17th International Conference on Space Operations, Dubai, United Arab Emirates, 6 - 10 March 2023.

SpaceOps-2023, ID 336

An innovative AI-based framework for on-ground anomaly detection and root cause analysis

Luca Manca* and Ilaria Bloise.†
AIKO S.r.l., Turin, Italy, 10123

Andreas Spörl‡
Deutsches Zentrum für Luft- und Raumfahrt e. V. (German Aerospace Center), 89081 Ulm, Germany

Kathrin Helmsauer§
Deutsches Zentrum für Luft- und Raumfahrt e. V. (German Aerospace Center), 82234 Weßling, Germany

Spacecraft health analysis is a critical task for mission success; detecting and announcing anomalies in spacecraft telemetry data in an automated 24/7 manner allows to enhance safety in operations and improves the timely response to critical failures. Our work proposes an innovative framework that uses state-of-the-art deep learning algorithms for anomaly detection and root cause analysis to support operators and engineers in monitoring satellite status. The proposed tool can automatically identify anomalies within the nominal operational ranges, reducing the need for human inspection and mitigating false positives. It has been tested on real spacecraft telemetry data from the TET-1 satellite of the German Aerospace Center and has shown promising results with a high rate of correct anomaly detection and low rate of false positives.

1. Introduction

Spacecraft health analysis is one of the central tasks for mission success throughout the entire satellite life-cycle. During routine operations, operators and engineers usually face the challenge of monitoring a huge quantity of telemetry parameters (more than a thousand, depending on satellite and mission complexity) to control the S/C status. Even though the criticality of problems occurring during nominal operations is lower if compared with Launch and Early Orbit Phase (LEOP), solutions are still yet to be found in order to improve the efficiency of control operations.

Furthermore, the operational experience that is dedicated during routine operations is much lower than the one devoted to LEOP, which is the highest critical phase of the mission. In fact, in case of failures occurring during LEOP and not promptly detected and isolated, the risk of mission failure is extremely high. As a result, during routine operations, the team's experience may decrease as engineers with less experience are often assigned to this phase, while the more experienced engineers are devoted to the LEOP due to its criticality.(Figure 1).

Thus, detecting and raising alarms anomalies in spacecraft telemetry data in an automated 24/7 manner allows to enhance safety in operations and improves timely response to critical failure. Additionally, the capability of analyzing the detected anomalies and automatically finding the possible root cause (i.e. parameters that caused the anomaly) could help operators during the process of anomaly inspection and satellite recovery.

Traditional monitoring systems require extensive spacecraft domain knowledge and are mainly based on Out-Of-Limit (OOL) approaches and manual analysis of plots and graphs; consequently, they are able to target only a subset of anomalies (mostly those that fall outside the nominal operational ranges). Artificial Intelligence (AI) and Deep Learning (DL) methods have been widely studied for anomaly detection tasks demonstrating promising results and showing an increase in performances and metrics on complex datasets. However, in Space System domain this topic is still under research and robust demonstrations on real spacecraft missions are just beginning to find their way into operational

*Deep Learning Engineer, AIKO S.r.l., luca.manca@aikospace.com
†Head of Machine Learning, AIKO S.r.l., ilaria@aikospace.com
‡Project Manager, Quantum Computing Initiative, andreas.spoerl@dlr.de
§Data Scientist, German Space Operations Center, kathrin.helmsauer@dlr.de

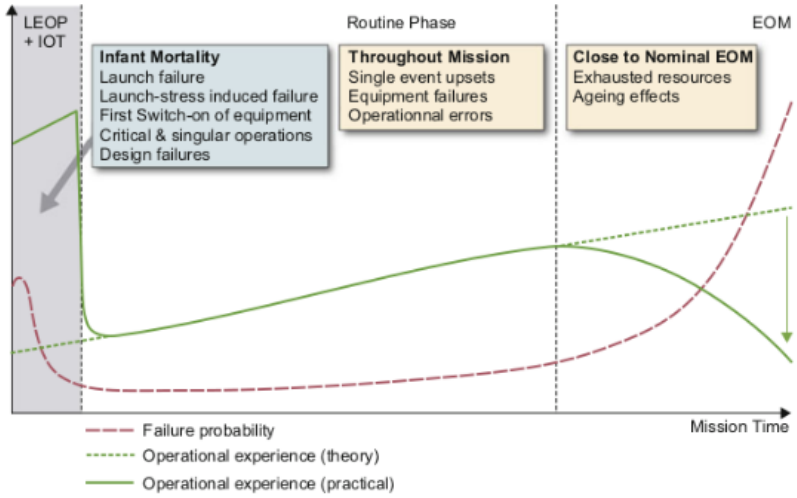

Fig. 1 Failure probability along mission time. [1]

scenarios. Moreover, one of the main drawbacks that are currently stalling the involvement of AI algorithms in real Space scenarios is a large amount of raised alarms due to false positives which could puzzle subsystem engineers and jeopardize the mission goal.

The main contributions of the proposed framework are: i) a DL-based tool with the possibility of choosing between univariate and multivariate approaches for anomaly detection; ii) a robust anomaly pruning system for false-positive mitigation; iii) a diagnosis method for root cause analysis.

2. Background and related work

In the area of spacecraft operations DL have shown great potential especially in the operation of the upcoming mega-constellations. The amount of data that is produced cannot be monitored and analysed manually by one operator. Autonomous telemetry (TM) analysis and data mining powered by AI systems is mandatory for managing large numbers of spacecraft at once. Overall, an AI-based tool can support the operator in their day-to-day tasks by autonomously preparing data, detecting anomalies and unusual behaviour, visualize the telemetry, and detect inter-dependencies of different states which are not obvious to the naked eye.

