Unsupervised Anomaly Detection for Satellite Telemetry based on Change-Point Detection

Adriano Puglisi
Dep. of Computer, Control, and Management Engineering Antonio Ruberti
Sapienza Università di Roma, Rome, Italy
puglisi@diag.uniroma1.it

Laurent Oudre
Université Paris-Saclay, Université Paris Cité, ENS Paris-Saclay, CNRS, SSA, INSERM, Centre Borelli
F-91190, Gif-sur-Yvette, France
laurent.oudre@ens-paris-saclay.fr

***Abstract*—Monitoring satellite telemetry for anomalies is a critical task in spacecraft operations, yet existing unsupervised detectors evaluated on the recent ESA Anomaly Detection Benchmark achieve near-zero event-wise precision despite high recall, making them operationally unusable on long telemetry streams. We propose a label-free anomaly detection pipeline that decomposes the task into regime-change detection and anomaly decision, operating in the time-frequency domain through the Short-Time Fourier Transform. Candidate changepoints are identified by applying the Pruned Exact Linear Time algorithm with an automatic penalty selection strategy based on knee detection over a logarithmic sweep. A per-dimension gain analysis with an elbow criterion retains only the channel-frequency components that genuinely contribute to each detected regime change, and a final gain gate calibrated from unlabeled telemetry suppresses residual false positives. A semi-supervised extension further reranks candidates with a Random Forest classifier trained on benchmark annotations. Evaluated on the lightweight channel subset of Mission 1 of the benchmark, the unsupervised variant achieves an event-wise F-score competitive with the classical unsupervised baselines, while the semi-supervised variant approaches the performance of the best forecasting pipeline without requiring manual mission-specific threshold engineering.[1]**

***Index Terms*—anomaly detection, satellite telemetry, changepoint detection, PELT, STFT**

I. Introduction

Monitoring satellite telemetry for anomalies is a critical and continuous task performed by spacecraft operations engineers in mission control centres worldwide. Modern satellites generate time series data across hundreds of channels, each recording physical measurements, status flags, counters, and telecommands at varying sampling rates. Detecting anomalies in such data early and reliably is essential for ensuring the safe and uninterrupted operation of missions.

Current operational practice relies primarily on limit checking, alarming when a measurement exceeds predefined nominal bounds, supplemented by manual inspection for anomalies that do not trigger simple threshold violations. While effective for well-understood failure modes, this approach scales poorly with the growing complexity of satellite systems and is ill-suited to detecting subtle, multivariate, or previously unseen anomalies. The need for more sophisticated, data-driven anomaly detection methods has been recognised by all major space agencies, including ESA, NASA, CNES, DLR, and JAXA, several of which have released telemetry datasets and machine learning solutions in recent years [1].

Although several agencies have recently released telemetry datasets, real anomaly annotations remain extremely sparse and operationally heterogeneous, making objective evaluation and supervised calibration difficult across missions. This difficulty is compounded by well-documented flaws in the evaluation protocols and benchmarks commonly used in the field, which can produce misleadingly optimistic results [2]. The recently released ESA Anomaly Detection Benchmark (ESA-ADB) [3] addresses this gap by providing a curated, annotated dataset of real telemetry from three ESA missions together with a rigorous evaluation pipeline.

The benchmark results show that virtually all unsupervised algorithms included in ESA-ADB, including classical outlier detectors such as Histogram-Based Outlier Score and Isolation Forest [4], achieve near-zero event-wise precision despite high recall, resulting in a near-zero $F_{0.5}$ score. This behaviour reflects a broader limitation of many unsupervised time-series anomaly detection methods, which tend to flag large portions of long telemetry streams as anomalous, making them operationally unusable regardless of their recall [5]. In this paper we propose a label-free anomaly detection framework that models anomalies as regime changes in the time-frequency domain. The approach combines changepoint detection with a sequence of data-driven filtering stages designed to suppress spurious detections and reduce false alarms.

Two variants are considered, an unsupervised version named uCPAD (unsupervised Change-Point Anomaly Detector) operating without anomaly annotations and a semi-supervised extension using a lightweight classifier named sCPAD (semisupervised Change-Point Anomaly Detector). Experimental results on ESA-ADB are reported and compared against existing baselines.

The remainder of the paper is organised as follows. Section II reviews related work on time-series anomaly detection and changepoint detection for satellite telemetry. Section III provides an overview of the proposed approach and summarises its main contributions. Section IV presents the methodology in detail. Section V reports the experimental results and benchmark comparison. Finally, Section VI con-

[1]Code available at https://github.com/Adrianoski/DSAA

cludes the paper.

II. Related Work

Research on anomaly detection for satellite telemetry can broadly be divided into two categories. The first consists of operational and mission-specific approaches developed within the space telemetry community, typically based on thresholding, forecasting, or highly specialised pipelines adapted to a given spacecraft. The second consists of general-purpose time-series anomaly detection and changepoint detection methods developed in the broader machine learning and signal processing literature.

