn from anomalies in the training set | MLEs, SOEs | Some types of anomalies are repeatable in satellite telemetry with varying frequency. The algorithm should be able to incorporate this knowledge. |

| | | | | |
|---|---|---|---|---|
| R6 | should | use non-target channels and telecommands as auxiliary data | MLEs, SOEs | An algorithm should be able to use all sources of information for optimal performance while focusing on anomaly detection for the target channels. |
| R7 | should | memorize specific rare nominal events | MLEs, SOEs | Only the first occurrence of a novel rare nominal event class may be alarmed. Subsequent occurrences should not be alarmed after feedback from SOEs. |
| R8 | should | natively handle irregular time series data | MLEs | Varying sampling frequencies and data gaps are common in satellite telemetry. Standard resampling and interpolation methods make algorithms unaware of this fact and may lead to many incorrect detections (see Section 5.3) |
| R9 | should | be able to run on a single high-end PC with a modern GPU | MLEs, SOEs | SOEs rarely have access to high-performance computing or cloud infrastructure, so ML algorithms are often initially deployed on a separate PC in a control room. Other arguments are data privacy, security, and integrity. |

## 4. Quality metrics for anomaly detection in satellite telemetry

The proper selection of metrics is fundamental for the reliability of any benchmark. Despite a long history of research in the domain of time series anomaly detection (TSAD), there is still no consensus on the universal metric or set of metrics for this task. The most popular sample-wise, point-adjust, or simple event-wise protocols (introduced in the seminal work by NASA [2]) are recently widely criticized for being overoptimistic and not reliable enough, and different new metrics are proposed covering different anomaly detection aspects [22], [23], [24], [25], [26], [27]. Therefore, it is no wonder that this topic sparked long discussions and workshops between SOEs and MLEs when creating the ESA-ADB.

Our quality evaluation pipeline is primarily targeted at practical aspects of mission control at ESOC, so SOEs had a key contribution in defining prioritized aspects of the quality assessment. The responsibility of MLEs was to select proper metrics to assess each aspect.

The most important aspect repeated unequivocally by all SOEs was the problem with too many false positives (low precision) in existing anomaly detection systems. It decreases adoption and trust in such automated systems among SOEs. At the same time, SOEs underlined that the precise alignment between detections and anomaly annotations (being the primary focus of many recent metrics [25], [28], [29]) is of lower importance because any overlap is enough to trigger manual anomaly investigation process which encompasses refinement of the overlap. Based on this, MLEs selected the recent corrected event-wise F0.5-score for anomaly detection in time series [22] as the most important metric in the benchmark (and in the related Kaggle competition). The metric (Equation 1) is a harmonic mean of the corrected event-wise precision and the event-wise recall (Equation 2), with precision having 2 times higher importance than recall,

$$F_{0.5_{e_{corr}}} = \frac{1.25 \cdot Pr_{e_{corr}} \cdot Rec_e}{0.25 \cdot Pr_{e_{corr}} + Rec_e} \tag{1}$$

$$Rec_e = \frac{TP_e}{TP_e + FN_e}, \qquad Pr_{e_{corr}} = \frac{TP_e}{TP_e + FP_e} \cdot \left(1 - \frac{FP_t}{N_t}\right) \tag{2}$$

where $TP_e$, $FN_e$, and $FP_e$ are the numbers of event-wise true positives, false negatives, and false positives, respectively; $FP_t$ is the number of nanoseconds with false positives and $N_t$ is the number of nominal nanoseconds. The correction factor using the time-wise false positive rate is the main improvement over the simple event-wise approach used by NASA [2] and is used to penalize excessively long detections. In summary, the metric promotes solutions with a small number of false positives and compact detections (not necessarily covering the whole anomalous segment, it is enough to detect at least one sample) as summarized in Figure 1.

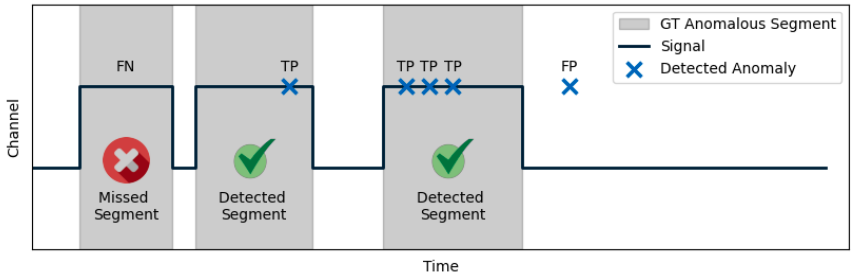


Figure 1. Examples of event-wise detections. The first anomaly from the left counts as a single event-wise false negative $FN_e$. The second and third anomalies both count as single event-wise true positives $TP_e$ (the number of correctly detected blue samples does not matter). There is one time-wise false positive $FP_t$ that also counts as a single event-wise false positive $FP_e$ and additionally increases the false positive rate in the correction factor.