In the following sections, we discuss the shortcomings of DL approaches developed in the field of anomaly detection and root cause analysis.

A. Deep learning for anomaly detection

A review focusing on DL for anomaly detection is presented by Pang et. al. [2]. They review 12 different modelling perspectives for using DL techniques for anomaly detection and discuss challenges in anomaly detection and how the presented DL methods tackle those. Generally, the area of anomaly detection comes with its own problem areas and complexities such as unknownness of the shape of possible anomalies, heterogeneous anomaly classes, rarity, imbalance of anomalies as well as diverse types of anomalies. From this Pang et. al. derive the six challenges and problem areas: low anomaly detection recall rate, anomaly detection in high-dimensional and/or not independent data, data-efficient learning of normality/abnormality, noise-resilient anomaly detection, detection of complex anomalies and anomaly explanation.

With focus on the aerospace sector, several surveys and investigations have been undertaken focusing on data mining (DM), forecasting, and anomaly detection of spacecraft telemetry and Fault Detection, Isolation and Recovery (FDIR) with the help of various AI methods. Ibrahim et. al. [3] explore deep learning methods based on Long-Short Term Memory (LSTM) algorithms for anomaly detection. An extensive survey on data mining and ML methods for TM analysis and anomaly detection as well as a first approach on fault prediction is presented by Hassanien et.al. [4]. Furthermore, Abdelghafar et. al. [5] give an overview of the development status and research activities in intelligent health monitoring systems for space missions based on data mining. Yairi et. al. [6] propose a data-driven anomaly

detection based on relevance vector regression and principal component analysis. Moreover, they explore the possibility of hybrid approaches by using dynamic Bayesian networks but also data visualization capabilities of data-mining techniques to assist operators understanding the health status of the spacecraft. Hundman et. al. [7] present a study investigating the detection of anomalies using LSTMs, using a labelled data set from the Soil Moisture Active Passive satellite and the Mars Science Laboratory Rover Curiosity. Besides anomaly detection, the study also focuses on TM prediction and introduces a dynamic error thresholding technique. An issue that is often encountered in deep learning is a high rate of false positives. To counteract this, their approach uses anomaly pruning, where a threshold is placed based on historic data and anomalies are re-classified as nominal based on the residual or distance towards the threshold. Doudkin et. al. [8] present an ensemble of neural networks (ENN) for TM forecasting. Delande et. al. [9] proposed a tool, called Nostradamus, based on a one-class SVM, which has been in operation for more than 6 years. O'Meara et. al. [10] developed an Automated Telemetry Health Monitoring System (ATHMoS), a novel algorithm for statistical outlier detection which can be runned on a parameter by parameter basis, as well as on a multi-parameter level. Finally, Tuli et. al. [11] developed a Transformer-based architecture for anomaly detection and diagnosis in multivariate time-series data. Additionally, they adopted an error thresholding technique named Peak-Over-Threshold (POT).

B. Deep learning for root cause analysis

Amiruddin et. al. [12] give an overview of neural networks used in fault diagnosis and detection and its implementation in engineering related systems. Mansell and Spencer [13] present a data-driven fault detection and isolation architecture that uses one class support vector machines (SVM) to provide a running time series of fault signals to an LSTM neural network for isolation. Gau et. al. [14] use a combination of principal component analysis (PCA) and SVM for fault detection and diagnosis in spacecraft systems. Firstly, PCA is used to extract features from input data and reduce the input data to low dimensional feature vectors. Then the method uses a binary SVM to detect whether there is a fault or not. If the fault is detected, a multi-class SVM is used to identify the fault type. Yairi et. al. [6] explain the machine-learning-/data-mining-based approach to the anomaly and fault detection issues for spacecraft systems and introduce several specific methods based on this concept. Nozari et. al. [15] suggest a model-free framework for fault detection and isolation of satellite reaction wheels.

3. Algorithm trade-off

This section describes the overall trade-off that led to the identification of the approaches selected to perform anomaly detection and root cause analysis. Firstly, an introduction of the different ML techniques is provided, followed by a short trade-off between different state-of-the-art DL architectures. Furthermore, the reconstruction-based anomaly detection technique is detailed, together with univariate and multivariate approaches. Finally, a brief overview of root cause analysis techniques is given.

A. Anomaly detection

Supervised vs unsupervised vs semi-supervised learning

In the literature there are three main approaches for solving anomaly detection tasks: supervised, unsupervised and semi-supervised learning techniques.

- Supervised learning: in supervised learning the algorithm solves a classification task by learning if a data point is nominal or not. For using this approach, one needs to have access to a sufficient number of nominal and anomalous examples. While having access to nominal telemetry data is usually effortless in spacecraft operations, the same is not true for anomalous behavior, where usually only few examples are available. This makes it hard to approach anomaly detection applications related to spacecraft domain with a supervised approach.
- Unsupervised learning: in unsupervised settings the algorithm does not have access to input-output examples. In other words, it does not know if a data point is nominal or anomalous. This task is usually tackled by means of clustering-based machine learning algorithms and it is also referred to as "Outlier anomaly detection". In fact, an outlier is defined as an observation which deviates from other observations (abnormal sample). No labels are provided but the algorithm tries to find the datapoints which do not conform to the majority of the observation. However, as mentioned in the previous point, since in spacecraft operations the availability of nominal telemetries is usually known, this information can be used to tackle the problem in a Semi-Supervised Learning approach.
- Semi-Supervised Learning: by definition, semi-supervised learning techniques take advantage of large amounts of unlabeled data as well as a small number of labeled samples. In our scenario, the labeled examples are the

nominal periods of data as provided by the operators' engineers. This task is usually referred to as "Novelty anomaly detection". In fact, by definition a novelty is an observation which deviates from known normal observations. Under this setting, the algorithm learns from known nominal data patterns (that have been labeled), and flags new observations that deviate significantly from known nominal patterns. Due to ease availability of nominal data points, this is thought to be the approach that best fits the application of spacecraft anomaly detection.