A. Anomaly Detection in Satellite Telemetry

The dominant operational approach to anomaly detection in satellite telemetry remains limit checking, in which scalar thresholds are defined per channel by domain experts and an alarm is raised whenever a measurement leaves its predefined nominal range. While simple and interpretable, this strategy requires continuous expert maintenance, and is blind to anomalies that manifest as changes in the joint behaviour of several channels rather than as out-of-range values in any single one.

Data-driven approaches have been investigated as a scalable alternative, the majority of which are forecasting- or reconstruction-based deep models trained on nominal operational behaviour. One of the most widely adopted baselines is Telemanom [1], proposed by NASA engineers, which trains an LSTM network to predict each telemetry channel from its recent history and flags intervals where the prediction error exceeds a nonparametrically estimated threshold. Similar approaches include prediction-error models based on bidirectional and Bayesian LSTM variants [6], deep clustering methods for nominal operational patterns [7], and Transformer-based architectures designed to model long-range temporal dependencies in telemetry [8].

A recurring limitation across this family of methods is their dependence on nominal training sets and threshold calibration procedures that are often mission-specific. As a consequence, portability across spacecraft, payloads, and operational regimes remains difficult.

Progress in telemetry anomaly detection has long been limited by the scarcity of realistic public benchmarks. The SAT benchmark [9] provides annotated telemetry from an ESA nanosatellite for single-channel anomaly detection, while the ESA Anomaly Detection Benchmark (ESA-ADB) [3] currently provides the most comprehensive operational evaluation framework, combining multi-channel telemetry from ESA missions with hierarchical event-based metrics, including the corrected event-wise F_β-score.

The ESA-ADB evaluation highlights an important distinction between unsupervised and semi-supervised approaches. Classical unsupervised detectors such as Pearson Correlation Coefficient (PCC), Histogram-Based Outlier Score (HBOS), Isolation Forest (iForest), and K-Nearest Neighbours (KNN) achieve high recall but near-zero event-wise precision, indicating extensive false-positive activity on long telemetry streams. In contrast, forecasting-based pipelines with additional thresholding and post-processing stages achieve higher results.

B. Time Series Anomaly Detection

Beyond the satellite domain, time-series anomaly detection (TSAD) is a broad and rapidly expanding field spanning statistics, signal processing, and machine learning. The dominant paradigm in recent years has been reconstruction- and forecasting-based deep learning, in which a model is trained to reproduce nominal behaviour and anomalies are identified through large reconstruction or prediction errors.

Representative methods include OmniAnomaly [10], a stochastic recurrent architecture designed to model distributions of multivariate temporal sequences, and USAD [11], which accelerates adversarial autoencoder training for anomaly detection. Transformer-based architectures and attention mechanisms have also been widely explored due to their ability to capture long-range dependencies in multivariate sequences.

Despite their architectural sophistication, these methods share a common assumption that the training data are predominantly nominal. Deep reconstruction models are also known to suffer from over-generalisation, whereby sufficiently expressive networks may reconstruct anomalous patterns almost as accurately as nominal ones.

The reliability of conclusions drawn from TSAD benchmarks has itself been questioned in recent years. Schmidl et al. [5] evaluate 71 anomaly detection algorithms across nearly one thousand datasets and report that no single approach consistently dominates across all settings, with simple statistical methods often matching or outperforming complex deep architectures under rigorous evaluation protocols.

More critically, Wu and Keogh [2] argue that several widely used anomaly detection benchmarks contain systematic flaws, including trivial anomalies, inaccurate labels, and run-to-failure bias, which can produce misleadingly optimistic evaluations. These concerns have motivated the development of more operationally realistic telemetry benchmarks such as ESA-ADB [3].

C. Changepoint Detection

Changepoint detection (CPD) consists in identifying the time instants at which the statistical properties of a signal change abruptly [12], [13]. Applications span finance, industrial monitoring, biosignal analysis, and fault detection. In satellite telemetry, anomalous events such as hardware failures, latch-ups, or unexpected responses to telecommands often appear as abrupt transitions in the statistical properties of one or more channels, making changepoint detection particularly suitable for this domain. Truong et al. [12] organise the CPD literature around three main elements: the search strategy, the segment cost function, and the penalty controlling the number of detected changepoints. In penalised offline CPD, the signal is segmented by minimising an objective composed of a data fidelity term, measuring how well each segment fits a statistical model, and a penalty term proportional to the number of