Additionally, SOEs mentioned 4 following aspects to assess, ordered by their priorities from the most important to the least important:

- **Precise identification of affected channels** (following the requirement R4 from Table 3) – this aspect enforced the need for separate annotations for each channel in ESA-ADB (as presented in Figure 2).
- **Avoiding separate repeated alarms for the same anomaly** – they may be considered false alarms in some situations and they negatively affect the trustworthiness of the algorithm.
- **Precise detection timing** – it is desirable to precisely detect anomaly start time for faster reaction and easier anomaly identification.
- **Overlap and proximity of the detection** – it may save some SOEs' work if an algorithm identifies the whole time span of the anomaly.

MLEs designed new metrics to assess each aspect separately [11] using a hierarchical evaluation pipeline – aspects with lower priority are assessed only if there is no statistically significant difference for the previous, higher priority aspect.

## 5. Case studies of particularly interesting events in the dataset

In this section, some of the most interesting case studies of events included in the ESA-ADB are presented, which caused particularly long debates between ML engineers and SOEs – on how to annotate, assess, or detect specific fragments.

### *5.1. Case study 1 – multi-segment anomalies*

Some anomalies and rare nominal events are characterized by a series of short disturbances in the signal as presented in Figure 2. It is a problematic situation in terms of event-wise metrics because such cases would have relatively higher weight in the overall score if we treat each disturbance as a separate event. According to SOEs, all these disturbances correspond to the same underlying event with a common root cause and should be assessed as such. Thus, MLEs decided to mark each disturbance as a separate segment in annotations, but also assign the same event ID to all of them. This way, all segments are treated collectively as a single event when calculating metrics (i.e., it is enough to detect any segment to count it as the event-wise true positive)

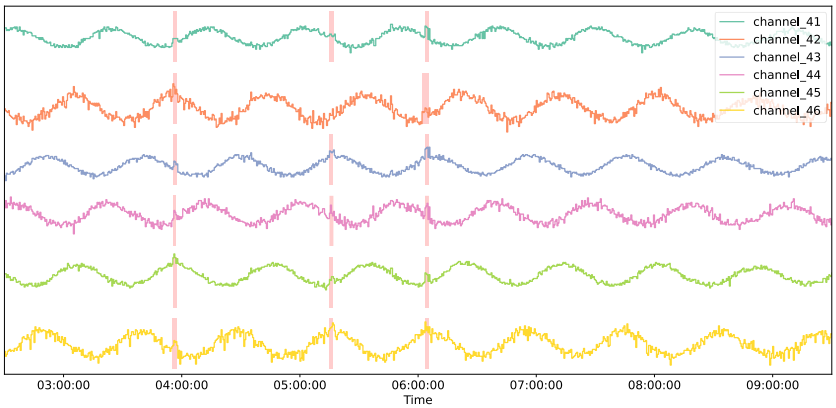


Figure 2. Annotation of the rare nominal event id_155 (highlighted with light red boxes) from Mission1. Although the annotations are scattered into 3 separate segments, they all correspond to the same underlying event with a common root cause. Each channel is annotated individually to accommodate cases where anomalies may not appear in certain channels, such as in channel_42. The Y-axis is omitted, as channels are normalized and vertically shifted for improved visualization.

### *5.2. Case study 2 – excessively long anomalies and mode changes*

Excessively long anomalies, i.e., lasting longer than a few days, are common in channels characterized by high inertia of measured values, e.g., thermal or attitude measurements. The same problem occurs for commanded changes of spacecraft modes when a new rare nominal state is activated for a few days. It is not desirable to annotate such long events in the benchmarking dataset for many reasons:

- to maintain a realistic ratio of anomalous to nominal samples in the dataset – this is one of the main flaws of existing anomaly detection benchmarks [30],
- to avoid the triviality of the dataset – the longer the anomaly, the higher the chance that a random detector hits it. This is the second common flaw of anomaly detection benchmarks [30],
- to avoid overlaps between annotations for different events – it complicates results analysis and metrics calculation,
- it is often hard to precisely identify the real end of a long anomaly,
- it is enough to alarm SOEs when a long anomaly starts, continuing the alarm brings more harm than benefit.

Hence, SOEs and MLEs agreed that the maximum length of a single annotation should not exceed 2 weeks. For longer anomalies, only the most representative part at the beginning should be annotated. For longer mode changes, only the precise moments of mode switches should be annotated.