Deep learning models

In terms of DL models suitable for time-series anomaly detection, Table 1 provides pros and cons of the state-of-the-art architectures.

Table 1 Deep learning state-of-the-art models [16].

Networks	Advantages	Limitations
Autoencoder	No prior data knowledge needed, can fuse multi-sensory data and compress data, easy to combine with classification and regression methods.	Needs a lot of data for training, cannot determine what information is relevant, not as efficient in reconstructing as GANs, it introduces a deterministic bias.
Convolutional Neural Network (CNN)	Outperforms Artificial Neural Networks (ANN) on many tasks (e.g., image recognition), would be less complex and saves memory compared to ANNs, automatically detects the important features without any human supervision.	Hyperparameter tuning is non-trivial, easy to overfit, high computational cost, needs a massive amount of training data.
Recurrent Neural Network (RNN)	Models time sequential dependencies.	Gradient vanishing and exploding problems, cannot process very long sequences if using *tanh* or *relu* as an activation function.
Deep Belief Network (DBN)	Has a layer-by-layer procedure for learning the top-down, generative weights, no requirement for labelled data when training, robustness in classification.	High computational cost.
Generative Adversarial Network (GAN)	A good approach to train a classifier in a semi-supervised way, does not introduce any deterministic bias compared to autoencoders, can be used to address the class imbalance issue, where the NN is biased because most of the data belongs to a single label.	The training is unstable due to the requirement of a Nash equilibrium, the original GAN is hard to learn to generate discrete data.
Transformer	State-of-the-art model in the Natual Language Processing (NLP) domain, self-attention mechanism, non-sequential process to allow parallel computation.	Hard to control its attention, requires a huge amount of training samples, still not widely used in time-series analysis.

LSTM-based models as well as Transformer-based architectures were selected to be explored during the activity. This choice was made with the aim of comparing innovative research (as for Transformer) with more stable models and implementation (as for LSTM models).

It is however understood that, more than the actual deep learning model implemented, the focus must be given to the actual pipeline for identifying anomalous data. Considering the semi-supervised learning framework, the approach most widely used is called "reconstruction-based anomaly detection".

Reconstruction-based anomaly detection

The main concept of a reconstruction-based anomaly detection is to let the model learn the distribution of available normal data (Figure 2) and then using the trained model for time-series prediction to reconstruct the data. (Figure 3).

The error between the reconstructed data and the original one is used to determine if a behaviour is anomalous. A high error is a strong indication of the presence of outlier data. To correctly classify the anomaly it is crucial to tune a

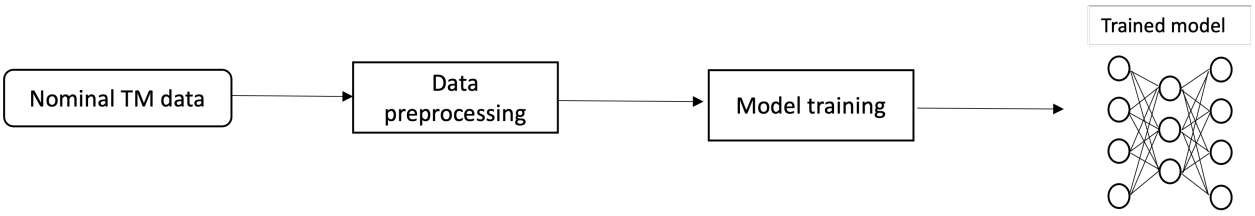

Fig. 2 Training pipeline schema.

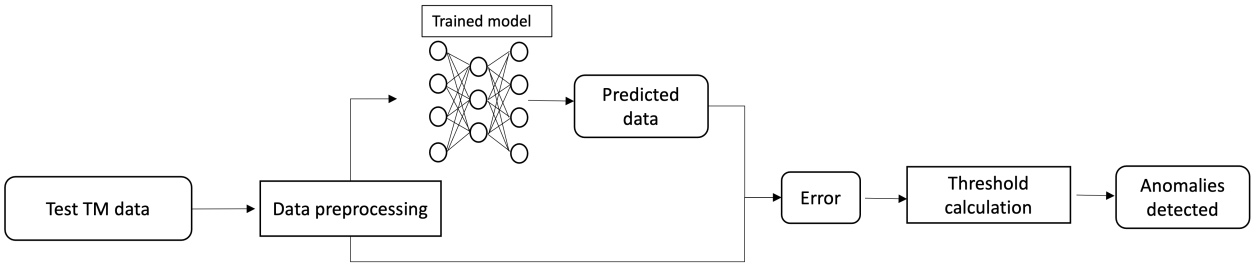

Fig. 3 Inference pipeline schema.

proper threshold beyond which the alarm indicating the abnormality is triggered. Therefore, the rules on the error can be summarized as:

- Error higher than the threshold: anomalous data;
- Error lower than the threshold: nominal data .

According to this general process, one of the main constraints is the identification of nominal TM data that can be used for model training.

The reconstruction-based anomaly detection method needs a threshold selection and tuning. The importance of this method lies in achieving robustness against model uncertainty in combination with sensitivity to small faults so that it is possible to reduce the number of false positives in the detected anomalies as as much as possible. In fact, with a very low threshold many anomalies will be detected, and this would imply an unjustified effort by the operators to investigate the nature of the problem raised. On the other hand, having a high threshold value could lead to missing true positives and therefore jeopardize the mission goal if the loss of anomaly is of high severity for the spacecraft health status.