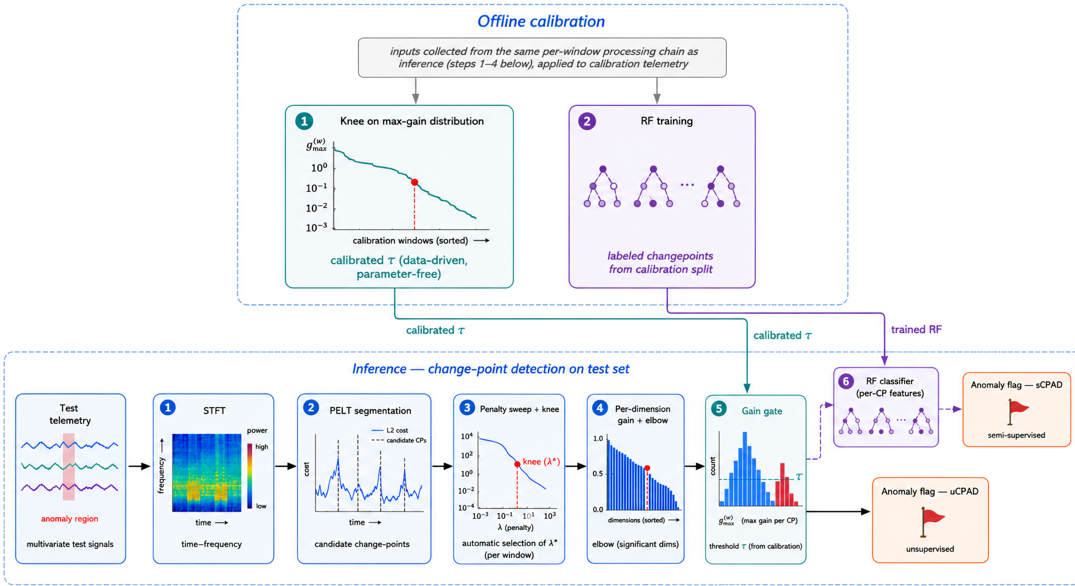


Fig. 1: Overview of the proposed change-point anomaly detection pipeline. The per-window processing chain (steps 1 to 4 in the inference panel, STFT, PELT segmentation, per-window penalty-sweep knee for λ^*, and per-dimension elbow filter) is applied to the calibration set and at the test set during inference. The offline calibration phase (top panel) persists two artefacts consisting in the gain-gate threshold τ, obtained as the knee of the distribution of per-window maximum gains $\{g_{\max}^{(w)}\}$, and, only for the semi-supervised variant *sCPAD*, a Random Forest model trained on labeled changepoints from the calibration split. At inference, the unsupervised variant *uCPAD* flags windows whose maximum gain exceeds τ (solid orange flag); the semi-supervised variant *sCPAD* additionally re-ranks surviving candidates with the trained Random Forest (purple dashed branch).

changepoints. The penalty discourages over-segmentation by balancing segmentation accuracy against model complexity.

Among exact search methods, the Pruned Exact Linear Time (PELT) algorithm proposed by Killick et al. [14] formulates changepoint detection as a penalised optimisation problem and minimises the segmentation objective while pruning suboptimal candidates during dynamic programming optimisation, yielding linear expected computational complexity under mild assumptions. Formally, for a signal $X = (X_1, \ldots, X_T)$ segmented by changepoints $0 = t_0 < t_1 < \cdots < t_K < t_{K+1} = T$, PELT minimises the penalised segmentation objective

$$\min_{K, t_1, \ldots, t_K} \left[\sum_{k=0}^{K} \mathcal{C}\big(X_{t_k:t_{k+1}}\big) + \lambda K \right], \tag{1}$$

where $\mathcal{C}(\cdot)$ denotes the segment cost and $\lambda > 0$ penalises the number of changepoints where larger values of λ yield fewer changepoints. The segment cost function determines the type of changepoints that can be detected. Classical least-squares costs detect shifts in mean, while variance-based costs target changes in volatility. More generally, kernel-based costs can detect arbitrary changes in the underlying distribution even when low-order moments remain unchanged [15]. Equivalently, rather than designing a cost tailored to a specific type of change, one may transform the signal so that a classical cost becomes sufficient; frequency-domain representations are a common instance of this strategy. This approach was explored by Oudre et al. [16], who applied changepoint detection to Short-Time Fourier Transform (STFT) representations for the segmentation of accelerometer signals during locomotion analysis, performing segmentation on time-frequency representations rather than on the raw temporal signal to capture structural changes in spectral content. A major practical challenge in penalised CPD is the selection of the penalty parameter λ, which controls the trade-off between segmentation fidelity and segmentation complexity. Small penalties tend to produce over-segmentation with many changepoints, whereas excessively large penalties may suppress meaningful structural changes. Classical penalty-selection strategies include Bayesian Information Criterion (BIC)-type penalties, slope heuristics, and cross-validation procedures. Many of these approaches rely either on strong distributional assumptions or on the availability of annotated signals. Jung et al. [17] propose a supervised penalty-learning strategy in which the penalty parameter is optimised to reproduce expert-annotated changepoints in physiological signals collected under semi-controlled conditions. The approach demonstrates that penalty selection has a strong influence on segmentation quality, but it relies fundamentally on labelled breakpoint annotations.