### *5.3. Case study 3 – irregular sampling rates*

Satellite telemetry collected at the ground segment is characterized by varying sampling frequencies across time and channels. These variations of rates may be detected as anomalies by algorithms trained with uniformly resampled data (which is the case for the majority of widely used algorithms). However, they should not be annotated as anomalies because they are inherent features of this type of real-life data and they should not be alarmed to operators. Also, it is very easy to detect such events by monitoring distances between samples, so we decided to not annotate them (with the exception where the distance between samples is extensively long and related to known communication problems – see Table 2)

### *5.4. Case study 4 – anomalies with low operational impact*

SOEs were surprised that ML algorithms detected several genuine anomalies missing in their anomaly reports. Some of these anomalies, like the one presented in Figure 3, are subtle and hard to spot manually in a wider context. Relatedly, they usually have low practical significance for SOEs. This led to the discussion of whether an anomaly should be annotated if it has no operational impact and nobody is bothered by it. These kinds of events meet the

definition of anomaly from Table *2* as they are atypical, rare, unplanned, and unwanted changes in the telemetry, but they may not need to be alarmed in practice (depending on the mission phase or opinion of SOEs). However, for the purposes of the benchmark, MLEs decided to annotate these anomalies to avoid 1) subjective factors in the annotation and 2) penalizing algorithms for detecting such genuine anomalies.

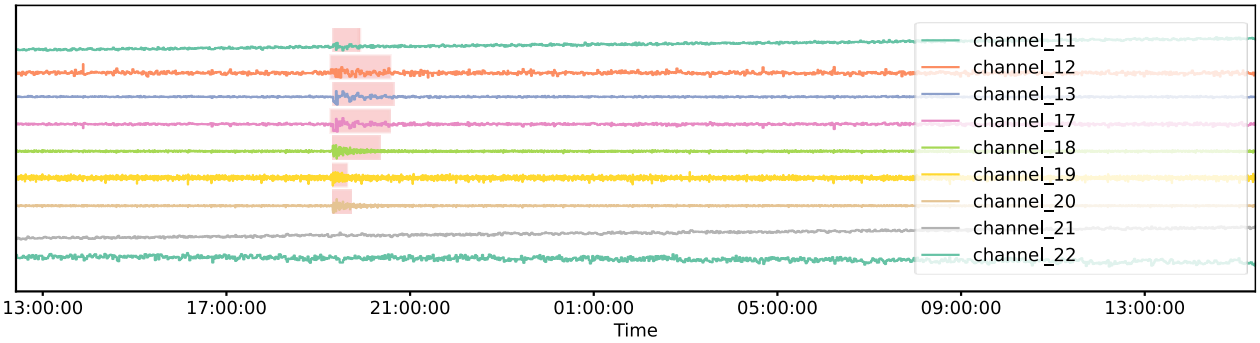


Figure 3. Annotation of anomaly id_631 (highlighted with light red boxes) from Mission2. The Y-axis is omitted, as channels are normalized and vertically shifted for improved visualization.

## 6. Conclusions

The paper highlights the importance of close cooperation between domain experts (i.e., SOEs) and MLEs when creating real-life benchmarks such as ESA-ADB. It is impossible to ensure the practical utility of the benchmark without understanding the real needs of domain experts and translating them into the proper requirements and metrics for ML algorithms. For example, it would not make sense to use the popular range-based metrics in ESA-ADB because SOEs do not really care about the overlap of detections and ground truth. On the other hand, MLEs should oversee the process and draw domain experts' attention to important technical aspects of the evaluation, i.e., the need for online streaming prediction (R3 from Table 3) or the importance of the proper balance between precision and recall of algorithms.

The effectiveness of the cooperation strongly depends on the common language and nomenclature that should be agreed already at the initial stages of a project. It is also a great chance to exchange knowledge between experts from different domains, leading to optimal solutions at later project stages and in similar future activities. In the making of the ESA-ADB, SOEs learned how to structure the data and anomaly reports for ML algorithms, what metrics should be used to assess their requirements, and what is the state-of-the-art in terms of advanced algorithms for their use case. They were also surprised that ML algorithms spotted several anomalies missing in their anomaly reports. SOEs realized that it is sometimes difficult even for them to determine whether a flagged event is an anomaly or not, and whether it should be annotated or not. ML algorithms can be an important tool in such cases, especially if providing some level of interpretability or explainability – explored in detail in our other SpaceOps paper [14]. MLEs learned how to design better algorithms for anomaly detection in spacecraft telemetry, what are the main problems to resolve in the future, and how to design the Kaggle challenge to engage the wider ML community. This common understanding is the key component on the way to real adoption and operationalization of advanced ML-based approaches in space operations.

