ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

**DEPARTMENT OF**
**ELECTRICAL, ELECTRONIC, AND INFORMATION ENGINEERING**
**"GUGLIELMO MARCONI" - DEI**

**SECOND CYCLE DEGREE**

**ELECTRONICS FOR INTELLIGENT SYSTEMS, BIG-DATA AND INTERNET OF THINGS**
**INGEGNERIA ELETTRONICA**

# HETEROGENEOUS DATASET INSPECTION AND SEMI-SUPERVISED METHODS FOR ANOMALY DETECTION

**Supervisor**

**Prof. Mauro Mangia**

**Co-Supervisor**

**Lorenzo Capelli**

**Defended by**

**Mahsa Jahangiri Mollahajlouie**

---

**Graduation Session / March / 2026**

**Academic Year 2025/2026**

# Abstract

With the increasing complexity of modern space missions, continuous monitoring of spacecraft health has become a critical challenge. Satellite telemetry data, typically collected as high-dimensional multichannel time series with heterogeneous sampling and complex temporal dependencies, represents the primary source for monitoring spacecraft subsystems. However, traditional rule-based or threshold-based monitoring approaches often fail to detect subtle anomalies.

This thesis investigates semi-supervised anomaly detection in spacecraft telemetry using the European Space Agency Anomaly Detection Benchmark (ESA-ADB) dataset. As a first step, a structural analysis of telemetry data is conducted to better understand signal characteristics and subsystem relationships. This analysis includes examination of temporal signal behaviour, cross-channel correlations, power spectral density (PSD), and sampling interval patterns across mission subsystems. These analyses provide insights into telemetry heterogeneity and support informed channel selection for anomaly detection experiments.

Building on this analysis, two advanced deep learning algorithms reported as effective for telemetry anomaly detection in the ESA context are implemented and evaluated: Telemanom, a forecasting-based approach using Long Short-Term Memory (LSTM) networks, and DC-VAE (Dilated Convolutional Variational Autoencoder), a reconstruction-based generative model designed to learn latent representations of normal telemetry behaviour. Experiments are performed within the TimeEval benchmarking framework in a Docker-based environment to ensure reproducibility and consistent evaluation.

The experimental study demonstrates how telemetry structure and subsystem dynamics influence anomaly detection behaviour. The results highlight the importance of combining structural data analysis with machine-learning-based detection models to interpret telemetry behaviour and support reliable spacecraft health monitoring.



**Keywords**: Satellite Telemetry, Semi-Supervised Anomaly Detection, Deep Learning, Telemanom, DC-VAE, ESA-ADB, TimeEval

# Acknowledgements

I would like to express my deepest gratitude to my supervisor, **Professor Mauro Mangia**, for his guidance and continuous support throughout the development of this thesis. His insights and encouragement were invaluable to me during this journey.

I am also deeply grateful to my co-supervisor, **Lorenzo Capelli**, for his constant availability, helpful suggestions, and technical support during the various stages of this work.

I would like to extend my thanks to **Francesco Brasini** for his support with the remote computing infrastructure and system management, which greatly facilitated the execution of the experimental phase of this research.

On a personal note, I am deeply grateful to my **family**. Although they have been physically far from me during this journey, their unwavering love, encouragement, and belief in me have always been my greatest source of strength and motivation.

I would also like to thank my friends **Amirhossein**, **Atena**, and **Mahnaz** for their continuous support and for sharing the memorable moments of this journey. A very special thanks goes to **Saul** for his constant presence and encouragement throughout this work.

# Table of Contents

# List of Figures

# List of Tables

# List of Acronyms

| | |
|---|---|
| ADB | Anomaly Detection Benchmark |
| ADTQC | Anomaly Detection Timing Quality Curve |
| AFF | Anomaly-Focused F-score |
| BPTT | Backpropagation Through Time |
| DC-VAE | Dilated Convolutional Variational Autoencoder |
| ELBO | Evidence Lower Bound |
| ESA | European Space Agency |
| ESA-ADB | European Space Agency Anomaly Detection Benchmark |
| EW | Event-Wise |
| FN | False Negative |
| FP | False Positive |
| GPU | Graphics Processing Unit |
| GT | Ground Truth |
| KL | Kullback–Leibler Divergence |
| LSTM | Long Short-Term Memory |
| MAD | Median Absolute Deviation |
| PSD | Power Spectral Density |
| TP | True Positive |
| TSAD | Time Series Anomaly Detection |
| VAE | Variational Autoencoder |

# Chapter 1: Introduction

## 1.1 Context and Motivation

Modern spacecraft generate large volumes of multichannel telemetry data that must be continuously monitored to ensure operational safety and reliability [1]. Satellite telemetry signals originate from heterogeneous subsystems, exhibit different sampling frequencies, contain noise and long-term drifts, and often reflect complex interdependencies between physical components [1]. Detecting abnormal behaviour in such data is a critical task for mission control centres.

Traditional rule-based monitoring approaches rely on predefined thresholds and manual inspection. However, the increasing dimensionality and duration of modern missions make purely manual analysis impractical. This motivates the use of data-driven anomaly detection methods capable of modelling normal behaviour and identifying deviations automatically [1].

The European Space Agency Benchmark for Anomaly Detection in Satellite Telemetry (ESA-ADB) was introduced to provide a realistic and reproducible evaluation setting for anomaly detection methods on real satellite telemetry data. The benchmark includes multiple missions, hierarchical subsystem organisation, target and non-target channels, and event-based evaluation metrics aligned with operational needs [1]. This benchmark forms the experimental foundation of this thesis.

## 1.2 Problem Statement

Despite the large number of anomaly detection approaches proposed for multivariate time series, their behaviour on real spacecraft telemetry remains insufficiently understood. Satellite telemetry data exhibit structural and temporal properties that differ significantly from many commonly used benchmark datasets.

In ESA Mission 1, telemetry signals are organised into hierarchical subsystems and vary in sampling frequency, temporal scale, and noise characteristics [1]. Moreover, anomalies are sparse and often event-based rather than point-wise. These characteristics introduce several challenges:

- High *dimensionality* and subsystem structure
- Variable sampling rates and temporal spans

- Sparse anomaly labels
- Strong inter-channel dependencies
- Concept drift across mission phases

Such properties influence anomaly detection models differently depending on their modelling paradigm. Forecasting-based methods rely on temporal predictability and short-term sequence consistency, while reconstruction-based methods depend on learning stable representations of normal behaviour across channels. In subsystem-structured telemetry, heterogeneous dynamics and varying temporal scales can affect prediction stability and reconstruction fidelity in distinct ways.

In addition, evaluation protocols for anomaly detection in operational settings require more than point-wise detection accuracy. Event-based metrics, channel-aware scoring, and mission-level aggregation significantly influence how performance is interpreted. Therefore, a controlled and reproducible evaluation framework is necessary to systematically assess forecasting and reconstruction approaches under realistic telemetry constraints.

## 1.3 Research Objectives

The primary objective of this thesis is to investigate and compare forecasting-based and reconstruction-based anomaly detection approaches on ESA telemetry data under a reproducible evaluation setting.

More specifically, this work aims to:

- Perform a structured exploratory analysis of ESA telemetry signals to characterise their statistical, spectral, and structural properties.
- Establish a forecasting-based baseline using Telemanom, adapted to the ESA-ADB framework.
- Evaluate a reconstruction-based approach, namely DC-VAE (Dilated Convolutional Variational Autoencoder), on selected subsystems and full-channel configurations.
- Analyse robustness across subsystems and assess scalability when increasing the number of channels.
- Investigate the feasibility of integrating memory mechanisms into DC-VAE and evaluate their impact relative to the baseline reconstruction model.
- Compare all approaches under consistent event-based and channel-aware metrics within the TimeEval framework.

These objectives define a progression from baseline modelling to extended architectural exploration.

## 1.4 Thesis Contributions

The main contributions of this thesis are as follows:

- A structured exploratory analysis of ESA Mission 1 telemetry data, including statistical, correlation-based, and spectral characterisation of subsystem behaviour and anomaly patterns.
- The development of a reproducible experimental pipeline integrating ESA-ADB with the TimeEval framework, including containerised execution and remote server deployment.
- The establishment of a forecasting-based baseline through the systematic evaluation of Telemanom under ESA-specific configurations.
- A subsystem-level and full-channel analysis of DC-VAE (Dilated Convolutional Variational Autoencoder), including robustness and scalability assessment across varying telemetry configurations.
- An exploratory extension of the DC-VAE architecture through the integration of memory-aware mechanisms, evaluated in comparison with the baseline reconstruction model.

Together, these contributions provide an empirical assessment of forecasting and reconstruction paradigms for anomaly detection in realistic satellite telemetry settings.

## 1.5 Thesis Structure

The remainder of this thesis is organised as follows.

Chapter 2 reviews the theoretical foundations of anomaly detection for multivariate time series and presents the evaluation principles relevant to satellite telemetry analysis. It also introduces the TimeEval benchmarking framework, which serves as the experimental backbone of this work.

Chapter 3 presents a structured exploratory data analysis of ESA Mission 1 telemetry. This chapter characterises subsystem organisation, inter-channel relationships, and statistical and spectral properties of the signals, providing insight into the structural complexity of the dataset.

Chapter 4 describes the implementation and experimental pipeline. It details the system environment, containerised execution setup, integration with ESA-ADB, and the

experimental protocol adopted to ensure reproducibility and consistency across configurations.

Chapter 5 integrates the algorithmic descriptions and the complete experimental analysis. It first establishes a forecasting-based baseline using Telemanom and then evaluates the reconstruction-based DC-VAE model across subsystem-specific and full-channel settings. The chapter further discusses robustness and scalability aspects and includes an exploratory assessment of memory-aware extensions within the same evaluation framework.

The thesis concludes with a synthesis of findings, a discussion of limitations, and directions for future research in satellite telemetry anomaly detection.

# Chapter 2 — Background Review & Evaluation Framework

## 2.1 Fundamentals of Anomaly Detection

### 2.1.1 Definition of Anomalies

An anomaly is commonly defined as an observation that deviates significantly from the expected behaviour of the underlying data distribution. In the context of time series analysis, anomalies correspond to temporal patterns that differ from normal system dynamics in magnitude, structure, or evolution over time. According to Chandola et al. (2009), an anomaly can be broadly described as "a pattern in the data that does not conform to the expected normal behavior" [2].

In time series data, anomalies are typically categorised into three main types. **Point anomalies** refer to individual observations that deviate from normal values at a specific timestamp. **Contextual anomalies** occur when a data point is anomalous only within a particular temporal or environmental context, even if it may appear normal in isolation. **Collective anomalies** describe sequences of observations that are individually normal but jointly form an anomalous pattern over a time interval [3].

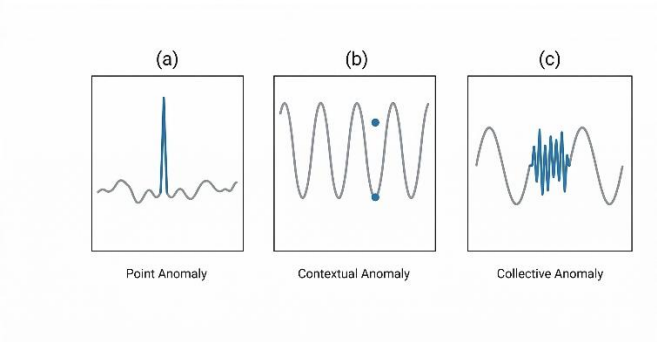


*Figure 2.1 — Illustration of common anomaly types in time series data: (a) point anomaly, (b) contextual anomaly, and (c) collective anomaly.*

In satellite telemetry, anomalies are often collective and event-based rather than isolated point deviations. Telemetry signals originate from interdependent subsystems, and abnormal behaviour may manifest as coordinated deviations across multiple channels over a temporal window. As highlighted in the ESA-ADB benchmark, anomaly detection in

spacecraft telemetry therefore requires models capable of capturing multivariate dependencies and temporal structure under realistic operational constraints [1].

### 2.1.2 Learning Paradigms in Anomaly Detection

Anomaly detection methods can be broadly categorised according to the availability of labelled data during training. The three principal learning paradigms are supervised, semi-supervised, and unsupervised anomaly detection [2].

In **supervised anomaly detection**, both normal and anomalous instances are available during training. The problem is typically formulated as a binary or multi-class classification task, where the model learns explicit decision boundaries between normal and abnormal patterns. While supervised methods can achieve strong performance when sufficient labelled anomalies are available, their applicability is limited in domains where anomalies are rare or difficult to label reliably [1], [2].

In **unsupervised anomaly detection**, no labelled data are assumed to be available. Models attempt to identify deviations from the inherent structure of the dataset, often relying on assumptions such as anomaly rarity or distributional separability. Clustering-based, density-based, and reconstruction-based approaches frequently fall into this category [1], [2].

**Semi-supervised anomaly detection** assumes access only to labelled normal data during training. The model learns a representation of normal behaviour and flags deviations at inference time. This setting is particularly relevant in safety-critical systems, where anomalous behaviour is rare, diverse, and costly to label. As discussed in the ESA-ADB benchmark, satellite telemetry datasets typically contain sparse and event-based anomaly annotations, making fully supervised training impractical in realistic operational scenarios [1].

In the context of this thesis, the considered anomaly detection models operate under a semi-supervised setting, where normal telemetry data are used for model training and anomalies are identified as deviations from learned normal patterns.

## 2.2 Modelling Paradigms for Time Series Anomaly Detection

### 2.2.1 Forecasting-Based Approaches

Forecasting-based anomaly detection methods model the temporal dynamics of a time series by predicting future observations from historical data. In such approaches, the prediction error is used to define an anomaly score, and anomalies are detected when this score exceeds a predefined threshold [4]. By formulating anomaly detection as a predictive modelling task,

these methods rely on deviations between expected and observed behaviour. For a given time step t, the anomaly score can be defined as:

$$s_t = |x_t - \hat{x}_t|$$

**(Equation 2.1)**

where x_t is the observed value and x̂_t is the predicted value.

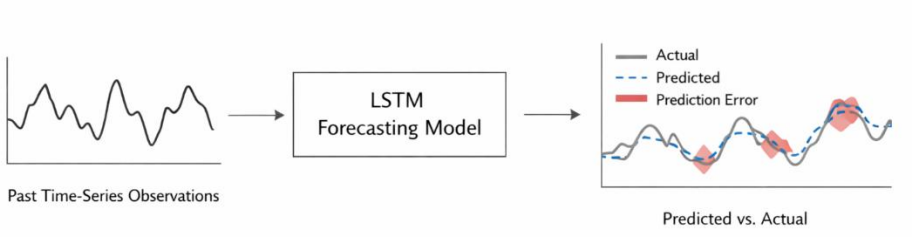


*Figure 2.2 — Forecasting-based anomaly detection pipeline. Past observations are used to train an LSTM forecasting model, and anomalies are identified based on the prediction error between observed and predicted values.*

In multivariate settings, forecasting models aim to capture interdependencies among correlated channels in order to learn the joint evolution of the system [4]. Recurrent neural networks (RNNs) are particularly suitable for sequential modelling; however, standard RNNs suffer from vanishing and exploding gradient problems when learning long-range temporal dependencies. Long Short-Term Memory (LSTM) networks, introduced by Hochreiter and Schmidhuber [5], address this limitation through gated memory mechanisms that enable stable learning over extended time intervals.

The internal structure of an LSTM cell and its gating mechanisms are illustrated in Figure 2.3.

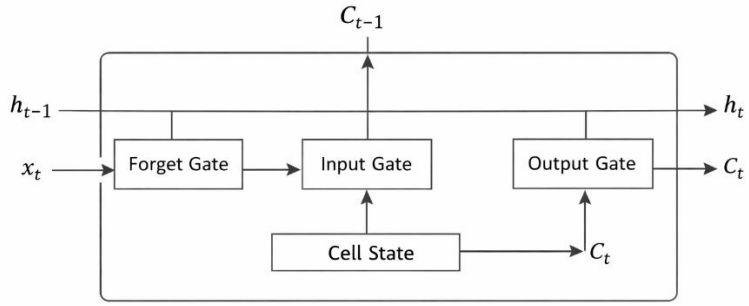


*Figure* 2.3 — *Internal architecture of an LSTM cell showing the forget, input, and output gating mechanisms.*

These properties make LSTM architectures well-suited for telemetry forecasting tasks. In spacecraft anomaly detection, the Telemanom approach employs LSTM models trained on normal operational data to predict future telemetry values. Anomalies are identified based

on deviations between predicted and observed values combined with a dynamic thresholding mechanism [3].

Forecasting-based methods are particularly effective when the system exhibits stable and predictable temporal dynamics. However, their performance may degrade under regime shifts, heterogeneous subsystem behaviour, or highly non-stationary conditions, which can affect predictive stability in complex telemetry environments.

### 2.2.2 Reconstruction-Based Approaches

Reconstruction-based anomaly detection methods learn a compact representation of normal data and attempt to reconstruct the input signal from this learned representation. An anomaly is identified when the reconstruction error exceeds a predefined threshold, indicating that the observed pattern deviates from the learned normal structure [4]. For reconstruction-based models, the anomaly score is typically defined as:

$$s_t = ||x_t - \hat{x}_t||$$

**(Equation 2.2)**

where x_t is the original input and x̃_t is the reconstructed output.

The general architecture of a Reconstruction-based model is illustrated in Figure 2.4.

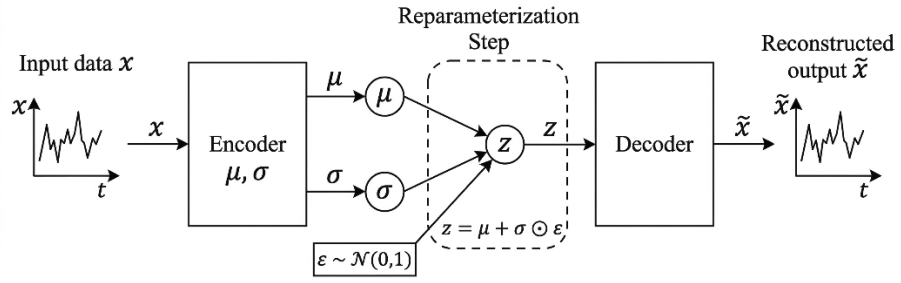


*Figure 2.4 — Architecture of a Variational Autoencoder (VAE).*

*The encoder maps the input to the latent distribution parameters μ and σ,*

*and a latent variable z is sampled using the reparameterization trick before reconstruction by the decoder.*

Unlike forecasting approaches, reconstruction-based models do not explicitly predict future values but instead evaluate how well the current observation conforms to the learned data manifold.

Autoencoders are a prominent class of reconstruction-based models. They consist of an encoder that maps the input into a lower-dimensional latent space and a decoder that

reconstructs the original input from this latent representation. By training on normal data, the model learns to efficiently represent typical system behaviour while failing to accurately reconstruct anomalous patterns [4].

Variational Autoencoders (VAEs) extend this idea by introducing a probabilistic latent variable model. Instead of learning a deterministic encoding, VAEs approximate the posterior distribution of latent variables and optimise a variational lower bound on the data likelihood [6]. This probabilistic formulation enables the modelling of uncertainty in the latent space and provides a principled reconstruction likelihood that can be used for anomaly scoring.

### 2.2.3 Conceptual Comparison of Forecasting and Reconstruction

#### *2.2.3.1 Sensitivity to Temporal Dynamics*

Forecasting-based methods are tightly coupled to *temporal predictability*: they learn to extrapolate future values from past observations, so their anomaly signal is primarily driven by *unexpected temporal deviations* (prediction residuals) and how thresholds react to them [4]. This can be advantageous when abnormal behaviour manifests as clear temporal disruptions, but it may become fragile under regime shifts or changes in operating conditions that alter the prediction landscape. In telemetry-focused forecasting pipelines such as Telemanom, this sensitivity is addressed through dynamic thresholding on prediction errors, aiming to adapt the decision boundary to changing error distributions [3].

#### *2.2.3.2 Handling of Subsystem Heterogeneity*

Reconstruction-based methods focus on learning a representation of normal patterns and measuring how well observations conform to this learned structure [4]. Reconstruction-based approaches evaluate how well an observation conforms to the learned representation of nominal data. In multivariate settings, this allows them to capture structural consistency across channels rather than relying solely on one-step temporal predictability. As a result, anomalies that disrupt the learned manifold structure may become detectable through increased reconstruction error. However, as highlighted in [7], optimizing an autoencoder purely for reconstruction accuracy can implicitly suppress features that are informative for anomaly detection. When the model is trained to minimise distortion, it may retain representations that reconstruct both nominal and certain anomalous patterns with comparable fidelity—particularly if anomalies are subtle or share statistical characteristics with normal data. This leads to an inherent trade-off: improving reconstruction quality does not necessarily improve detection effectiveness, and in some cases may even reduce anomaly separability. The relationship between reconstruction objectives and detection performance is therefore not monotonic, but governed by the balance between information preservation and discriminative sensitivity, as explicitly analysed in [7]. In spacecraft

telemetry, heterogeneity is further amplified by subsystem organisation, mixed channel behaviours, and non-target channels included for realism, which motivates careful channel grouping and evaluation design [1].

#### *2.2.3.3 Implications for Satellite Telemetry Data*

For satellite telemetry, the choice between forecasting- and reconstruction-based approaches is influenced not only by model architecture, but also by the characteristics of the data and by operational evaluation requirements. Telemetry streams are multivariate and subsystem-structured, and evaluation is typically performed at the event level, reflecting operational decision-making processes rather than purely point-wise accuracy [1]. This setting naturally favours semi- or unsupervised methods that can be assessed using event-based metrics aligned with mission operations.

Forecasting-based pipelines such as Telemanom model nominal temporal dynamics using LSTM predictors and combine prediction residuals with adaptive thresholding mechanisms to identify deviations [3]. In this formulation, anomalies are detected as departures from learned temporal patterns in the prediction space.

Reconstruction-based pipelines, by contrast, learn a compressed representation of nominal behaviour and detect anomalies through deviations in reconstruction error. In multivariate settings, this enables the modelling of cross-channel structure rather than relying exclusively on one-step temporal predictability. However, as analysed in recent work on autoencoder-based compression and anomaly detection, optimisation for reconstruction distortion does not necessarily translate into improved detection performance, and improvements in detection may come at the cost of reconstruction quality [4], [7]. The relationship between reconstruction objectives and detection effectiveness must therefore be interpreted with care.

Finally, studies on multivariate anomaly detection highlight that modelling correlated signals and the choice of aggregation strategy across channels can significantly affect detection behaviour, reinforcing the importance of consistent evaluation protocols when comparing different paradigms [8].

## 2.3 Deep Learning Architectures for Multivariate Telemetry

### 2.3.1 Recurrent Neural Networks and LSTM

#### *2.3.1.1 Sequential Modelling*

Recurrent Neural Networks (RNNs) are designed to process time-varying inputs by maintaining internal states that evolve across discrete time steps. In principle, recurrent

networks use feedback connections to store representations of recent input events in the form of activations, enabling short-term memory over sequential data [5]. However, conventional training algorithms such as Backpropagation Through Time (BPTT) suffer from vanishing and exploding gradients, making it difficult to learn long-range temporal dependencies [5].

Long Short-Term Memory (LSTM) was introduced to overcome the error backflow problems of conventional recurrent networks by enforcing constant error flow through the internal states of special units [5]. This is achieved through an architecture built around memory cells containing a self-connected linear unit, often referred to as the constant error carousel (CEC), which allows error signals to propagate without vanishing or exploding. The architecture further includes multiplicative input and output gate units that regulate access to this constant error flow. The input gate controls when information is written to the memory cell, while the output gate controls when the stored information is exposed to other units. In this way, the gates protect stored contents from perturbation and prevent irrelevant memory contents from influencing the network output [5]. By design, this structure enables the network to learn to bridge long time intervals and to model long-range temporal dependencies that are difficult to capture with standard BPTT-trained recurrent networks, where error signals tend to either decay exponentially or blow up [5].

In multivariate telemetry settings, where multiple channels evolve jointly over time, LSTM architectures provide a natural framework for capturing temporal dependencies across channels while maintaining internal state representations that evolve consistently with system dynamics.

#### *2.3.1.2 Role in Forecasting-Based Anomaly Detection*

In forecasting-based anomaly detection pipelines, LSTM models are used to predict telemetry values by learning from normal command and telemetry sequences [3]. For each time step, a predicted value is generated and the prediction error is computed as the absolute difference between the observed and predicted telemetry value [3].

To determine whether values are nominal, thresholds are applied to smoothed prediction errors, and values corresponding to smoothed errors above the threshold are classified as anomalies [3]. In Telemanom, LSTM-based predictions are combined with a non-parametric dynamic thresholding approach to automatically assess prediction errors and identify anomalous regions [3].

Under this paradigm, anomaly detection is performed in the residual space defined by prediction error rather than directly in the raw signal space. When trained on nominal data, LSTMs capture and model normal behavior of the system, and deviations from predicted behaviour indicate potential anomalies [3].

## 2.3.2 Variational Autoencoders

### *2.3.2.1 Latent Variable Modelling*

Variational Autoencoders (VAEs) are generative models that introduce latent random variables to describe observed data [6]. Instead of learning a purely deterministic mapping between inputs and outputs, VAEs assume that observed data x are generated from latent variables z through a parameterised likelihood model p(x given z).

In this framework, a prior distribution is defined over the latent variables, typically a standard normal distribution. Because the true posterior distribution p(z given x) is generally intractable, VAEs introduce an approximate posterior distribution q(z given x), parameterised by an encoder network. The model is trained by maximising a variational lower bound on the marginal likelihood of the data, commonly referred to as the Evidence Lower Bound (ELBO) [6]. This objective balances two terms: a reconstruction term and a regularisation term based on Kullback–Leibler divergence between the approximate posterior and the prior. The training objective maximises the Evidence Lower Bound (ELBO), defined as:

$$L(x) = E_{q(z|x)}[log\ p(x|z)] - KL(q(z|x)\ ||\ p(z))$$

**(Equation 2.3)**

### *2.3.2.2 Probabilistic Reconstruction*

In contrast to classical autoencoders that minimise a deterministic reconstruction error, VAEs perform probabilistic reconstruction. The decoder models the conditional likelihood p(x given z), allowing reconstructed outputs to be interpreted as samples from a learned data distribution rather than fixed point estimates [6].

The optimisation objective maximises the expected log-likelihood of the data under the approximate posterior while regularising the latent representation to remain close to the prior distribution. This probabilistic formulation enables structured latent representations and provides a principled mechanism for modelling uncertainty [6].

In anomaly detection settings, samples that have low likelihood under the learned generative model or produce high reconstruction-related discrepancy can be interpreted as deviations from normal behaviour.

## 2.3.3 Dilated Convolutional Networks

### *2.3.3.1 Receptive Field Expansion*

Dilated convolutions modify standard convolution by introducing gaps between filter elements, so that the filter effectively spans a larger input region than its actual length would

suggest. In this formulation, the convolution operates over an expanded area by skipping input values at fixed intervals, enabling a wider receptive field without increasing the filter size [9].

Unlike standard convolutions, dilated convolutions permit the receptive field to expand exponentially as network depth increases, without a proportional rise in computational cost. In particular, stacking dilated convolutional layers allows the network to achieve very large receptive fields using relatively few layers, while still preserving the input resolution and maintaining computational efficiency [9].

By doubling the dilation factor at each layer (e.g., 1, 2, 4, …), the receptive field increases exponentially, allowing the model to incorporate information from long temporal contexts. As stated in WaveNet, *"Exponentially increasing the dilation factor results in exponential receptive field growth with depth"* [9].

#### *2.3.3.2 Long-Range Dependency Modelling*

Dilated causal convolutions are introduced in WaveNet to address the challenge of modelling long-range temporal dependencies in raw audio signals. The architecture is specifically designed to provide very large receptive fields, enabling the model to incorporate information from distant time steps without relying on recurrent connections [9].

Unlike recurrent architectures, causal convolutional models eliminate explicit recurrence. As a result, they can be trained more efficiently, particularly for long sequences, while still capturing extended temporal structure through stacked dilated layers [9].

In multivariate telemetry settings, this property makes dilated convolutional networks attractive for modelling long-range temporal correlations across channels, while preserving resolution and computational tractability.

## 2.4 Selected Algorithms in the ESA-ADB Context

### 2.4.1 Overview of Telemanom

Telemanom is an LSTM-based anomaly detection framework developed for spacecraft telemetry data [3]. The approach treats anomaly detection as a forecasting problem, where a recurrent neural network is trained on nominal telemetry sequences to predict future values based on past observations. In the original formulation, the method is described as an unsupervised anomaly detection approach that leverages LSTMs to model high-volume telemetry streams by learning normal command and telemetry behaviour [3]. The overall Telemanom anomaly detection workflow is illustrated in Figure 2.5.

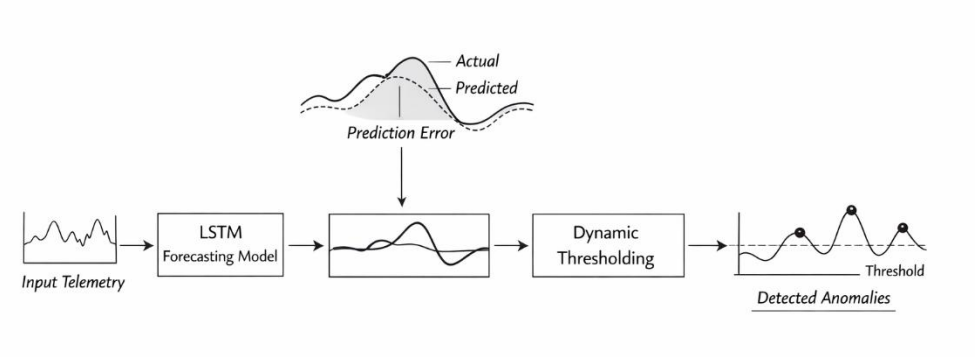


*Figure 2.5 — Simplified Telemanom anomaly detection pipeline based on LSTM forecasting and dynamic thresholding of prediction errors.*

For each time step, the model produces a predicted value and computes the prediction error as the absolute difference between the observed and predicted signal [3]. Anomalies are then identified by applying thresholds to smoothed prediction errors, and values whose smoothed errors exceed the threshold are classified as anomalous [3]. The approach further incorporates a non-parametric dynamic thresholding mechanism combined with pruning to refine anomaly regions [3].

Within the ESA-ADB benchmark framework, which provides real satellite telemetry data organised into missions and subsystems with event-based evaluation metrics [1], Telemanom serves as a forecasting-based baseline. It detects anomalies in the residual space defined by prediction errors, making it particularly suited for telemetry channels exhibiting structured temporal evolution.

### 2.4.2 Overview of DC-VAE (Dilated Convolutional Variational Autoencoder)

The Dilated Convolutional Variational Autoencoder (DC-VAE) is a reconstruction-based anomaly detection model that combines probabilistic latent variable modelling with dilated convolutional architectures. The model builds upon the Variational Autoencoder framework, in which observed data are assumed to be generated from latent variables through a parameterised conditional likelihood model [6].

In the VAE formulation, a prior distribution is defined over latent variables, and an approximate posterior is learned via an encoder network. The model is trained by maximising a variational lower bound on the marginal likelihood of the data, balancing a reconstruction term and a regularisation term based on Kullback–Leibler divergence [6].

To enhance temporal modelling capacity, DC-VAE replaces standard convolutional layers with dilated convolutions. As described in WaveNet, dilated convolutions allow the receptive field to grow exponentially with depth while maintaining computational efficiency [9].

Stacking dilated layers enables the network to capture long-range temporal dependencies without resorting to recurrent connections [9].

In reconstruction-based anomaly detection, models such as autoencoders are trained to reconstruct normal data patterns, and reconstruction error is used as an anomaly score. Samples that produce significantly higher reconstruction errors than nominal data are considered anomalous [4].

Within the ESA-ADB framework, which provides annotated real satellite telemetry from multiple missions and evaluates algorithms through a hierarchical benchmarking pipeline [1], reconstruction-based approaches can serve as alternatives to forecasting-based methods such as Telemanom. By modelling temporal structure through dilated convolutions and latent probabilistic representations, DC-VAE enables the analysis of multivariate telemetry signals under semi-supervised settings consistent with the benchmark design.

### 2.4.3 ESA-ADB Requirements and Algorithm Selection

ESA-ADB is built on top of the TimeEval framework and extends it with new algorithms and evaluation mechanisms tailored to satellite telemetry benchmarking. The benchmark design is driven by operational constraints and by a set of technical requirements defined for telemetry anomaly detection algorithms [1].

#### *2.4.3.1 ESA-ADB Requirements (R1–R9)*

In ESA-ADB, candidate algorithms are screened against **nine technical requirements (R1–R9)**. The paper clarifies that some requirements are treated as mandatory ("shall"), while others are recommended ("should"), and Table 5 summarises how preselected algorithms fulfil each requirement using scores {0, 0.5, 1} [1].

Below is a concise interpretation of the requirements **as described and exemplified in the ESA-ADB paper**:

- **R1 — Thresholding support:** algorithms should provide a *dedicated thresholding mechanism* (the paper notes that some methods only partially fulfil this when they "do not provide dedicated thresholding mechanisms") [1].
- **R2 — Online detection feasibility:** algorithms should support *online detection*, i.e., operate without using "future samples"; partial fulfilment may occur when online use is possible "but with a large computational overhead" [1].
- **R3 — (As evaluated in Table 5):** requirement assessed in Table 5 as part of the ESA-ADB screening process [1].
- **R4 — Training data anomalies:** algorithms should handle the realistic case where *anomalies exist in training data* (the paper notes partial fulfilment when methods "handle anomalies in training data but cannot learn from them") [1].