III. Overview of the Proposed Approach

The proposed approach reformulates anomaly detection in satellite telemetry as a two-stage problem combining changepoint detection and anomaly decision. Rather than assigning anomaly scores directly to individual samples, the method first identifies candidate regime changes and subsequently determines whether these changes correspond to anomalous behaviour. This decomposition is motivated by the observation that many telemetry anomalies manifest primarily through abrupt changes in the statistical or spectral structure of one or more channels. An overview of the complete pipeline is shown in Fig. 1. The telemetry signal is first transformed into the time-frequency domain through the Short-Time Fourier Transform (STFT). Changepoints are then detected on the resulting multivariate representation using the Pruned Exact Linear Time (PELT) algorithm [14]. The proposed uCPAD variant combines automatic penalty selection through a logarithmic sweep with gain-based filtering in order to suppress spurious detections and reduce false positives. The semi-supervised extension sCPAD augments the pipeline with a lightweight Random Forest classifier trained on annotated changepoints.

The main contributions of this work are threefold:

(i) a label-free penalty-selection strategy for penalised changepoint detection based on automatic knee detection over a logarithmic penalty sweep;

(ii) a per-dimension gain analysis with elbow-based filtering for conservative and interpretable anomaly decisions in the time-frequency domain;

(iii) a data-driven gain gate calibrated exclusively from unlabeled telemetry in order to suppress spurious changepoints and reduce false positives.

IV. Methodology

The proposed method applies changepoint detection to satellite telemetry in a label-free manner. Given a multivariate telemetry signal

$$x_t = \left(x_t^{(1)}, \dots, x_t^{(C)}\right) \in \mathbb{R}^C$$

where C denotes the number of telemetry channels and $x_t^{(c)}$ the value of channel c at time t, the pipeline processes the test set through five stages comprising *(i)* preprocessing and standardisation, *(ii)* time-frequency feature extraction, *(iii)* penalty selection via sweep, *(iv)* per-dimension gain analysis with elbow filtering, and *(v)* anomaly flagging. By separating regime-change detection from the anomaly decision, the pipeline operates more conservatively than point-wise outlier detectors, suppressing false positives on long telemetry streams.

A. Preprocessing and Standardisation

The raw telemetry channels are loaded at the sampling rate of $0.033\,\mathrm{Hz}$ (one sample every $30\,\mathrm{s}$). Each channel c is standardised to zero mean and unit variance:

$$\tilde{x}_t^{(c)} = \frac{x_t^{(c)} - \mu^{(c)}}{\sigma^{(c)}}, \tag{2}$$

where $\mu^{(c)}$ and $\sigma^{(c)}$ are computed over *all* samples of the calibration set without any label filtering. No anomaly annotations are used during preprocessing.

The standardised telemetry stream is partitioned into non-overlapping fixed windows of $W = 5$ days, yielding 512 windows on each split. Each window is processed independently. Windows containing only missing values are discarded. This decomposition allows the algorithm to operate in a streaming fashion while limiting the computational cost of the PELT optimisation to manageable temporal segments.

B. Time-Frequency Feature Extraction

Within each window, the multivariate signal $\{\tilde{x}_t^{(c)}\}_{c=1}^{C}$ is transformed into the time-frequency domain via the STFT whose coefficients are computed independently for each telemetry channel and the resulting per-frequency power time series are concatenated into a feature matrix $\mathbf{X} \in \mathbb{R}^{T \times D}$, where T denotes the number of STFT frames and $D = F \times C$ the total number of dimensions, with F the number of retained frequency bins and C the number of channels.

Each dimension of $\mathbf{X}$ therefore corresponds to a specific channel-frequency pair. This representation, originally proposed in [16], captures structural changes in the spectral content of telemetry signals that may not be visible directly in the raw time domain. The STFT is computed with `nperseg` $= 128$ samples and $75\,\%$ inter-segment overlap using a Hann window. At the telemetry sampling rate of $0.033\,\mathrm{Hz}$ this yields $F = 65$ frequency bins spanning the full available band up to the Nyquist frequency $f_s/2 \approx 0.0167\,\mathrm{Hz}$, all of which are retained.

C. Unsupervised Penalty Selection via Sweep

The core of the method is the application of PELT [14] to the multivariate STFT feature matrix $X = (X_1, \dots, X_T)$, $X_t \in \mathbb{R}^D$, minimising the penalised segmentation objective of Eq. (1) with $\mathcal{C}(\cdot)$ the squared-error (ℓ_2) segment cost and $\lambda > 0$ controlling the segmentation granularity. The penalty λ is selected automatically through a logarithmic penalty sweep over sixty log-spaced values, ranging from large penalties (few changepoints) to small ones (increasingly fine segmentations). The exact bounds are scaled by the per-window signal variance and dimensionality, so they adapt automatically to each window without manual tuning. For each λ, PELT is applied to $\mathbf{X}$ and the number of detected changepoints $K(\lambda)$ is recorded. The optimal penalty λ^* is then selected as the point of maximum curvature, the *knee*, of the curve $K(\lambda)$, identified through a discrete second-order difference criterion over the logarithmic sweep. PELT is configured with a squared-error (ℓ_2) segment cost, a minimum segment length of two STFT frames, and a candidate-step parameter of two. If the selected knee corresponds to $K = 0$, the window is considered structurally homogeneous and no changepoints are reported.