## Acknowledgments

This work was supported by the European Space Agency under contract number 4000137682/22/D/SR "ESA Anomalies Dataset for International AI Anomaly Detection Benchmark". The authors thank the ESA workforce involved in the project.

## References

[1] J. Martinez, D. Alessandro, S. Bruno, and F. Jörg, "DrMUST - a Data Mining Approach for Anomaly Investigation," in *SpaceOps 2012 Conference*, in SpaceOps Conferences. Stockholm: American Institute of Aeronautics and Astronautics, Jun. 2012. doi: 10.2514/6.2012-1275109.

[2] K. Hundman, V. Constantinou, C. Laporte, I. Colwell, and T. Soderstrom, "Detecting Spacecraft Anomalies Using LSTMs and Nonparametric Dynamic Thresholding," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Jul. 2018, pp. 387–395. doi: 10.1145/3219819.3219845.

[3] J. Martinez and A. Donati, "Novelty Detection with Deep Learning," in *2018 SpaceOps Conference*, Marseille: American Institute of Aeronautics and Astronautics, May 2018. doi: 10.2514/6.2018-2560.

[4] B. Ruszczak *et al.*, "Machine Learning Detects Anomalies in OPS-SAT Telemetry," in *Computational Science – ICCS 2023*, J. Mikyška, C. de Mulatier, M. Paszynski, V. V. Krzhizhanovskaya, J. J. Dongarra, and P. M. A. Sloot, Eds., in Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2023, pp. 295–306. doi: 10.1007/978-3-031-35995-8_21.

[5] J. Careless, "Calian Awarded CSA Satellite Anomaly Detection Contract," SpaceQ Media Inc. Accessed: Apr. 01, 2025. [Online]. Available: http://spaceq.ca/calian-awarded-csa-satellite-anomaly-detection-contract/

[6] S. Fuertes, B. Pilastre, and S. D'Escrivan, "Performance assessment of NOSTRADAMUS & other machine learning-based telemetry monitoring systems on a spacecraft anomalies database," in *2018 SpaceOps Conference*, in SpaceOps Conferences. , American Institute of Aeronautics and Astronautics, 2018. doi: 10.2514/6.2018-2559.

[7] C. O'Meara, L. Schlag, L. Faltenbacher, and M. Wickler, "ATHMoS: Automated Telemetry Health Monitoring System at GSOC using Outlier Detection and Supervised Machine Learning," in *SpaceOps 2016 Conference*, Daejeon, Korea: American Institute of Aeronautics and Astronautics, May 2016. doi: 10.2514/6.2016-2347.

[8] T. Yairi, N. Takeishi, T. Oda, Y. Nakajima, N. Nishimura, and N. Takata, "A Data-Driven Health Monitoring Method for Satellite Housekeeping Data Based on Probabilistic Clustering and Dimensionality Reduction," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 53, no. 3, pp. 1384–1401, Jun. 2017, doi: 10.1109/TAES.2017.2671247.

[9] G. De Canio, J. Eggleston, J. Fauste, A. M. Palowski, and M. Spada, "Development of an actionable AI roadmap for automating mission operations," in *2023 SpaceOps Conference*, Dubai, United Arab Emirates: American Institute of Aeronautics and Astronautics, Mar. 2023. Accessed: Feb. 12, 2024. [Online]. Available: https://star.spaceops.org/user_manudownload.php?doc=303__bm05ydei.pdf

[10] G. De Canio, E. Ntagiou, F. Antonello, and J. Eggleston, "Artificial Intelligence for mission operations automation roadmap: the European Space Operations Centre updates and vision," in *2025 SpaceOps Conference*, Montreal, Canada: Canada Space Agency, 2025.

[11] K. Kotowski *et al.*, "European Space Agency Benchmark for Anomaly Detection in Satellite Telemetry," Jun. 06, 2024, *arXiv*.

[12] G. De Canio *et al.*, "Advancing satellite health monitoring and control with AI," in *2025 SpaceOps Conference*, Montreal, Canada: Canada Space Agency, 2025.

[13] N. Policella *et al.*, "Innovating Space Operations with AI: The AISHGO Project," in *2025 SpaceOps Conference*, Montreal, Canada: Canada Space Agency, 2025.

[14] K. Kotowski *et al.*, "Towards Explainable and Secure AI for Space Mission Operations," in *2025 SpaceOps Conference*, Montreal, Canada: Canada Space Agency, 2025.