Nowadays, the most used technique for thresholding purposes is the Out-Of-Limit (OOL) method which consists of defining a nominal range with lower and upper thresholds. This method is heavily used for satellites. When one or more parameters exceed an upper or lower threshold it will trigger an alarm. Since satellites are becoming sophisticated and complex, this approach might show room for improvement with a huge number of parameters to monitor due to the following considerations:

- it requires a significant amount of domain knowledge and expertise from operators to set up consistent boundaries over all the parameters;
- it requires expertise for each abnormal behavior expected to evaluate the ones that exceed the fixed boundaries;
- it is not suitable for detecting anomalies inside the nominal range, as a consequence it is not robust enough to detect the various types of anomalies.

Concerning the error thresholding technique, two approaches have been widely developed and used among the community providing promising results. These are known as Non-Parametric Dynamic Threshold (NDT) [7] method and Peak Over Threshold (POT) method [17].

Univariate vs. multivariate approach

Concerning dataset modelling, two different approaches can be adopted: univariate and multivariate. Univariate means that a single model is trained for each telemetry parameter, while in a multivariate approach, a single model is trained for the subset of telemetries.

Table 2 Univariate vs. multivariate approach.

	Advantages	**Disadvantages**
Univariate	The training of every model is usually easier, and it allows the tracking of each channel independently.	In case many parameters need to be analyzed, a high number of models need to be trained, leading to a high computational time required during training.
Multivariate	It allows to have a single model for a subset of parameters, catching also the relations between all of them.	The training is usually more challenging and model complexity increases.

In terms of anomaly detection techniques, there is no difference between univariate and multivariate models, in the sense that in both cases, the reconstruction-based anomaly detection can be applied.

To thoroughly evaluate the pros and cons of both approaches, we have decided to implement both the techniques.

B. Root cause analysis

Concerning root cause analysis, far less resources are available compared to anomaly detection. It is in fact known that, especially in spacecraft operations, it is a topic that is nearly unexplored within AI/ML. Additionally, the studies performed in this area (as [14]), are more related to fault detection and classification, which is a slightly different topic compared to root cause analysis: In anomaly classification, the goal is to detect an anomaly and classify it. In root cause analysis on the other hand, the goal is to find the cause that lead to the fault propagation. In other words, one possible outcome, which has been the focus of this work, has been to try to find a subset of telemetries that might be the cause of the anomalies, and those that the operators must consider first when investigating a particular anomaly.

4. Methods

A. High-level system overview

The pipeline that we developed for anomaly detection and root cause analysis is depicted in Figure 4.

As can be seen, there is a unique flow that is in charge of identifying anomalies and then, for each detected anomaly, performing the root cause analysis.

The developed framework is divided into two main steps: training and inference. During training, depicted in the lower part of the figure, nominal TM data are selected and used to train the deep learning models. During inference, represented in the upper part of the figure, test TM data, that might contain both nominal as well as anomalous behavior, are fed into the trained models to detect potential anomalies together with their related root cause candidates.

B. Detailed system description

1. Pre-processor

As it can be expected when dealing with spacecraft telemetry time-series, since each parameter is usually sampled with differing resolutions over time, dataset pre-processing is required. As an example, a telemetry might be sampled with a nominal sampling rate of 30s, while it can increase during ground station contacts or when requested from the ground. In fact, to feed a DL model with time series telemetry data, it is necessary that the input data is sampled with the same sampling rate. As an example, if the input data is constituted by a single telemetry parameter, this parameter needs to be sampled with the same sampling rate for all the time periods to be analyzed; the same holds if the input data is characterized by multiple telemetry data: in this case the sampling rate needs to be the same also across all the involved parameters. To handle this issue, during pre-processing, data is resampled and interpolated according to a pre-defined sampling rate which is identified common between all the involved parameters. Though resampling and interpolation may be essential for certain machine learning tasks, they can also introduce certain complications. These include altering the true nature of the data and potentially sacrificing the granularity of the original data.

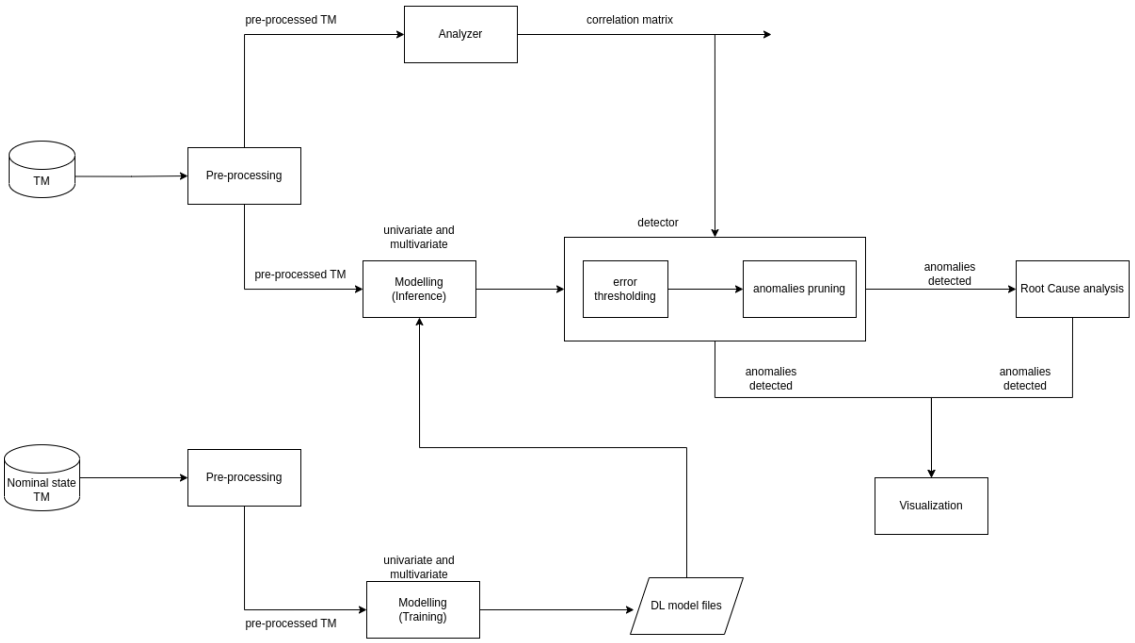

Fig. 4 High-level system overview.