- **R5 — Affected channels reporting:** algorithms should provide a *list of affected channels* for multivariate anomalies; partial fulfilment may require "additional mechanisms or modifications … to provide a list of affected channels" [1].
- **R6 — (As evaluated in Table 5):** requirement assessed in Table 5 as part of the ESA-ADB screening process [1].
- **R7 — Rare nominal events:** algorithms should be able to *learn/memorise rare nominal events*; the paper states that "none of the preselected algorithms are able to explicitly learn rare nominal events (R7)" [1].
- **R8 — Varying sampling rates:** algorithms should handle *varying sampling rates / irregular gaps* without relying on naive resampling assumptions; the paper states that "none of the preselected algorithms … handle varying sampling rates (R8)" [1].
- **R9 — Practical runtime:** algorithms "should be possible to run in a reasonable time on a single high-end PC"; in ESA-ADB this is effectively treated as mandatory in practice [1].

Table 2.1 reports the ESA-ADB requirements analysis for a set of preselected algorithms, where each requirement is scored as not fulfilled (0), partially fulfilled (0.5), or fully fulfilled (1). This screening step motivates the choice of baseline and reference methods used throughout this thesis.

| | Algorithm | **R1** | **R2** | R3 | R4 | R5 | R6 | R7 | R8 | **R9** | Included in ESA-ADB |
|---|---|---|---|---|---|---|---|---|---|---|---|
| UNSUPERVISED | COPOD[60] | 1 | 0.5 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | NO |
| | HBOS[33] | 1 | 1 | 0 | 1 | 0.5 | 0 | 0 | 0 | 1 | YES |
| | iForest[30] | 1 | 1 | 1 | 1 | 0.5 | 0 | 0 | 0 | 1 | YES |
| | Windowed iForest[30] | 1 | 1 | 1 | 1 | 0.5 | 0 | 0 | 0 | 0.5 | SUBSETS |
| | k-Means[73] | 1 | 1 | 1 | 1 | 0.5 | 0 | 0 | 0 | 0.5 | NO |
| | KNN[34] | 1 | 1 | 1 | 1 | 0.5 | 0 | 0.5 | 0 | 0.5 | SUBSETS |
| | LOF[72] | 1 | 1 | 1 | 1 | 0.5 | 0 | 0 | 0 | 0.5 | NO |
| | Matrix Profile[36,37] | 1 | 1 | 0.5 | 0 | 0.5 | 0 | 0.5 | 0 | 1 | NO |
| | PCC[32] | 1 | 0.5 | 1 | 1 | 0.5 | 0 | 0 | 0 | 1 | YES |
| | Torsk[74] | 0.5 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0.5 | NO |
| SEMI-SUPERVISED | DAE[76] | 0.5 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0.5 | NO |
| | DC-VAE[28]* | 0.5 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0.5 | NO |
| | DC-VAE-ESA* | 1 | 1 | 1 | 0.5 | 1 | 1 | 0 | 0 | 1 | YES |
| | GlobalSTD* | 1 | 1 | 0 | 0.5 | 1 | 0 | 0 | 0 | 1 | YES |
| | Hybrid KNN[77] | 1 | 1 | 1 | 0 | 0.5 | 0 | 0.5 | 0 | 0.5 | NO |
| | LSTM-AD[78] | 0.5 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0.5 | NO |
| | OmniAnomaly[26] | 0.5 | 1 | 1 | 0 | 0.5 | 0 | 0 | 0 | 0.5 | NO |
| | RobustPCA[75] | 0.5 | 0.5 | 1 | 0.5 | 0.5 | 0 | 0 | 0 | 1 | NO |
| | Telemanom[2] | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0.5 | NO |
| | Telemanom-ESA* | 1 | 1 | 1 | 0.5 | 1 | 1 | 0 | 0 | 1 | YES |

*Table 2.1 — Analysis of Preselected Algorithms According to ESA-ADB Requirements [1].*

#### *2.4.3.2 Rationale for Selecting Telemanom-ESA and DC-VAE-ESA*

Because it is infeasible to include all existing TSAD algorithms, ESA-ADB selects a subset based on substantive arguments and the requirement screening described above. In Table 5 of the ESA-ADB paper, the adapted methods **Telemanom-ESA** and **DC-VAE-ESA** are among the algorithms marked as included in the benchmark and show strong coverage of the requirement set compared to their non-adapted counterparts [1].

In particular, the requirement-analysis table highlights that Telemanom-ESA and DC-VAE-ESA are evaluated as fully satisfying several key operational criteria (including online use and affected-channel reporting), which supports their use as representative baselines for forecasting-based and reconstruction-based detection in the ESA-ADB context [1].

#### *2.4.3.3 Alignment with ESA-ADB Requirements*

The experimental design adopted in this thesis follows the requirement-driven benchmark philosophy of ESA-ADB. By selecting algorithms that were screened and analysed under the R1–R9 criteria in [1], the study ensures consistency with the benchmark's operational assumptions, including online feasibility, semi-supervised applicability, and compatibility with hierarchical event-based evaluation. This alignment guarantees that the subsequent implementation and experimental analysis remain coherent with the benchmark framework rather than relying on externally imposed modelling assumptions.

## 2.5 Evaluation Metrics for Satellite Telemetry

### 2.5.1 Point-Wise Performance Metrics

Point-wise evaluation measures anomaly detection performance at the level of individual timestamps by comparing predicted anomaly labels against ground-truth labels for each time step. In anomaly detection literature, performance is commonly summarised using Precision, Recall, and F1-score.

Precision is defined as the proportion of predicted anomalous samples that are truly anomalous , while Recall measures the proportion of true anomalous samples that are correctly detected [4]. The F1-score combines Precision and Recall through their harmonic mean, providing a single balanced measure of detection performance [4]. Let TP denote True Positives, FP False Positives, and FN False Negatives.

$$Precision = TP / (TP + FP)$$

**(Equation 2.4)**

$$Recall = TP / (TP + FN)$$

**(Equation 2.5)**

$$F_1\_Score = 2 \times (Precision \times Recall) / (Precision + Recall)$$

**(Equation 2.6)**

Although these metrics are widely used and provide intuitive interpretability, they may not fully reflect operational relevance in satellite telemetry contexts. In ESA-ADB, anomalies are sparse and often event-based rather than point-wise, which motivates complementing point-wise evaluation with event-based and hierarchical scoring mechanisms [1].

## 2.5.2 Event-Based Evaluation

### *2.5.2.1 Event-Level Detection*

In ESA-ADB, anomaly detection performance is not evaluated solely at the level of individual timestamps. Instead, the benchmark introduces event-wise and hierarchical evaluation mechanisms to better reflect operational requirements. As described in [1], anomalies are sparse and often event-based rather than point-wise, which motivates the use of event-oriented scoring procedures.

The benchmark evaluates detections with respect to annotated anomaly events rather than isolated anomaly points. The conceptual difference between point-wise evaluation and event-based anomaly evaluation is illustrated in Figure 2.6.

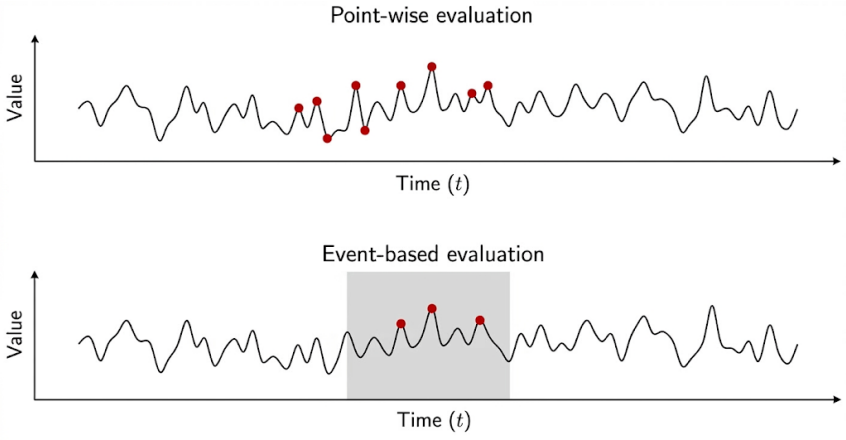


*Figure 2.6 — Illustration of point-wise and event-based anomaly evaluation.*

*Point-wise metrics evaluate detections at individual timestamps, whereas event-based evaluation considers whether a detection occurs within the temporal window of an anomaly event.*

In particular, the ESA-ADB framework assesses algorithms using a hierarchical evaluation pipeline that aggregates detection results across events and channels [1]. This approach allows performance to be interpreted at multiple levels, including event-wise and mission-level aggregation.

#### *2.5.2.2 Operational Relevance*

The ESA-ADB benchmark was designed in close cooperation with spacecraft operations engineers to ensure that evaluation metrics reflect real operational needs [1]. The paper explicitly states that results of typical anomaly detection algorithms assessed in the proposed hierarchical evaluation pipeline indicate that new approaches are necessary to address operators' needs [1].

Event-based evaluation is therefore aligned with operational practice, where anomalies correspond to meaningful system events rather than isolated samples. This design ensures that performance measures capture the practical usefulness of anomaly detection systems in mission control contexts [1].

### 2.5.3 Channel-Aware and Mission-Level Aggregation

#### *2.5.3.1 Multichannel Evaluation*

Satellite telemetry is inherently multivariate: anomalies may affect only a subset of channels, propagate across correlated signals, or appear differently depending on subsystem structure. For this reason, ESA-ADB goes beyond point-wise scoring and uses a hierarchical evaluation pipeline with channel-aware components, enabling performance to be assessed in a way that reflects multichannel detection behaviour [1]. In such settings, evaluation must consider not only whether an anomaly is detected in time, but also whether the detector correctly localises the affected channels and avoids spreading detections across unrelated signals.

More generally, multivariate anomaly detection studies emphasise that modelling correlated signals and the way detection scores are aggregated across channels can substantially influence the final detection behaviour and measured performance [8]. This motivates using channel-aware evaluation in addition to event-based scoring when interpreting results on realistic telemetry datasets.

#### *2.5.3.2 Performance Interpretation in Operational Settings*

In ESA-ADB, anomaly detection is evaluated in a way that reflects how satellite telemetry is monitored in mission control centres, where anomaly detection is a daily operational practice of spacecraft operations engineers (SOEs) [1]. For this reason, the benchmark does not limit evaluation to isolated point detections, but introduces a hierarchical evaluation pipeline that compares algorithms across aspects with different operational priorities [1].

Within this hierarchy, higher-level objectives such as correct identification of affected subsystems and channels are prioritised over purely point-wise accuracy [1]. Aggregation across channels and subsystems is therefore an explicit component of the evaluation design, allowing results to be interpreted at the level of operational decision-making rather than individual timestamps.

By incorporating channel-aware and mission-level aggregation, ESA-ADB enables performance interpretation that is aligned with real telemetry monitoring practice, where the usefulness of a detector depends not only on detection sensitivity, but also on localisation, interpretability, and manageable false alarms [1].

## 2.6 The TimeEval Benchmarking Framework

Reliable comparison of anomaly detection algorithms requires a controlled and reproducible experimental setting. ESA-ADB is built on top of the TimeEval benchmarking framework to enable systematic and reproducible evaluation of anomaly detection algorithms on real satellite telemetry data [1].

The benchmark separates dataset definition, algorithm execution, and metric computation to ensure fair and reproducible comparison across methods [1]. This structured design reduces experimental variability and standardises the evaluation process.

Furthermore, the evaluation pipeline aggregates results hierarchically across channels, subsystems, and missions [1]. Such hierarchical aggregation enables performance interpretation at multiple operational levels and supports consistent comparison of different anomaly detection paradigms under identical experimental conditions.

# Chapter 3 — Exploratory Data Analysis

## 3.1 ESA Telemetry Dataset Overview

### 3.1.1 ESA-AD and ESA-ADB Context

The experimental framework adopted in this thesis is based on the European Space Agency Anomaly Detection dataset (ESA-AD) and its associated benchmark suite, ESA-ADB [1]. ESA-AD consists of real satellite telemetry collected from three operational ESA missions and was designed to provide a realistic, large-scale environment for anomaly detection research. ESA-ADB extends this dataset by defining standardised evaluation procedures, train–test splits, and performance metrics, thereby enabling reproducible and comparable experimentation across different modelling approaches [1].

The dataset is characterised by its scale and operational authenticity. Across the missions, ESA-AD contains more than 1.5 billion telemetry samples organised into hundreds of channels representing sensor measurements, system states, and operational parameters [1]. Mission 1, which constitutes the focus of the present work, includes 76 telemetry channels grouped into multiple channel groups and four subsystems reflecting functional divisions of the spacecraft [1]. The data span several years of operation and include both nominal and anomalous mission phases.

Each telemetry channel is provided as a time series indexed by timestamps and stored as a structured data object. In addition to raw telemetry signals, the dataset includes anomaly annotations specifying the start and end times of labelled anomalous intervals. These annotations were generated through a combination of domain-expert validation and iterative refinement, ensuring reliable event-level ground truth suitable for benchmarking anomaly detection methods [1].

A key characteristic of ESA-AD is the low anomaly density. Anomalous segments represent only a small fraction of the overall timeline, typically below two percent of total samples [1]. This strong class imbalance reflects realistic operational conditions in which abnormal behaviour is rare relative to nominal system functioning. Furthermore, anomalies are defined as time intervals rather than isolated points, and multiple short segments may correspond to a single operational event [1]. This event-based annotation structure has direct implications for evaluation methodology.

For privacy and confidentiality reasons, ESA-AD has undergone anonymisation procedures. Channel names, subsystem identifiers, anomaly classes, and physical units have been renamed. The time axis has been shifted and scaled, and signal values have been normalised

within channel groups while preserving structural relationships between signals [1]. These transformations ensure that mission-sensitive information remains protected without affecting the statistical properties relevant for anomaly detection research.

Overall, ESA-AD and ESA-ADB provide a large-scale, multivariate, hierarchically organised telemetry dataset with sparse, event-based anomaly annotations. This structural complexity makes it particularly suitable for investigating the behaviour of time-series anomaly detection models under realistic space mission conditions.

### 3.1.2 Mission 1 Structural Statistics

Mission 1 forms the core experimental basis of this thesis. It comprises 76 telemetry channels organised into four subsystems and multiple channel groups, providing a structured multivariate telemetry environment for subsystem-level exploratory analysis [1]. The telemetry signals represent a heterogeneous mixture of physical sensor readings, system status indicators, and operational parameters. The dataset includes both densely sampled signals capturing fast subsystem dynamics and more slowly varying channels reflecting gradual system evolution. This variation in sampling density introduces temporal heterogeneity that must be considered during exploratory analysis and modelling design. The mission data are partitioned chronologically into training and testing segments to preserve temporal causality. This setup reflects realistic operational conditions in which models are trained on historical telemetry and evaluated on later mission periods [1]. Taken together, Mission 1 exhibits high dimensionality, hierarchical subsystem organisation, and substantial temporal heterogeneity. These properties directly motivate the subsystem-level exploratory analysis conducted in this chapter and influence the modelling choices discussed in later chapters.

### 3.1.3 Target vs Non-Target Channels

Within ESA-AD, telemetry channels are categorised into target and non-target signals [1]. Target channels are those for which anomaly annotations are provided and against which detection performance is evaluated. These channels correspond to measurements considered operationally relevant for anomaly monitoring and constitute the primary focus of benchmarking procedures [1].

Non-target channels, in contrast, are not annotated for anomalies and are not directly used for performance evaluation [1]. They may include auxiliary signals such as status indicators, counters, or contextual telemetry that provide complementary system information but are not intended to serve as primary anomaly carriers. While non-target channels may exhibit irregular behaviour or extreme values, such behaviour is not labelled as anomalous within the benchmark framework [1].

The distinction between target and non-target channels has important methodological implications. In multivariate modelling settings, non-target channels can still contribute contextual information that improves the predictive or reconstructive modelling of target signals. However, evaluation metrics are computed exclusively on target channels, meaning that improvements in detection performance must ultimately manifest in those signals [1].

From a structural perspective, the coexistence of target and non-target channels within the same subsystem introduces additional heterogeneity. Some subsystems contain only target channels, while others include a mixture of both types. This structural diversity affects correlation patterns, variance distribution, and inter-channel dependencies, and therefore plays a central role in the exploratory analysis presented in the following sections.

### 3.1.4 Subsystem Organisation and Channel Grouping

The telemetry architecture of ESA Mission 1 is organised hierarchically, reflecting the functional structure of the spacecraft. At the highest level, channels are grouped into subsystems, each corresponding to a major operational domain such as power management, thermal regulation, attitude control, or communication. In Mission 1, the 76 telemetry channels are distributed across four subsystems, enabling structured analysis aligned with spacecraft functionality [1].

Within each subsystem, channels are further organised into channel groups. These groups cluster signals that share similar physical characteristics, operational roles, or statistical behaviour. Channel grouping facilitates structured data handling and supports the design of modelling strategies that exploit intra-group similarities while preserving inter-group distinctions [1]. This organisation also reflects practical engineering considerations, as signals within a group often originate from related sensors or control modules.

The hierarchical organisation serves multiple purposes. From a data management perspective, it reduces complexity by partitioning a high-dimensional telemetry space into smaller, functionally coherent subsets. From an analytical standpoint, it enables subsystem-level investigation of correlation patterns, sampling behaviour, and spectral properties. Such partitioning is particularly relevant in large-scale multivariate anomaly detection scenarios, where modelling all channels jointly may obscure subsystem-specific dynamics.

Importantly, subsystem boundaries do not necessarily imply statistical homogeneity. Channels within the same subsystem may differ in sampling rate, variance magnitude, or temporal dynamics. Similarly, channel groups may exhibit internal coherence while remaining weakly coupled to other groups. Consequently, the subsystem and channel-group organisation provides a structural prior for analysis, but does not eliminate the need for detailed exploratory investigation.

This hierarchical structuring forms the foundation for the subsystem-level analyses presented in the following sections, where each subsystem is examined in terms of its statistical characteristics, inter-channel dependencies, and spectral behaviour.

### 3.1.5 Dataset Partitioning (Training / Validation / Test)

To ensure reproducibility and realistic evaluation conditions, the ESA-AD dataset is partitioned chronologically into training and testing segments as defined in the ESA-ADB benchmark specification [1]. For Mission 1, the first portion of the timeline is allocated to model development, while the latter portion is reserved exclusively for performance evaluation. This temporal split preserves causality by preventing future information from influencing model training.

Unlike random shuffling strategies commonly used in conventional machine learning pipelines, chronological splitting preserves the natural temporal order of telemetry data. This prevents information leakage from future observations into the training set and more closely reflects operational deployment conditions, where anomaly detection models are trained on historical data and subsequently applied to future mission telemetry.

The sparse and event-based nature of anomaly annotations further reinforces the importance of proper partitioning. Since anomalous intervals are rare and unevenly distributed across the mission timeline, maintaining the original temporal structure helps ensure that performance metrics reflect operational detectability rather than artefacts introduced by artificial resampling.

Overall, the chronological training–validation–test division defined in ESA-ADB establishes a principled experimental framework for evaluating anomaly detection methods under realistic mission conditions [1].

### 3.1.6 Anonymisation and Normalisation Procedure

To comply with confidentiality and security requirements, the ESA-AD dataset has undergone a structured anonymisation process prior to public release [1]. This procedure ensures that mission-sensitive information is not disclosed while preserving the statistical and structural properties necessary for anomaly detection research.

The anonymisation process includes the systematic renaming of missions, subsystems, channels, anomaly classes, and physical units [1]. As a result, identifiers no longer reveal spacecraft-specific or engineering-sensitive details. In addition, the temporal axis has been shifted and scaled, so that the dataset timeline begins at an artificial reference date. This transformation maintains the relative temporal ordering of events while preventing reconstruction of the original mission chronology.

Signal values have also been normalised within channel groups [1]. This normalisation procedure rescales telemetry measurements to a common numerical range, typically within the interval ([0,1]), while preserving intra-group relationships and dependencies. By performing normalisation at the channel-group level rather than globally across all channels, the benchmark retains structural coherence within related signals while reducing scale disparities that could reveal physical magnitudes.

Importantly, these anonymisation and normalisation steps are designed to be statistically reversible in principle and do not alter the intrinsic temporal dynamics, correlation structures, or anomaly distributions of the data [1]. Consequently, the dataset remains suitable for rigorous anomaly detection experimentation, even though absolute physical interpretations of signal magnitudes are abstracted.

The anonymised and normalised structure of ESA-AD therefore represents a compromise between operational confidentiality and research transparency, enabling realistic large-scale benchmarking without exposing mission-critical details.

## 3.2 Methodological Framework for Subsystem Analysis

### 3.2.1 Channel Shape and Length Inspection

Before applying any statistical or spectral analysis, each subsystem was first examined at the raw data level to verify the structural integrity of the telemetry channels. This preliminary inspection focused on two fundamental properties: the effective time-series length of each channel and the qualitative behaviour of its initial temporal segment.

The length of each channel was extracted and compared within each subsystem to identify potential heterogeneity in acquisition duration or sampling density. Differences in channel length may arise due to subsystem-specific logging configurations or irregular data acquisition patterns. Since several subsequent analyses rely on consistent temporal alignment, this step provides early identification of channels requiring careful handling.

| Subsystem | Channels (Grouped by Length) | Samples | Count | Structural Characterisation |
|---|---|---|---|---|
| 1 | 1–3 | 10,513,336 | 3 | Multi-length (heterogeneous) |
| 1 | 4–11 | 3,589,376 | 8 | |
| 1 | 61–63 | 487,448 | 3 | |
| 1 | 67–69 | 4,810,995 | 3 | |
| 3 | 53–56, 70–73, 75–76 | ~14,6M | 10 | Mostly homogeneous |
| 3 | 74 | 19,286,353 | 1 | High-rate outlier |
| 5 | 41–46 | 15,381,169 | 6 | Fully homogeneous |
| 6 | 12,13,16–20,25–28,34–37 | 14,258,506 | 15 | Multi-regime (heterogeneous) |
| 6 | 14,15,21–24,29–33,38–40 | 5,272,893 | 16 | |
| 6 | 50–52 | 4,974,682 | 3 | |
| 6 | 57–60 | 14,643,410 | 4 | |
| 6 | 64–66 | 4,810,995 | 3 | |

*Table 3.1 — Number of samples per channel group within each subsystem.*

In addition to numerical length inspection, the first fixed number of samples (e.g., 5,000 points) of each channel were plotted to visually assess their raw temporal behaviour. This qualitative inspection allows the identification of constant-valued channels, abrupt saturations, strong periodic components, or anomalous discontinuities.

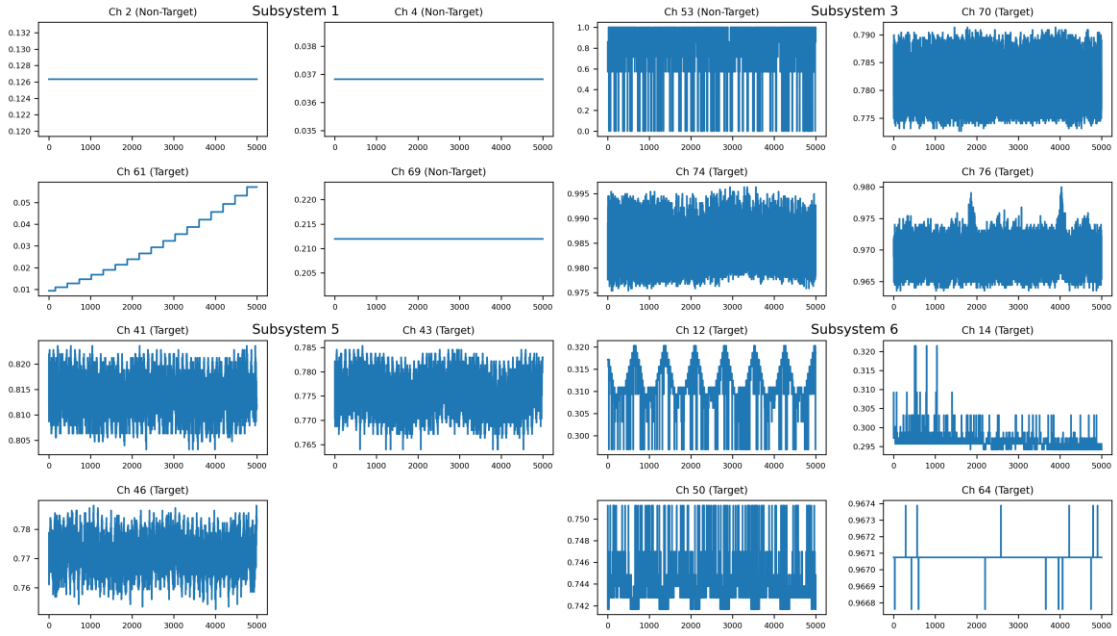


*Figure 3.1 — Representative time-domain windows (first 5,000 samples) across selected channels from each subsystem of Mission 1. The figure highlights structural diversity in amplitude range, temporal density, and signal morphology.*

Visual inspection at this stage does not aim to detect anomalies, but rather to understand baseline signal morphology. Establishing this structural baseline ensures that later statistical and spectral analyses are grounded in verified signal behaviour rather than artefacts introduced by formatting or acquisition inconsistencies.

### 3.2.2 Sampling Frequency Estimation

Following the inspection of channel length and structural grouping, the next step consisted of estimating the effective sampling frequency of each telemetry channel. While channel length provides information about the total number of recorded samples, it does not directly indicate the temporal resolution at which the signals were acquired. Therefore, sampling frequency estimation was necessary to characterise the acquisition regimes within and across subsystems.

For each channel, the average time difference between consecutive timestamps was computed, and the sampling frequency was derived as the reciprocal of this mean time interval. This procedure provides an empirical estimate of the effective sampling rate, accounting for the actual time indexing of the telemetry data rather than relying solely on the number of samples.

The results revealed that sampling frequency is not uniform across all subsystems. In certain subsystems, channels share identical or nearly identical temporal resolutions, indicating acquisition from a common logging configuration. In others, distinct sampling regimes coexist within the same subsystem, reflecting heterogeneous measurement sources or instrumentation modules. Such differences were particularly evident in subsystems exhibiting multiple channel-length groups.

| Subsystem | Channel Group | Dominant Δt (sec) | Sampling Frequency fs (Hz) | Notes |
|---|---|---|---|---|
| Subsystem 1 | 1–3 | 90 | 0.0111 | Moderate-rate telemetry channels |
| Subsystem 1 | 4–11, 67–69 | 180 | 0.0056 | Lower-rate housekeeping measurements |
| Subsystem 1 | 61–63 | 900 | 0.0011 | Very slow monitoring parameters |
| Subsystem 3 | 53–56, 70–73, 75–76 | 30 | 0.0333 | Stable sampling regime |
| Subsystem 3 | 74 | 30 | 0.0333 | Multi-rate behaviour: additional 3 s and 6 s intervals occur in early mission segments |
| Subsystem 5 | 41–46 | 30 | 0.0333 | Homogeneous subsystem with uniform sampling |
| Subsystem 6 | 47–49, 57–60 | 30 | 0.0333 | Higher-rate telemetry signals |
| Subsystem 6 | 12–40, 50–52 | 90 | 0.0111 | Intermediate-rate telemetry group |
| Subsystem 6 | 64–66 | 180 | 0.0056 | Lower-rate telemetry channels |

*Table 3.2 — Estimated sampling frequency per channel group within each subsystem.*

Sampling frequency estimation in *Table 3.2* reveals that telemetry acquisition is organised into a small number of dominant temporal regimes across the dataset. Most channels operate at sampling intervals of approximately **30 s, 90 s, or 180 s**, while a small group of slowly varying parameters in Subsystem 1 is sampled at approximately **900 s** intervals. These regimes likely correspond to different categories of onboard measurements, ranging from rapidly varying telemetry streams to slowly changing health or status indicators.

An additional observation concerns **Channel 74 in Subsystem 3**. Although its dominant sampling interval aligns with the **30 s regime**, inspection of timestamp differences reveals additional **3 s and 6 s intervals** concentrated in the early portion of the mission timeline. This behaviour indicates a **multi-rate temporal acquisition pattern**, which is analysed in more detail in the following subsection.

The presence of multiple sampling frequencies has direct analytical implications. First, it affects the interpretable frequency range in spectral analysis, since the Nyquist limit depends on the sampling rate. Second, it introduces challenges in multivariate modelling, as signals sampled at different temporal resolutions may require alignment, resampling, or careful

window selection. Third, heterogeneous sampling regimes may indicate underlying structural separation between physical measurement families.

Overall, sampling frequency estimation complements channel-length inspection by revealing the temporal granularity of telemetry acquisition. Together, these two structural dimensions provide a foundational understanding of subsystem organisation before proceeding to correlation and spectral analyses.

### 3.2.3 Time-Domain Behaviour Analysis

Following the verification of channel length and sampling frequency, a qualitative and quantitative examination of the time-domain behaviour of each channel was conducted. This step aims to characterise the intrinsic temporal dynamics of telemetry signals prior to correlation and spectral analysis.

Representative time-domain windows of selected channels are illustrated in Figure 3.1. The figure highlights the diversity of signal morphologies observed across the subsystems, ranging from nearly constant telemetry parameters to highly dynamic oscillatory signals.

For each subsystem, a fixed-length window from the initial segment of the mission timeline (e.g., the first 5,000 samples) was visualised to provide a consistent basis for comparison across channels. The objective of this inspection was not anomaly detection, but structural understanding. Specifically, the analysis focused on identifying amplitude range, smoothness, oscillatory behaviour, discrete level switching, trend components, and potential saturation effects.

Clear structural differences emerged between subsystems. In some cases, channels exhibit smooth, continuous variations with gradual oscillatory patterns, characteristic of analog physical measurements such as thermal or electrical telemetry. In other cases, signals display discrete, step-like transitions between limited value levels, indicating digital status or control signals. Such differences are particularly evident in subsystems containing both target and non-target channels.

Time-domain inspection also revealed variability in signal volatility. Certain channels show low-amplitude fluctuations around stable means, suggesting slowly varying physical processes. Others exhibit more dynamic behaviour with higher variance and visible periodic components. These observations provide early indications of spectral structure and inter-channel dependencies explored in subsequent sections.

No major structural discontinuities were observed within the inspected initial windows of the analysed channels, indicating that the dataset is temporally stable in its early segments. This stability supports the validity of later statistical and frequency-domain analyses.

Overall, time-domain behaviour analysis establishes a qualitative understanding of signal morphology and functional diversity within and across subsystems. This step bridges raw data inspection and more formal correlation and spectral characterisation, ensuring that subsequent analytical conclusions are grounded in observable temporal dynamics.

### 3.2.4 Correlation Structure Analysis

After establishing the temporal characteristics of individual channels, the next step consisted of analysing inter-channel dependencies within each subsystem through correlation analysis. While time-domain inspection reveals individual signal behaviour, correlation structure provides insight into collective dynamics and potential functional coupling between telemetry signals.

For each subsystem, pairwise linear correlations were computed between channels using a fixed-length temporal window to ensure comparability. The resulting correlation matrices were visualised as heatmaps in the subsystem-specific analyses presented later in this chapter, with correlation coefficients constrained to the interval [-1, 1] for consistent interpretation.

The correlation patterns vary significantly across subsystems. In some cases, channels form tightly coupled clusters with strong positive correlations, suggesting that they measure related physical quantities or originate from the same sensing module. Such clusters often correspond to channels with similar sampling regimes and spectral behaviour.

In contrast, other subsystems exhibit weak or near-zero correlations between channels, indicating structural independence or functional separation. This behaviour is particularly evident in subsystems containing a mixture of digital status signals and continuous measurement channels. Discrete switching signals typically show limited linear correlation with analog telemetry, reflecting their fundamentally different signal dynamics.

Additionally, subsystems with heterogeneous sampling frequencies or channel lengths often display block-wise correlation structures, where high correlation is observed within groups sharing similar acquisition regimes, and weaker correlation appears across groups. This reinforces the structural grouping suggested by the channel-length and sampling analyses.

Beyond identifying redundancy, correlation analysis can serve two methodological purposes. First, it provides early evidence of potential dimensionality reduction opportunities in highly redundant channel groups. Second, it highlights subsystems where multivariate modelling may benefit from exploiting strong inter-channel dependencies.

Overall, correlation structure analysis reveals the internal organisation of telemetry signals within each subsystem and establishes a structural foundation for subsequent spectral and statistical characterisation.

### 3.2.5 Power Spectral Density (PSD) Analysis

To complement time-domain and correlation analyses, the spectral characteristics of each channel were examined through Power Spectral Density (PSD) estimation. While time-domain inspection reveals qualitative signal behaviour and correlation analysis highlights inter-channel dependencies, PSD analysis provides insight into the frequency content and periodic structure of telemetry signals.

For each channel, the PSD was estimated using Welch's method, which offers a robust and noise-reduced approximation of the signal's frequency distribution. The frequency axis was defined according to the previously identified sampling frequency of each channel, ensuring correct interpretation of spectral limits.