D. Per-Dimension Gain Analysis and Elbow Selection

Once the global changepoints $\{t_k^*\}$ are obtained by applying PELT to the multivariate feature matrix $\mathbf{X}$ using λ^*, the contribution of each individual dimension is analysed separately.

For each dimension d, the *gain* is defined as the reduction in ℓ_2 cost induced by the global segmentation:

$$g_d = \mathcal{C}(\mathbf{X}^{(d)}) - \sum_k \mathcal{C}\left(\mathbf{X}^{(d)}_{t_{k-1}^*:t_k^*}\right), \tag{3}$$

where $\mathbf{X}^{(d)}$ denotes the d-th column of $\mathbf{X}$.

Importantly, the changepoints are computed once globally on the full multivariate representation $\mathbf{X}$. The gain analysis then evaluates, independently for each univariate dimension, how well these global changepoints explain the local signal structure. A large gain therefore indicates that the global segmentation meaningfully explains the behaviour of that specific channel-frequency component, whereas near-zero gains suggest that the detected changepoints are locally spurious.

Rather than applying a manually tuned threshold on g_d, the gains are sorted in decreasing order and an elbow criterion is applied automatically to identify the subset of dimensions that contribute most strongly to the detected regime changes. The elbow is determined geometrically after normalising the sorted gains to the interval $[0, 1]$ on both axes. The selected point corresponds to the maximum distance from the straight line connecting the largest and smallest gains. Only dimensions with positive gain located before the elbow are retained. The corresponding STFT frequency bins are mapped back to the set of selected frequencies used for the final anomaly decision.

This conservative filtering stage is designed to suppress weak or isolated changepoints and prioritise precision over recall, consistent with the operational requirements of satellite telemetry monitoring [3].

E. Gain Gate

As an additional safeguard against spurious detections, a gain gate is calibrated directly from the telemetry. The complete pipeline (STFT extraction, penalty sweep, PELT segmentation, and gain computation) is applied to all windows indiscriminately, without using anomaly annotations at any stage. For each window in which PELT detects at least one changepoint, the maximum per-dimension gain $g_{\max} = \max_d g_d$ is recorded. The gain gate threshold τ is selected automatically as the knee of the sorted gain distribution, applying the same geometric elbow criterion used for the per-dimension selection of Section IV-D. This parameter-free criterion adapts to the actual shape of the calibration distribution without requiring a contamination assumption or a manually chosen percentile.

$$\tau = \mathrm{Knee}\Big(\{g_{\max}^{(w)}\}_{w \in \mathcal{W}_{\text{calibration}}}\Big) \,. \tag{4}$$

At test time, windows whose maximum gain does not exceed τ are suppressed even if PELT reports changepoints. This gate encodes the intuition that changepoints producing gains comparable to the bulk of nominal fluctuations are unlikely to correspond to meaningful anomalies.

F. Anomaly Flagging

The final anomaly decision combines the three previous filtering stages: global changepoint detection through the penalty sweep, per-dimension validation through gain analysis, and conservative suppression through the gain gate.

A test window is flagged as anomalous if and only if:

1) The penalty sweep yields $K^* \geq 1$ changepoints
2) At least one dimension survives the per-dimension gain elbow selection of Section IV-D
3) The gain gate threshold of Section IV-E is exceeded

All samples belonging to a flagged window are marked as anomalous in the output binary mask. Consecutive flagged windows are merged into a single predicted event (*run*), which is then compared against the ground-truth anomaly intervals using the corrected event-wise F_β-score of ESA-ADB with $\beta = 0.5$, as described in Section V.

G. Per-CP Classifier (Semi-Supervised Variant)

The unsupervised pipeline can be refined by appending a Random Forest re-ranker that filters, on a per-changepoint basis, the candidates surviving the gain gate. This is the only component using ESA-ADB annotations, turning the overall method into a *semi-supervised* variant. For each candidate t_k^* we extract a 17-dimensional descriptor, reusing quantities already computed by the unsupervised pipeline, covering local gain statistics, contributing channel–frequency dimensions, adjacent-segment geometry, and window-level features (λ^*, knee index, total changepoints). The Random Forest comprises 300 trees, with `class_weight=balanced`, and is fitted on the calibration split. The test split is never seen at any stage. At inference, a window is flagged if at least one of its candidates exceeds the calibrated threshold. The results reported in Tables I and II use a fixed random seed of 42.

V. RESULTS

We evaluate the proposed approach on the public ESA-ADB benchmark [3], using the *lightweight* protocol that restricts experiments to channels 41–46 of Mission 1 over an 84-month label-free calibration period and an 84-month evaluation period containing 29 `Anomaly` events and 36 Rare Events; this is the official subset on which ESA-ADB baselines are reported (*e.g.* Suppl. Tab. 9), guaranteeing direct comparability without re-implementation bias. We adopt the corrected event-wise F_β-score with $\beta = 0.5$ as the primary metric, weighing precision higher than recall to reflect the operational priority of minimising false alarms.