2. Analyzer

The analyzer performs dataset exploration in order to extract useful and interesting information from the preprocessed telemetry data. The main outputs are the correlation matrices that compute the correlation indices between a particular set of parameters. In addition to the primary outputs, supplementary information is also extracted, such as information regarding the presence of data gaps (period of time when telemetry could not be downlinked) and various statistical information, including the mean and standard deviation of the telemetry and the primary sampling rates used for each telemetry.

3. Modelling

Inside the modelling component, it is possible to perform model training and inference. The model solves a time-series forecasting problem and the idea is to let it learn the nominal telemetry behavior so that, during inference, by comparing the model prediction to the original points, it is possible to detect potential anomalous behavior.

Two different approaches are adopted:

- univariate: one model trained for a single telemetry parameter;
- multivariate: one model trained for a subset of telemetry parameter.

For the univariate approach, an LSTM-based model has been implemented and tested, while for the multivariate one, both the LSTM and the transformer-based model has been tested.

Concerning the window length of the prediction for univariate approach, the model is feed with telemetries with a period equal to three orbits and it predicts one orbit in advance. Thus, since TET-1 has an orbital period approximately equal to 90 minutes and since the sampling rate chosen for resampling and interpolation has been set equal to 1m, the model is fed with time-series of 270 points to predict the future 90 points of telemetries.

4. Detector

The detector module is composed of two main steps:

1) Error threshold, that receives the model predictions as well as the original telemetries as input, it computes the error between them and, with a thresholding method, it outputs the set of anomalies detected.
2) Anomalies pruning, that receives all the detected anomalies and prunes the potential false positives.

Concerning the error thresholding approach, two techniques have been explored: Non-Dynamic Error Thresholding (NDT) [7] and Peak-Over-Threshold (POT) [17].

The detected anomalies are then input into the anomalies pruning step where the potential false positives are pruned.

5. Anomalies pruning

The methodology behind this approach is to loop through the time period analyzed during processing and compute an anomaly pruning score, which is then threshold to remove potential false positives. Different from the processor, where the detection only takes the error trend of a singular parameter into account, here the pruning considers the behavior at a system level, using as inputs all the anomalies detected previously.

Figure 5 shows the schema behind this approach:

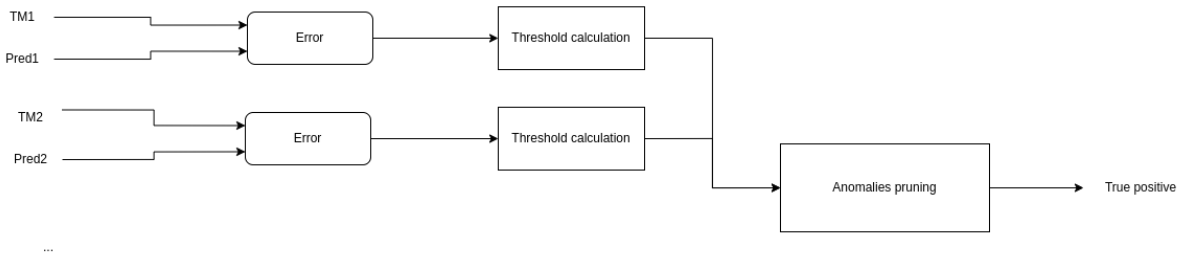

Fig. 5 Anomalies pruning approach.

The anomaly pruning score is computed with Equation 1:

$$AP_{score} = \frac{\sum_{i=1}^{N} a_i \cdot \frac{t_i}{W_t}}{N_{params}} \cdot \frac{\sum_{i,j=1}^{N} a_i \cdot a_j \cdot corr_{i,j}}{\binom{a_{tot}}{2}} \tag{1}$$

where:

- a_i is 1 if an anomaly is detected for parameter i, else is 0;
- t_i is the duration of anomaly a_i;
- W_t is the length of the sliding window considered;
- N_{params} is the total amount of parameters analysed;
- $corr_{i,j}$ is the correlation value between $param_i$ and $param_j$;
- a_{tot} is the total amount of anomalies detected in a given time window.

Two main contributions accounted in this equation are:

1) Number of anomalies weighted by their time length: anomalies that last for a longer time should have a higher contribution, thus be less inclined to pruning;
2) Correlations between the anomalous parameters: usually a "real" anomaly will manifest also in the parameters that are correlated to it.

The pruning score computed is then threshold using a simple outlier detection approach: firstly, the mean and standard deviation of the function is computed, then all the values below one standard deviation. Additionally, an anomaly severity (or score) is also added in the following way:

- Low severity: if $\mu(AP_{score}) + \sigma(AP_{score}) < AP_{score} < \mu(AP_{score}) + 2\sigma(AP_{score})$
- Medium severity: if $\mu(AP_{score}) + 2\sigma(AP_{score}) < AP_{score} < \mu(AP_{score}) + 3\sigma(AP_{score})$
- High severity: if $AP_{score} > \mu(AP_{score}) + 3\sigma(AP_{score})$

The advantage of using such a pruning score lies in the fact that with this technique it is possible to analyse the anomaly detection problem from a system perspective and not only from a parameter-level analysis, thus possibly reducing the total number of false positive that the algorithm might detect. The main downside of this approach is that, especially in cases where the parameters being analyzed come from different subsystems and thus, they might not be highly correlated, the 2^{nd} contribution might become less relevant. In cases like that, there might be the possibility that anomalies which involve only one or a few parameters might be erroneously pruned.