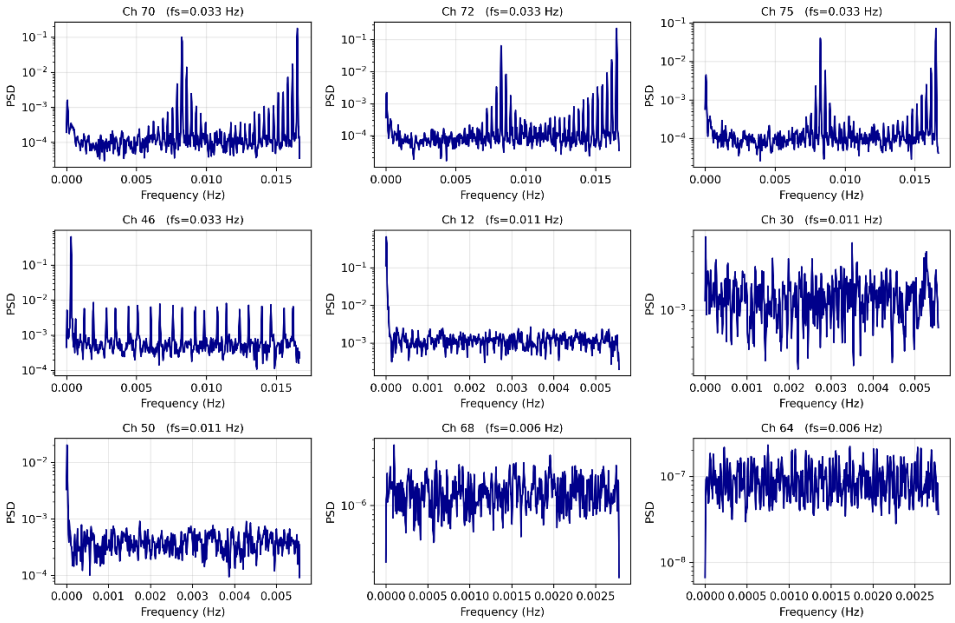


*Figure 3.2 — Representative PSD plots for each subsystem, grouped by structural category.*

The spectral analysis reveals substantial variation across channels. Some signals exhibit dominant low-frequency components associated with slow-varying physical processes such as thermal regulation or long-period operational cycles. In contrast, other channels display broader spectral profiles without pronounced peaks, indicating non-periodic behaviour or discrete switching dynamics often associated with digital telemetry signals.

In subsystems characterised by heterogeneous sampling frequencies, the spectral bandwidth differs accordingly. Channels with higher sampling rates exhibit extended frequency ranges and may reveal higher-frequency components that remain inaccessible in slower-sampled signals. This confirms the structural distinctions previously identified in the sampling frequency analysis.

Spectral coherence within channel groups further reinforces the correlation findings. Channels belonging to the same structural or physical groups often exhibit similar spectral envelopes, suggesting the presence of related dynamic processes. Conversely, channels with distinct spectral signatures often correspond to structurally separate acquisition regimes.

Overall, PSD analysis provides a frequency-domain perspective on subsystem organisation, revealing periodicity patterns, bandwidth differences, and structural signal families.This spectral characterisation complements the structural analysis and provides a frequency-domain perspective prior to statistical feature extraction.

### 3.2.6 Statistical Feature Extraction

Following structural, temporal, and spectral analysis, a set of descriptive statistical features was extracted from each telemetry channel to quantify its distributional characteristics. While previous sections focused on qualitative and relational properties of the signals, this step provides a numerical summary of channel-level behaviour.

For each channel, first-order and higher-order statistics were computed, including mean, standard deviation, variance, minimum, maximum, range, skewness, kurtosis, and median absolute deviation (MAD). These features capture central tendency, dispersion, asymmetry, tail behaviour, and robustness to outliers.

| Subsystem | Channel | Mean | Std | Range | Skew | Kurtosis |
|---|---|---|---|---|---|---|
| 1 | 1 | 0.138 | 0.005 | 0.341 | 46.04 | 2205.13 |
| 1 | 4 | 0.117 | 0.084 | 0.334 | 0.739 | -0.271 |
| 1 | 61 | 0.529 | 0.354 | 0.999 | -0.126 | -1.498 |
| 3 | 53 | 0.831 | 0.151 | 1.000 | -4.09 | 19.51 |
| 3 | 70 | 0.780 | 0.007 | 0.796 | -64.02 | 6834.22 |
| 5 | 41 | 0.810 | 0.016 | 0.982 | -31.68 | 1528.83 |
| 5 | 46 | 0.767 | 0.016 | 0.960 | -25.02 | 1120.47 |
| 6 | 12 | 0.247 | 0.044 | 0.874 | 4.40 | 68.35 |
| 6 | 47 | 0.002 | 0.012 | 1.000 | 80.55 | 6486.27 |
| 6 | 64 | 0.967 | 0.018 | 0.999 | -53.15 | 2823.28 |

*Table 3.3 — Representative Statistical Features of Selected Telemetry Channels.*

The extracted statistics reveal substantial diversity across subsystems. Channels with near-zero variance and negligible standard deviation typically correspond to constant or quasi-constant telemetry, often associated with stable configuration parameters or digital states. In contrast, channels with higher variance and broader ranges reflect more dynamic physical processes.

Skewness and kurtosis further differentiate signal types. Highly skewed or heavy-tailed distributions suggest asymmetric operational regimes or rare excursions from nominal

states. Conversely, approximately symmetric distributions with moderate kurtosis are characteristic of stable continuous measurements.

In subsystems exhibiting heterogeneous channel groups, statistical features often exhibit similar patterns according to acquisition regime and signal type. Channels sharing similar sampling frequencies and spectral characteristics often display comparable variance and distributional profiles. This reinforces the structural organisation identified in earlier sections.

Statistical feature extraction therefore provides a compact quantitative representation of telemetry behaviour, bridging raw signal inspection and algorithmic modelling. These features not only characterise subsystem diversity but also highlight channels with distinctive statistical signatures that may influence anomaly detection performance in later chapters.

## 3.3 Structural Characterisation of Subsystems

### 3.3.1 Subsystem 1 — Mixed Target and Non-Target Dynamics

Subsystem 1 represents a structurally heterogeneous telemetry group characterised by the coexistence of active target channels and predominantly inactive or low-variance non-target channels. As identified in Section 3.2, the subsystem comprises channels with four distinct acquisition lengths, indicating multiple logging regimes within a single functional grouping. This multi-length structure already suggests that Subsystem 1 cannot be treated as a statistically uniform entity.

From a compositional perspective, Subsystem 1 includes target channels (61–63) alongside non-target channels (1–11 and 67–69). The target channels exhibit noticeable temporal variability and measurable amplitude fluctuations, while several non-target channels display near-constant behaviour with negligible variance. This dichotomy introduces intrinsic heterogeneity at both structural and dynamical levels.

Time-domain inspection confirms this mixed behaviour (see Figure 3.1). Target channels show continuous variation and observable dynamic patterns, whereas a subset of non-target channels appears effectively flat over the inspected windows. Such disparity directly affects both statistical descriptors and correlation patterns. Channels with near-zero variance contribute limited information content and may distort higher-order statistics such as skewness and kurtosis.

Correlation analysis reveals limited global coupling across the entire subsystem. While some local dependencies exist among channels sharing similar acquisition lengths, the overall correlation matrix does not exhibit a strongly cohesive block structure. This indicates that

Subsystem 1 is not dominated by a single tightly coupled physical process, but rather contains loosely related telemetry signals with varying functional roles.

Spectral analysis further reinforces this interpretation (see Figure 3.2). Target channels demonstrate identifiable low-frequency components consistent with gradual physical dynamics, whereas several non-target channels exhibit minimal spectral energy beyond baseline noise levels. The coexistence of spectrally active and spectrally weak channels within the same subsystem highlights its mixed structural nature.

Statistical feature profiling further supports these observations. Target channels in Subsystem 1 exhibit finite variance and moderate amplitude ranges, reflecting continuously varying telemetry signals. For instance, several non-target channels display noticeably larger standard deviations than quasi-static configuration channels. In contrast, several non-target channels present extremely small standard deviation and compressed value ranges, indicating nearly constant telemetry behaviour typical of configuration or system state parameters. This divergence in statistical properties reinforces the interpretation of Subsystem 1 as a mixed-dynamics subsystem containing both active sensor measurements and stable system-state signals.

Overall, Subsystem 1 can be characterised as a heterogeneous structure combining active measurement signals and quasi-static auxiliary channels across multiple acquisition regimes. This structural diversity has important implications for multivariate anomaly detection, as modelling assumptions of uniform temporal behaviour or shared statistical profiles may not fully hold across the entire subsystem.

### 3.3.2 Subsystem 3 — Mixed Telemetry Dynamics with a Multi-Rate Channel

Subsystem 3 presents a largely coherent telemetry group with one temporally irregular channel embedded within it. Most channels in this subsystem exhibit stable sampling behaviour and comparable signal morphology, while Channel 74 deviates from this pattern due to its heterogeneous acquisition regime. This configuration creates a subsystem that is largely homogeneous in terms of signal dynamics but contains a distinct temporal outlier.

From a compositional perspective, Subsystem 3 contains both non-target channels (53–56) and target channels (70–76). Unlike Subsystem 1, where several channels display near-constant behaviour, most signals in Subsystem 3 exhibit continuous temporal variation and measurable amplitude fluctuations. The channels represent active telemetry streams rather than quasi-static status indicators.

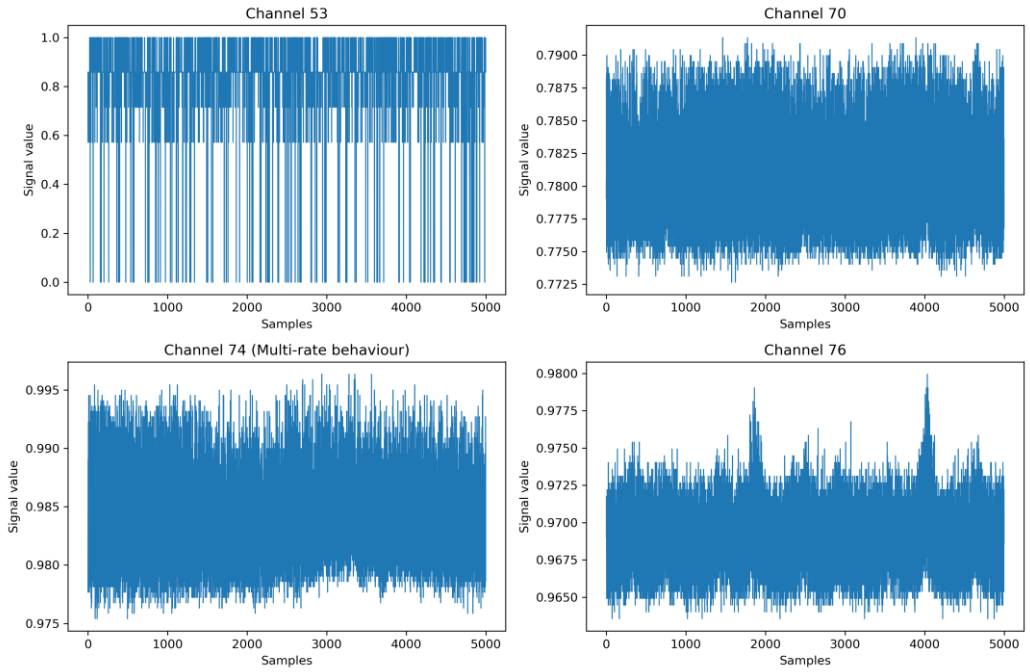


*Figure 3.3 — Representative time-domain windows of Subsystem 3 channels.*

Visual inspection of representative time-domain segments shows that the majority of channels share similar oscillatory behaviour and comparable amplitude ranges. These signals display gradual fluctuations typical of continuously measured physical quantities. Their temporal morphology remains consistent across the subsystem, suggesting measurement of related system processes.

Correlation analysis further supports the structural coherence of the subsystem. Most channels form a moderately correlated block, indicating shared system dynamics or common operational influences.

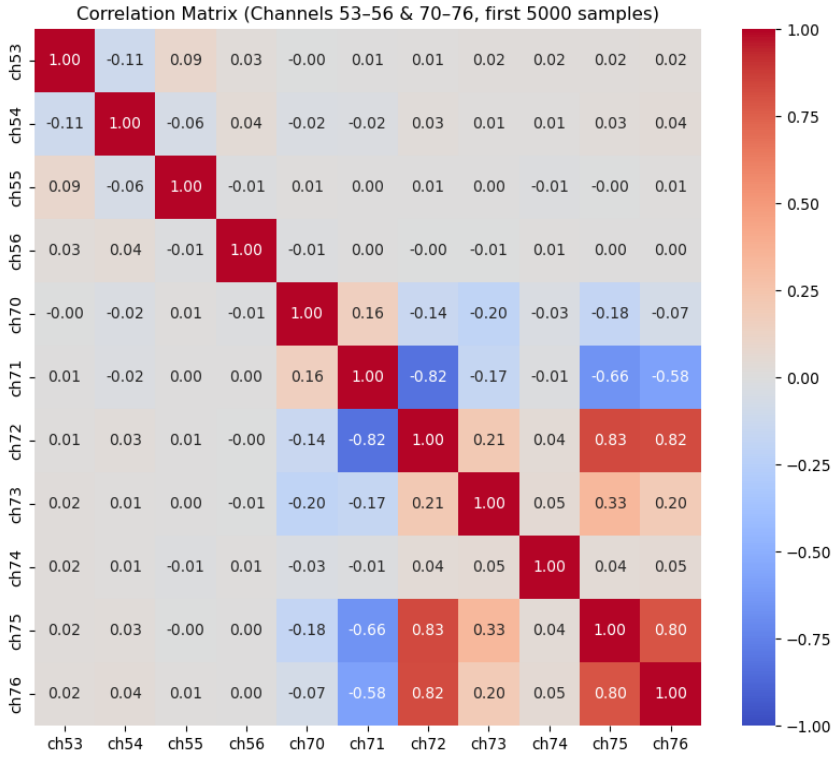


*Figure 3.4 — correlation heatmap of Subsystem 3.*

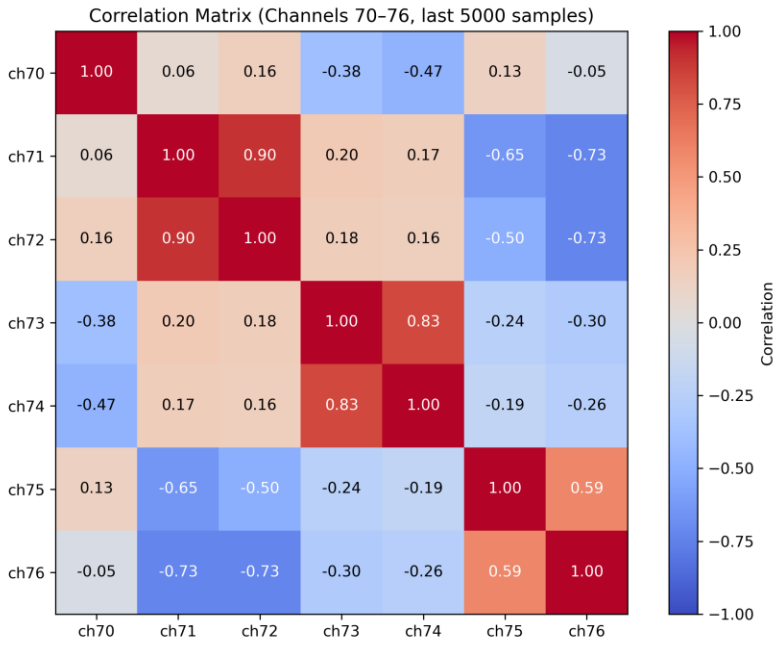


*Figure 3.5 — Correlation structure for channels sharing the same sampling regime (tail segment).*

Spectral inspection reveals that the dominant frequency content of these channels is concentrated in the low-frequency region. This behaviour is consistent with slow system dynamics and suggesting that most channels operate under similar acquisition conditions and sampling rates.

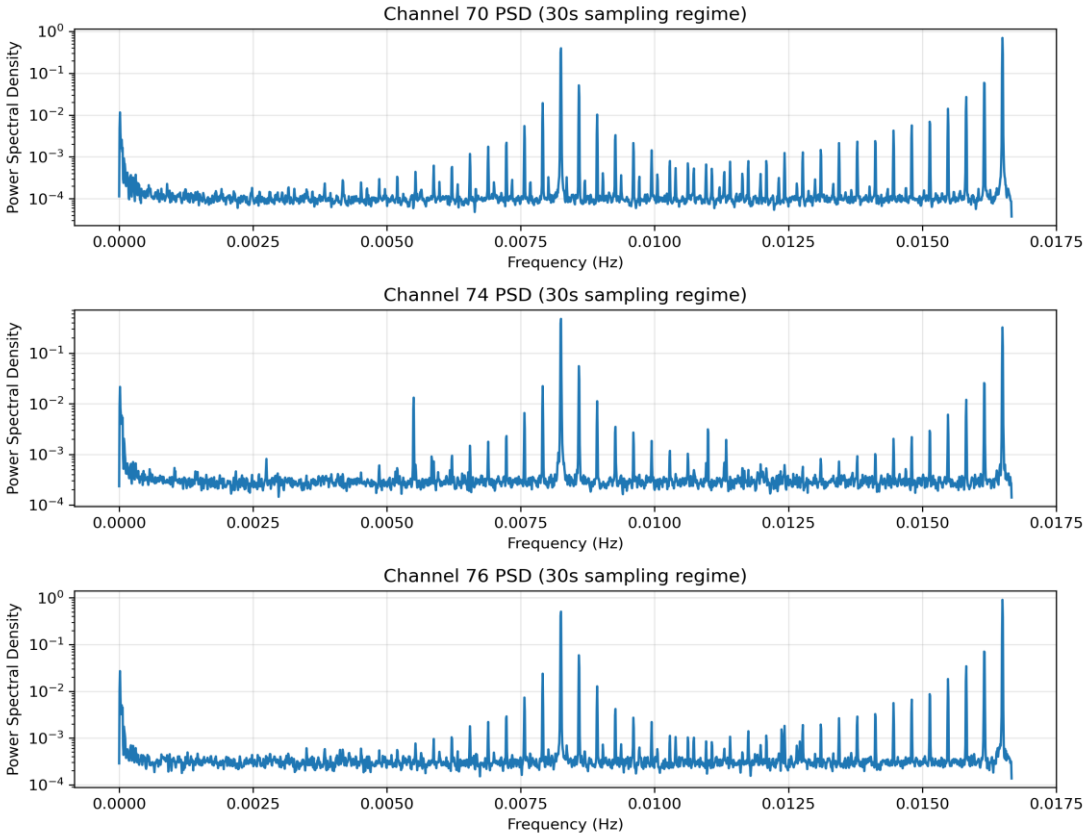


*Figure 3.6— PSD comparison across representative Subsystem 3 channels.*

While the dominant frequency peaks appear at similar locations across Channels 70, 74, and 76, Channel 74 exhibits slightly higher spectral power in several components. This behaviour suggests stronger amplitude fluctuations rather than fundamentally different underlying dynamics.

Although the subsystem is largely homogeneous, one channel—Channel 74—exhibits markedly different temporal acquisition behaviour. This channel therefore requires a more detailed examination.

### *3.3.2.1 Channel 74 — Multi-Rate Temporal Acquisition Behaviour*

A detailed inspection of Channel 74 reveals that its sampling structure differs substantially from the remaining channels of Subsystem 3. Rather than following a single consistent

sampling interval, the channel exhibits multiple acquisition regimes throughout the mission timeline.

Analysis of timestamp differences between consecutive samples reveals three dominant sampling intervals: approximately 30 seconds, 6 seconds, and 3 seconds. The 30-second interval represents the dominant regime, accounting for roughly 70% of all observations. Shorter intervals of 6 seconds and 3 seconds occur less frequently and are primarily concentrated in the early portion of the mission timeline.

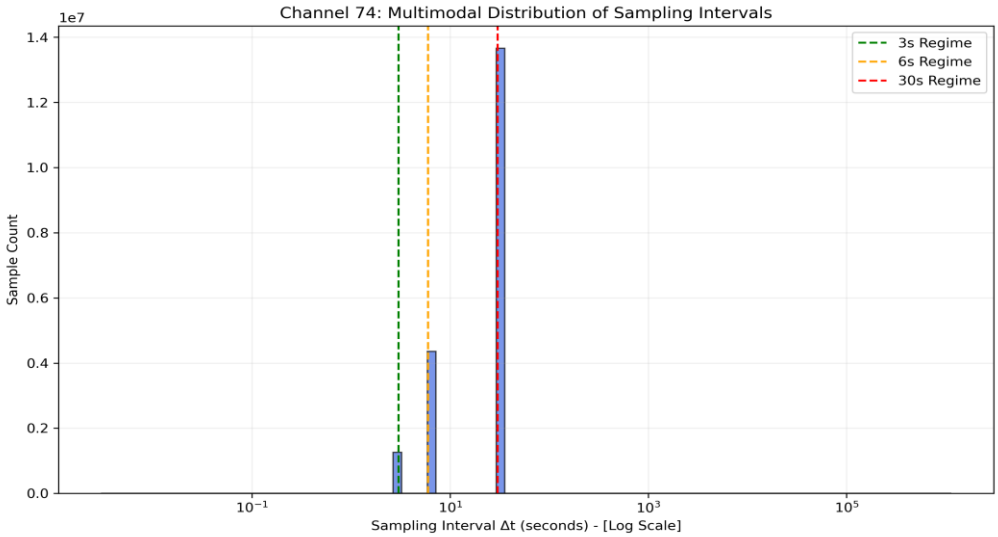


*Figure 3.7 — Log-scaled distribution of Δt for Channel 74.*

Temporal inspection of these sampling intervals indicates that the higher-frequency regimes appear mainly during the initial operational phase of the mission, after which the channel appears to stabilise to the 30-second sampling configuration that persists throughout most of the dataset.

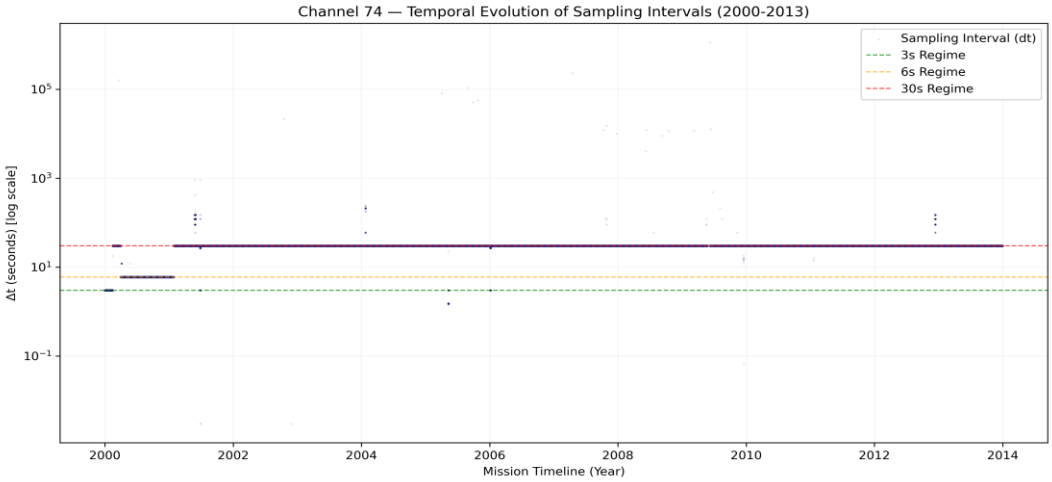


*Figure 3.8 — Temporal evolution of sampling intervals for Channel 74.*

This behaviour contrasts with the rest of Subsystem 3, where channels such as Channel 76 maintain a nearly constant sampling interval of approximately 30 seconds across the entire mission timeline. Consequently, Channel 74 represents a temporal structural outlier within the subsystem rather than a typical member of the dominant telemetry group.

From a modelling perspective, the presence of multiple acquisition regimes implies that naive estimation of sampling frequency using global averages can produce misleading values. For instance, averaging all timestamp differences yields an apparent sampling interval of approximately 22.9 seconds, which does not correspond to any actual acquisition regime present in the data.

Instead, the dominant regime of 30 seconds provides the most representative sampling configuration for the channel and aligns with the acquisition behaviour observed across the rest of the subsystem.

| Channel | Channel Type | Dominant Δt (sec) | Secondary Δt (sec) | Sampling Behaviour |
|---|---|---|---|---|
| 53 | Non-Target | ~30 | – | Uniform sampling |
| 54 | Non-Target | ~30 | – | Uniform sampling |
| 55 | Non-Target | ~30 | – | Uniform sampling |
| 56 | Non-Target | ~30 | – | Uniform sampling |
| 70 | Target | ~30 | – | Uniform sampling |
| 71 | Target | ~30 | – | Uniform sampling |
| 72 | Target | ~30 | – | Uniform sampling |
| 73 | Target | ~30 | – | Uniform sampling |
| **74** | Target | **30** | **6, 3** | **Multi-rate sampling (early mission transition)** |
| 75 | Target | ~30 | – | Uniform sampling |
| 76 | Target | ~30 | – | Uniform sampling |

*Table 3.4 — Dominant sampling regimes of subsystem 3 channels.*

As shown in Table 3.4, the majority of channels in Subsystem 3 operate under a stable 30-second acquisition regime. Channel 74 is the only channel in the subsystem that exhibits multiple sampling intervals.

Overall, Subsystem 3 can therefore be characterised as a structurally coherent telemetry group with one temporally heterogeneous channel. While the majority of signals follow a stable acquisition configuration, Channel 74 introduces a multi-rate temporal structure that reflects changes in measurement configuration during the early mission phase.

This combination of largely uniform signals and a temporally heterogeneous channel highlights an important structural property of the dataset: even within apparently homogeneous subsystems, individual channels may exhibit distinct acquisition characteristics that may need to be considered during subsequent modelling and anomaly detection analyses.

### 3.3.3 Subsystem 5 — Spectrally Coherent Target Subsystem

Subsystem 5 appears to be the most structurally homogeneous subsystem among those analysed. All channels within this group share identical acquisition length and consistent sampling characteristics, indicating a unified logging configuration and coherent subsystem-level measurement regime. Unlike Subsystem 1 and Subsystem 6, no multi-length or multi-rate behaviour is observed here.

From a compositional perspective, Subsystem 5 consists exclusively of target channels (41–46). This structural purity eliminates the mixed-dynamics effect observed in Subsystem 1 and allows the subsystem to function as a controlled environment for analysing anomaly detection behaviour without interference from quasi-static auxiliary signals.

Time-domain inspection reveals consistent oscillatory patterns across channels, with comparable amplitude ranges and smooth temporal evolution. The signals appear structurally similar, suggesting that the channels may measure closely related physical quantities or tightly coupled subsystem dynamics.

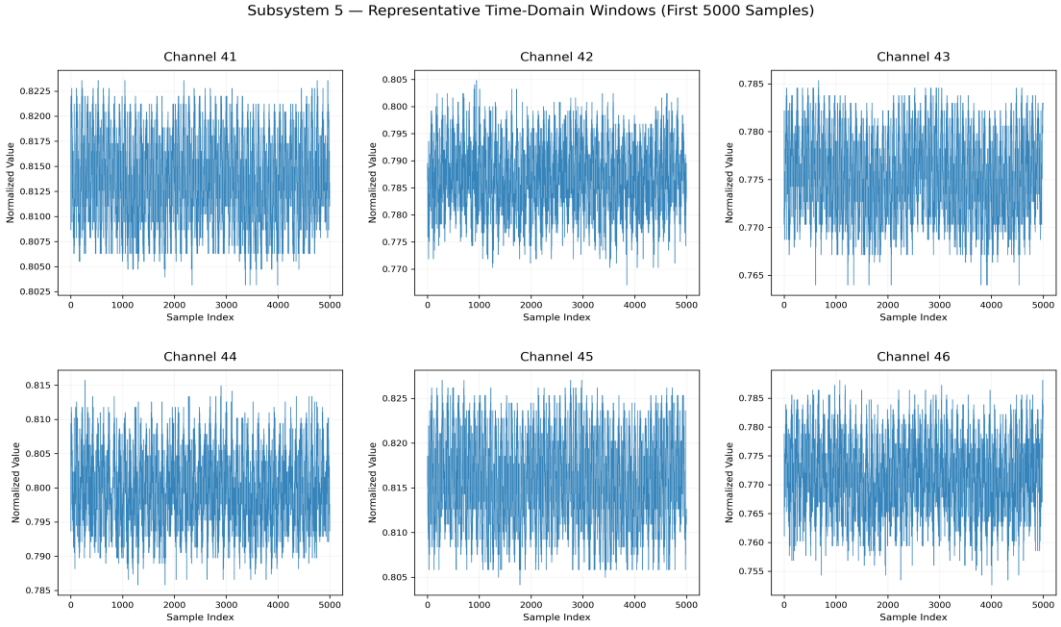


*Figure 3.9 — Representative time-domain windows (first 5,000 samples) for channels 41–46 in Subsystem 5. All channels display consistent oscillatory behaviour and similar amplitude ranges, indicating strongly coherent subsystem dynamics. The absence of quasi-static or structurally divergent signals further confirms the homogeneous nature of this target-only telemetry group.*

Correlation analysis indicates strong internal coherence. The correlation matrix exhibits pronounced block structure with high positive correlations across most channel pairs. This indicates substantial redundancy and strong interdependence within the subsystem.

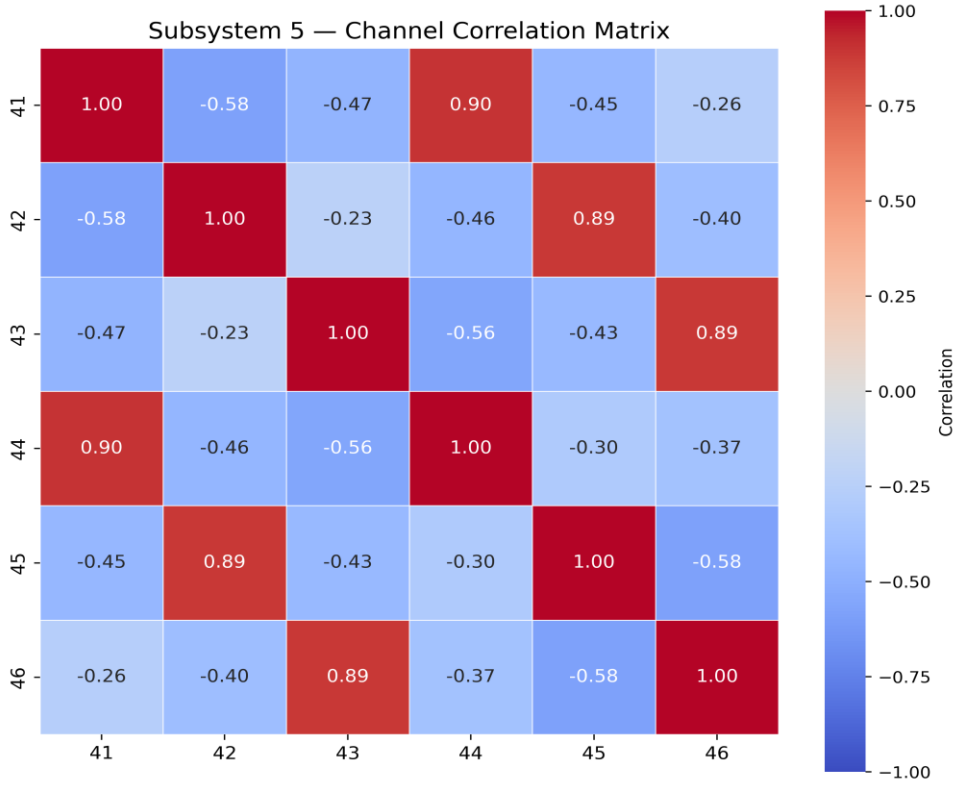


*Figure 3.10— Correlation matrix of channels 41–46 in Subsystem 5. The heatmap shows strong internal coupling, with several channel pairs exhibiting high positive correlations. This pronounced block structure confirms the coherent behaviour of the subsystem and indicates that the channels measure closely related physical processes.*

Spectral analysis reinforces this interpretation. Channels display highly similar power spectral density profiles, characterised by dominant low-frequency components and overlapping spectral envelopes. The close alignment of spectral peaks across channels suggests the presence of common periodic processes governing the subsystem dynamics.