Precision and recall are defined as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \qquad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

where TP, FP, and FN denote true positives, false positives, and false negatives, respectively. The F_β-score is defined as

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}}. \tag{5}$$

Following the benchmark, we report two evaluations. The *anomaly-only* evaluation (Table I, mirroring Supplementary Table 9 of [3]) treats only the 29 annotated anomalies as positive; detections falling exclusively on rare nominal events are excluded from FP_e. The *all-events* evaluation (Table II, mirroring Table 2 of [3]) merges the 29 anomalies and 36 rare nominal events into a single positive ground-truth class of 65 events.

TABLE I: Anomaly-only evaluation on the Mission1 lightweight channel subset. Methods are grouped according to their level of supervision and calibration specificity. Boldface marks the best value in each column within the generic methods category.

Method	P	R	$F_{0.5}$
Generic unsupervised methods			
uCPAD	**0.233**	0.241	**0.234**
GlobalSTD5	0.205	0.241	0.211
PCC	<0.001	0.310	<0.001
HBOS	<0.001	0.379	<0.001
iForest	<0.001	0.414	<0.001
Window iForest	<0.001	**0.552**	<0.001
KNN	<0.001	0.448	<0.001
GlobalSTD3	<0.001	0.310	<0.001
Generic semi-supervised methods			
sCPAD	**0.814**	0.207	**0.513**
Telemanom-ESA	0.074	**0.931**	0.090
DC-VAE-ESA STD5	0.021	0.241	0.026
DC-VAE-ESA STD3	0.001	0.310	0.001
Post-processed forecasting baseline			
Telemanom-ESA-Pruned	0.999	0.862	0.968

TABLE II: All-events evaluation on the Mission1 lightweight channel subset, where anomalies and rare nominal events are merged into a single positive class. Methods are grouped according to their level of supervision and calibration specificity. Boldface marks the best value in each column within the generic methods category.

Method	P	R	$F_{0.5}$
Generic unsupervised methods			
uCPAD	**0.515**	0.431	**0.495**
GlobalSTD5	0.288	0.169	0.253
GlobalSTD3	0.001	0.431	0.001
KNN	<0.001	**0.754**	<0.001
Window iForest	<0.001	0.738	<0.001
iForest	<0.001	0.585	<0.001
HBOS	<0.001	0.585	<0.001
PCC	<0.001	0.554	<0.001
Generic semi-supervised methods			
sCPAD	**0.791**	0.431	**0.678**
Telemanom-ESA	0.148	**0.894**	0.178
DC-VAE-ESA STD5	0.063	0.338	0.075
DC-VAE-ESA STD3	0.002	0.554	0.003
Post-processed forecasting baseline			
Telemanom-ESA-Pruned	0.999	0.424	0.786

A. Anomaly-only Evaluation

In the anomaly-only evaluation (Table I), the proposed uCPAD approach achieves the highest $F_{0.5}$ score among all generic unsupervised methods included in ESA-ADB, while also attaining higher precision than classical unsupervised baselines such as PCC, HBOS, iForest, and KNN. These methods exhibit moderate recall but near-zero precision, indicating extensive false-positive activity over long telemetry streams. In contrast, uCPAD operates more conservatively by detecting changepoint-bounded regime transitions, resulting in a more balanced precision–recall trade-off. The semi-supervised extension sCPAD further improves precision by filtering false-positive changepoints. This reduces the number of false alarms and increases the $F_{0.5}$ score from 0.234 to 0.513, corresponding to an improvement of approximately 119% over the unsupervised variant. Among the generic semi-supervised methods, sCPAD achieves the best overall precision and $F_{0.5}$ score.

Telemanom-ESA attains the highest recall among the generic semi-supervised approaches, but at the expense of a lower precision due to a large number of false alarms.

B. All-events Evaluation

In the all-events evaluation (Table II), where rare nominal events are merged with anomalies into a single positive class, uCPAD again achieves the highest $F_{0.5}$ score among all generic unsupervised methods included in ESA-ADB. The inclusion of rare events particularly benefits changepoint-based predictions, since several detections previously excluded under the anomaly-only protocol now become valid true positives. As a result, uCPAD attains a more balanced precision–recall trade-off than the remaining classical unsupervised baselines.

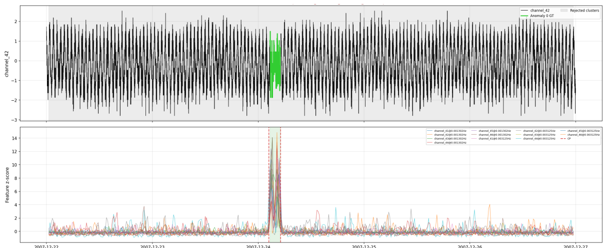
As in the anomaly-only setting, the semi-supervised extension sCPAD provides a further substantial improvement, with $F_{0.5}$ increasing from 0.495 to 0.678. This gain is driven both by a sharper rejection of spurious detections and by a wider coverage of true events. Since the classifier is trained using both Anomaly and Rare Event labels as positive examples, it correctly preserves changepoints associated with rare operational regimes. Among the generic semi-supervised methods, sCPAD achieves the highest precision and $F_{0.5}$ score.