6. Root cause analysis

The detected anomalies are further processed with the objective of finding a set of possible root cause candidates.

A sliding window approach is used to loop over the anomalies and, for each of them, the error between model predictions and original telemetry is retrieved. The parameters not involved in the anomaly are ranked by the error and marked as the possible root cause candidates.

Thus, the process can be described with the following steps:

- Loop on the entire time period using a sliding window
- If an anomaly is encountered
    1) Identify nominal parameters (the ones not involved in the anomaly in the considered time window)
    2) Rank the nominal parameters based on the prediction error

7. Visualizer

The results are uploaded in InfluxDB, an open-source time-series database and then displayed into Grafana, an open-source analytics and interactive web application, to be analysed in more depth.

The original telemetry streams are compared with the model prediction, and, with the information of the raised alerts, it is possible to perform further investigation to understand if the detected anomalies might be false positives or not.

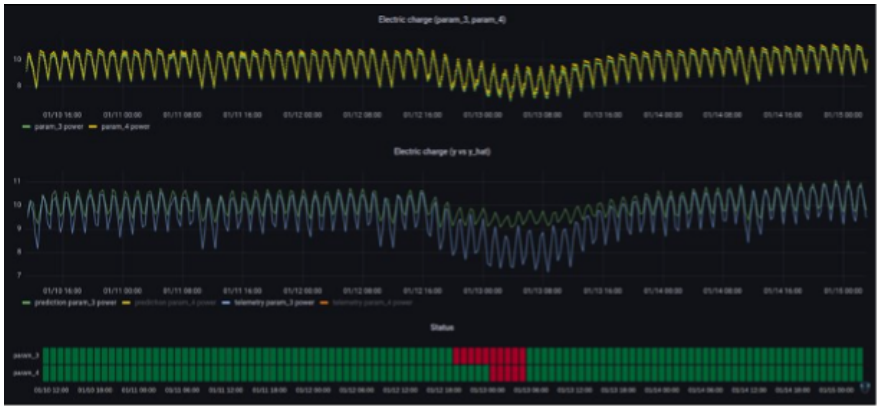

Fig. 6 Grafana dashboard example.

5. Experiments

A. Dataset

Dataset used under this research come from TET-1, a satellite operated by the German Aerospace Center (DLR/GSOC). TET-1 available dataset comprises of a set of telemetry points covering the time period that goes from 01/01/2013 up to 31/12/2013. A set of time windows where the data is identified as nominal is known, while for the other periods no information was available. Table 3 provides a summary regarding the known nominal periods of the dataset.

Table 3 TET-1 available data and event information.

Event	Timestamp
Nominal	24/01/2013 - 04/02/2013
Nominal	18/02/2013 - 02/03/2013
Nominal	29/04/2013 - 12/05/2013
Nominal	11/07/2013 - 24/07/2013
Nominal	02/08/2013 - 14/08/2013
Nominal	26/08/2013 - 13/09/2013
Nominal	29/10/2013 - 11/11/2013

B. Experiments details

Concerning model experiments, three main architectures have been trained, one for the univariate configuration and two for the multivariate.

Table 4 Experiments details.

Experiments	Approach	Model	Error thresholding	Parameters
exp1	Univariate	LSTM-based	NDT	17
exp2	Multivariate	LSTM-based	NDT	13
exp3	Multivariate	Transformer-based	POT	13

Concerning the multivariate approach, some TMs are discarded since they are sampled with a completely different sampling rate with respect to the other, making it impossible to be modelled with a unique model because of strong deviations and interpolation to be applied to these different parameters with the risk of changing a lot the nature of the parameters itself.

C. Results and assessment

To perform algorithm assessment and verify the correctness of the detected anomalies, the following methodology has been adopted (Figure 7).

In a first step, the obtained DL results are compared to the official anomaly reports written by the satellite operators. If they found a match, the anomaly was confirmed. Unfortunately, the anomaly reports often don't show the full picture as they are written immediately when a new anomaly occurs. An anomaly above a certain criticality level has to be mitigated as soon as possible and the root cause analysis and/or further steps to mitigate the problem in the future, which might take more time, will then be tracked in an offical quality assurance report. Hence, no match does not imply that there was no anomaly.

If there was no match, we have then checked the list of executed telecommands (TCs). Common issues here can be a running experiment and/or a reboot, which cause unusual behavior in the satellite telemetry. In this case, the anomaly is marked as a novelty due to TCs.

Finally, as a last step, the most resource intensive option is used: get an operator to have a detailed look at the plots. If they can confirm that the behavior is anomalous, the anomaly is confirmed. Otherwise, it is marked as a false positive.

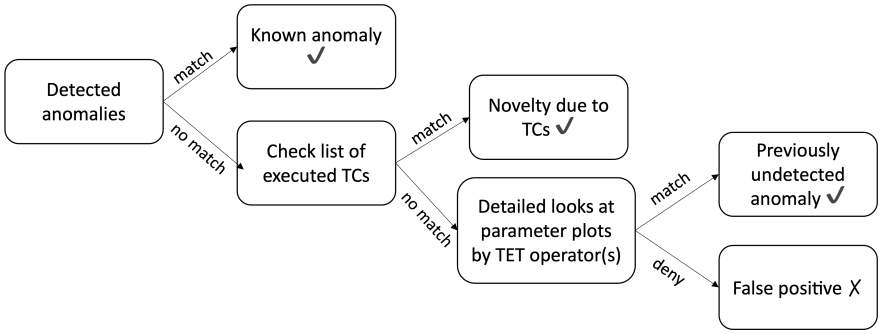

Fig. 7 Assessment flowchart.