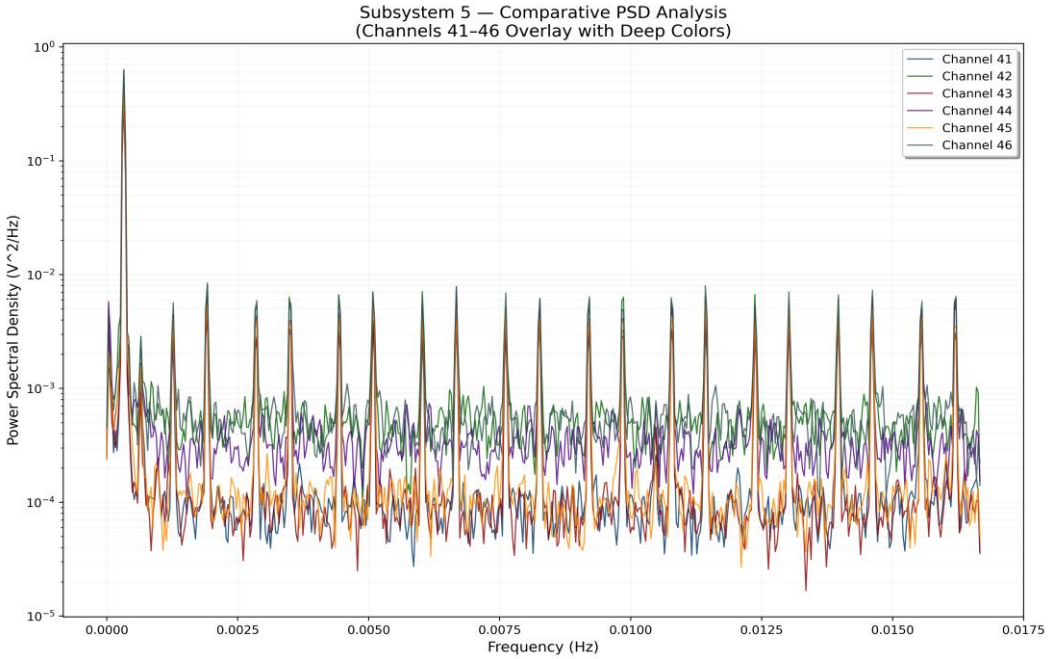


*Figure 3.11 — Overlay of PSD estimates for channels 41–46 in Subsystem 5. The spectra show consistent low-frequency dominance and closely aligned peaks across channels, indicating coherent subsystem dynamics.*

Statistical feature profiles further support the homogeneous characterisation. Variance, amplitude range, and higher-order distributional metrics remain within comparable ranges across channels. No extreme dispersion outliers or quasi-constant signals are observed.

| Channel | Mean | Std | Range | Skewness | Kurtosis |
|---|---|---|---|---|---|
| 41 | 0.810 | 0.0157 | 0.982 | -31.68 | 1528.83 |
| 42 | 0.784 | 0.0162 | 0.966 | -26.46 | 1191.48 |
| 43 | 0.770 | 0.0180 | 0.951 | -30.09 | 1224.31 |
| 44 | 0.795 | 0.0221 | 0.976 | -29.04 | 1010.50 |
| 45 | 0.812 | 0.0161 | 1.000 | -29.34 | 1394.99 |
| 46 | 0.767 | 0.0161 | 0.960 | -25.02 | 1120.47 |

*Table 3.5 — Statistical features of channels 41–46 in Subsystem 5.*

The statistical characteristics reported in Table 3.5 further support the homogeneous nature of Subsystem 5. The mean values remain tightly clustered between approximately 0.767 and 0.812, indicating that all channels operate within a similar measurement range and likely reflect closely related subsystem variables. Likewise, the standard deviation values are consistently low (≈0.015–0.022), suggesting comparable signal variability and the absence of strongly fluctuating outlier channels.

The amplitude ranges also remain highly consistent across channels (≈0.95–1.00), reinforcing the observation that the subsystem exhibits stable dynamic behaviour with

similar signal envelopes. Although skewness and kurtosis values appear large in magnitude, they remain relatively consistent across all channels, indicating similar distributional shapes rather than isolated statistical anomalies.

Overall, the close alignment of central tendency, dispersion, and higher-order statistical descriptors across channels confirms that Subsystem 5 forms a statistically coherent telemetry group. This level of homogeneity provides a controlled environment for anomaly detection experiments, as variations observed during model evaluation are less likely to originate from intrinsic channel heterogeneity and are more likely related to anomaly detection behaviour

Overall, subsystem 5 can therefore be described as a highly homogeneous, tightly coupled, spectrally coherent target subsystem. It represents a structural archetype in which multivariate modelling is expected to benefit from strong inter-channel redundancy and consistent temporal dynamics. This structural clarity contrasts sharply with the heterogeneous regimes observed in other subsystems and provides a useful baseline for interpreting anomaly detection performance in later chapters.

### 3.3.4 Subsystem 6 — Heterogeneous Multi-Cluster Subsystem

Subsystem 6 appears to be the most structurally complex subsystem analysed in this study. Unlike Subsystem 5, which exhibits full homogeneity, and Subsystem 3, which contains a single structured outlier, Subsystem 6 is characterised by multiple acquisition regimes, distinct channel-length groups, and heterogeneous sampling frequencies. This diversity establishes it as a strongly multi-regime telemetry environment.

As identified in Section 3.2, channels within Subsystem 6 are distributed across several distinct length groups, indicating separate logging configurations or measurement families within the same functional subsystem. This multi-length structure is not marginal but substantial, with clearly separated acquisition regimes coexisting in parallel.

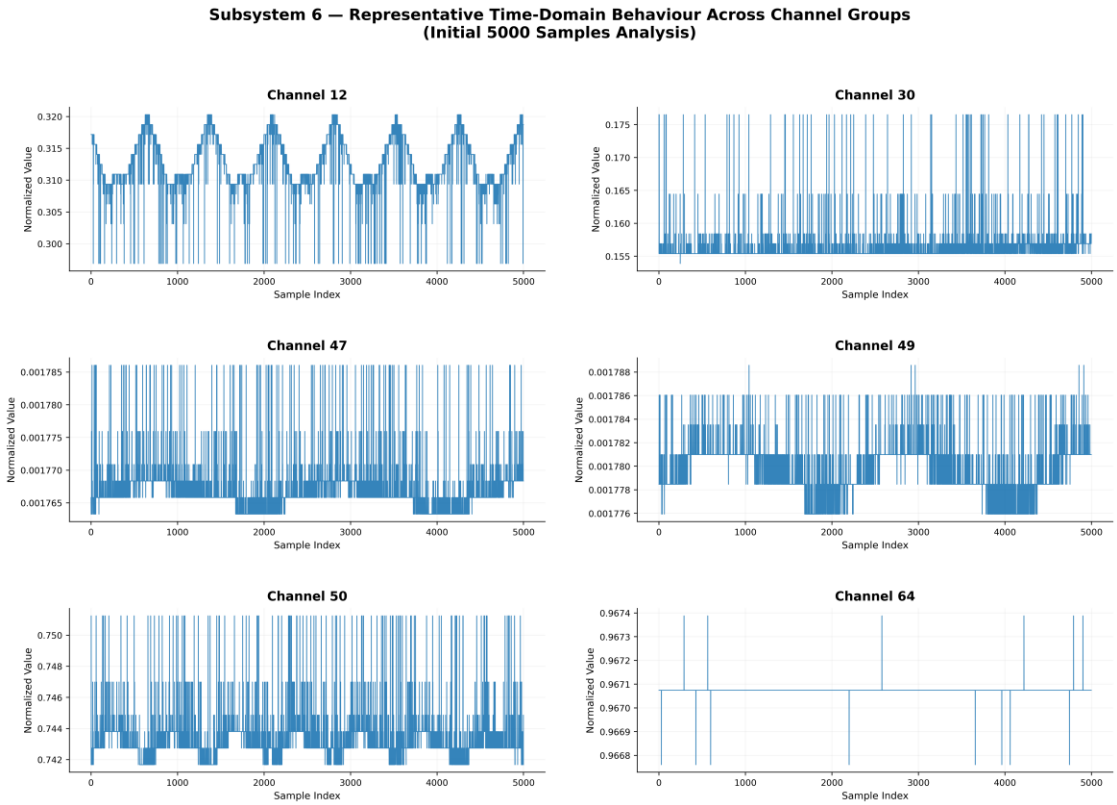


*Figure 3.12 — Representative time-domain grid showing channels from different length groups within Subsystem 6.*

The time-domain inspection highlights the pronounced heterogeneity of Subsystem 6. Unlike the homogeneous behaviour observed in Subsystem 5, the channels in this subsystem exhibit markedly different temporal patterns and amplitude regimes. Channels 12 and 30 display continuous fluctuations with moderate variability, suggesting that these channels may measure gradually evolving physical quantities. In contrast, Channels 47 and 49 show sparse spike-like activations around an otherwise stable baseline, indicating event-driven or threshold-triggered telemetry behaviour.

Channels 50 and 64 present yet another regime. Channel 50 exhibits step-like variations with occasional abrupt transitions, while Channel 64 remains largely stable with only rare deviations from its baseline level. These differences reflect distinct acquisition regimes and functional roles among the channels.

Overall, the coexistence of continuous signals, event-like spikes, and quasi-static channels within the same subsystem suggests that Subsystem 6 represents a heterogeneous multi-cluster telemetry environment. Such structural diversity has important implications for anomaly detection, as models operating on this subsystem must handle signals with very different temporal characteristics and sampling regimes.

Correlation analysis reflects this structural segmentation. Instead of a single cohesive block, the correlation matrix displays partial clustering aligned with acquisition groups. Stronger correlations tend to occur within channels sharing similar lengths and sampling rates, while weaker or less structured dependencies appear across groups.

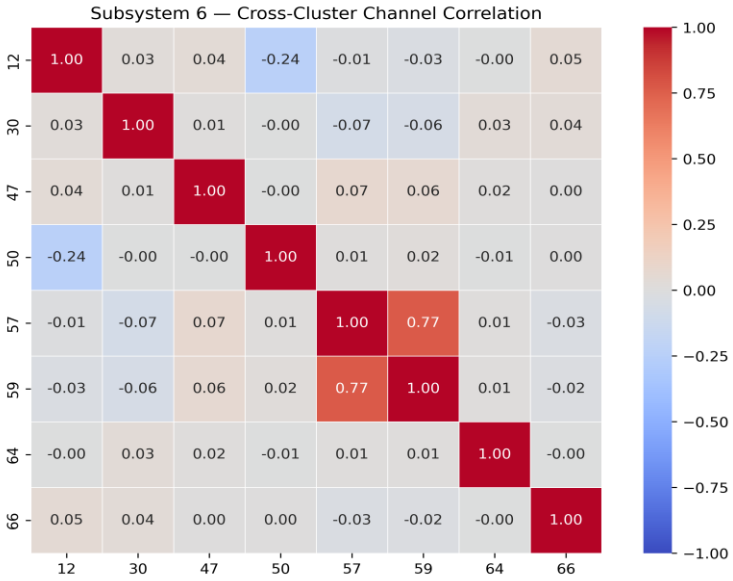


*Figure 3.13 — Cross-cluster correlation heatmap in Subsystem 6.*

The representative cross-cluster correlation matrix highlights the structural heterogeneity of Subsystem 6. Channels drawn from different acquisition groups exhibit generally weak or moderate correlation, confirming that the subsystem does not behave as a single coherent measurement block but rather consists of multiple loosely coupled telemetry clusters.

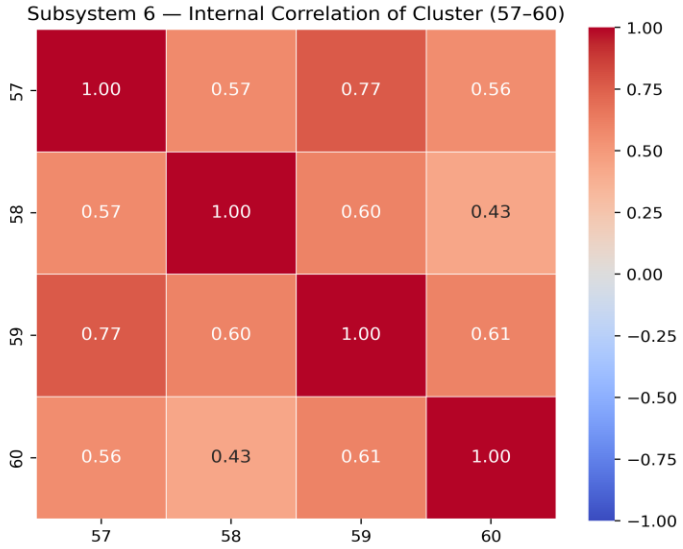


*Figure 3.14 — Correlation heatmap of channels 57–60 within Subsystem 6.*

The matrix reveals consistently positive correlations across all channel pairs, with coefficients ranging approximately from 0.43 to 0.77. The strongest relationship is observed between Channels 57 and 59 ($\rho \approx 0.77$), while the remaining pairs also exhibit moderate to strong positive dependence. This pattern indicates that channels 57–60 form a coherent internal cluster, likely reflecting measurements of closely related physical processes or shared subsystem dynamics within the broader heterogeneous structure of Subsystem 6.

Spectral analysis further differentiates the internal families. Channels with higher sampling frequencies display broader spectral bandwidths and potentially reveal higher-frequency components. Lower-rate channels, constrained by their sampling regime, exhibit narrower spectral envelopes. This suggests that spectral diversity in Subsystem 6 is largely related to acquisition heterogeneity.

| Sampling regime | Channel group | Representative |
|---|---|---|
| 30 s (fs ≈ 0.033 Hz) | 47–52 / 57–60 | **47** |
| 90 s (fs ≈ 0.011 Hz) | 12–40 / 50 | **12** |
| 180 s (fs ≈ 0.0055 Hz) | 64–66 | **64** |

*Table 3.6 — Representative channels for each sampling regime in Subsystem 6.*

The figure compares the spectral profiles of channels sampled under different acquisition regimes. Channel 47 (fs ≈ 0.033 Hz) exhibits the widest observable spectral bandwidth, while Channel 12 (fs ≈ 0.011 Hz) shows a more restricted frequency range.

Channel 64 (fs ≈ 0.0055 Hz), sampled at the lowest rate, presents the narrowest spectral envelope. This progressive contraction of the observable frequency range reflects the direct influence of sampling frequency on spectral resolution and confirms that the spectral diversity of Subsystem 6 originates from heterogeneous acquisition regimes.

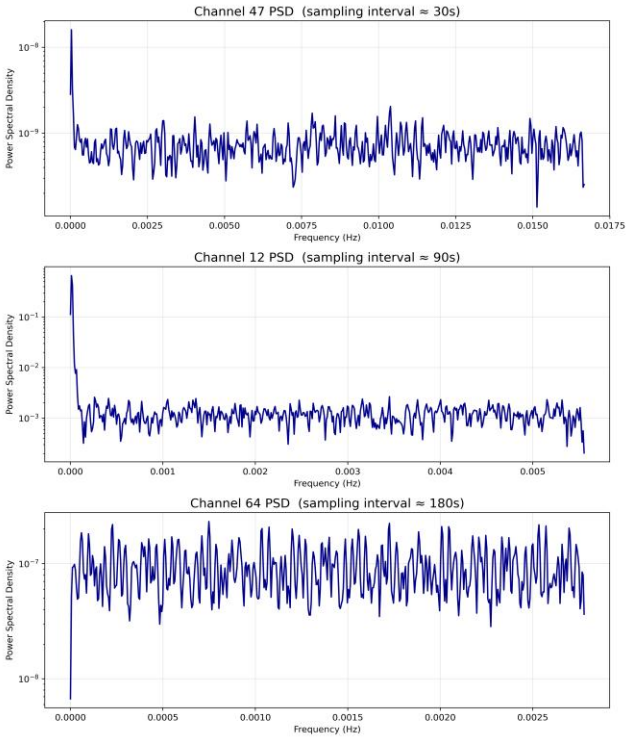


*Figure 3.15 — Power spectral density comparison across representative channels of Subsystem 6.*

The spectral analysis further highlights the heterogeneous acquisition structure of Subsystem 6. The power spectral density (PSD) plots reveal clear differences in the observable frequency bandwidth across channels sampled under different regimes. Channel 47, sampled with the highest temporal resolution (sampling interval ≈ 30 s), exhibits the widest spectral range, extending up to approximately 0.017 Hz. This broader bandwidth allows higher-frequency components of the signal dynamics to be observed.

In contrast, Channel 12, acquired at a lower sampling rate (≈ 90 s), displays a narrower spectral envelope, with the observable frequency range limited to approximately 0.005 Hz. The reduction in spectral bandwidth is consistent with the lower Nyquist frequency imposed by the slower sampling interval.

The effect becomes even more pronounced for Channel 64, sampled at the lowest rate within the subsystem (≈ 180 s). The corresponding PSD is restricted to a much smaller frequency domain, with spectral content observable only up to approximately 0.0028 Hz. As expected, lower sampling frequency constrains the maximum resolvable signal frequency and compresses the spectral representation.

Overall, the progressive reduction of spectral bandwidth from Channel 47 to Channel 64 directly reflects the heterogeneous acquisition regimes present within Subsystem 6. These results indicate that the spectral diversity observed in this subsystem is largely driven by differences in sampling configuration rather than intrinsic signal variability alone. Such

acquisition heterogeneity must therefore be carefully considered when designing anomaly detection models operating on this subsystem.

Statistical feature profiles are correspondingly diverse. Variance, amplitude range, and distributional characteristics vary across groups, indicating that Subsystem 6 does not represent a unified dynamical process but rather a composite of structurally distinct telemetry families. To illustrate the statistical diversity present within Subsystem 6, Table 3.7 reports representative statistical descriptors for selected channels from different structural families identified during the exploratory analysis.

| Channel | Family | Mean | Std | Range | Skewness | Kurtosis |
|---|---|---|---|---|---|---|
| 12 | Continuous telemetry | 0.2465 | 0.0442 | 0.874 | 4.40 | 68.35 |
| 47 | Digital / quasi-binary | 0.0019 | 0.0124 | 1.000 | 80.55 | 6486.27 |
| 50 | Continuous moderate | 0.7522 | 0.0302 | 0.937 | -16.01 | 347.11 |
| 57 | Correlated cluster | 0.7879 | 0.0348 | 0.865 | -3.39 | 72.65 |
| 64 | Near-saturated telemetry | 0.9667 | 0.0182 | 0.999 | -53.15 | 2823.28 |

*Table 3.7 — Representative statistical features of selected channel families in Subsystem 6.*

The statistical descriptors confirm the heterogeneous composition of Subsystem 6. Continuous telemetry channels such as Channel 12 exhibit moderate variance and positively skewed distributions typical of analogue sensor measurements. In contrast, channels such as Channel 47 display extreme skewness and kurtosis values, reflecting quasi-binary or rarely changing telemetry behaviour. Additional channel families, including the correlated cluster represented by Channel 57 and the near-saturated telemetry group represented by Channel 64, exhibit distinct statistical signatures. These differences indicate that Subsystem 6 contains multiple structurally distinct signal families rather than a single coherent dynamical process.

Overall, Subsystem 6 can be characterised as a strongly heterogeneous multi-regime subsystem. Its internal structural segmentation suggests that multivariate modelling strategies must account for acquisition diversity and partial clustering rather than assuming uniform temporal behaviour. Compared to other subsystems, subsystem 6 provides one of the most challenging structural configurations for anomaly detection due to its inherent complexity and internal variability.

## 3.4 Cross-Subsystem Structural Comparison

### 3.4.1 Homogeneous vs Heterogeneous Subsystems

A primary structural distinction emerging from the exploratory analysis concerns the degree of internal homogeneity within each subsystem. Although all subsystems belong to the same mission-level telemetry framework, their internal statistical and dynamical coherence varies substantially.

Subsystem 5 represents the most homogeneous subsystem observed in this analysis. All channels share identical acquisition length, comparable sampling frequency, strong inter-channel correlation, and closely aligned spectral characteristics. No quasi-static or structurally distinct signals are embedded within this group. As a result, Subsystem 5 behaves as a tightly coupled multivariate entity, where internal redundancy and largely uniform dynamics dominate.

Subsystem 3 can be characterised as mostly homogeneous with a structured outlier. The majority of channels exhibit coherent temporal and spectral behaviour; however, the presence of a single high-rate channel (Channel 74) introduces a degree of heterogeneity. While this outlier does not fragment the subsystem entirely, it creates a measurable asymmetry in acquisition regime and frequency-domain content.

Subsystem 1 represents a mixed-dynamics heterogeneous configuration. Here, structural diversity arises from the coexistence of active target channels and quasi-static non-target channels, combined with multiple acquisition lengths. Rather than forming a single cohesive dynamical block, the subsystem contains signals with substantially different variance levels and information content.

Subsystem 6 appears to exhibit the strongest structural heterogeneity among the analysed subsystems. Multiple acquisition regimes, several channel-length groups, and block-wise correlation patterns indicate the presence of internally coherent clusters rather than a single unified subsystem behaviour. In this case, heterogeneity is not marginal but appears to be a defining characteristic of the subsystem.

Taken together, the four subsystems form a structural spectrum ranging from full homogeneity (Subsystem 5) to strong multi-cluster heterogeneity (Subsystem 6), with Subsystems 3 and 1 occupying intermediate positions. This spectrum provides a structured framework for interpreting variability in modelling behaviour observed in later chapters. Table 3.8 summarises the main structural characteristics identified across the analysed subsystems.

| Subsystem | Structural Type | Sampling Regimes | Correlation Structure | Spectral Behaviour |
|---|---|---|---|---|
| 1 | Mixed dynamics | Multiple | Weak global coupling | Heterogeneous |
| 3 | Mostly homogeneous with outlier | Mostly uniform + one high-rate channel | Moderate coherence | Slight spectral asymmetry |
| 5 | Fully homogeneous | Single regime | Strong internal correlation | Highly aligned spectra |
| 6 | Multi-cluster heterogeneous | Multiple regimes | Clustered correlation blocks | Strong spectral diversity |

*Table 3.8 — Structural comparison of Mission-1 Subsystems.*

The comparative overview highlights a clear structural gradient across the analysed subsystems. While Subsystem 5 exhibits strong internal coherence and uniform acquisition conditions, Subsystem 6 demonstrates pronounced structural heterogeneity with multiple sampling regimes and clustered channel behaviour. Subsystems 1 and 3 occupy intermediate positions within this structural spectrum.

### 3.4.2 Sampling Regimes and Temporal Granularity

Beyond internal homogeneity, a second defining dimension across subsystems is the variation in sampling regimes and temporal granularity. While all telemetry signals are time-indexed, their effective sampling densities and acquisition lengths differ considerably across subsystems, influencing both observable dynamics and analytical treatment.

Subsystem 5 exhibits largely uniform temporal granularity. All channels share identical acquisition length and sampling frequency, resulting in aligned time-series representations and consistent Nyquist limits across signals. This temporal uniformity simplifies comparative analysis and multivariate modelling, as no resampling or alignment adjustments are typically required.

Subsystem 3 presents limited temporal asymmetry. Most channels share a common acquisition regime, but Channel 74 operates at a higher effective resolution, resulting in a slightly extended observable frequency bandwidth and finer temporal detail. Although this does not fundamentally disrupt subsystem coherence, it introduces a subtle granularity mismatch that may affect spectral comparability and multivariate alignment.

Subsystem 1 demonstrates moderate heterogeneity in temporal coverage. Multiple channel-length groups coexist, and several channels exhibit substantially shorter acquisition spans. This implies that certain telemetry streams were logged under different operational

conditions or instrumentation policies, introducing partial temporal misalignment within the subsystem.

Subsystem 6 appears to display the strongest temporal granularity divergence among the analysed subsystems. Multiple acquisition regimes with clearly separated sample counts indicate structurally distinct logging configurations. As a result, the subsystem contains signals with different effective bandwidths and temporal resolutions, directly impacting spectral range and potential model receptive fields.

Table 3.9 summarises the dominant sampling regimes and acquisition structures observed across the analysed subsystems.

| Subsystem | Channels | Dominant Sampling Interval | Approx. Sampling Frequency | Acquisition Length Structure |
|---|---|---|---|---|
| 1 | 1–11, 61–63, 67–69 | 3 s, 180 s, 90 s, 900 s | 0.333 Hz, 0.0056 Hz, 0.011Hz, 0.0011 Hz | Multiple length groups |
| 3 | 53–56, 70–76 | Mostly 30 s (Ch 74 higher-rate segment) | ≈0.033 Hz | Mostly uniform with one outlier |
| 5 | 41–46 | 30 s | ≈0.033 Hz | Fully uniform |
| 6 | 12–40, 47–52, 57–60, 64–66 | 30 s, 90 s, 180 s | 0.033 Hz, 0.011 Hz, 0.0056 Hz | Strongly multi-regime |

*Table 3.9 — Dominant sampling regimes across Mission-1 subsystems.*

The comparison indicates that temporal granularity varies substantially across subsystems, ranging from fully aligned acquisition regimes in Subsystem 5 to strongly heterogeneous multi-rate logging configurations in Subsystems 1 and 6.

From a modelling perspective, differences in temporal granularity influence both frequency-domain interpretation and multivariate learning strategies. Subsystems with uniform sampling generally permit more straightforward joint modelling, whereas multi-regime subsystems require careful handling of temporal alignment and bandwidth disparities. Consequently, sampling structure emerges as a critical axis of structural differentiation within Mission 1.

### 3.4.3 Correlation Archetypes Across Subsystems

Inter-channel correlation patterns constitute another major dimension of structural differentiation across subsystems. While time-domain behaviour reflects individual signal dynamics, correlation analysis reveals the degree of collective coupling and redundancy within each subsystem.

Subsystem 5 exhibits the strongest internal coherence among the analysed subsystems. Its correlation matrix exhibits a dense block of high positive coefficients, indicating that most channels evolve in a largely coordinated manner. This pattern reflects a tightly coupled dynamical process in which channels likely measure related physical quantities or dependent system variables.

Subsystem 3 displays a predominantly cohesive correlation structure with a minor structural deviation. The majority of channels form a moderately to strongly correlated block, while Channel 74 exhibits comparatively weaker or slightly differentiated behaviour due to its distinct sampling regime. The overall structure remains largely integrated, but with identifiable asymmetry.

Subsystem 1 presents a more fragmented correlation landscape. Active target channels show moderate internal dependencies, whereas quasi-static or low-variance non-target channels contribute weak or less structured correlations. As a result, the correlation matrix does not form a single dominant block but instead reflects mixed dynamical roles within the subsystem.

Subsystem 6 exhibits a multi-cluster correlation pattern. Rather than forming one cohesive block, channels appear to organise into partially independent clusters, each characterised by stronger intra-group correlation and weaker inter-group coupling. This block-wise structure aligns with previously identified acquisition regimes and sampling groups, indicating that structural segmentation extends beyond temporal granularity into dynamical interaction patterns.

These correlation archetypes define a continuum from tightly coupled systems (Subsystem 5) to loosely structured multi-cluster environments (Subsystem 6). Such structural differences are particularly relevant for multivariate anomaly detection models, as modelling strategies exploiting inter-channel redundancy may perform differently depending on the degree of internal correlation coherence.

### 3.4.4 Spectral Families and Periodicity Patterns

Spectral analysis reveals distinct frequency-domain patterns across subsystems. Subsystem 5 exhibits highly coherent spectral envelopes, with channels sharing dominant low-frequency components and overlapping periodic structures. This consistency suggests that the channels are influenced by closely related subsystem dynamics.

Subsystem 3 shows largely similar low-frequency behaviour among most channels, with Channel 74 extending into a slightly broader observable bandwidth due to its higher sampling rate. The subsystem remains spectrally cohesive but includes a structured frequency outlier.

Subsystem 1 presents mixed spectral characteristics. Target channels display identifiable low-frequency activity, while several non-target channels contain very limited spectral energy beyond baseline levels. This divergence reinforces its mixed-dynamics classification.

Subsystem 6 demonstrates multiple spectral families aligned with its acquisition regimes. Channels within the same acquisition group share comparable spectral bandwidths, while cross-group spectral envelopes differ significantly.

Overall, spectral behaviour broadly reflects the structural heterogeneity observed across subsystems: coherent subsystems exhibit aligned periodicity, whereas heterogeneous subsystems contain multiple frequency families.

### 3.4.5 Digital vs Continuous Telemetry Signals

A final structural distinction across subsystems concerns the nature of the telemetry signals themselves: continuous measurement signals versus discrete or quasi-digital telemetry.

Continuous signals, typically associated with continuously varying physical measurements, exhibit smooth temporal evolution, finite variance, and identifiable low-frequency spectral components. These signals appear to dominate in Subsystems 3 and 5, where gradual oscillatory behaviour and coherent spectral envelopes are generally observed.

In contrast, several channels—particularly within Subsystem 1—display near-constant values or abrupt level transitions. Such behaviour is often characteristic of digital status indicators, configuration flags, or rarely changing control variables. These signals often exhibit very low variance, compressed value ranges, and limited meaningful spectral structure.

Subsystem 6 contains a mixture of both behaviours, but with clearer grouping that aligns with the acquisition regimes identified earlier. Some channel clusters demonstrate continuous dynamics, while others show more discrete or sparsely varying characteristics.

This distinction between digital and continuous telemetry contributes to structural heterogeneity beyond sampling or correlation differences. Continuous signals tend to align well with spectral and correlation-based modelling assumptions, whereas digital or quasi-static channels may require separate treatment or contribute limited dynamical information.

Recognising this behavioural divide provides a more nuanced perspective on subsystem composition and further clarifies the structural diversity present within Mission 1.

## 3.5 Structural Implications for Anomaly Detection Modelling

The exploratory analysis conducted in this chapter indicates that Mission 1 does not represent a statistically uniform telemetry environment. Instead, it comprises subsystems

that span a spectrum from tightly coupled, homogeneous structures to strongly heterogeneous, multi-regime configurations. These structural differences are not superficial; they shape the operational landscape in which anomaly detection models must operate.

Subsystem-level analysis revealed variability in acquisition regimes, temporal granularity, inter-channel correlation, spectral coherence, and signal type. Some subsystems exhibit consistent periodic behaviour and strong multivariate redundancy, while others contain quasi-static channels, discrete telemetry signals, or multiple internally segmented channel families. Such diversity implies that modelling assumptions valid in one subsystem may not transfer directly to another.

Importantly, the presence of homogeneous structures suggests favourable conditions for multivariate learning approaches that exploit inter-channel coherence. Conversely, heterogeneous subsystems introduce structural fragmentation that may challenge models that rely on uniform temporal dynamics or shared statistical properties. Similarly, the coexistence of continuous measurement signals and quasi-digital channels affects the interpretability and stability of residual-based anomaly scoring.

Therefore, anomaly detection performance in subsequent chapters must be interpreted in light of the structural constraints identified here. Differences in model behaviour across subsystems cannot be attributed solely to algorithmic design choices; they are inherently linked to the underlying telemetry organisation. Taken together, the analysed subsystems form a structural continuum ranging from highly coherent telemetry groups to heterogeneous multi-regime environments composed of partially independent channel families. This structural perspective provides the necessary context for Chapter 4 and Chapter 5, where forecasting-based and reconstruction-based approaches are implemented and evaluated. The results of those experiments will therefore be interpreted not only as outcomes of the algorithms themselves, but also in relation to the structural characteristics of the telemetry data analysed in this chapter.

# Chapter 4 — Implementation & Experimental Pipeline

## 4.1 Experimental Environment

All experiments in this thesis were conducted within a controlled remote computing environment to ensure reproducibility, consistency, and alignment with the ESA-ADB benchmark protocol. Given the computational demands of multivariate telemetry processing and deep learning-based anomaly detection, a structured server-based setup was adopted rather than local execution. The experimental infrastructure combines remote hardware resources, containerised execution, and a benchmark-aligned evaluation framework. This section describes the hardware platform, software stack, and containerisation strategy that form the backbone of the experimental pipeline.

### 4.1.1 Hardware Infrastructure

All experiments were executed on remote server infrastructure provided by the hosting laboratory. During the initial phase of the project, experiments were conducted on the *Sheldon* server. Subsequently, the computational environment was migrated to the *Bernadette* server following infrastructure reorganisation. Both servers provided sufficient computational resources for training forecasting-based and reconstruction-based anomaly detection models on multichannel satellite telemetry data. The migration did not alter the experimental configuration or evaluation protocol, as all experiments were executed within containerised environments. The use of remote server infrastructure ensured stable computational resources across experiments, isolation from local machine variability, centralised storage for datasets and experiment outputs, and reliable support for long-running training procedures. Secure remote access was used for model training, evaluation, and result aggregation. The controlled server-based setup allowed systematic experimentation under consistent runtime conditions.

### 4.1.2 Software Stack

The experimental pipeline was built on a Python-based ecosystem integrated with the official ESA-ADB benchmark framework. Core components included Python for model implementation and experiment control, the TimeEval framework for structured benchmarking and evaluation, Docker for containerised execution, and Git for version control and branch-based experiment management. The TimeEval framework was used to

standardise dataset handling, metric computation, and experiment configuration. This ensured that both forecasting-based and reconstruction-based models were evaluated under identical conditions. Version control was managed using a structured Git workflow, enabling controlled experimentation across branches (e.g., baseline, memory-augmented variants, and subset-based configurations).

### 4.1.3 Containerisation and Execution Setup

To guarantee reproducibility and alignment with the official ESA-ADB benchmark, experiments were executed within Docker containers provided by the ESA-ADB repository. Containerisation ensured dependency isolation, consistent runtime environments, controlled algorithm execution, and fair comparison across models. Each experiment followed the predefined ESA-ADB structure, including dataset configuration, algorithm registration, and evaluation scripts. Custom algorithm extensions (e.g., DC-VAE variants and memory-aware modifications) were integrated into this framework while preserving the benchmark execution protocol. Container-based execution also simplified resource management and prevented conflicts between algorithm dependencies. This approach ensured that all reported results are reproducible within the same benchmark environment.