Telemanom-ESA-Pruned achieves the highest overall performance in both benchmarks; however, the ESA-ADB paper attributes its advantage to additional dynamic thresholding and pruning procedures applied during post-processing [3]. Notably, the pruning threshold was not derived automatically but set manually based on visual inspection of smoothed training errors, a choice the benchmark authors themselves describe as highly subjective and probably suboptimal [3]. By contrast, uCPAD and sCPAD calibrate their operating point from the empirical gain distribution observed in unlabeled calibration data, without any manual threshold engineering.

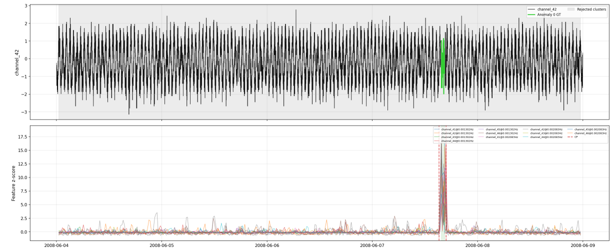
C. Ablation and Robustness

We first assess the contribution of the gain gate by disabling it ($\tau = 0$) and then varying its threshold. Without the gate, both variants flag all 512 test windows, causing the all-events $F_{0.5}$ to fall from 0.495 to 0.000 for uCPAD and from 0.678 to 0.250 for sCPAD. The gain gate is therefore responsible for suppressing the widespread false-positive behaviour produced by the unfiltered changepoints.

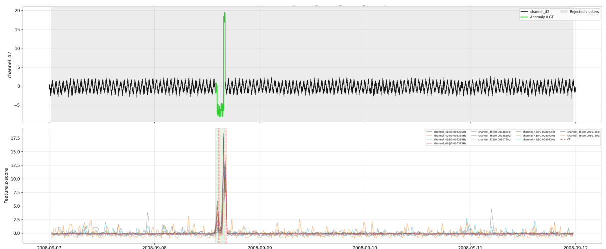
The automatically selected knee lies at the 85.4th percentile. For sCPAD, it matches the best result in the evaluated grid, while for uCPAD it remains within 0.031 of the best test-set operating point. Performance varies smoothly around the selected threshold, indicating that the result does not depend on a narrowly tuned value. Finally, across 20 runs varying both



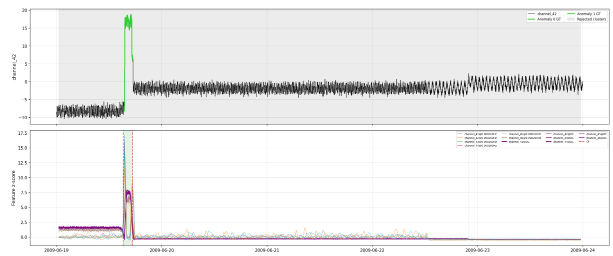
(a) Anomaly Event 1



(b) Anomaly Event 2



(c) Anomaly Event 3



(d) Anomaly Event 4

Fig. 2: Qualitative examples of anomaly detection in the time-frequency domain across four Mission1 anomaly events. For each case, the detected changepoint aligns with a ground-truth anomaly interval characterised by an abrupt change in spectral energy.

the stratified train and validation split and the Random Forest seed, sCPAD obtained an event-wise $F_{0.5}$ of 0.534 ± 0.025 in the anomaly-only setting and 0.712 ± 0.028 in the all-events setting. The single-seed results reported in Tables I and II fall within, or below, the corresponding empirical distributions.

D. Recall Ceiling and Limitations

The relatively low recall of both proposed variants reflects a property of the upstream changepoint stage rather than a limitation of the classifier itself. The proposed approach produces a compact set of changepoint-bounded intervals that capture anomalies, but remains less sensitive to subtle, gradual, or very short deviations whose contribution to the segmentation cost remains limited. Because the classifier only filters changepoints already produced by the upstream detector, it cannot recover events missed during the segmentation stage. This conservative behaviour nevertheless reduces the false-alarm rate compared with most unsupervised ESA-ADB baselines and contributes to the more balanced precision–recall trade-off achieved by uCPAD and sCPAD. Improving recall while preserving the calibration-light nature of the proposed approach remains an important direction for future work. A qualitative example of a successful detection is shown in Fig. 2, where the changepoint detector identifies an abrupt spectral regime transition corresponding to a ground-truth anomaly event.

TABLE III: Gain-gate ablation and threshold sensitivity under the all-events evaluation. The bold row is selected without labels.