In total, there were 18 anomalies reported by the operators in 2013. Sometimes, two of those anomalies happened so close together that the DL model used detected one longer timeframe including two anomalies instead of two short timeframes, one for each anomaly. Hence, the number of detected anomalies according to the anomaly reports is higher than the number of detections.

In the results shown below, the anomalies caused by TCs are considered as true positives since, from a model perspective, it is correct that those are detected. In fact, since the model does not have the knowledge of the TCs that are sent to the spacecraft, it would be impossible to detect them. Surely, in a real scenario, operators may not be alerted by the system following a command and this must be properly taken into consideration for future steps.

The assessment has been performed considering only the high and medium severity anomalies detected.

1. Exp1 - Univariate - LSTM-based model

The overall validation metrics for the medium and high-risk detections are reported in Table 5.

Table 5 Confusion matrix, univariate approach, LSTM-based model.

	Positive	**Negative**
Positive	39 (TP)	3 (FN)
Negative	1 (FP)	- (TN)

Out of the 39 positive detections, 12 correspond to anomalies as reported in 15 operators' anomaly reports and 27 were expected novelties caused by TCs, e. g. a commanded experiment or software updates. Out of the 18 anomaly reports for 2013, 15 were covered by positive detections and 3 were missed.

2. Exp2 - Multivariate - LSTM-based model

The overall validation metrics for the medium and high-risk detections are reported in Table 6

Table 6 Confusion matrix, multivariate approach, LSTM-based model.

	Positive	**Negative**
Positive	37 (TP)	6 (FN)
Negative	0 (FP)	- (TN)

Out of the 37 positive detections, 10 correspond to anomalies as reported in 12 operators' anomaly reports and 27 were expected novelties caused by TCs, e. g. a commanded experiment or software updates. Out of the 18 anomaly reports for 2013, 12 were covered by positive detections and 6 were missed.

3. Exp3 - Multivariate - Transformer-based model

The overall validation metrics for the medium and high-risk detections are reported in Table 7

Table 7 Confusion matrix, multivariate approach, Transformer-based model.

	Positive	**Negative**
Positive	33 (TP)	8 (FN)
Negative	0 (FP)	- (TN)

Out of the 33 positive detections, 12 correspond to anomalies as reported in 12 operators' anomaly reports and 21 were expected novelties caused by TCs, e. g. a commanded experiment or software updates. Out of the 18 anomaly reports for 2013, 10 were covered by positive detections and 8 were missed.

As a summary, the achieved results are promising. All but one of the detections were actual anomalous behavior, either as an anomaly or as a novelty caused by TCs. Our analysis of the medium and high-risk detections did not yield any previously undetected anomalies as every detection has either been recorded in an anomaly report or caused by TCs. We assume that an analysis of the low-risk detections would yield results here.

6. Future work

This section describes the possible future technical improvements to boost the performance in the next steps forward.

First of all, the number of telemetries monitored by the tool can be increased. While during the research, the analysis has been limited to only the 17 parameters that are most monitored from the operators, it is important to extend this analysis in future to consider a higher number of telemetries. Of particular interest can be the case in which a complete spacecraft subsystem (e.g., Attitude Determination and Control System (ADCS), thermal, power etc.) is modelled by the proposed technology. The scalability of the framework is crucial for future applications such as constellations.

Secondly, the robustness of the anomalies pruning step can be enhanced. The idea behind the overall approach proposed is to reason at a system level and keep only the most severe anomalies, those that are common to a set of parameters correlated to each other. However, this approach might be lacking:

- If the telemetries under consideration are only a tiny fraction of the spacecraft parameters, and those are just slightly correlated to each other, the anomalies pruning assumption might fall. The subcase considered during the work (taking into account only the most relevant parameters - 17) is already close to the validity boundary, since many more telemetries characterize TET-1 spacecraft.
- The equation which is set up to compute the anomaly's severity and then used to threshold the possible false positives, is lacking from an explainability component. While it hides some physical elements (e.g., telemetries correlations), it is still quite challenging to understand why an anomaly is pruned without properly analyzing the telemetries.

Furthermore, TCs should be included in the analysis. As shown in the obtained results, many anomalies are due to TC sent to the spacecraft which led to a slight change in the telemetries' trend. From an algorithm point of view, those anomalies are legitimate: Since the nominal telemetry behavior has changed, it is correct to identify them as anomalous. However, from an operational point of view, those are surely false positives, and should not be marked as anomalous.

Finally, the root cause analysis technique applied must be enhanced and properly assessed. The root cause analysis process developed in this work represents a first attempt into the possibility of automatically investigating the anomalies detected to output a list of potential root cause candidates. This can be seen as an extremely useful output from the operators since it can allow them to automatically shortlist the telemetries to look at when anomalies appear and faster the anomalies investigation process. However, the proposed approach can be enhanced considering the following points:

- When looking at an anomaly, it makes much more sense to limit the root cause analysis timespan only to the period that preceded the anomaly. As it has been implemented here instead, the entire anomaly period (from which the telemetry gets anomalous till when all of them return nominal) is considered.
- From an operational point of view, instead of considering the model errors as leader discriminative when identifing anomalies' root cause candidates, it might be most relevant to look at variations in the correlations between the telemetries, before the anomaly appeared.