## 4.2 Integration with the ESA-ADB Benchmark Framework

The experimental pipeline adopted in this thesis is built upon the official ESA-ADB benchmark framework provided by KPLabs, aligned with the benchmark protocol described in [1]. The ESA-ADB benchmark provides a structured and reproducible environment for evaluating anomaly detection methods on real satellite telemetry data, including predefined dataset organisation, channel roles, and event-based evaluation metrics. Integrating the experiments within this framework ensured methodological consistency with prior ESA-ADB evaluations and preserved comparability with existing benchmark results [1]. This alignment allows the reported results to be interpreted within the broader context of satellite telemetry anomaly detection research.

### 4.2.1 Official ESA-ADB Repository and Benchmark Alignment

All experiments were implemented using the official ESA-ADB repository provided by KPLabs, aligned with the benchmark specification described in [1]. The repository provides preprocessed mission datasets, structured dataset metadata, algorithm registration interfaces, container-based execution templates, and standardised evaluation scripts within a unified experimental framework. This infrastructure enforces strict separation between training and testing data and implements a consistent metric computation pipeline. As a result, forecasting-based and reconstruction-based approaches are evaluated under

identical benchmark conditions. Using the official repository ensures reproducibility of experiments, fair comparison across algorithms, alignment with benchmark-defined evaluation criteria, and compatibility with event-based scoring. No modifications were made to the core evaluation logic of the benchmark; all adaptations were confined to algorithm configuration and controlled experimental design parameters.

### 4.2.2 Dataset Configuration and Mission Selection

All experiments in this thesis were conducted on ESA Mission 1. Mission 1 contains multichannel telemetry data organised into hierarchical subsystems, including both target and non-target channels. Target channels correspond to signals for which anomaly labels are provided and are therefore used for quantitative evaluation. Non-target channels are included to model contextual subsystem dynamics but are excluded from metric computation. The benchmark structure preserves subsystem grouping, which enables both subsystem-level analysis and full-channel evaluation. This structural organisation plays a central role in the experimental design adopted in later chapters.

### 4.2.3 Subsystem and Channel Selection Strategy

To analyse model behaviour under controlled complexity, experiments were conducted at two structural levels: subsystem-based configurations and full target-set configurations.

The subsystem-level experiments allow detailed investigation of model stability, convergence behaviour, and channel-wise heterogeneity under reduced multivariate complexity. In contrast, the full-channel configuration assesses scalability and robustness under increased subsystem diversity. This dual-level evaluation strategy enables controlled comparison between local subsystem performance and mission-level generalisation.

## 4.3 TimeEval-Based Evaluation Setup

To ensure consistent and reproducible evaluation across forecasting-based and reconstruction-based models, all experiments were conducted within the TimeEval benchmarking framework integrated into ESA-ADB. TimeEval provides a structured interface for dataset handling, algorithm execution, and metric computation under controlled experimental conditions. This setup ensures strict separation between training and testing data, standardised metric computation, and fair comparison between models with different anomaly detection paradigms.

### 4.3.1 Training–Testing Protocol

All experiments follow the predefined training–testing splits provided by the ESA-ADB benchmark [1]. The dataset is partitioned into temporally ordered segments to prevent leakage of future information into the training phase. For forecasting-based models (e.g., Telemanom [3]), the model is trained to predict future values based on historical sequences. Anomaly scores are computed from the residual between predicted and observed values, following the residual-based scoring principle described in [3]. For reconstruction-based models (e.g., DC-VAE), the model is trained to reconstruct normal behaviour. Anomaly scores are derived from reconstruction error, consistent with autoencoder-based anomaly detection approaches described in [10]. Training is conducted exclusively on the designated training portion of the dataset, while evaluation metrics are computed only on the testing partition. No anomaly labels are used during model training, ensuring a semi-supervised anomaly detection setting as defined in [2]. This protocol prevents temporal leakage, ensures consistent comparison across models, and preserves alignment with the ESA-ADB benchmark evaluation rules.

### 4.3.2 Event-Based and Channel-Aware Metrics

Satellite telemetry anomaly detection differs from conventional point-wise anomaly classification. In operational settings, detecting complete anomaly events with limited delay is more critical than identifying isolated anomalous points. The ESA-ADB benchmark therefore adopts event-based and channel-aware metrics, as formally defined in [1]. The primary evaluation metrics used in this thesis include AFF Precision, AFF Recall, AFF $F_{0.5}$ score, Event-Wise $F_{0.5}$, and ADTQC (Anomaly Detection Temporal Quality Criterion), all formally defined within the ESA-ADB benchmark specification [1]. The AFF (Anomaly F-score Framework) metrics evaluate detection performance while accounting for event duration and overlap rather than simple point-wise matches. This reflects operational requirements in satellite monitoring systems [1]. The ADTQC metric quantifies temporal alignment between detected and ground-truth anomaly intervals, penalising delayed or fragmented detections. This metric is particularly relevant for time-sensitive spacecraft operations [1]. Using these metrics ensures that evaluation reflects temporal coherence, event completeness, detection delay, and channel-specific performance. This approach avoids misleading conclusions that could arise from purely point-wise accuracy metrics.

### 4.3.3 Aggregation Strategy and Mission-Level Evaluation

Performance in ESA telemetry is evaluated at multiple hierarchical levels. First, metrics are computed at the individual channel level for all target channels. This allows detailed analysis of channel-specific behaviour, heterogeneity, and subsystem sensitivity. Second, channel-level metrics are aggregated to produce subsystem-level and mission-level performance

indicators. This aggregation enables comparison between subsystem-restricted experiments, full target-set configurations, and forecasting versus reconstruction paradigms. The dual-level aggregation strategy aligns with the hierarchical telemetry structure described in [1], where subsystems represent coherent physical or functional units. By analysing both channel dispersion and aggregated performance, the evaluation framework captures local model stability, cross-channel heterogeneity, and scalability to higher multivariate complexity. This hierarchical evaluation design plays a central role in interpreting the experimental results presented in Chapter 5.

### 4.3.4 Integration with the TimeEval Framework

All experiments were orchestrated using the TimeEval framework provided within the ESA-ADB benchmark repository. TimeEval served as the central experiment management system, handling dataset selection, parameter configuration, algorithm execution, and metric computation. Datasets were accessed through the benchmark-defined *Dataset Manager*, ensuring consistent collection selection and fixed train–test partitions. Algorithm configurations were defined via structured parameter grids, allowing controlled variation of batch size, prediction window length, and training epochs without altering evaluation logic. Performance metrics, including ESAScores, Channel-Aware F-score, and ADTQC, were computed automatically through the TimeEval evaluation pipeline. This integration eliminated manual metric calculation and ensured alignment with benchmark-defined scoring protocols. The use of TimeEval guaranteed that all experimental comparisons were conducted under a consistent and reproducible evaluation framework.

## 4.4 Experimental Design and Control Variables

The experimental design adopted in this thesis follows a controlled comparison philosophy. Rather than performing unrestricted hyperparameter search, specific variables were systematically varied while keeping all other parameters fixed. A full hyperparameter grid search was intentionally avoided in order to preserve interpretability of experimental trends and to prevent overfitting to benchmark-specific configurations. The objective was not to optimise for maximal performance at any cost, but to analyse model behaviour under controlled and comparable conditions. This approach enables clear attribution of performance differences to individual experimental factors. The primary goal of this design is not only to measure performance, but also to analyse model behaviour under controlled structural and optimisation changes.

### 4.4.1 Forecasting vs Reconstruction Paradigm Setup

Two fundamentally different anomaly detection paradigms were evaluated: forecasting-based detection (Telemanom [3]) and reconstruction-based detection (DC-VAE, based on variational autoencoder principles [6]). Forecasting-based models learn temporal predictability by estimating future values from historical context. Anomalies are detected when prediction residuals exceed a dynamic threshold, following the methodology introduced in [3]. Reconstruction-based models instead learn a compressed latent representation of normal behaviour. Anomalies are identified when reconstruction error deviates significantly from the learned distribution, consistent with autoencoder-based anomaly detection strategies described in [10] and [6]. By evaluating both paradigms under identical benchmark conditions, the experimental setup enables direct comparison between predictive consistency and reconstruction fidelity in subsystem-structured telemetry.

### 4.4.2 Hyperparameter Configuration and Batch Variations

To analyse optimisation behaviour and model sensitivity, selected hyperparameters were systematically varied while keeping architectural structure fixed. The controlled variables included the number of training epochs, batch size, prediction window size for forecasting models, and subsystem versus full-channel configuration. For DC-VAE experiments, architectural components (encoder, decoder, latent dimension) remained unchanged. Only optimisation-related parameters such as batch size and epoch count were modified to study convergence stability and gradient variance effects. For Telemanom, prediction window size was varied to examine the influence of forecasting horizon on precision–recall balance and temporal stability, consistent with forecasting-based anomaly detection principles. This design isolates optimisation-related effects from architectural factors, allowing interpretation of performance trends independent of structural model changes.

### 4.4.3 Subsystem-Level vs Full-Channel Experiments

Satellite telemetry is organised into hierarchical subsystems with heterogeneous temporal and statistical properties [1]. To investigate scalability and robustness, experiments were conducted at two structural levels: 1. Subsystem-restricted configurations 2. Full target-channel configurations

Subsystem-level experiments enable controlled analysis under reduced multivariate dimensionality, supporting investigation of convergence behaviour, channel-wise dispersion, and local structural heterogeneity. Full-channel experiments introduce increased subsystem diversity and inter-channel variability, allowing assessment of scalability, sensitivity to heterogeneous dynamics, and generalisation beyond relatively homogeneous subsystems.

By comparing subsystem-restricted and full-channel results, the experimental design isolates the effect of multivariate complexity on anomaly detection performance.

## 4.5 Reproducibility and Engineering Considerations

In addition to methodological rigour, particular attention was given to engineering stability and reproducibility throughout the experimental process. The ESA-ADB benchmark framework relies on containerised execution, remote computational infrastructure, and long-running deep learning experiments. Under such conditions, maintaining consistent and deterministic behaviour across runs becomes as important as algorithmic performance itself.

### 4.5.1 Docker and Execution Challenges

All experiments were conducted within Docker containers configured according to the official ESA-ADB benchmark structure. Containerisation ensured that dependency versions, library configurations, and runtime environments remained fixed across experimental iterations. This isolation proved essential when migrating experiments between servers, as it prevented environment-related inconsistencies from affecting results. During early experimentation phases, several execution failures were traced to directory path misconfigurations and container mounting inconsistencies. These issues were resolved by aligning the project structure strictly with the benchmark-prescribed hierarchy and validating execution entry points prior to large-scale runs. Once stabilised, the container-based workflow ensured identical runtime conditions for both Telemanom and DC-VAE experiments. Within the ESA-ADB benchmark architecture, TimeEval orchestrates algorithm execution by launching each method inside its corresponding Docker container. Consequently, Docker was not used as an independent execution tool, but as an integral component of the TimeEval evaluation pipeline. This design ensured consistent runtime isolation while preserving benchmark-defined experiment control.

### 4.5.2 Resource Constraints and Runtime Management

Training deep learning models on multichannel satellite telemetry data is computationally intensive and sensitive to resource constraints. Batch size and epoch configurations were therefore selected not only from an optimisation perspective but also with regard to memory stability and runtime feasibility. Larger batch sizes increased GPU memory pressure, whereas smaller batch sizes influenced convergence smoothness and gradient stability. Consequently, batch-size variation served both as an experimental variable and as a mechanism to maintain stable execution. Long-duration experiments were executed on remote servers with continuous monitoring to prevent incomplete training cycles and corrupted output artefacts.

### 4.5.3 Debugging, Stability, and Determinism

During initial Telemanom experiments, certain runs produced invalid anomaly score outputs containing undefined (NaN) values. Investigation revealed that these instabilities originated from training dynamics and residual propagation under specific parameter settings. To preserve result integrity, anomaly score files were systematically inspected, incomplete runs were discarded, and stable configurations were identified through controlled re-execution. Only validated experimental outputs were included in the quantitative analyses presented in Chapter 5. Dataset partitions and evaluation procedures were kept fixed throughout the study to ensure deterministic behaviour and consistent comparison across configurations.

### 4.5.4 Version Control and Experiment Traceability

All experimental configurations were managed under structured version control. Separate branches were maintained for baseline models, memory-augmented variants, subsystem-restricted studies, and full-channel evaluations. Each experimental run generated timestamped result directories, enabling traceability between parameter settings and reported performance metrics. This structured workflow ensured that experimental findings were reproducible and that performance variations could be clearly attributed to controlled configuration changes rather than unintended environmental effects.

## 4.6 Summary of Experimental Methodology

This chapter has defined the experimental framework used to evaluate forecasting- and reconstruction-based anomaly detection methods on ESA satellite telemetry data. The experimental environment was formally specified in terms of hardware infrastructure, software stack, containerised execution, and integration with the TimeEval benchmark framework. A consistent training–testing protocol was adopted following the benchmark-defined data partitions. Performance was assessed using event-based and channel-aware metrics to ensure mission-relevant evaluation beyond point-wise accuracy. Experimental variables were carefully controlled to isolate the effects of optimisation parameters, forecasting horizon, subsystem selection, and architectural extensions such as memory augmentation. Engineering stability and reproducibility were treated as core methodological constraints. Containerised execution, controlled parameter variation, and structured version management ensured that reported performance differences arise from intentional experimental modifications rather than environmental inconsistencies. With the experimental design formally established, the next chapter presents the algorithmic implementations and quantitative results obtained under this framework.

# Chapter 5 – Experimental Results

## 5.1 Telemanom Evaluation

### 5.1.1 Baseline: Vanilla Telemanom on ESA Mission 1

As a first step in the evaluation phase, the original Telemanom implementation was executed on the ESA Mission 1 dataset in order to establish a baseline reference. The experiment was conducted in a semi-supervised setting using a fixed validation split and a reduced training duration (1 epoch) to verify training stability.

However, during execution, the model exhibited numerical instability, with the training loss becoming *NaN* from the initial batches onward. Inspection of the logs revealed that this behaviour originated from the internal normalisation step of the original Telemanom implementation, where division by near-zero standard deviation values within sliding windows produced undefined values. Although the raw channel statistics did not contain missing values or zero variance globally, the window-based preprocessing resulted in unstable training dynamics.

This outcome indicates that the vanilla Telemanom implementation does not directly generalise to ESA multivariate telemetry without adaptation. In contrast to the vanilla implementation, the Telemanom-ESA variant introduces several dataset-specific extensions, including configurable LSTM layer dimensions, explicit multichannel input and target handling, a minimum error threshold parameter, and optional dynamic score thresholding. These modifications are designed to improve numerical robustness and adapt the model to ESA's multivariate telemetry structure. Consequently, further experiments were conducted using the customised Telemanom-ESA variant, which incorporates dataset-specific adjustments to improve numerical robustness and multichannel handling.

### 5.1.2 Telemanom-ESA: Subset-Based Evaluation (Channels 41–46)

Following the numerical instability observed in the vanilla implementation, the customised Telemanom-ESA variant was evaluated on ESA Mission 1. The first set of experiments focused on a reduced subset of target channels (Channels 41–46) in order to analyse model behaviour under a controlled multichannel configuration. Restricting the evaluation to a limited group of channels allows a clearer interpretation of convergence behaviour, detection stability, and channel-wise performance before extending the analysis to the full Mission 1 target set.

### *5.1.2.1 Global Results (Epoch 1 vs 10)*

The initial experiment was conducted with a single training epoch in order to verify convergence behaviour and assess the baseline predictive capability of the model. The global performance metrics for both training durations (1 and 10 epochs) are jointly reported in Table 5.1, enabling a direct quantitative comparison between minimal and extended training. The visual comparison of global performance metrics between 1 and 10 training epochs is illustrated in Figure 5.1, indicating limited sensitivity to extended training within the controlled subset configuration.

| Algorithm | Dataset | Epoch | AFF F0.5 | AFF Precision | AFF Recall | EW F0.5 | ADTQC |
|---|---|---|---|---|---|---|---|
| Telemanom-ESA | 3_months | 1 | 0.709987 | 0.733232 | 0.630085 | 0.137598 | 0.19082 |
| Telemanom-ESA | 3_months | 10 | 0.693571 | 0.728488 | 0.581988 | 0.083444 | 0.19082 |
| Time (min) | Train | 1, 10 | 1.004356, 6.397221 | | | | |
| Time (min) | Execute | 1, 10 | 74.202672, 66.203848 | | | | |

*Table 5.1 — Global results for Telemanom-ESA on ESA Mission 1 (Subset 41–46), Epoch 1 vs Epoch 10.*

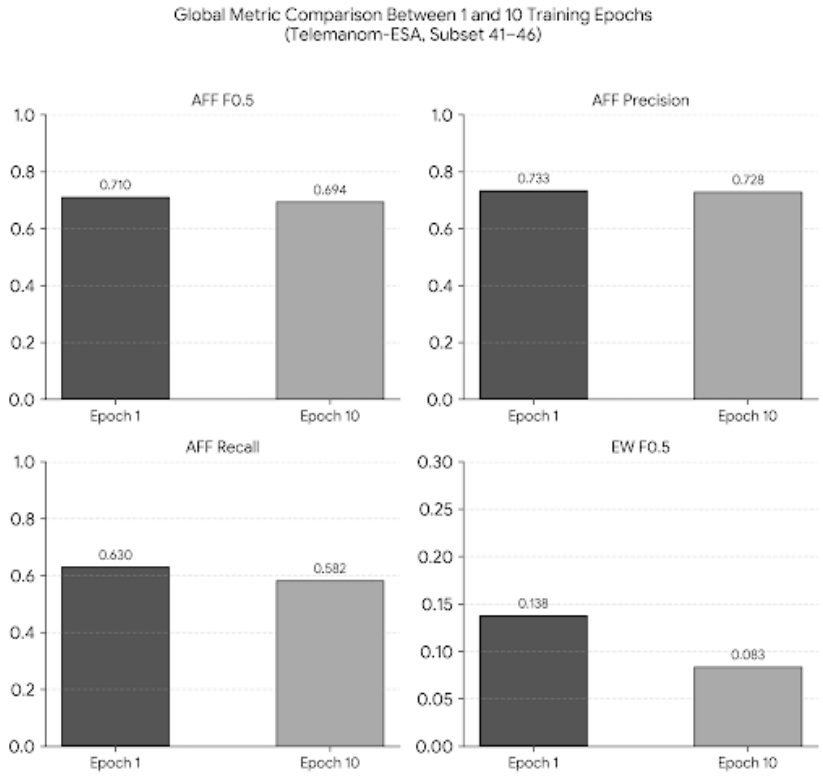


*Figure 5.1 — Global performance comparison between 1 and 10 training epochs for Telemanom-ESA on ESA Mission 1 (Subset 41–46).*

### *5.1.2.2 Channel-Wise Analysis*

Channel-wise performance for the 1-epoch configuration is reported in Table 5.2, enabling a more granular analysis of detection behaviour across individual telemetry channels. The results indicate strong anomaly detection capability at the channel level, with AFF F_{0.5} scores consistently above 0.79 and reaching up to approximately 0.89. Event-wise recall values remain high across channels, suggesting that the predictive formulation successfully captures temporal deviations preceding anomaly segments. The Global ADTQC metric further confirms that detection delays remain limited, indicating timely alarm generation.

Although the global comparison in Table 5.1 already shows that extending training from 1 to 10 epochs does not produce substantial structural improvements, a finer-grained analysis is required to determine whether local channel behaviour changes. The channel-wise results for the 10-epoch configuration are therefore reported in Table 5.3, following the 1-epoch breakdown presented above.

| Channel – Subsystem 5 | Precision_AFF | Recall_AFF | F0.5_Score_AFF | EW_Recall |
|---|---|---|---|---|
| **Channel 41** | 0.885621 | 0.786228 | 0.863781 | 0.862069 |
| **Channel 42** | 0.876927 | 0.819832 | 0.864880 | 0.892857 |
| **Channel 43** | 0.876392 | 0.793835 | 0.858535 | 0.928571 |
| **Channel 44** | 0.897498 | 0.843233 | 0.886094 | 0.892857 |
| **Channel 45** | 0.810128 | 0.733632 | 0.793579 | 0.892857 |
| **Channel 46** | 0.857661 | 0.791152 | 0.843479 | 0.925926 |

*Table 5.2 — Channel-wise results for Telemanom-ESA on ESA Mission 1 (Subset 41–46), Epoch 1.*

| Channel – Subsystem 5 (Epoch 10) | Precision_AFF | Recall_AFF | F0.5_Score_AFF | EW_Recall |
|---|---|---|---|---|
| **Channel 41** | 0.829456 | 0.689607 | 0.797126 | 0.758621 |
| **Channel 42** | 0.852715 | 0.695973 | 0.815962 | 0.857143 |
| **Channel 43** | 0.812814 | 0.719293 | 0.792214 | 0.928571 |
| **Channel 44** | 0.834670 | 0.737255 | 0.813181 | 0.857143 |
| **Channel 45** | 0.868142 | 0.741107 | 0.839367 | 0.857143 |
| **Channel 46** | 0.769529 | 0.637489 | 0.738920 | 0.888889 |

*Table 5.3 — Channel-wise results for Telemanom-ESA on ESA Mission 1 (Subset 41–46), Epoch 10.*

### *5.1.2.3 Training Duration Interpretation*

A comparative analysis between the two configurations reveals that increasing the number of epochs does not lead to substantial structural performance improvements. As shown in Table 5.1, global AFF F_0.5, precision, recall, and ADTQC metrics remain within a narrow

range across the two training durations. At the channel level (Tables 5.2 and 5.3), minor fluctuations in precision and recall are observed, yet these variations do not translate into a systematic improvement trend. In several channels, slight recall gains are counterbalanced by moderate precision variations, resulting in marginal net changes in F-score. Importantly, no consistent upward pattern emerges when extending training from 1 to 10 epochs.

Although this behaviour may suggest that the model reaches a relatively stable operating regime early in training, the present analysis does not formally establish convergence. Rather, it indicates that, under the current configuration and subset selection, additional optimisation does not materially alter the predictive characteristics of the model.

From an experimental design perspective, this observation reduces the necessity of extensive hyperparameter exploration with respect to training duration in this controlled subset scenario. Since increasing the number of epochs does not produce systematic performance gains, subsequent analysis shifts towards parameters that are more intrinsically related to the predictive nature of the model. In particular, the prediction horizon constitutes a structurally meaningful factor, as it directly influences temporal anticipation capability and alarm stability.

Overall, the subset-based evaluation indicates that Telemanom-ESA provides stable and competitive predictive anomaly detection performance on ESA Mission 1 under the considered configuration. However, the limited sensitivity to training duration within a restricted channel setup raises an important question regarding generalisation. It remains to be examined whether comparable behaviour is preserved when the model is exposed to the full set of target channels. Accordingly, the next section extends the evaluation to the complete Mission 1 target configuration in order to assess scalability and robustness under increased multivariate complexity.

### 5.1.3 Full Target Set Evaluation (Generalisation Study)

The evaluation was subsequently extended from the controlled subset configuration (Channels 41–46) to the complete Mission 1 target set in order to assess the generalisation capability of Telemanom-ESA under increased multivariate complexity. The global performance metrics for the full configuration (Epoch 1) and for the subsystem 5 are reported in Table 5.4.

### 5.1.3.1 Subset vs Full Comparison

| Configuration | Dataset | AFF F0.5 | AFF Precision | AFF Recall | EW F0.5 | Global ADTQC |
|---|---|---|---|---|---|---|
| Subset 41–46 | 3_months | 0.709987 | 0.733232 | 0.630085 | 0.137598 | 0.190825 |
| Full Target Set (58) | 3_months | 0.474172 | 0.514701 | 0.360596 | 0.007288 | 0.593230 |
| Time (min) | Train | Subset, Full | 1.004356, 1.104479 | | | |
| Time (min) | Execute | Subset, Full | 74.202672, 224.240539 | | | |

*Table 5.4 — Global comparison between Subset 41–46 and the Full Target Set (Epoch 1).*

A direct comparison with the subset-based results reveals a clear performance degradation. While the subset configuration achieved an Anomaly AFF F_{0.5} score of approximately 0.71, the full target configuration decreases to approximately 0.47. Precision and recall both decline, and the Global ADTQC score increases substantially (from ≈0.19 to ≈0.59), indicating a larger detection delay under the complete multichannel scenario. This behaviour suggests that the predictive model, although stable in a restricted subsystem, encounters increased difficulty when exposed to a more heterogeneous set of telemetry channels.

### 5.1.3.2 Channel Dispersion Analysis

To better characterise this degradation, channel-wise summary statistics of the Anomaly AFF F_{0.5} scores were computed for both configurations. The corresponding mean, median, and standard deviation values are reported in Table 5.5.

| Configuration | Mean F0.5 | Median F0.5 | Std F0.5 |
|---|---|---|---|
| Subset 41–46 | **0.782095** | **0.762439** | 0.073103 |
| Full Target Set (58) | 0.540032 | 0.541849 | **0.152259** |

*Table 5.5 — Summary statistics of channel-wise AFF F0.5 (Subset vs Full Target Set).*

The subset configuration exhibits a high mean F_{0.5} (≈0.78) with low dispersion (std ≈0.07), indicating consistent performance across the six channels. In contrast, the full target configuration shows a substantially lower mean (≈0.54) and a markedly higher standard deviation (≈0.19), revealing strong performance heterogeneity. This increase in dispersion indicates that the model does not degrade uniformly; rather, certain channels remain well modelled, while others experience significant performance loss.

### 5.1.3.3 Performance Distribution

To further investigate this variability, the five best-performing and five worst-performing channels (based on F_{0.5}) were identified for the full configuration. These results are summarised in Table 5.6.

| Bottom 5 Channels | F0.5 | Top 5 Channels | F0.5 |
|---|---|---|---|
| Channel 16 | 0.000000 | Channel 15 | 0.751024 |
| Channel 24 | 0.000000 | Channel 43 | 0.767458 |
| Channel 32 | 0.000000 | Channel 62 | 0.811315 |
| Channel 33 | 0.000000 | Channel 63 | 0.812284 |
| Channel 40 | 0.000000 | Channel 61 | 0.814506 |

*Table 5.6 — Top 5 and Bottom 5 channels (Full Target Set, Epoch 1).*

The presence of low-performing channels suggests structural differences across telemetry streams. These differences may arise from varying noise levels, anomaly sparsity, temporal dynamics, or subsystem-specific operational regimes. Importantly, the degradation is not universal; a subset of channels continues to achieve high F_{0.5} values, indicating that the predictive formulation remains effective when temporal patterns are sufficiently regular.

### 5.1.3.4 Structural Heterogeneity Discussion

A complementary distributional analysis was also performed by grouping channels according to performance ranges. The results are reported in Table 5.7.

| Configuration | Total Channels | F0.5 < 0.4 | 0.4 ≤ F0.5 < 0.6 | F0.5 ≥ 0.6 |
|---|---|---|---|---|
| Subset 41–46 | 6 | 0 (0.00%) | 0 (0.00%) | 6 (100.00%) |
| Full Target Set | 58 | 5 (8.62%) | 31 (53.45%) | 22 (37.93%) |

*Table 5.7 — Distribution of channel-wise AFF F0.5 ranges (Full Target Set, Epoch 1). (Counts and percentages for F0.5 < 0.4, 0.4–0.6, ≥0.6)*

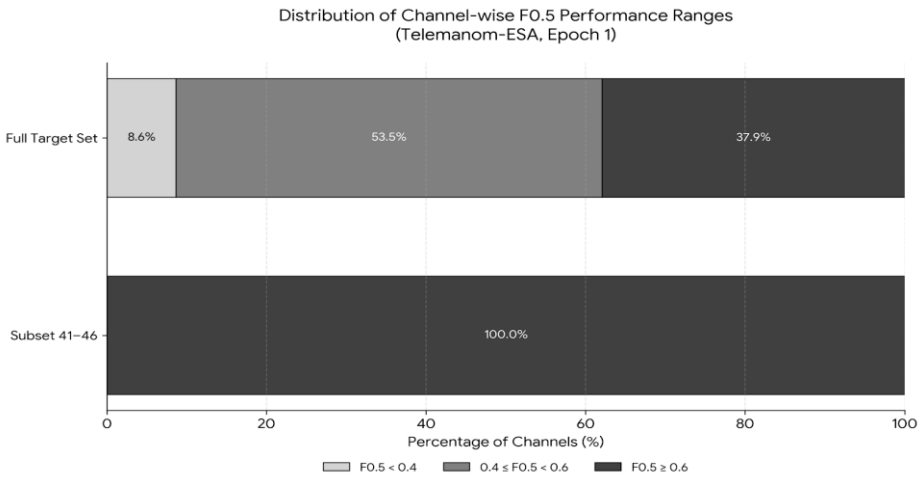


*Figure 5.2 — Distribution of channel-wise AFF F0.5 ranges for Subset 41–46 and the Full Target Set (Epoch 1).*

The visual comparison highlights the performance dispersion observed when extending from a restricted subset to the full multichannel configuration.

The distribution reveals that only approximately 38% of channels achieve $F_{0.5} \geq 0.6$, while more than half fall within the moderate range (0.4–0.6), and a small but non-negligible fraction drops below 0.4. This contrasts sharply with the subset configuration, where all channels achieved $F_{0.5} \geq 0.6$.

Together, these results indicate that the subset-based evaluation provided an optimistic but locally consistent performance estimate. When evaluated across the full target set, Telemanom-ESA demonstrates increased sensitivity to channel-specific characteristics, leading to broader performance dispersion and higher detection delay.

From an interpretative standpoint, this behaviour should not be interpreted as numerical instability, but rather as sensitivity to cross-subsystem heterogeneity. Rather, it reflects the intrinsic structural heterogeneity of the Mission 1 telemetry channels. The subset 41–46 appears to represent a relatively homogeneous subsystem, whereas the full configuration integrates multiple subsystems with differing statistical and temporal properties.

Consequently, the full target evaluation provides a more realistic estimate of operational performance and highlights the importance of analysing predictive anomaly detectors under diverse multichannel conditions.

### 5.1.4 Prediction Window Sensitivity Analysis

To further investigate the behaviour of Telemanom-ESA, a systematic sensitivity analysis was conducted on the prediction window size using the optimised configuration for Subset 41–46 (Epoch 1, layers = 80). The objective of this study was to assess whether short-term

forecasting horizons (e.g., 5–10) provide measurable advantages over moderate horizons (20–30), and to evaluate the stability of the model at channel level.

#### *5.1.4.1 Global Metrics*

Table 5.8 reports the global metrics for prediction windows ranging from 5 to 30. In parallel, Table 5.9 summarises the channel-wise distribution of the AFF F0.5 scores in terms of mean, median, standard deviation, minimum, and maximum values across the six channels.

| prediction_window_size | AFF_F0.5 | AFF_precision | AFF_recall | EW_F0.5 | ADTQC |
|---|---|---|---|---|---|
| 5 | 0.765498 | 0.785411 | **0.695016** | 0.138229 | **0.190825** |
| 10 | 0.709987 | 0.733232 | 0.630085 | 0.137598 | **0.190825** |
| 20 | **0.768192** | **0.800828** | 0.660522 | 0.137636 | 0.219822 |
| 25 | 0.760027 | 0.783240 | 0.679477 | **0.167975** | 0.219822 |
| 30 | 0.720349 | 0.734642 | 0.668335 | 0.131451 | **0.190825** |

*Table 5.8 — Global metrics for different prediction window sizes (Subset 41–46).*

| Window | Mean | Median | Std | Min | Max |
|---|---|---|---|---|---|
| 5 | **0.8696** | **0.8760** | 0.0361 | 0.8008 | 0.9089 |
| 10 | 0.8517 | 0.8612 | 0.0316 | 0.7936 | 0.8861 |
| 20 | 0.8533 | 0.8513 | **0.0280** | 0.8135 | 0.8849 |
| 25 | 0.8536 | 0.8773 | **0.0472** | **0.7765** | 0.8900 |
| 30 | 0.8483 | 0.8622 | 0.0332 | 0.7878 | 0.8753 |

*Table 5.9 — Channel-wise distribution statistics of AFF F0.5 across prediction window sizes.*

At global level, the variation across window sizes is moderate rather than dramatic. The highest AFF F0.5 score is obtained at window size 20 (0.768), closely followed by window sizes 5 (0.765) and 25 (0.760). Window 10 produces a noticeable drop (0.709), while window 30 slightly underperforms compared to the others (0.720).

Precision peaks at window 20 (0.801), whereas recall reaches its maximum at window 5 (0.695). This indicates a mild precision–recall trade-off as the forecasting horizon increases. Importantly, the ADTQC metric remains identical for window sizes 5, 10, and 30, suggesting comparable temporal alignment of detections despite differences in classification balance.

The Event-Wise F0.5 metric shows limited sensitivity to window size, with window 25 yielding the highest value (0.168). However, the differences remain small, reinforcing the observation that prediction window size does not induce structural performance shifts within this range.

Overall, the global analysis suggests that moderate window sizes (20–25) offer a slightly more balanced behaviour, but none of the configurations exhibits a clear dominance.

#### *5.1.4.2 Channel_ Level Stability*

The channel-wise statistics provide deeper insight into model robustness.

Window size 5 achieves the highest mean (0.8696) and median (0.8760) AFF F0.5 values, indicating strong overall consistency across channels. Window 20 exhibits the lowest standard deviation (0.0280), implying the most uniform behaviour across the subsystem. In contrast, window 25 presents the highest variance (0.0472) and the lowest minimum value (0.7765), suggesting greater sensitivity to channel-specific characteristics.