Percentile	τ	Windows	uCPAD	sCPAD
Gate off	0.00	512	0.000	0.250
50.0	139.22	256	0.188	0.485
60.0	148.81	205	0.237	0.503
69.9	171.74	154	0.294	0.513
75.0	188.07	128	0.336	0.555
79.9	204.95	103	0.399	0.604
85.0	223.16	77	0.478	0.678
85.4 (knee)	**225.14**	**75**	**0.495**	**0.678**
89.8	262.04	52	0.526	0.650
94.9	365.29	26	0.497	0.497

VI. Conclusions

We presented a label-free changepoint detection approach for anomaly detection in satellite telemetry, based on PELT and evaluated on the ESA Anomaly Detection Benchmark. The pipeline requires neither labelled anomalies nor mission-specific threshold tuning, detecting anomalies as changepoint-bounded regime transitions in the time-frequency domain. On the Mission 1 lightweight subset, the unsupervised variant uCPAD achieves the highest $F_{0.5}$ among all generic unsupervised baselines in ESA-ADB, with a more balanced precision–recall trade-off and conservative false-alarm behaviour; the semi-supervised sCPAD raises $F_{0.5}$ to 0.513 (anomaly-only) and 0.678 (all-events). The main limitation is the recall, in fact, gradual or weak deviations are often missed during segmentation and cannot be recovered downstream. Future work will target more sensitive multivariate cost functions and adaptive penalty calibration.

Acknowledgements

The work of Adriano Puglisi is supported by Thales Alenia Space and Regione Lazio through the fellowship "AI-based Digital Industrial Processes for Space Systems Production Workflow" (CUP B83C23001060009; SIGEM 22066DP000000035).

REFERENCES

[1] 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*, pp. 387–395, 2018.

[2] R. Wu and E. J. Keogh, "Current time series anomaly detection benchmarks are flawed and are creating the illusion of progress," *IEEE transactions on knowledge and data engineering*, vol. 35, no. 3, pp. 2421–2429, 2021.

[3] K. Kotowski, C. Haskamp, J. Andrzejewski, B. Ruszczak, J. Nalepa, D. Lakey, P. Collins, A. Kolmas, M. Bartesaghi, J. Martinez-Heras, and G. D. Canio, "European space agency benchmark for anomaly detection in satellite telemetry," 2025.

[4] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *2008 eighth ieee international conference on data mining*, pp. 413–422, IEEE, 2008.

[5] S. Schmidl, P. Wenig, and T. Papenbrock, "Anomaly detection in time series: a comprehensive evaluation," 2022.

[6] P. Benecki, S. Piechaczek, D. Kostrzewa, and J. Nalepa, "Detecting anomalies in spacecraft telemetry using evolutionary thresholding and lstms," in *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, pp. 143–144, 2021.

[7] M. A. Obied, F. F. Ghaleb, A. E. Hassanien, A. M. Abdelfattah, and W. Zakaria, "Deep clustering-based anomaly detection and health monitoring for satellite telemetry," *Big Data and Cognitive Computing*, vol. 7, no. 1, p. 39, 2023.

[8] Y. Hou, H. Li, Y. Wang, L. Wang, and Z. Xu, "Satellite anomaly detection based on improved transformer method," in *2023 24th Asia-Pacific Network Operations and Management Symposium (APNOMS)*, pp. 322–325, IEEE, 2023.

[9] B. Ruszczak, K. Kotowski, D. Evans, and J. Nalepa, "The ops-sat benchmark for detecting anomalies in satellite telemetry," *Scientific Data*, vol. 12, no. 1, p. 710, 2025.

[10] Y. Su, Y. Zhao, C. Niu, R. Liu, W. Sun, and D. Pei, "Robust anomaly detection for multivariate time series through stochastic recurrent neural network," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2828–2837, 2019.

[11] J. Audibert, P. Michiardi, F. Guyard, S. Marti, and M. A. Zuluaga, "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.

[12] C. Truong, L. Oudre, and N. Vayatis, "Selective review of offline change point detection methods," *Signal processing*, vol. 167, p. 107299, 2020.

[13] S. Aminikhanghahi and D. J. Cook, "A survey of methods for time series change point detection," *Knowledge and information systems*, vol. 51, no. 2, pp. 339–367, 2017.

[14] R. Killick, P. Fearnhead, and I. A. Eckley, "Optimal detection of changepoints with a linear computational cost," *Journal of the American Statistical Association*, vol. 107, no. 500, pp. 1590–1598, 2012.

[15] S. Arlot, A. Celisse, and Z. Harchaoui, "A kernel multiple change-point algorithm via model selection," *Journal of machine learning research*, vol. 20, no. 162, pp. 1–56, 2019.

[16] L. Oudre, A. Lung-Yut-Fong, and P. Bianchi, "Segmentation of accelerometer signals recorded during continuous treadmill walking," in *2011 19th European Signal Processing Conference*, pp. 1564–1568, IEEE, 2011.

[17] S. Jung, L. Oudre, C. Truong, E. Dorveaux, L. Gorintin, N. Vayatis, and D. Ricard, "Adaptive change-point detection for studying human locomotion," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 2020–2024, IEEE, 2021.