7. Conclusions

The work presented represents a first attempt into the direction of having a tool capable of supporting the operator during their daily activities. Artificial intelligence and deep learning algorithms for anomaly detection and root cause analysis can have an enormous impact on the operators. Such an anomaly detection algorithm can provide the operators a concise list of anomalous behavior during the past day. Compared to the current practice, where operators manually inspect the parameter plots, this can save a significant amount of time that can instead being spent in identifying the root cause and ideally prevent such anomalies in the future. Additionally, the algorithm can prevent future outages by notifying the operator of small but significant changes in the satellite's telemetry, that the operators might not have seen when inspecting the plots. Again, the operators can inspect this behavior and intervene long before a high priority anomaly might have caused an outage of the satellite.

Acknowledgments

The work presented here was carried in the frame of the DL4SPACE project, funded by the European Space Agency (ESA Contract No. 4000135426/21/D/AH). The authors would like to thank the ESA's staff and in particular Gabriele De Canio and Patrick Fleith for their insightful comments, support, and guidance throughout.

References

[1] Pasquier, H., Cruzen, C., Schmidhuber, M., and Lee, Y., *Space Operations: Inspiring Humankind's Future*, 2019. https://doi.org/10.1007/978-3-030-11536-4.

[2] Pang, G., Shen, C., Cao, L., and Hengel, A., "Deep Learning for Anomaly Detection: A Review," *ACM Computing Surveys*, Vol. 54, 2021, pp. 1–38. https://doi.org/10.1145/3439950.

[3] Ibrahim, S. K., Ahmed, A., Zeidan, M. A. E., and Ziedan, I. E., "Machine Learning Methods for Spacecraft Telemetry Mining," *IEEE Transactions on Aerospace and Electronic Systems*, Vol. 55, No. 4, 2019, pp. 1816–1827. https://doi.org/10.1109/TAES.2018.2876586.

[4] Hassanien, A. E., Darwish, A., and Abdelghafar, S., "Machine learning in telemetry data mining of space mission: basics, challenging and future directions," *Artificial Intelligence Review*, Vol. 53, 2020. https://doi.org/10.1007/s10462-019-09760-1.

[5] Abdelghafar, S., Darwish, A., and Hassanien, A. E., *Intelligent Health Monitoring Systems for Space Missions Based on Data Mining Techniques*, 2020, pp. 65–78. https://doi.org/10.1007/978-3-030-20212-5_4.

[6] Yairi, T., Kawahara, Y., Fujimaki, R., Sato, Y., and Machida, K., "Telemetry-mining: a machine learning approach to anomaly detection and fault diagnosis for space systems," *2nd IEEE International Conference on Space Mission Challenges for Information Technology (SMC-IT'06)*, 2006, pp. 8 pp.–476. https://doi.org/10.1109/SMC-IT.2006.79.

[7] Hundman, K., Constantinou, V., Laporte, C., Colwell, I., and Söderström, T., "Detecting Spacecraft Anomalies Using LSTMs and Nonparametric Dynamic Thresholding," *CoRR*, Vol. abs/1802.04431, 2018. URL http://arxiv.org/abs/1802.04431.

[8] Doudkin, A., Marushko, Y., Owsiński, J., and Pawlowski, T., "Spacecraft Telemetry Time Series Forecasting With Ensembles of Neural Networks," *2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, Vol. 2, 2019, pp. 752–756. https://doi.org/10.1109/IDAACS.2019.8924252.

[9] P. Delandea, M. B. M. Z. A. B., P.-B. Lambertb, "AI for Satellite Anomaly Detection: On-Ground Operational Feedback and Development of On-Board Experiments," *Proceedings of the 73rd International Astronautical Congress (IAC)*, Vol. 2, 2022.

[10] O'Meara, C., Schlag, L., Faltenbacher, L., and Wickler, M., "ATHMoS: Automated Telemetry Health Monitoring System at GSOC using Outlier Detection and Supervised Machine Learning," 2016. https://doi.org/10.2514/6.2016-2347.

[11] Tuli, S., Casale, G., and Jennings, N. R., "TranAD: Deep Transformer Networks for Anomaly Detection in Multivariate Time Series Data," *CoRR*, Vol. abs/2201.07284, 2022. URL https://arxiv.org/abs/2201.07284.

[12] Mohd Amiruddin, A., Zabiri, H., Taqvi, S. A. A., and Tufa, L. D., "Neural network applications in fault diagnosis and detection: an overview of implementations in engineering-related systems," *Neural Computing and Applications*, Vol. 32, 2020. https://doi.org/10.1007/s00521-018-3911-5.

[13] Mansell, J., and Spencer, D., "Data-driven Fault Detection and Isolation for Small Spacecraft," 2019.

[14] Gao, Y., Yang, T., Xing, N., and Xu, M., "Fault detection and diagnosis for spacecraft using principal component analysis and support vector machines," *2012 7th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, 2012, pp. 1984–1988. https://doi.org/10.1109/ICIEA.2012.6361054.

[15] Abbasi Nozari, H., Castaldi, P., Banadaki, H., and Simani, S., "Novel Non-Model-Based Fault Detection and Isolation of Satellite Reaction Wheels Based on a Mixed-Learning Fusion Framework," *IFAC-PapersOnLine*, Vol. 52, 2019, pp. 194–199. https://doi.org/10.1016/j.ifacol.2019.11.222.

[16] Ran, Y., Zhou, X., Lin, P., Wen, Y., and Deng, R., "A Survey of Predictive Maintenance: Systems, Purposes and Approaches," , 2019. https://doi.org/10.48550/ARXIV.1912.07383, URL https://arxiv.org/abs/1912.07383.

[17] Rosso, G., "Extreme Value Theory for Time Series using Peak-Over-Threshold method," , 2015. https://doi.org/10.48550/ARXIV.1509.01051, URL https://arxiv.org/abs/1509.01051.