Window 30 shows a slight degradation in both mean and median compared to smaller windows, although without instability. Window 10 performs consistently but remains below the performance of window 5 and 20.

The key observation is that the ranking of channels remains relatively stable across window sizes, and no window configuration introduces structural collapse in any specific channel. This reinforces the conclusion that the subsystem is intrinsically well-structured and that Telemanom-ESA operates within a stable regime for prediction windows between 5 and 30.

#### *5.1.4.3 Design Implications*

The sensitivity study indicates that the prediction window size influences the precision–recall trade-off but does not substantially alter the overall detection capability within Subset 41–46. The relatively limited variance observed across configurations suggests that, within the examined range, detection performance may be more strongly associated with the intrinsic structure of the channels and the model's internal thresholding dynamics than with the exact forecasting horizon. However, this observation should be interpreted as range-specific rather than universally generalisable.

From a design perspective, window size 20 represents a reasonable compromise, providing the highest global AFF F0.5 and the lowest channel-wise variance. Window size 5, however, achieves the highest mean and median channel performance and may be preferred when prioritising subsystem-level robustness.

These findings confirm that moderate forecasting horizons are sufficient for stable operation and that aggressive tuning of the prediction window does not yield disproportionate performance gains.

While Telemanom provides a forecasting-based perspective on anomaly detection, it remains fundamentally dependent on prediction consistency across channels. To complement this approach, the following section investigates a reconstruction-based model, namely DC-VAE, enabling a different interpretation of anomaly behaviour in ESA telemetry.

## 5.2 DC-VAE Baseline Evaluation

### 5.2.1 Experimental Setup

In order to analyse the influence of the training batch size on model behaviour, a controlled comparison was performed using two configurations: a batch size of 16 and a batch size of 64. This comparison allows the investigation of how optimisation dynamics affect anomaly detection performance in the DC-VAE model.

The objective of this analysis is to examine the influence of batch size on the behaviour of the anomaly detection model from multiple perspectives. In particular, the study evaluates how different batch configurations affect overall detection performance at the global level, the stability and consistency of metrics across individual channels, and the model's ability to reconstruct temporally coherent anomaly events as reflected in event-wise recall. By analysing these dimensions jointly, the impact of batch size can be interpreted not only in terms of aggregated performance scores, but also in relation to channel-specific robustness and event-level detection quality.

### 5.2.2 Global Performance (Subset 41–46)

The global comparison is first presented in Table 5.10, followed by detailed channel-wise results in Tables 5.11 and 5.12.

| Configuration | Batch Size | Epoch | AFF_F0.5 | AFF_Precision | AFF_Recall | EW_F0.5 | ADTQC |
|---|---|---|---|---|---|---|---|
| DC-VAE | 16 | 1 | 0.43353 | 0.52076 | 0.25960 | 0.000014 | 0.64144 |
| DC-VAE | 64 | 1 | 0.46026 | 0.51760 | 0.31894 | 0.000389 | 0.92023 |
| Train_Time (min) | 16, 64 | | 12.85, 6.95 | | | | |
| Execute_Time (min) | 16, 64 | | 218.96, 58.99 | | | | |

*Table 5.10 — Global performance comparison of DC-VAE for Batch Size 16 and 64 (Subset 41–46, Epoch 1).*

As shown in Table 5.10, increasing the batch size from 16 to 64 is associated with changes across several evaluation metrics. At the global level, **Anomaly_AFF_F0.50** increases from 0.4335 to 0.4603, reflecting a slightly more favourable precision–recall balance under the larger batch configuration. This variation is accompanied by an increase in recall, from 0.2596 to 0.3189, suggesting that the model detects a higher proportion of anomalous points when trained with a larger batch size.

The **Global_ADTQC_Anomaly_Total** metric also increases, from 0.6414 to 0.9202, indicating a difference in the temporal alignment between detected and reference anomaly

intervals. In addition, the execution time decreases when using Batch 64, meaning that the larger batch configuration does not introduce additional computational overhead in this setting.

### 5.2.3 Channel-Level Behaviour

Increasing the batch size reduces gradient variance during training, resulting in smoother parameter updates and a more stable global optimisation process. While this often improves the consistency of aggregated global metrics, it does not necessarily guarantee better performance at the individual channel level.

| Channel | AFF_Precision | AFF_Recall | AFF_F0.5 | EW_Recall |
|---|---|---|---|---|
| 41 | 0.637410 | 0.394807 | 0.567648 | 0.275862 |
| 42 | 0.551710 | 0.296749 | 0.470808 | 0.285714 |
| 43 | 0.521360 | 0.242423 | 0.423827 | 0.464286 |
| 44 | 0.702425 | 0.583925 | 0.675028 | 0.250000 |
| 45 | 0.622820 | 0.490997 | 0.591081 | 0.285714 |
| 46 | 0.552730 | 0.261216 | 0.451873 | 0.296296 |

*Table 5.11 — Channel-wise metrics for DC-VAE (Batch Size 16, Subset 41–46, Epoch 1).*

| Channel | AFF_Precision | AFF_Recall | AFF_F0.5 | EW_Recall |
|---|---|---|---|---|
| 41 | 0.571136 | 0.370067 | 0.515156 | 0.068966 |
| 42 | 0.501362 | 0.312067 | 0.447119 | 0.000000 |
| 43 | 0.497787 | 0.188715 | 0.374966 | 0.000000 |
| 44 | 0.510485 | 0.333880 | 0.461648 | 0.035714 |
| 45 | 0.551745 | 0.489073 | 0.537957 | 0.035714 |
| 46 | 0.548997 | 0.447602 | 0.525203 | 0.000000 |

*Table 5.12 — Channel-wise metrics for DC-VAE (Batch Size 64, Subset 41–46, Epoch 1).*

A more detailed inspection at channel level reveals a more nuanced behaviour. Under the Batch-16 configuration, performance varies across channels, with Channel 44 achieving the highest AFF F0.5 score, while Channel 45 also demonstrates relatively strong detection capability. Channel 43, however, shows weaker AFF recall despite maintaining comparatively higher event-wise recall. When the batch size increases to 64, the performance distribution across channels becomes more uniform, and channels such as 45 and 46 exhibit a more balanced precision–recall behaviour. At the same time, event-wise recall decreases for several channels, suggesting that larger batch sizes promote more stable but less reactive detection dynamics.

This suggests that smaller batches may introduce stronger stochastic updates, which can occasionally increase event-wise recall but reduce global stability.

### 5.2.4 Event-Level Limitations

| Channel | F0.5 (Batch 16) | F0.5 (Batch 64) | Difference (16 − 64) |
|---|---|---|---|
| 41 | 0.567648 | 0.515156 | +0.052492 |
| 42 | 0.470808 | 0.447119 | +0.023689 |
| 43 | 0.423827 | 0.374966 | +0.048861 |
| 44 | 0.675028 | 0.461648 | +0.213380 |
| 45 | 0.591081 | 0.537957 | +0.053124 |
| 46 | 0.451873 | 0.525203 | −0.073330 |

*Table 5.13 — Direct comparison of channel-wise AFF F0.5 scores between Batch Size 16 and 64 (Subset 41–46).*

Table 5.13 directly compares the AFF_F0.50 scores across channels under the two batch-size configurations. The comparison shows that channels 41, 42, and 43 achieve higher scores when trained with Batch 64, whereas channel 44 performs better under Batch 16. Channel 45 exhibits a slight decrease in performance with the larger batch size, while channel 46 remains comparatively stable across both configurations.

These variations suggest that the influence of batch size is not uniform across channels. Rather than producing a consistent improvement, its effect appears to depend on channel-specific characteristics, including anomaly density, the temporal regularity of anomaly intervals, and the stability of reconstruction in latent space. This channel-dependent behaviour may suggest that batch size interacts with the underlying data structure rather than acting as a purely global optimisation parameter.

### 5.2.5 Baseline Interpretation

It is important to note that batch size does not directly increase or decrease model accuracy. Instead, it affects the variance of the gradient estimates during optimisation. Smaller batch sizes introduce higher gradient noise, leading to more oscillatory parameter updates. This behaviour can increase sensitivity to local patterns in the data, which may improve detection in certain channels but also result in less stable global behaviour.

In contrast, larger batch sizes reduce gradient variance, producing smoother optimisation dynamics and more stable decision boundaries. This often leads to more consistent global metrics. However, the reduced stochasticity may limit sensitivity to subtle or rare anomalies at the channel level. Therefore, the observed differences should be interpreted as a trade-off between stability and sensitivity rather than a strict improvement or degradation.

## 5.3 Event-Level Structural Analysis of Batch Size Effects

### 5.3.1 Motivation for Channel Selection

Two representative channels were selected in order to illustrate the structural effects of batch size on anomaly score dynamics. Channel 43 exhibits noticeable structural differences between the Batch-16 and Batch-64 configurations in terms of activation density and temporal distribution of anomaly scores. Channel 45, on the other hand, demonstrates clear anomaly detection behaviour under both configurations, enabling a direct comparison between sensitivity and stability effects introduced by the different training regimes.

Together, they provide complementary insight into how batch size modifies score dynamics and event formation.

### 5.3.2 Channel 43 – Structural Differences in Activation Patterns

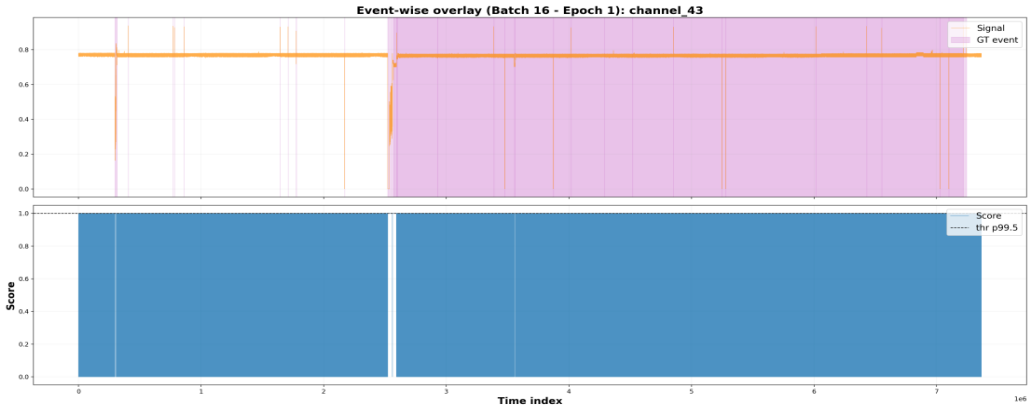


*Figure 5.3 — Event-wise overlay for Channel 43 (Batch 16, Epoch 1).*

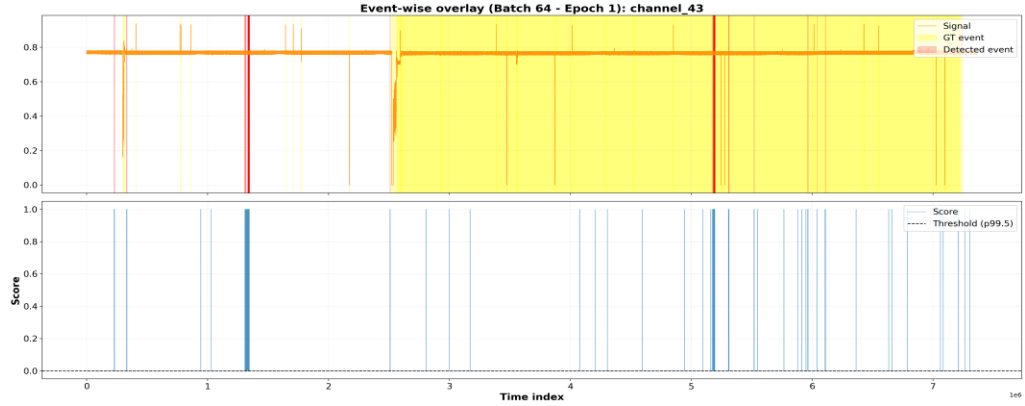


*Figure 5.4 — Event-wise overlay for Channel 43 (Batch 64, Epoch 1).*

In Channel 43, the dominant ground-truth anomaly occupies a large continuous time region. Both batch configurations detect this region, but their internal detection structures differ significantly.

#### *5.3.2.1 Batch 16 Behaviour*

Under the Batch 16 configuration, the anomaly score shows many frequent activations over time. Several small peaks exceed the representative threshold (p99.5), and before the merging step, the detections appear fragmented into multiple short segments. Although the final merged result covers the ground-truth anomaly interval, it is formed from many small activations rather than a single continuous block.

This pattern suggests that a smaller batch size makes the model more sensitive to local variations in reconstruction error. As a result, the detector reacts to minor fluctuations, producing dense and closely spaced anomaly signals that are later combined during post-processing.

#### *5.3.2.2 Batch 64 Behaviour*

In contrast, the Batch 64 configuration produces a more structured activation pattern. The anomaly score contains fewer isolated spikes, and the detected regions appear more concentrated in time. Instead of many small activations, the detections tend to form more coherent temporal blocks.

The anomaly interval is still identified, but it is represented through more compact and consolidated activation segments. This behaviour suggests that the larger batch size results in a more stable decision boundary, with reduced sensitivity to small local fluctuations in reconstruction error.

Overall, Channel 43 illustrates the difference between the two training regimes: a smaller batch size increases local responsiveness, while a larger batch size promotes temporal consolidation and smoother detection patterns.

### 5.3.3 Channel 45 – Sensitivity versus Stability Trade-off

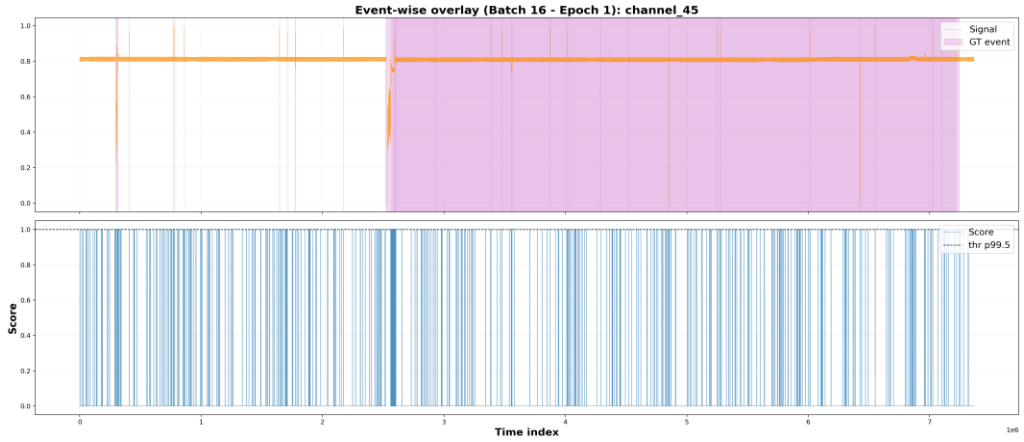


*Figure 5.5 — Event-wise overlay for Channel 45 (Batch 16, Epoch 1).*

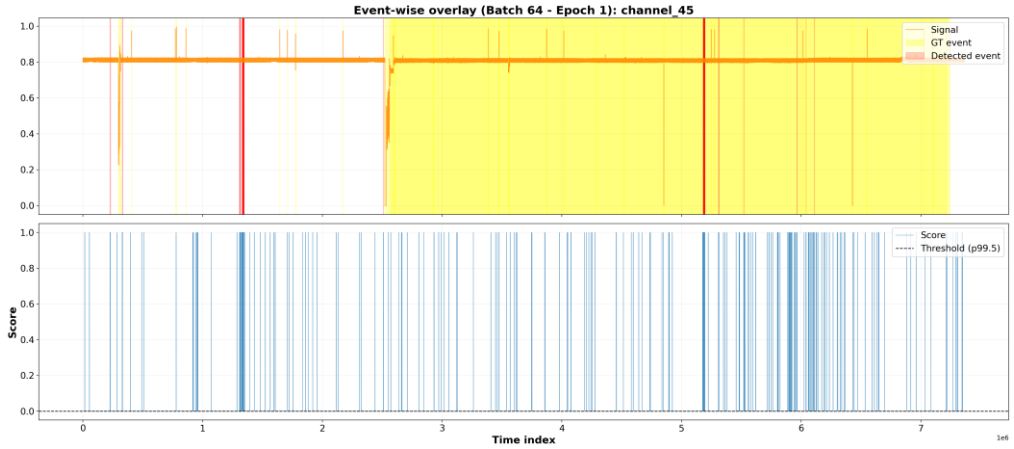


*Figure 5.6 — Event-wise overlay for Channel 45 (Batch 64, Epoch 1).*

Channel 45 provides an even clearer illustration of the batch-size trade-off.

#### *5.3.3.1 Batch 16 Behaviour*

Under the Batch-16 configuration, the anomaly score exhibits dense activation patterns across extended time intervals. Numerous local peaks exceed the percentile threshold, producing a highly reactive detection behaviour. While this configuration increases sensitivity to local reconstruction variations, it also results in more fragmented anomaly score dynamics at the raw signal level.

#### *5.3.3.2 Batch 64 Behaviour*

In contrast, the Batch-64 configuration produces a more structured anomaly score profile. Score activations become sparser and more temporally consolidated, resulting in compact

event detections with fewer scattered activations outside the main anomaly interval. Compared to the smaller batch configuration, this behaviour reflects increased temporal stability in the anomaly detection process.

### 5.3.4 Comparative Interpretation

Taken together, Channels 43 and 45 demonstrate that batch size does not simply "improve" or "degrade" detection performance. Instead, it alters the structural properties of the anomaly score:

- **Batch 16** increases local sensitivity, producing dense and reactive score behaviour.
- **Batch 64** enhances temporal stability, leading to more consolidated event-level detections.

Smaller batch sizes introduce noisier gradient updates during training, which can make the model more responsive to subtle local deviations. Larger batch sizes smooth the optimisation trajectory, resulting in a more stable global decision structure.

Importantly, this structural shift explains differences observed in global metrics. Improvements under Batch 64 should not be interpreted as purely higher accuracy, but rather as a change in decision boundary stability and event consolidation behaviour.

### 5.3.5 Implications for Subsystem-Level Monitoring

From a subsystem monitoring perspective, the results suggest:

- If detecting subtle local irregularities is critical, smaller batch training may increase sensitivity.
- If reducing fragmented detections and improving temporal coherence is more important, larger batch training may be preferable.

Channels 43 and 45 thus provide empirical evidence that batch size acts as a structural regulariser in the anomaly detection pipeline, influencing how reconstruction errors translate into event-level decisions.

## 5.4 Effect of Training Epochs on DC-VAE Performance (Subset 41–46)

### 5.4.1 Experimental Setup and Computational Constraints

In this section, the effect of increasing the number of training epochs from 1 to 10 is analysed for Subset 5.

During experimentation, training with 10 epochs and batch size 64 resulted in GPU memory limitations. To ensure stable execution and successful completion of training, the batch size was reduced from 64 to 16. This modification was introduced solely for computational feasibility and does not alter the conceptual comparison between different epoch configurations.

Therefore, the comparison in this section is conducted between:

- Epoch = 1, Batch = 16
- Epoch = 10, Batch = 16

This ensures methodological consistency when analysing the effect of training duration.

### 5.4.2 Global Performance Comparison (Epoch 1 vs Epoch 10)

| Configuration | Batch Size | Epoch | AFF_F0.5 | AFF_Precision | AFF_Recall | EW_F0.5 | ADTQC |
|---|---|---|---|---|---|---|---|
| DC-VAE | 16 | 1 | 0.43353 | 0.520758 | 0.259607 | 0.000014 | 0.64144 |
| DC-VAE | 16 | 10 | 0.49652 | 0.567842 | 0.330496 | 0.000567 | 0.82850 |

*Table 5.14 — Global performance comparison of DC-VAE (Epoch 1 vs Epoch 10, Batch Size 16, Subset 41–46).*

Increasing the number of training epochs from 1 to 10 leads to a clear improvement across all global evaluation metrics. The AFF F0.5 score rises from 0.4335 to 0.4965, while both precision and recall increase simultaneously. In addition, the ADTQC metric improves substantially, indicating better temporal consolidation of detected anomaly intervals.

- AFF_F0.5 increases from **0.4335 to 0.4965**
- Precision increases from **0.5208 to 0.5678**
- Recall increases from **0.2596 to 0.3305**
- ADTQC increases from **0.6414 to 0.8285**

The simultaneous increase in both precision and recall indicates that the model does not merely become more conservative or more permissive. Instead, it learns a more consistent and discriminative reconstruction boundary.

The improvement in ADTQC further suggests better temporal consolidation of anomaly detections, meaning fewer fragmented or unstable predictions over time.

Overall, the global results indicate that the model is not yet fully converged at epoch 1 and benefits significantly from extended training.

### 5.4.3 Channel-Level Behaviour

| Channel | AFF_Precision | AFF_Recall | AFF_F0.5 | EW_Recall |
|---|---|---|---|---|
| 41 | 0.637410 | 0.394807 | 0.567648 | 0.275862 |
| 42 | 0.551710 | 0.296749 | 0.470808 | 0.285714 |
| 43 | 0.521360 | 0.242423 | 0.423827 | **0.464286** |
| 44 | **0.702425** | **0.583925** | **0.675028** | 0.250000 |
| 45 | 0.622820 | 0.490997 | 0.591081 | 0.285714 |
| 46 | 0.552730 | 0.261216 | 0.451873 | 0.296296 |

*Table 5.15 — Channel-wise results for DC-VAE (Epoch 1, Batch Size 16, Subset 41–46).*

| **Channel** | **Precision_AFF** | **Recall_AFF** | **F0.5_Score_AFF** | **EW_Recall** |
|---|---|---|---|---|
| **41** | 0.684719 | 0.497706 | 0.636859 | 0.323077 |
| **42** | 0.596248 | 0.451159 | 0.560216 | **0.523810** |
| **43** | 0.696844 | 0.536546 | 0.657554 | 0.281250 |
| **44** | 0.571454 | 0.408291 | 0.529161 | 0.349206 |
| **45** | **0.715242** | **0.599940** | **0.688768** | 0.328125 |
| **46** | 0.667449 | 0.484218 | 0.620489 | 0.241935 |

*Table 5.16 — Channel-wise results for DC-VAE (Epoch 10, Batch Size 16, Subset 41–46).*

The impact of increased training epochs is consistently visible across individual channels.

Notable improvements include:

- Channel 43: F0.5 increases from 0.4238 to 0.6576
- Channel 45: F0.5 increases from 0.5911 to 0.6888
- Channel 42: Significant increase in event-wise recall

These improvements indicate that additional training epochs allow the model to better capture stable anomaly structures, particularly in channels with more consistent deviation patterns.

In several channels, both precision and recall increase simultaneously. This suggests that the model's decision boundary becomes sharper and more representative of the underlying anomaly distribution, rather than simply shifting towards over-detection.

### 5.4.4 Quantitative Delta Analysis (Epoch 10 – Epoch 1)

To better quantify the effect of increasing the number of training epochs, the metric variation between Epoch 1 and Epoch 10 was computed for each channel as:

Delta = Metric at Epoch 10 − Metric at Epoch 1

Table 5.17 and Figure 5.7 represent Channel-wise Delta analysis (Epoch 10 – Epoch 1), showing how precision, recall, F0.5, and event-wise recall evolve with increased training duration.

| Channel | Δ Precision | Δ Recall | Δ F0.5 | Δ EW_Recall |
|---|---|---|---|---|
| 41 | +0.047309 | +0.102899 | +0.069211 | +0.047215 |
| 42 | +0.044538 | +0.154410 | +0.089408 | +0.238096 |
| 43 | +0.175485 | +0.294123 | +0.233727 | −0.183035 |
| 44 | −0.130971 | −0.175634 | −0.145866 | +0.099206 |
| 45 | +0.092423 | +0.108943 | +0.097687 | +0.042411 |
| 46 | +0.114719 | +0.223280 | +0.168617 | −0.054361 |

*Table 5.17 — Channel-wise delta analysis (Epoch 10 – Epoch 1) for precision, recall, F0.5, and event-wise recall.*

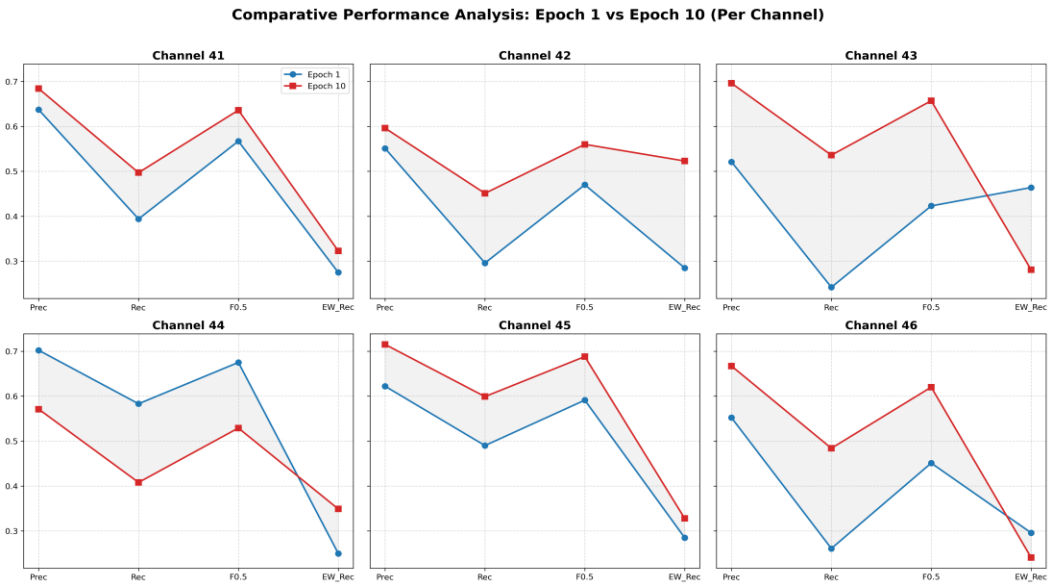


*Figure 5.7 — Channel-wise delta analysis (Epoch 10 – Epoch 1) for precision, recall, F0.5, and event-wise recall.*

A positive Delta value indicates performance improvement after longer optimisation, whereas a negative Delta value indicates degradation.

As shown in Table 5.17 and Figure 5.7, increasing the number of training epochs leads to improvements in most channel-level metrics.

More precisely, the following trends can be observed:

- **Precision_AFF** increases in the majority of channels, suggesting improved stability of the anomaly decision boundary.
- **Recall_AFF** shows noticeable growth, particularly in Channels 42, 43, 45, and 46, indicating enhanced sensitivity to anomalous segments.
- **F0.5_AFF** improves in five out of six channels, confirming that extended training strengthens the balance between precision and recall.
- **Channel 44** represents an exception, where several metrics decrease slightly, possibly indicating early saturation or mild overfitting effects.

Importantly, the Delta analysis demonstrates that the observed global improvement is not driven by a single channel but is distributed across multiple sensors, reflecting a more general enhancement of the learned representation.

## 5.4.5 Event-Level Behaviour Under Extended Training- Channels 43 and 44

### *5.4.5.1 Channel 43 – Consistent Metric and Event-Level Improvement*

To further validate the convergence analysis, event-wise anomaly overlays were compared between Epoch 1 and Epoch 10 for Channel 43.

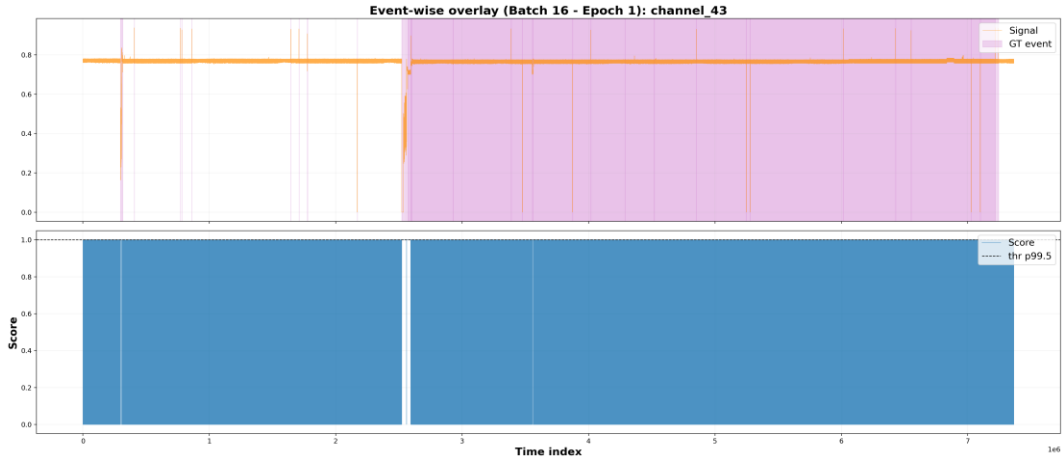


*Figure 5.8 — Event-wise anomaly overlay for Channel 43 at Epoch 1 (Batch Size 16).*

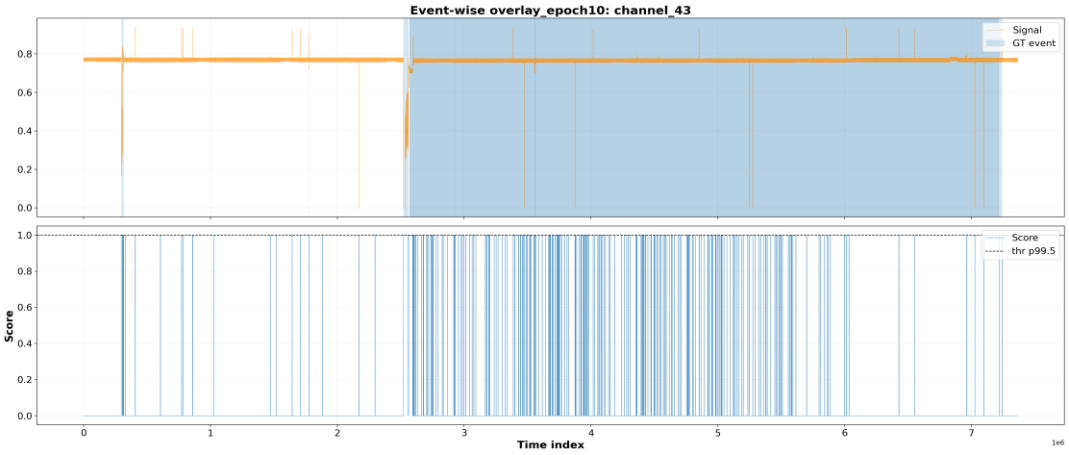


*Figure 5.9 — Event-wise anomaly overlay for Channel 43 at Epoch 10 (Batch Size 16).*

Figure 5.8 illustrates the behaviour at Epoch 1. The anomaly score appears largely saturated, with broad activation across extended time ranges. This behaviour suggests insufficient optimisation and unstable reconstruction boundaries.

In contrast, Figure 5.9 shows the model output after 10 training epochs. The anomaly activations become more structured and concentrated around specific temporal segments. The score distribution is less uniformly saturated and better aligned with anomalous regions.

This qualitative transition supports the quantitative delta analysis: the improvement in Precision_AFF and F0.5_AFF is reflected in more selective and discriminative event detection.

#### *5.4.5.2 Channel 44 – Metric Degradation versus Event-Level Stabilisation*

Channel 44 presents a more nuanced behaviour. Unlike Channel 43, several aggregate metrics (Precision_AFF, Recall_AFF, and F0.5_AFF) decrease when moving from Epoch 1 to

Epoch 10. From a purely metric-based perspective, this could be interpreted as performance degradation.

However, the event-wise overlays reveal a different dynamic.