[15] K. Kotowski, C. Haskamp, B. Ruszczak, J. Andrzejewski, and J. Nalepa, "Annotating Large Satellite Telemetry Dataset For ESA International AI Anomaly Detection Benchmark," in *Proceedings of the 2023 conference on Big Data from Space*, Vienna: Publications Office of the European Union, Nov. 2023, pp. 341–344. doi: 10.2760/46796.

[16] B. Ruszczak, K. Kotowski, J. Andrzejewski, C. Haskamp, and J. Nalepa, "OXI: An online tool for visualization and annotation of satellite time series data," *SoftwareX*, vol. 23, Jul. 2023, doi: 10.1016/j.softx.2023.101476.

[17] K. Kotowski *et al.*, "Spacecraft Anomaly Challenge on ESA dataset." Kaggle, 2025. [Online]. Available: https://kaggle.com/competitions/esa-adb-challenge

[18] A. Blázquez-García, A. Conde, U. Mori, and J. A. Lozano, "A Review on Outlier/Anomaly Detection in Time Series Data," *ACM Comput. Surv.*, vol. 54, no. 3, p. 56:1-56:33, Apr. 2021, doi: 10.1145/3444690.

[19] P. Fleith, "Controlled Anomalies Time Series (CATS) Dataset." Feb. 2023. doi: 10.5281/zenodo.7646897.

[20] D. Lakey and T. Schlippe, "A Comparison of Deep Learning Architectures for Spacecraft Anomaly Detection," in *IEEE AeroConf 2024*, Big Sky, Montana, Mar. 2024. doi: 10.5281/zenodo.10829339.

[21] European Cooperation for Space Standardization, "ECSS-E-ST-10-06C – Technical requirements specification." European Space Agency, Mar. 06, 2009. Accessed: Feb. 28, 2024. [Online]. Available: https://ecss.nl/standard/ecss-e-st-10-06c-technical-requirements-specification/

[22] M. El Amine Sehili and Z. Zhang, "Multivariate Time Series Anomaly Detection: Fancy Algorithms and Flawed Evaluation Methodology," in *Performance Evaluation and Benchmarking*, R. Nambiar and M. Poess, Eds., Cham: Springer Nature Switzerland, 2024, pp. 1–17. doi: 10.1007/978-3-031-68031-1_1.

[23] L. Herrmann, M. Bieber, W. J. C. Verhagen, F. Cosson, and B. F. Santos, "Unmasking overestimation: a re-evaluation of deep anomaly detection in spacecraft telemetry," *CEAS Space J*, Feb. 2024, doi: 10.1007/s12567-023-00529-5.

[24] J. Nalepa, M. Myller, J. Andrzejewski, P. Benecki, S. Piechaczek, and D. Kostrzewa, "Evaluating algorithms for anomaly detection in satellite telemetry data," *Acta Astronautica*, Jun. 2022, doi: 10.1016/j.actaastro.2022.06.026.

[25] A. Huet, J. M. Navarro, and D. Rossi, "Local Evaluation of Time Series Anomaly Detection Algorithms," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Washington DC USA: ACM, Aug. 2022, pp. 635–645. doi: 10.1145/3534678.3539339.

[26] S. Sørbø and M. Ruocco, "Navigating the metric maze: a taxonomy of evaluation metrics for anomaly detection in time series," *Data Min Knowl Disc*, Nov. 2023, doi: 10.1007/s10618-023-00988-8.

[27] W.-S. Hwang, J.-H. Yun, J. Kim, and B. G. Min, "Do you know existing accuracy metrics overrate time-series anomaly detections?," in *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*, Virtual Event: ACM, Apr. 2022, pp. 403–412. doi: 10.1145/3477314.3507024.

[28] J. Paparrizos, P. Boniol, T. Palpanas, R. S. Tsay, A. Elmore, and M. J. Franklin, "Volume under the surface: a new accuracy evaluation measure for time-series anomaly detection," *Proc. VLDB Endow.*, vol. 15, no. 11, pp. 2774–2787, Jul. 2022, doi: 10.14778/3551793.3551830.

[29] V. Jacob, F. Song, A. Stiegler, B. Rad, Y. Diao, and N. Tatbul, "Exathlon: a benchmark for explainable anomaly detection over time series," *Proc. VLDB Endow.*, vol. 14, no. 11, pp. 2613–2626, Oct. 2021, doi: 10.14778/3476249.3476307.

[30] R. Wu and E. Keogh, "Current Time Series Anomaly Detection Benchmarks are Flawed and are Creating the Illusion of Progress," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 3, pp. 2421–2429, Mar. 2023, doi: 10.1109/TKDE.2021.3112126.