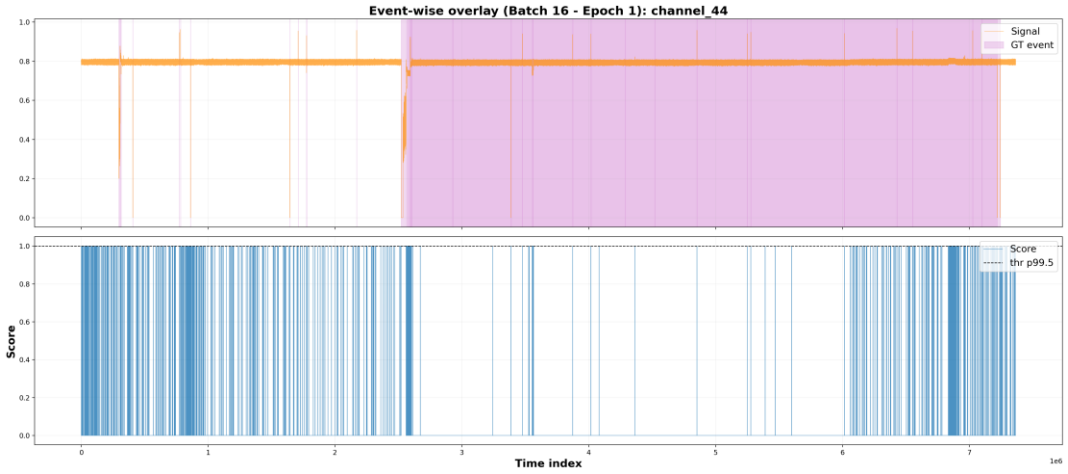


*Figure 5.10 — Event-wise anomaly overlay for Channel 44 at Epoch 1 (Batch Size 16).*

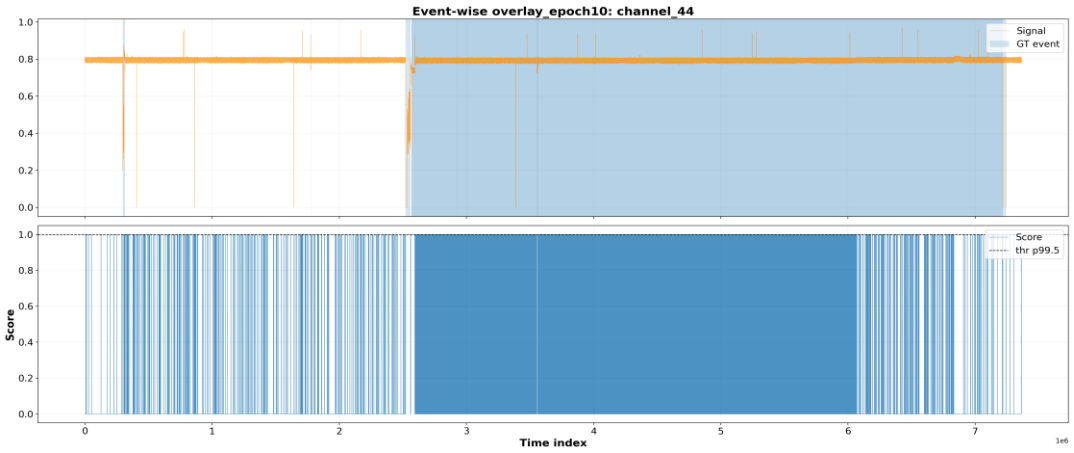


*Figure 5.11 — Event-wise anomaly overlay for Channel 44 at Epoch 10 (Batch Size 16).*

At Epoch 1 (Figure 5.10), the anomaly score exhibits dense and frequent activations, including numerous short-lived spikes. This behaviour suggests over-sensitivity and unstable decision boundaries, where many segments are flagged without strong temporal consolidation.

At Epoch 10 (Figure 5.11), anomaly activations become temporally smoother and more consolidated. Although fewer segments are flagged overall, the detections appear more structurally coherent and less noisy. The score distribution indicates improved stability, even if certain aggregate metrics decrease.

This behaviour highlights an important trade-off:

- Short training (Epoch 1) may inflate some metrics through aggressive activation.
- Extended training (Epoch 10) may reduce over-activation, improving structural consistency at the cost of certain recall-related measures.

Therefore, Channel 44 demonstrates that metric variation alone does not fully capture behavioural improvement. While delta values suggest partial degradation, the event-level perspective indicates improved boundary stability and reduced noise.

This contrast between Channels 43 and 44 reinforces a central thesis insight: convergence effects are not uniformly expressed across all sensors. Some channels benefit through simultaneous metric and structural improvement, while others exhibit a precision–recall or sensitivity–stability trade-off under extended optimisation.

### 5.4.6 Convergence Interpretation

From an optimisation perspective, increasing the number of training epochs allows the DC-VAE model to refine its latent representation and stabilise reconstruction parameters. With only one epoch, the optimisation process remains in an early stage, where the reconstruction boundaries are still unstable and the variance estimation across channels is not fully adapted to the underlying telemetry distribution. This behaviour is consistent with the widespread activation patterns observed in several channels at Epoch 1.

By extending training to 10 epochs, the optimisation trajectory progresses closer to convergence. The reconstruction error distribution becomes more structured, anomaly scores are less saturated, and decision boundaries become more temporally coherent. As demonstrated in the channel-level delta analysis and the event-wise overlays, this leads to improved discrimination between normal and anomalous behaviour in most channels.

Importantly, these improvements are not the result of architectural modifications. The DC-VAE structure, hyperparameters, and detection mechanism remain unchanged. The only difference lies in training duration. Therefore, the observed performance gains can be attributed to optimisation convergence rather than model redesign.

### 5.4.7 Validity and Scope of the Comparison

The epoch analysis presented in this section isolates the effect of extended optimisation under a fixed batch size configuration. Because both experiments were conducted with batch size = 16, the observed differences can be attributed to training duration rather than to gradient-scale effects or batch-related variance dynamics.

It is important to emphasise that this comparison evaluates convergence behaviour, not architectural modifications. The DC-VAE structure remains unchanged across experiments; only the number of optimisation iterations differs.

Therefore, the improvements observed at Epoch 10 reflect enhanced convergence stability and improved reconstruction calibration rather than structural capacity expansion.

However, this analysis does not assess the interaction between batch size and epoch count. Such interaction effects are examined separately in the batch size comparison section.

### 5.4.8 Final Interpretation

The comparison between Epoch 1 and Epoch 10 demonstrates that extended optimisation substantially improves the stability and discriminative capability of the DC-VAE model in Subset 5.

At the global level, the increase in F0.5 and ADTQC indicates stronger anomaly consolidation and improved temporal coherence. At the channel level, most sensors exhibit simultaneous gains in precision and recall, confirming that the improvement is not dominated by a single channel but rather distributed across the subsystem.

However, the behaviour of Channel 44 highlights that longer training does not universally improve every metric. While event-level alignment may become more structured, certain precision–recall components can slightly degrade. This suggests that extended optimisation enhances convergence stability but may also introduce mild channel-specific trade-offs.

Overall, the results indicate that one epoch is insufficient for stable convergence on this dataset. Ten epochs allow the model to develop a more calibrated latent representation and more reliable anomaly boundaries, without altering the architectural structure.

## 5.5 Cross-Subsystem Structural Analysis

### 5.5.1 Structural Heterogeneity in Subsystem 3

#### *5.5.1.1 Motivation for Isolating Channel 74*

As discussed in Chapter 3 (Raw Data Analysis), Subsystem 3 exhibits internal structural heterogeneity, with Channel 74 demonstrating markedly different statistical and dynamical characteristics compared to the remaining target channels (70–73, 75, 76). The present section does not repeat the full raw-data analysis; instead, it examines the modelling implications of that structural divergence within the DC-VAE framework.

Subsystem 3 consists of non-target channels (53–56) and target channels (70–76). In Chapter 3, Channel 74 was identified as structurally distinct in terms of sampling frequency,

correlation structure, and spectral behaviour. While the majority of channels in this subsystem operate at approximately 30-second intervals (≈0.033 Hz), Channel 74 is sampled at a significantly higher rate (≈3-second interval, ≈0.333 Hz), resulting in substantially denser temporal representation and broader spectral support.

Moreover, correlation analysis showed that Channels 70–73, 75, and 76 form a coherent group with strong mutual correlation, suggesting shared physical dynamics. Channel 74, in contrast, exhibited weak correlation with the remaining channels, indicating that it likely represents a different operational process or timescale.

From a modelling perspective, this structural heterogeneity raises an important question: how does the inclusion of a temporally denser and weakly correlated channel influence latent representation learning and reconstruction stability in a multivariate variational autoencoder?

Including Channel 74 may introduce:

- scale imbalance due to differing sampling density,
- variance heterogeneity in the input space,
- potential bias in latent encoding toward high-frequency components.

Conversely, it may also provide complementary information that enhances anomaly discrimination.

To assess this systematically, two controlled experimental configurations are evaluated:

- Subsystem 3 including Channels 70–76 (with Channel 74)
- Subsystem 3 excluding Channel 74 (Channels 70–73, 75, 76)

This comparison enables isolation of Channel 74's modelling impact, allowing us to determine whether its structural divergence improves detection robustness or instead destabilises anomaly scoring.

### 5.5.2 Global Impact of Channel 74 on Subsystem 3 Performance

To evaluate the influence of Channel 74 on anomaly detection performance, a controlled comparison was conducted between two configurations of Subsystem 3:

- Channels 70–73, 75, 76 (excluding Channel 74)
- Channels 70–76 (including Channel 74)

Both experiments were performed under identical training settings (Epoch = 1, Batch size = 64), ensuring that any observed performance variation can be attributed solely to the inclusion of Channel 74.

#### *5.5.2.1 Global-Level Comparison (With vs Without Channel 74)*

| Configuration | AFF_F0.5 | AFF_Precision | AFF_Recall | EW_F0.5 | ADTQC |
|---|---|---|---|---|---|
| **Without 74** | 0.416841 | 0.525390 | 0.228228 | 0.000002 | 0.999050 |
| **With 74** | 0.406917 | 0.498041 | 0.234959 | 0.000002 | 0.999934 |

*Table 5.18 — Global performance comparison for Subsystem 3 with and without Channel 74 (Epoch 1, Batch Size 64).*

At the global level, the inclusion of Channel 74 results in a slight decrease in the primary performance metric. The AFF F0.5 score decreases from **0.4168** (without Channel 74) to **0.4069** (with Channel 74). This reduction is mainly driven by a noticeable drop in precision (from 0.5254 to 0.4980), while recall shows only a marginal increase (from 0.2282 to 0.2350).

The decline in precision indicates that the model generates more false positive activations when Channel 74 is included in the multivariate reconstruction space. Although recall improves slightly, the magnitude of this increase is insufficient to compensate for the precision degradation, resulting in an overall reduction in F0.5.

Event-wise F0.5 remains effectively unchanged (≈0.000002 in both cases), and ADTQC values are nearly identical and close to unity, indicating that temporal consistency does not meaningfully improve with the inclusion of Channel 74.

From a modelling perspective, this behaviour suggests that Channel 74 does not provide complementary anomaly information that enhances global detection capability. Instead, its structural heterogeneity—previously identified in the raw data analysis—may introduce variance imbalance or latent-space distortion. The higher sampling frequency and distinct dynamical profile of Channel 74 likely alter the joint reconstruction dynamics, slightly destabilising anomaly boundary calibration for the subsystem as a whole.

Therefore, at the global level, the inclusion of Channel 74 does not improve detection performance and may mildly deteriorate precision-dominated metrics.

### 5.5.3 Channel-Level Impact of Including Channel 74

To better understand the influence of Channel 74 on Subsystem 3, a channel-wise comparison was conducted between the two configurations: (i) excluding Channel 74 and (ii) including Channel 74 within the multivariate input space.

### *5.5.3.1 Channel-Wise Metric Comparison*

| **Channel** - sub 3 | **Precision_AFF** | **Recall_AFF** | **F0.5_Score_AFF** | **EW_Recall** |
|---|---|---|---|---|
| **Channel 70** | 0.523425 | 0.392344 | 0.490641 | 1.000000 |
| **Channel 71** | 0.515621 | 0.395610 | 0.486127 | 1.000000 |
| **Channel 72** | 0.475343 | 0.384794 | 0.453978 | 1.000000 |
| **Channel 73** | 0.518136 | 0.413527 | 0.493184 | 1.000000 |
| **Channel 75** | 0.619362 | 0.510930 | 0.594143 | 0.800000 |
| **Channel 76** | 0.568384 | 0.265408 | 0.462737 | 1.000000 |

*Table 5.19 — Channel-wise results for Subsystem 3 excluding Channel 74 (Epoch 1, Batch Size 64).*

| **Channel** | **Precision_AFF** | **Recall_AFF** | **F0.5_Score_AFF** | **Recall_EW** |
|---|---|---|---|---|
| 70 | 0.496565 | 0.377768 | 0.467182 | 1.000 |
| 71 | 0.478810 | 0.378810 | 0.454798 | 0.800 |
| 72 | 0.486603 | 0.367830 | 0.457084 | 1.000 |
| 73 | 0.470042 | 0.370954 | 0.446204 | 0.800 |
| 74 | **0.607417** | **0.506207** | **0.584062** | 0.800 |
| 75 | 0.490551 | 0.365560 | 0.459152 | 0.800 |
| 76 | 0.533535 | 0.265318 | 0.443804 | 1.000 |

*Table 5.20 — Channel-wise results for Subsystem 3 including Channel 74 (Epoch 1, Batch Size 64).*

When Channel 74 is excluded, the six remaining channels (70–73, 75, 76) exhibit relatively homogeneous behaviour. Precision_AFF values range approximately between 0.47 and 0.62, while Recall_AFF remains moderate but stable across channels. Channel 75 demonstrates the strongest performance in this configuration, achieving the highest F0.5 score (0.594), indicating a well-balanced precision–recall trade-off. Event-wise recall is also largely saturated at 1.0 for most channels, suggesting consistent event detection coverage.

When Channel 74 is introduced into the model, several subtle but systematic changes become visible. First, Channel 74 itself exhibits comparatively strong performance, with Precision_AFF = 0.607 and Recall_AFF = 0.506, confirming that it is individually well modelled. However, the inclusion of this structurally distinct channel results in mild degradation across several of the original channels. Precision and F0.5 scores for Channels 70–73 decrease slightly, while Channel 75 experiences a more noticeable drop in both precision and recall.

Importantly, these changes are not dramatic, but they are consistent. The global F0.5 score decreases from 0.4168 (without 74) to 0.4069 (with 74), and precision declines accordingly. This pattern suggests that the addition of Channel 74 introduces representational competition within the latent space. Because Channel 74 operates on a different temporal scale and variance structure, the shared reconstruction model must accommodate heterogeneous dynamics, which may slightly reduce discrimination strength for the slower channels.

Channel 76 remains largely unaffected in terms of recall, indicating that the influence is not uniform across all signals. Instead, the effect appears selective and related to dynamical similarity.

Overall, the channel-wise comparison supports the hypothesis that structural heterogeneity within a multivariate input space can influence anomaly reconstruction stability. While Channel 74 is individually informative, its inclusion slightly redistributes modelling capacity, leading to small but measurable performance shifts in other channels.

## 5.5.4 Event-Level Detection Behaviour in Subsystem 3 (Epoch 1, Batch 64)

### *5.5.4.1 Representative Channels Analysis (Channels 71 and 75)*

To further interpret the channel-wise quantitative metrics, event-wise anomaly overlays were examined for representative channels within Subsystem 3. Figures 5.12 and 5.13 illustrate the detection behaviour for Channels 71 and 75 under the configuration without Channel 74.

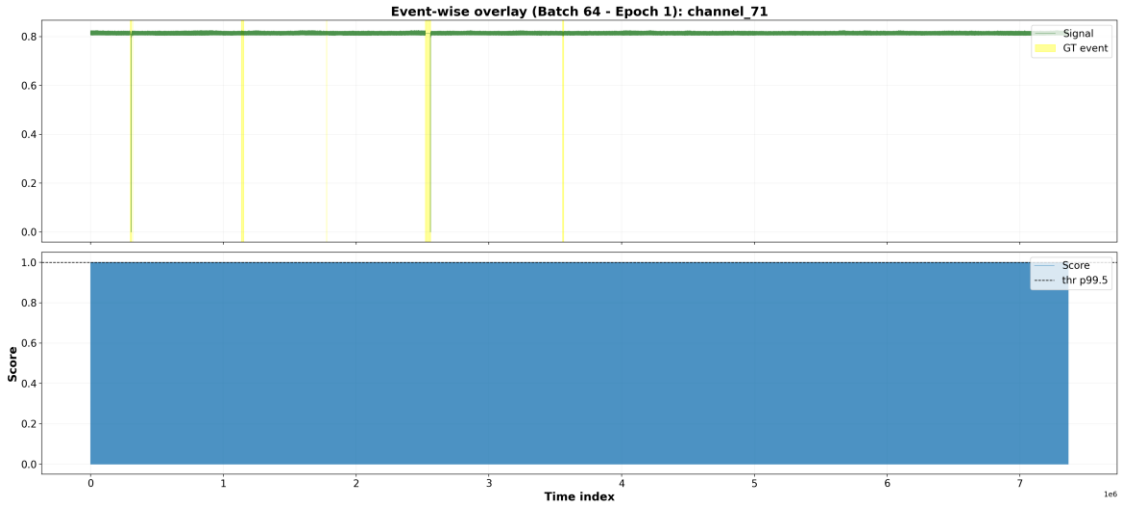


*Figure 5.12 — Event-wise anomaly overlay for Channel 71 in Subsystem 3 (Epoch 1, Batch Size 64, without Channel 74).*

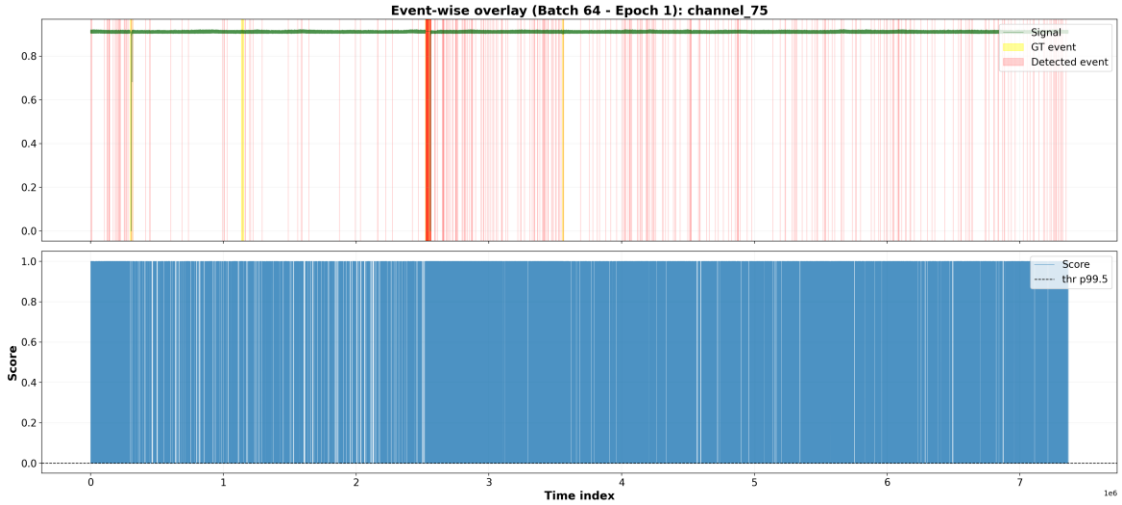


*Figure 5.13 — Event-wise anomaly overlay for Channel 75 in Subsystem 3 (Epoch 1, Batch Size 64, without Channel 74).*

In Channel 71 (Figure 5.12), the anomaly score exhibits near-continuous activation across the time axis. Although the ground-truth (GT) events appear as discrete and sparsely distributed intervals, the detected anomaly regions extend almost uniformly throughout the signal. This indicates that, at Epoch 1, the model fails to establish a sufficiently selective decision boundary, leading to score saturation and loss of temporal discrimination.

A more pronounced behaviour is observed in Channel 75 (Figure 2.13). In this case, not only does the model produce extended activation regions, but it also generates numerous false positive detections outside the GT intervals. The detected anomaly spans appear substantially broader than the ground-truth events, suggesting that reconstruction-based scoring is overly sensitive to minor deviations in this channel.

This qualitative inspection aligns with the modest F0.5 and precision values observed in the channel-wise results. The model, at this early training stage, tends to favour high coverage over precision, effectively labelling large portions of the signal as anomalous. Such behaviour is consistent with under-converged optimisation, where the reconstruction error distribution has not yet stabilised and threshold calibration remains poorly discriminative.

### 5.5.5 Interpretation of Structural Heterogeneity Effects

The analysis of Subsystem 3 reveals a structurally coherent yet temporally challenging set of channels for reconstruction-based anomaly detection. Although the channels exhibit strong mutual correlation and similar low-frequency dynamics, the model at Epoch 1 demonstrates limited discriminative capability at the event level. The anomaly score frequently saturates, leading to extended activation regions and reduced precision, particularly in channels such as 75.

The controlled comparison with and without Channel 74 further confirms that the subsystem's behaviour is influenced more by optimisation stability than by individual channel dominance. While Channel 74 introduces structural heterogeneity, its presence does not fundamentally alter the overall detection regime under limited training.

### 5.5.6 Concluding Remarks on Subsystem 3

Overall, the findings suggest that Subsystem 3 requires deeper optimisation or improved calibration to achieve selective event-level detection. The current configuration highlights the sensitivity of reconstruction-based models to temporal scale variations and score thresholding dynamics within moderately correlated multivariate subsystems.

Finally, it is worth noting that the hyperparameters used in this study were not specifically optimised for Subsystem 3. Given the structural heterogeneity observed across ESA subsystems, dedicated hyperparameter tuning at the subsystem level may further improve anomaly discrimination and reconstruction stability. This direction is therefore identified as a potential extension for future work.

## 5.6 Memory-Augmented DC-VAE

### 5.6.1 Baseline Reference: DC-VAE without Memory

The underlying DC-VAE model is derived from the approach proposed in [11].

This section reports the baseline performance of the DC-VAE algorithm without any memory mechanism, serving as the reference point for all subsequent comparisons. The evaluation is conducted on ESA Mission 1, channels 41 to 46, using standard TimeEval metrics: AFF Precision, AFF Recall, AFF F0.5 score, and Event-Wise Recall (EW Recall).

At a channel level, the baseline model achieves moderate AFF performance, with F0.5 scores ranging approximately from 0.37 to 0.54. Precision values remain relatively stable across channels, while recall varies more noticeably. This indicates that the model is able to identify anomalous deviations at individual time steps, but does so inconsistently across complete anomaly intervals.

A critical limitation of the baseline configuration emerges in the event-wise evaluation. For several channels (notably channels 42, 43, and 46), the EW Recall is exactly zero. This means that although isolated anomalous points are occasionally detected, they do not form temporally coherent segments that overlap sufficiently with annotated anomaly events. Consequently, the baseline DC-VAE behaves primarily as a point-wise anomaly detector and fails to reconstruct the temporal structure of anomalous events.

This limitation motivates the introduction of a memory mechanism aimed at improving temporal coherence and event-level detection.

| Algorithm | Dataset | AFF F0.50 | AFF Precision | AFF Recall | EW F0.50 | Global ADTQC |
|---|---|---|---|---|---|---|
| DC-VAE | 3 months | 0.46026 | 0.51760 | 0.31894 | 0.00039 | 0.92023 |

*Table 5.21 — Global performance of DC-VAE without memory on ESA Mission-1 (Channels 41–46, Epoch 1, Batch 64).*

| Channel | AFF Precision | AFF Recall | AFF F0.50 | EW Recall |
|---|---|---|---|---|
| 41 | 0.571136 | 0.370067 | 0.515156 | 0.068966 |
| 42 | 0.501362 | 0.312067 | 0.447119 | 0.000000 |
| 43 | 0.497787 | 0.188715 | 0.374966 | 0.000000 |
| 44 | 0.510485 | 0.333880 | 0.461648 | 0.035714 |
| 45 | 0.551745 | 0.489073 | 0.537957 | 0.035714 |
| 46 | 0.548997 | 0.447602 | 0.525203 | 0.000000 |

*Table 5.22 — Channel-wise performance of DC-VAE without memory on ESA Mission-1 (Channels 41–46).*

## 5.6.2 Memory-Augmented DC-VAE: Methodology and Experimental Design

This section is structured in two complementary parts. First, the **methodological design** of the memory mechanism is described in a self-contained manner, detailing how the memory is constructed and integrated into the DC-VAE framework. Second, the **experimental results** obtained using this mechanism are reported and analysed. This separation ensures conceptual clarity while keeping the narrative aligned with the experimental nature of Chapter 5.

### *5.6.2.1 Memory Construction and Integration (Methodology)*

This subsection describes how the latent-space memory was constructed and how it was embedded into the TimeEval/DC-VAE pipeline. The objective is to preserve the original DC-VAE training objective while enriching the decision stage with a compact, data-driven prior derived from recurring anomaly patterns.

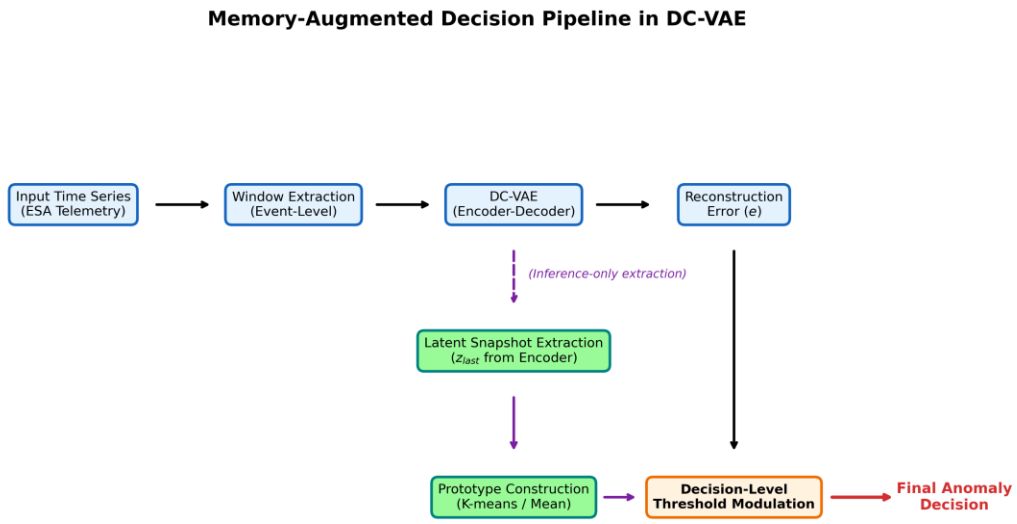


*Figure 5.14 — Decision-level integration of the latent memory module into the DC-VAE pipeline. Latent snapshots extracted from the encoder are used only for prototype construction and do not influence representation learning or reconstruction. The memory module operates exclusively at the decision stage by modulating anomaly thresholds based on similarity to learned prototypes.*

This subsection documents how the latent-space memory was constructed and how it was integrated into the TimeEval/DC-VAE implementation. The objective is to keep the DC-VAE training objective unchanged while enriching the decision stage with a compact prior derived from recurring anomalies.

### *5.6.2.2 Anomaly Characterisation and Class Selection*

To identify suitable candidates for memory construction, the ESA Mission 1 anomaly metadata was analysed for the selected channel subset (channels 41–46). Labels were joined with the anomaly-type taxonomy and filtered to confirmed anomalies. The resulting frequency analysis shows that **class_3 / Global** anomalies dominate the selected subset, indicating a highly recurring failure mode that is suitable for prototype-based memory.

| Class | Locality | Count |
|---|---|---|
| class_3 | Global | 174 |
| class_14 | Global | 31 |
| class_15 | Global | 30 |
| class_20 | Global | 24 |
| class_17 | Global | 17 |
| class_9 | Global | 12 |
| class_2 | Global | 7 |
| class_11 | Global | 6 |
| class_21 | Local | 6 |
| class_22 | Global | 6 |
| class_22 | Local | 6 |
| class_1 | Local | 1 |

*Table 5.23 — Distribution of anomaly classes in Subset 41–46 (ESA Mission-1). Note: Class_3 (Global) anomalies dominate the selected channels, motivating their use as the primary memory source.*

#### *5.6.2.3 Window Definition and Export*

After selecting the target anomaly class, all temporal windows corresponding to **class_3 / Global** anomalies occurring on channels 41–46 were exported. Although the metadata contains many channel-wise windows, these windows correspond to a smaller set of unique anomaly events (unique anomaly IDs) that manifest simultaneously across multiple channels. The exported windows serve as a reproducible interface between metadata selection and latent extraction.

| Event ID | Start Time | End Time | Channels Involved |
|---|---|---|---|
| id_107 | 2006-04-22T06:45:39.171Z | 2006-04-22T06:45:49.083Z | 41–46 |
| id_93 | 2004-05-15T02:22:12.261Z | 2004-05-15T02:22:25.173Z | 42–46 |
| … | … | … | … |

*Table 5.24 — Unique class_3 (Global) anomaly events used for latent memory construction (29 unique IDs, 174 channel-wise windows).*

- Total channel-wise windows: 174
- Unique anomaly events: 29
- Channels involved: 41–46
- Purpose: Latent prototype extraction for memory module

Extracted anomaly IDs:
{id_107, id_109, id_110, id_114, id_118, id_121, id_122, id_124,
id_129, id_130, id_132, id_140, id_142, id_149, id_150, id_160,
id_165, id_166, id_172, id_176, id_177, id_183, id_184, id_185,
id_186, id_187, id_83, id_90, id_93}

*Note: In total, 174 channel-wise windows corresponding to 29 unique anomaly IDs were extracted and used for latent memory construction.*

#### *5.6.2.4 Latent Extraction (One Vector per Anomaly Event)*

For each anomaly ID, a fixed-length window of telemetry samples (T = 256) ending at the anomaly start time was sliced from the training time series for the selected channels. Each window was passed through the **DC-VAE encoder** (in inference mode) and a single event embedding was recorded using the final latent vector of the sequence, **z_last ∈ R^J** (J = 2 in this experiment). This yields one latent vector per anomaly ID and avoids redundancy introduced by channel-wise duplication of global anomalies.

| Parameter | Value |
|---|---|
| Window length (T) | 256 |
| Latent dimension (J) | 2 |
| Channels | 41-46 |
| Unique anomaly IDs | 29 |

*Table 5.25 — Latent extraction configuration for memory construction (window length T, latent dimension J, selected channels, and number of unique anomaly IDs).*

#### *5.6.2.5 Prototype Construction*

To construct a compact and interpretable latent memory, two complementary prototype strategies were explored. The objective was to progressively increase the representational capacity of the memory while maintaining simplicity and stability.

##### *5.6.2.5.1 Mean Prototype (Single Reference)*

As a baseline memory representation, a mean prototype was first computed by averaging all extracted latent embeddings corresponding to the selected anomaly class. This results in a single latent reference vector that captures the central tendency of the anomaly distribution in latent space. While this representation is highly compact, it implicitly assumes that anomaly embeddings form a unimodal cluster.

*Note:* The distribution of distances between individual anomaly embeddings and the mean prototype is reported below to assess intra-class variability.

*Latent shape: (29, 2)*
*Prototype shape: (2,)*

| ID | Distance | ID | Distance | ID | Distance |
|---|---|---|---|---|---|
| id_107 | 0.91303277 | id_121 | 1.0049919 | id_142 | 1.8765147 |
| id_109 | 2.3532038 | id_122 | 0.4991074 | id_149 | 1.8992487 |
| id_110 | 1.3081064 | id_124 | 1.5127496 | id_150 | 0.8309541 |
| id_114 | 4.871944 | id_129 | 0.33747452 | id_160 | 4.5424566 |
| id_118 | 1.5062541 | id_130 | 3.5974758 | id_165 | 2.3971639 |
| id_132 | 0.38471368 | id_140 | 4.102321 | id_166 | 1.173488 |
| id_172 | 2.107302 | id_176 | 0.87773323 | id_177 | 2.5552409 |
| id_183 | 3.4167407 | id_184 | 1.5162135 | id_185 | 2.0332634 |
| id_186 | 0.44568533 | id_187 | 0.1759981 | id_83 | 1.535772 |
| id_90 | 0.5052257 | id_93 | 1.380035 | | |

*Table 5.26 — Euclidean distance of each anomaly embedding to the mean latent prototype.*

Distance_to_mean_prototype summary:

count: 29.000000

mean: 1.781393

std: 1.279112

min: 0.175998

50% : 1.512750

max 4.871944

*The distribution of distances to the mean prototype exhibits a mean value of 1.781393 with a standard deviation of 1.279112, indicating moderate dispersion among anomaly embeddings. The wide range (0.175998–4.871944) suggests the presence of both tightly clustered and highly divergent anomaly patterns in the latent space.*

#### *5.6.2.5.2 Multi-Prototype Memory via K-Means Clustering (K = 2)*

Inspection of the latent embeddings revealed that anomaly representations are not perfectly homogeneous in latent space. To explicitly model this intra-class variability, k-means clustering was applied to the set of event-level embeddings, and the resulting cluster centroids were used as memory prototypes.

In the **first refined configuration (Iteration 1)**, the number of clusters was set to **K = 2**, yielding a coarse partition of the anomaly space into two dominant latent modes. This choice represents the minimal extension beyond the mean prototype and serves as a conservative multi-prototype baseline.

| **K** | 2 |
|---|---|
| ID | cluster |
| 83, 90, 93107, 109, 110, 118121, 122, 124, 129130, 132, 142, 149150, 160, 165, 166176, 183, 186, 187 | 1 |
| 114, 140, 172, 177, 184, 185 | 0 |

*Table 5.27 — K-means clustering assignments for latent anomaly embeddings (K = 2)*

| ID | Distance Reduction | ID | Distance Reduction | ID | Distance Reduction |
|---|---|---|---|---|---|
| id_107 | 0.47328296 | id_121 | -0.5145407 | id_150 | -0.3085642 |
| id_109 | 0.30626702 | id_122 | -0.578866 | id_160 | 0.4095602 |
| id_110 | 0.65036756 | id_124 | 0.011204004 | id_165 | 0.26778102 |
| id_114 | 2.5402827 | id_129 | -0.5521528 | id_166 | -0.09499991 |
| id_118 | -0.23216295 | id_130 | 0.5024686 | id_172 | 0.3841288 |
| id_132 | -0.30357942 | id_140 | 1.8869338 | id_176 | 0.5753114 |
| id_142 | 0.33567345 | id_149 | 0.6736442 | id_177 | 1.5414374 |
| id_183 | 0.33700204 | id_184 | 0.4259678 | id_185 | 0.68435776 |
| id_186 | -0.18189728 | id_187 | -0.43824184 | id_83 | -0.31227982 |
| id_90 | -0.43830502 | id_93 | 0.032691002 | | |

*Table 5.28 — Distance reduction relative to the mean prototype after K=2 clustering.*

#### *5.6.2.5.3 Prototype Refinement with Increased Granularity (K = 3)*

To further evaluate whether a richer memory representation can better capture latent diversity, a **second configuration with K = 3** prototypes was subsequently explored. This setting increases the granularity of the memory and allows additional latent substructures to be represented.

| K | 3 |
|---|---|
| ID | Cluster |
| 114, 140, 172 , 177, 184, 185 | 2 |
| 107, 110, 118, 121, 122, 124,129, 130, 132, 142, 149, 150,166, 176, 186, 187, 83, 90, 93 | 1 |
| 109, 160, 165, 183 | 0 |

*Table 5.29 — K-means clustering assignments for latent anomaly embeddings (K = 3).*

| ID | Distance Reduction | ID | Distance Reduction | ID | Distance Reduction |
|---|---|---|---|---|---|
| id_107 | -0.11931133 | id_121 | -0.0711174 | id_150 | -0.63417506 |
| id_109 | 1.531759 | id_122 | -0.6882521 | id_160 | 3.13884 |
| id_110 | 0.13644314 | id_124 | 0.5817019 | id_165 | 1.5844221 |
| id_114 | 2.5402827 | id_129 | -0.29847538 | id_166 | 0.49946284 |
| id_118 | 0.35750866 | id_130 | 0.66675544 | id_172 | 0.3841288 |
| id_132 | -0.58388233 | id_140 | 1.8869338 | id_176 | 0.37839663 |
| id_142 | 0.6886896 | id_149 | 0.26712132 | id_177 | 1.5414374 |
| id_183 | 3.1606362 | id_184 | 0.4259678 | id_185 | 0.68435776 |
| id_186 | 0.1746355 | id_187 | -0.5985288 | id_83 | 0.262411 |
| id_90 | -0.05379194 | id_93 | 0.5998351 | | |

*Table 5.30 — Distance reduction relative to the mean prototype after K=3 clustering.*

The impact of increasing the number of prototypes from K = 2 to K = 3 is analysed in the subsequent refinement stage (Iteration 2), where both channel-level and event-wise performance metrics are reported.

#### *5.6.2.6 Integration into TimeEval/DC-VAE*

The final memory artefacts (prototype NumPy files) were placed inside the DC-VAE algorithm package so they can be loaded at inference time in a Docker-safe manner. Importantly, memory is incorporated as a **post-decoder decision refinement**: the encoder/decoder architecture and the training loss remain unchanged, and only the thresholding step is modulated using information derived from the current latent state.

*The memory prototypes were packaged within the DC-VAE algorithm container to ensure compatibility with TimeEval's Docker-based execution pipeline. . If the prototype files are not found, the algorithm automatically falls back to the baseline DC-VAE decision rule, ensuring identical behaviour to the no-memory configuration.*

### 5.6.3 Memory-Aware Scoring Formulation (DC-VAE + Latent Memory)

This subsection presents the first exploratory integration of latent memory into the DC-VAE framework, focusing on the scoring formulation and its initial empirical behaviour.

#### *5.6.3.1 Normalised reconstruction residual (baseline DC-VAE)*

Let $x_{t,c}$ denote the observed value at time index $t$ for target channel $c$. The DC-VAE provides the reconstruction mean $\mu_{t,c}$ and an estimated standard deviation $\sigma_{t,c}$, derived from the predicted log-variance.

For each time step $t$ and channel $c$, the normalised reconstruction residual is defined as:

$$r_{t,c} = \frac{|x_{t,c} - \mu_{t,c}|}{\sigma_{t,c}}$$

**(Equation 5.1)**

In the baseline DC-VAE, anomaly detection is performed by comparing this residual to a channel-specific threshold $\alpha_c$:

$$\text{anomaly}_{t,c} = \left(r_{t,c} > \alpha_c\right)$$

**(Equation 5.2)**

This rule is equivalent to the thresholding used in the original implementation:

$$x_{t,c} < \mu_{t,c} - \alpha_c^{\text{down}}\sigma_{t,c} \quad \text{or} \quad x_{t,c} > \mu_{t,c} + \alpha_c^{\text{up}}\sigma_{t,c}$$

**(Equation 5.3)**

#### *5.6.3.2 Latent Distance Computation*

For each input window ending at time $t$, the DC-VAE encoder produces a latent representation $z_t \in \mathbb{R}^J$.
In this work, $J = 2$, and the latent vector corresponding to the final time step is used:

$$z_t = z_t^{\text{last}}$$

**(Equation 5.4)**

A memory bank is defined as a set of latent prototypes $p_k{}_{k=1}^{K}$ obtained, for example, via k-means clustering.
The distance from the current latent vector to the memory is computed as the minimum Euclidean distance to the prototypes:

$$d_t = \min_k \|z_t - p_k\|_2$$

**(Equation 5.5)**

To ensure scale robustness, a normalised distance is defined as:

$$\tilde{d}_t = \frac{d_t}{\text{median}(d) + \varepsilon}$$

**(Equation 5.6)**

where $\varepsilon$ is a small constant used to avoid numerical instability.

#### *5.6.3.3 Threshold Modulation Strategy*

Rather than modifying the encoder–decoder architecture or retraining the DC-VAE, the memory mechanism is integrated at the decision level by modulating the anomaly detection threshold in a time-dependent manner.
For each channel $c$, a memory-aware threshold is defined as:

$$\alpha'_c(t) = \alpha_c\left(1 + \lambda \tilde{d}_t\right)$$

**(Equation 5.7)**

where $\lambda > 0$ controls the strength of the memory influence.

The resulting decision rule becomes:

$$\text{anomaly}_{t,c}^{\text{mem}} = \left(r_{t,c} > \alpha'_c(t)\right)$$

**(Equation 5.8)**

When $\tilde{d}_t$ is small, the latent representation is close to known memory prototypes, and the threshold remains close to the baseline value $\alpha_c$.

When $\tilde{d}_t$ is large, the latent state is far from previously observed patterns, and the threshold increases, making the detector more conservative and reducing false positives caused by unfamiliar latent behaviour.

#### *5.6.3.4 Alternative Variant*

If the desired behaviour is to increase sensitivity for latent states that are far from memory (i.e., treating novelty as higher anomaly likelihood), the modulation can be inverted:

$$\alpha'_c(t) = \alpha_c\left(1 - \lambda \tilde{d}_t\right)$$

**(Equation 5.9)**

In practice, $\alpha'_c(t)$ must be clipped to remain strictly positive.

The memory mechanism presented in this section should not be interpreted as a structural modification of DC-VAE, but rather as an exploratory, decision-level extension inspired by few-shot anomaly detection concepts proposed in recent ESA-related research. The objective was to empirically evaluate whether latent prototypes of recurring anomalies can improve

event-level detection without altering the original representation learning process. For this reason, the memory module was deliberately integrated after the decoder, at the scoring stage, in order to avoid increasing architectural complexity or modifying the core training dynamics of the model.

## 5.6.4 Experimental Results and Observations

The memory-aware scoring formulation described above was first evaluated in an exploratory setting to assess its feasibility and qualitative impact on anomaly detection behaviour. In this initial experiment, the memory mechanism was integrated at the decision level by modulating the channel-wise thresholds according to the normalised latent distance $\tilde{d}_t$, while keeping the DC-VAE architecture and training procedure unchanged.

### *5.6.4.1 Exploratory Results*

The results of this exploratory integration show that the memory-aware formulation is technically stable and can be executed within the TimeEval framework without introducing numerical or optimisation issues. However, from a performance perspective, the initial formulation leads to a noticeable reduction in recall across several channels when compared to the baseline DC-VAE.

This behaviour can be directly explained by the adopted threshold modulation strategy. As defined in **Equation 5.7**, the effective threshold $\alpha'_c(t)$ increases when the latent representation is far from the stored memory prototypes. While this mechanism successfully suppresses spurious detections caused by unfamiliar latent patterns, it also makes the detector overly conservative in regions of the latent space that are insufficiently represented by the memory. As a result, moderate but meaningful anomalies may fail to exceed the inflated threshold, leading to missed detections and fragmented anomaly segments.

In particular, although point-wise anomaly scores occasionally exceed the baseline threshold, the increased memory-aware threshold prevents these detections from forming temporally coherent sequences. This effect is reflected in the event-wise metrics, where improvements are not observed despite the presence of local anomaly peaks. These observations indicate that the initial memory formulation, while conceptually sound, applies the memory influence in an overly global and suppressive manner.

Importantly, this exploratory result does not invalidate the memory concept itself. Instead, it highlights a critical design insight: when the latent state resembles previously observed anomalous patterns, the detector should become more sensitive rather than more conservative. The findings of this exploratory phase therefore motivate a revision of the memory interaction strategy, leading to the corrected, similarity-driven threshold modulation introduced in the subsequent iteration.

### *5.6.4.2 Iteration 1 (K=2, λ=0.1)*

In the first refined iteration, the memory mechanism was redesigned to correct the shortcomings observed in the exploratory attempt. The key objective was to ensure that memory information increases sensitivity when the system behaviour resembles previously observed anomalies, rather than suppressing detections.

#### *5.6.4.2.1 Memory Construction*

Latent representations produced by the DC-VAE encoder for known anomalous segments are collected offline and clustered using k-means. This process results in a set of memory prototypes, each representing a characteristic anomalous pattern in the latent space.

#### *5.6.4.2.2 Similarity-Based Modulation*

During inference, for each latent vector produced by the encoder, the distance to the memory is computed as the minimum Euclidean distance to the stored prototypes. This distance is then transformed into a similarity score using an exponential decay function, yielding a bounded measure that is high when the current latent state closely matches a known anomaly pattern.

#### *5.6.4.2.3 Threshold Formulation*

In the baseline DC-VAE, anomaly detection relies on adaptive thresholds defined as functions of the reconstruction mean and variance, scaled by channel-wise coefficients learned during the alpha-selection phase. In this iteration, these thresholds are locally modulated using the memory similarity score. When similarity is high, the thresholds are reduced, increasing sensitivity to deviations that resemble known anomalies. When similarity is low, the method reverts smoothly to baseline behaviour.

#### *5.6.4.2.4 Channel-Level Results*

| Channel | Precision_AFF | Recall_AFF | F05_Score_AFF | Recall_EW |
|---|---|---|---|---|
| 41 | 0.613609 | 0.494646 | 0.585449 | 0.292308 |
| 42 | 0.509719 | 0.306529 | 0.450054 | 0.492063 |
| 43 | 0.586182 | 0.400895 | 0.536582 | 0.156250 |
| 44 | 0.508626 | 0.315606 | 0.453193 | 0.571429 |
| 45 | 0.529857 | 0.343146 | 0.477855 | 0.515625 |
| 46 | 0.526784 | 0.320004 | 0.466496 | 0.354839 |

*Table 5.31 — Channel-wise performance of Memory-Augmented DC-VAE (Iteration 1: K=2, λ=0.1).*

This corrected formulation leads to a consistent improvement over the baseline across all channels. AFF F0.5 scores increase to the range of approximately 0.45 to 0.59, and event-wise recall becomes non-zero for all channels, reaching values as high as 0.57. These results confirm that memory integration enables the model to aggregate anomaly evidence over time and detect coherent anomalous events. However, performance remains heterogeneous across channels, indicating that the memory influence is not yet fully optimised.

#### *5.6.4.3 Iteration 2 (K=3, λ=0.2)*

The second refined iteration further strengthens the memory mechanism to evaluate whether a richer latent representation and a stronger modulation effect can yield additional improvements. Two controlled modifications are introduced.

First, the number of memory prototypes used in k-means clustering is increased to K = 3, allowing the memory to represent a broader range of anomalous patterns. Second, the memory influence coefficient λ is increased to 0.2, amplifying the effect of the similarity-based threshold modulation during inference.

##### *5.6.4.3.1 Channel-Level Results*

| Channel | Precision_AFF | Recall_AFF | F0.5_Score_AFF | EW_Recall |
|---|---|---|---|---|
| **Channel 41** | 0.848806 | 0.825812 | 0.844105 | 0.241379 |
| **Channel 42** | 0.884530 | 0.788887 | 0.863590 | 0.250000 |
| **Channel 43** | 0.713444 | 0.637671 | 0.696883 | 0.285714 |
| **Channel 44** | 0.720927 | 0.622645 | 0.698864 | 0.250000 |
| **Channel 45** | 0.779534 | 0.743603 | 0.772073 | 0.285714 |
| **Channel 46** | 0.526498 | 0.252706 | 0.432730 | 0.444444 |

*Table 5.32 — Channel-wise performance of Memory-Augmented DC-VAE (Iteration 2: K=3, λ=0.2).*

##### *5.6.4.3.2 Performance Interpretation*

This iteration yields a substantial improvement across most channels. AFF Precision and Recall both increase significantly, with AFF F0.5 scores reaching values above 0.84 for channels 41 and 42, and remaining high for channels 43 to 45. Event-Wise Recall also improves markedly, with values ranging from approximately 0.24 up to 0.44, including channels that previously exhibited zero EW Recall in the baseline configuration.

Channel 46 shows a more nuanced behaviour, with lower point-wise recall but a notable increase in event-wise recall. This suggests that the refined memory mechanism prioritises temporal continuity for this channel, potentially reflecting channel-specific dynamics or noise characteristics.

#### *5.6.4.3.3 Global Metrics Comparison*

At subset level, global AFF F0.5 changes only slightly across configurations (and remains close to the baseline magnitude).The most visible effect of memory is therefore reflected more clearly in channel-level metrics and event-wise behaviour than in the aggregated global score.

| Configuration | Train Time (min) | Execute Time (min) | AFF F0.5 | AFF Precision | AFF Recall | EW F0.5 | Global ADTQC |
|---|---|---|---|---|---|---|---|
| Iteration 1 (K=2, λ=0.1) | 7.58 | 62.93 | 0.4317 | 0.5171 | 0.2599 | 0.000014 | 0.6646 |
| Iteration 2 (K=3, λ=0.2) | 6.93 | 57.73 | 0.4370 | 0.5255 | 0.2611 | 0.000030 | 0.6604 |

*Table 5.33 — Global performance comparison of Memory Iteration 1 and Iteration 2 on Subset 41–46.*

This apparent discrepancy between channel-level improvements and the relatively stable global score can be explained by the aggregation mechanism of global metrics. Since global AFF measures average performance across all channels, strong improvements in a subset of channels may be partially offset by marginal changes or weaker behaviour in others. As a result, the aggregated score reflects a smoothed average effect, whereas the channel-wise analysis reveals the localized impact of the memory mechanism more clearly. In other words, improvements concentrated in specific channels are diluted when averaged with channels exhibiting limited change, leading to modest variations in the global metric despite substantial localized gains.

### 5.6.5 Comparative Analysis and Discussion

In this section, the performance of the DC-VAE model is analysed and compared across three configurations:
(i) the baseline model without memory,
(ii) the model with memory mechanism – iteration 1, and
(iii) the model with memory mechanism – iteration 2.

All results are reported at channel level for channels 41 to 46 of ESA Mission-1, using consistent evaluation metrics: **AFF Precision**, **AFF F0.5 score**, and **Event-Wise Recall (EW Recall)**. This comparison allows us to isolate the effect of introducing memory and to assess how successive refinements of the memory mechanism influence detection quality.

| Channel | (Metric) | Baseline (NO Mem) | Iteration 1 (with Mem) | Iteration 2 (with Mem) |
|---|---|---|---|---|
| **41** | **Precision** | 0.5711 | 0.6136 | **0.8488** |
| | **EW-Recall** | 0.0689 | **0.2923** | 0.2413 |
| | **F0.5 Score** | 0.5151 | 0.5854 | **0.8441** |
| **42** | **Precision** | 0.5013 | 0.5097 | **0.8845** |
| | **EW-Recall** | 0.0000 | **0.4920** | 0.2500 |
| | **F0.5 Score** | 0.4471 | 0.4500 | **0.8635** |
| **43** | **Precision** | 0.4977 | 0.5861 | **0.7134** |
| | **EW-Recall** | 0.0000 | 0.1562 | **0.2857** |
| | **F0.5 Score** | 0.3749 | 0.5365 | **0.6968** |
| **44** | **Precision** | 0.5104 | 0.5086 | **0.7209** |
| | **EW-Recall** | 0.0357 | **0.5714** | 0.2500 |
| | **F0.5 Score** | 0.4616 | 0.4531 | **0.6988** |
| **45** | **Precision** | 0.5517 | 0.5298 | **0.7795** |
| | **EW-Recall** | 0.0357 | **0.5156** | 0.2857 |
| | **F0.5 Score** | 0.5379 | 0.4778 | **0.7720** |
| **46** | **Precision** | **0.5489** | 0.5267 | 0.5264 |
| | **EW-Recall** | 0.0000 | 0.3548 | **0.4444** |
| | **F0.5 Score** | **0.5252** | 0.4664 | 0.4327 |

*Table 5.34 — Per-channel comparison: Baseline DC-VAE vs Memory Iteration 1 vs Memory Iteration 2.*

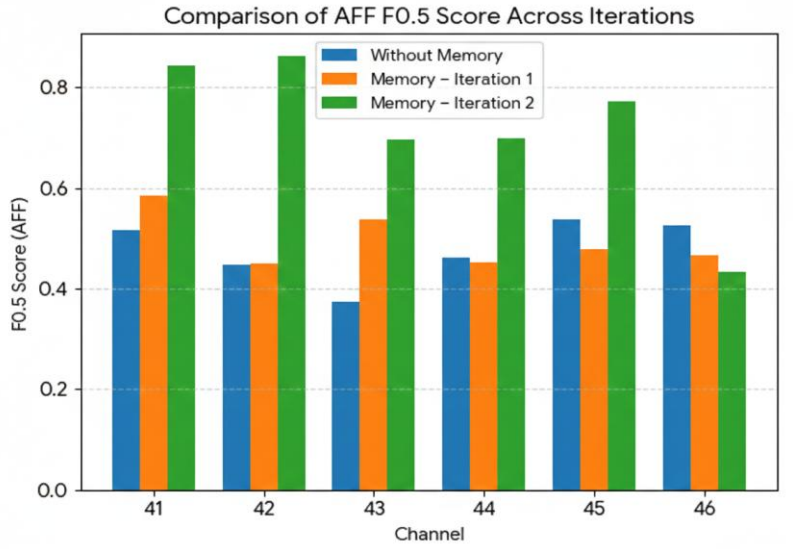


*Figure 5.15 — Comparison of AFF F0.5 per channel across baseline and memory iterations.*

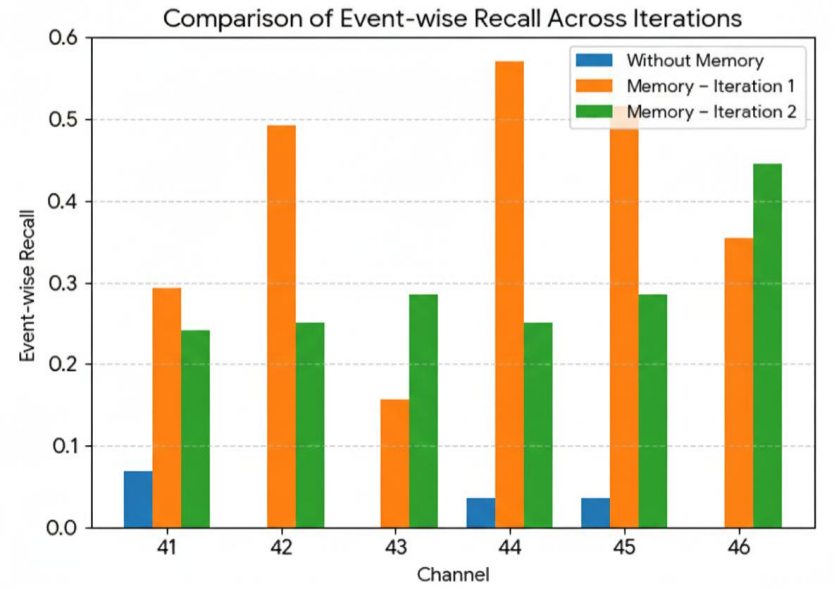


*Figure 5.16 — Comparison of Event-Wise Recall per channel across baseline and memory iterations.*

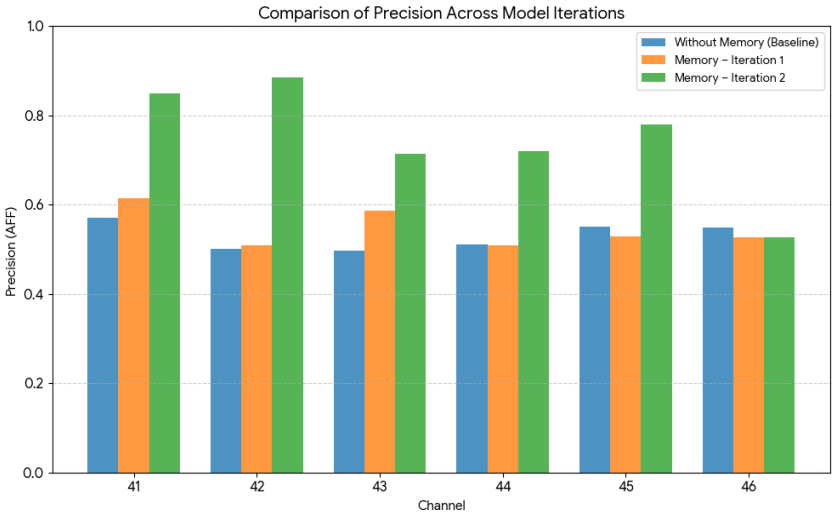


*Figure 5.17 — Comparison of AFF Precision per channel across baseline and memory iterations.*

### *5.6.5.1 Per-Channel Analysis*

- **Channels 41 & 42 (Peak Optimization):** These channels exhibit the clearest benefits of the proposed memory-augmented architecture. In the baseline configuration, detection performance is dominated by false alarms, largely due to

contamination of the training data with anomalous patterns. In Iteration 2, the introduction of refined memory prototypes leads to a substantial improvement in Precision, exceeding **85%** for both channels. This indicates that the memory mechanism successfully separates recurring ESA-specific anomaly signatures from background variability, resulting in more reliable point-wise detection.

- **Overcoming Detection Blindness (EW-Recall):** A major limitation of the baseline DC-VAE is its inability to detect anomalies as temporally coherent events. Channels 42, 43, and 46 exhibit an **EW-Recall of 0.00**, indicating a complete failure to recover anomaly intervals despite the presence of isolated anomaly scores. The integration of latent memory prototypes resolves this limitation by enabling temporal aggregation of detection evidence. As a result, event-wise recall increases substantially in subsequent iterations, reaching values of up to **57%**, demonstrating effective reconstruction of anomaly event structure.
- **Precision-Recall Trade-off:** A clear strategic shift is observed between Iteration 1 and Iteration 2. The first iteration favours sensitivity, leading to higher Recall at the expense of Precision. In contrast, Iteration 2 prioritises model maturity and error suppression through stronger memory modulation and a richer prototype set. This results in a marked increase in Precision and AFF F0.5 scores, which exceed **0.70** for most channels. Such a balance is particularly important for operational satellite telemetry monitoring, where false alarms carry a high operational cost.
- **The Case of Channel 46:** Channel 46 displays a distinct behaviour compared to the other sensors. While point-wise performance metrics remain moderate, the model achieves its highest **EW-Recall (0.44)** in the final iteration. This indicates that the memory mechanism captures long-term temporal dependencies specific to this channel, favouring event continuity over isolated point-wise accuracy. The observed trade-off suggests channel-specific dynamics rather than a general performance degradation.

#### *5.6.5.2 Discussion*

While the memory-augmented DC-VAE demonstrates clear improvements over the baseline in terms of temporal coherence and event-level detection, its behaviour remains partially channel-dependent. Channels characterized by recurring and structured anomaly patterns benefit most from memory integration, whereas channels with higher stochastic variability exhibit more moderate gains. This observation suggests that the effectiveness of latent memory is closely tied to the repeatability and structure of anomaly manifestations in the latent space.

The insights obtained from the memory-augmented DC-VAE provide a useful reference point for evaluating alternative anomaly detection approaches presented in the following sections, including algorithms that rely on explicit temporal modelling or sequence-based decision rules.

#### *5.6.5.3 Concluding Remarks*

While the baseline DC-VAE exhibits strong performance in global aggregate metrics, its behaviour reflects a reconstruction strategy that primarily captures the dominant system dynamics. The introduction of the decision-level memory module does not modify the representation learning process, but rather influences the thresholding stage by incorporating similarity to previously observed anomaly patterns.

The observed changes at channel level suggest that memory-based threshold modulation may improve the detection of recurring and structurally consistent anomalies in certain cases. At the same time, global AFF F0.5 values remain relatively stable, indicating that the memory mechanism does not uniformly increase overall performance, but rather influences the distribution of detection behaviour across channels.

The main contribution of this exploratory extension is therefore not a structural improvement of DC-VAE, but an empirical assessment of how latent prototypes can affect event-level coherence when applied at the decision stage. In particular, channels that previously exhibited fragmented detections become capable of forming more coherent anomaly intervals, while noisier channels show more moderate effects.

It is important to emphasise that the memory module is deliberately integrated after the decoder, without altering the encoder–decoder architecture or the training objective. As such, it should be interpreted as a post-hoc decision refinement mechanism rather than an architectural modification.

## 5.7 General Conclusions and Future Work

This thesis presented a systematic investigation of anomaly detection methods for ESA Mission-1 telemetry data, with particular emphasis on multi-level evaluation and structural interpretation. Both forecasting-based and reconstruction-based approaches were analysed under controlled experimental settings, including variations in training configuration, subsystem structure, and event-level behaviour.

The results demonstrate that global aggregate metrics alone are not sufficient to characterise anomaly detection performance in multivariate space telemetry. While stable subset-level scores can be achieved, channel-wise and event-level analyses reveal significant variability across signals. Factors such as training duration, batch configuration, and structural heterogeneity within subsystems influence detection behaviour in different and sometimes non-uniform ways. These findings reinforce the importance of combining quantitative metrics with structural and event-level inspection.

The exploratory memory extension introduced in this work was deliberately implemented at the decision stage, without modifying the encoder–decoder architecture or the training objective of DC-VAE. The results indicate that decision-level refinement using latent prototypes can influence event coherence in specific channels, while global aggregate metrics remain largely stable. These observations should be interpreted as empirical findings rather than architectural improvements.

A potential direction for future research concerns the placement of the memory mechanism within the DC-VAE pipeline. In this work, memory was applied only after the decoder, at the thresholding stage. Alternative configurations, such as incorporating memory within the latent representation or allowing it to interact with training dynamics, may lead to different behavioural trade-offs between stability and sensitivity. A systematic evaluation of such configurations is beyond the scope of this thesis and is left for future investigation.

# Chapter 6 — Conclusions and Final Remarks

## 6.1 Overview of the Thesis

As introduced in Chapter 1, the main objective of this thesis was to investigate how data-driven anomaly detection techniques behave when applied to spacecraft telemetry data. In particular, the study aimed to analyse the interaction between the structural properties of telemetry signals and the performance of modern anomaly detection algorithms.

The increasing complexity of modern spacecraft systems has led to the generation of large volumes of telemetry data that must be continuously monitored in order to ensure mission safety and operational reliability. Detecting abnormal behaviour within these telemetry streams is therefore a critical task for space missions. However, anomaly detection in spacecraft telemetry presents several challenges, including heterogeneous sensor measurements, varying sampling frequencies, correlated subsystem behaviour, and the scarcity of labelled anomalous events.

Within this context, this thesis investigated anomaly detection in spacecraft telemetry using the European Space Agency Anomaly Detection Benchmark (ESA-ADB). The work focused on the comparison between two different modelling paradigms: forecasting-based and reconstruction-based anomaly detection methods. These approaches were represented by the Telemanom and DC-VAE algorithms, respectively.

To address the research objectives defined in Chapter 1, the thesis combined three complementary components. First, a detailed exploratory analysis of the telemetry dataset was conducted to understand the structural and temporal characteristics of the signals. Second, a reproducible experimental pipeline was established using the TimeEval benchmarking framework. Third, the selected anomaly detection algorithms were applied and analysed in order to examine their behaviour on the ESA telemetry dataset.

Through this structured approach, the thesis connects data characteristics, algorithmic modelling strategies, and experimental evaluation in order to better understand anomaly detection in realistic spacecraft telemetry environments.

## 6.2 Understanding the Structure of Telemetry Data

One of the central outcomes of this work is the improved understanding of the structural properties of the ESA telemetry dataset. As discussed in Chapter 3, spacecraft telemetry cannot be treated as a simple collection of independent signals. Instead, it reflects the architecture of a complex engineering system in which multiple subsystems interact.

Each telemetry channel corresponds to measurements generated by sensors monitoring specific components or subsystems of the spacecraft. As a result, many channels exhibit correlated behaviour. The correlation analysis performed in this thesis revealed strong multivariate dependencies between several channels, which often reflect shared physical processes or operational relationships within the spacecraft.

These structural relationships have important implications for anomaly detection. In practical scenarios, anomalies may not appear as isolated deviations in a single channel but may instead manifest as coordinated changes across multiple signals belonging to the same subsystem.

In addition to correlation analysis, spectral analysis using Power Spectral Density (PSD) highlighted the heterogeneous temporal characteristics of the telemetry signals. Some channels display slow variations and long-term trends, while others exhibit faster dynamic behaviour. This diversity in temporal patterns illustrates the challenges faced by anomaly detection models that attempt to capture both long-term dependencies and short-term fluctuations.

Overall, the exploratory data analysis performed in this thesis demonstrates that spacecraft telemetry should be interpreted as a structured multivariate system. Understanding this structure is essential for interpreting the behaviour of anomaly detection algorithms and for designing more effective monitoring solutions.

## 6.3 Experimental Evaluation Using the TimeEval Framework

To analyse anomaly detection methods under controlled and reproducible conditions, this thesis employed the TimeEval benchmarking framework. TimeEval provides a standardized infrastructure for running anomaly detection experiments, including dataset management, evaluation procedures, and containerized execution environments.

The use of this framework ensured that experiments were conducted in a reproducible manner, allowing consistent comparison between different algorithms. Reproducibility is particularly important in anomaly detection research, where differences in experimental setup can significantly affect performance evaluation.

Within this framework, two representative anomaly detection algorithms were analysed. Telemanom represents a forecasting-based approach in which neural networks are used to predict future values of telemetry signals, and anomalies are detected when observed values deviate significantly from predicted behaviour. The Telemanom experiments therefore establish a forecasting-based baseline for anomaly detection performance on the ESA telemetry dataset.

In contrast, DC-VAE represents a reconstruction-based approach that learns a compact representation of normal system behaviour and detects anomalies through elevated reconstruction errors. The evaluation of the DC-VAE model demonstrates the applicability of reconstruction-based approaches for modelling complex telemetry behaviour across multiple subsystems.

Applying these algorithms within the TimeEval framework enabled a systematic comparison between forecasting-based and reconstruction-based anomaly detection strategies under consistent experimental conditions.

## 6.4 Key Insights from the Comparative Analysis

The comparative analysis performed in this thesis provides several insights that address the research objectives introduced in Chapter 1.

First, the structural organisation of telemetry channels significantly influences anomaly detection performance. Channels belonging to the same subsystem often share correlated dynamics, which can affect how algorithms interpret deviations from normal behaviour. Subsystems exhibiting strong structural coherence provide more stable modelling conditions, whereas heterogeneous subsystems containing mixed signal types or acquisition regimes introduce additional modelling complexity.

Second, the temporal characteristics of telemetry signals affect the suitability of different modelling approaches. Forecasting-based methods rely on accurate prediction of future signal values, which may be difficult for signals exhibiting irregular or multi-scale dynamics. Reconstruction-based methods instead focus on learning the overall structure of normal signal patterns and may therefore capture broader behavioural trends.

Third, the evaluation of anomaly detection algorithms must consider the event-based nature of anomalies in spacecraft telemetry. In operational scenarios, anomalies often occur as time intervals associated with system events rather than isolated point deviations. Consequently, evaluation metrics should reflect the temporal extent of anomalies rather than relying exclusively on point-wise measures.

Finally, the experiments indicate that anomaly detection behaviour varies substantially across subsystems. Differences in signal structure, sampling regimes, and inter-channel dependencies influence how forecasting-based and reconstruction-based approaches respond to abnormal system behaviour. These observations highlight the importance of integrating structural telemetry analysis with algorithmic modelling strategies.

## 6.5 Limitations and Future Research Directions

Although this thesis provides a structured investigation of anomaly detection in spacecraft telemetry, several limitations remain.

First, the experimental analysis focused primarily on selected telemetry channels from ESA Mission 1. Expanding the study to a larger subset of channels or additional missions could provide further insight into the generalisation of the observed behaviours.

Second, the algorithms analysed in this work were evaluated largely in their standard configurations within the benchmarking framework. Future research could explore modified architectures or hybrid models that combine forecasting and reconstruction principles in order to improve detection robustness.

Third, the exploratory analysis revealed strong structural relationships between telemetry channels that could potentially be exploited through more advanced modelling approaches. Graph-based learning techniques, for example, could incorporate subsystem topology or cross-channel dependencies directly into anomaly detection models.

In addition, this thesis explored the feasibility of integrating memory-aware mechanisms into the DC-VAE architecture in order to capture historical anomaly patterns and refine anomaly scoring behaviour. Further research could investigate more advanced memory-based or context-aware models for anomaly detection in telemetry data.

Finally, improving the interpretability of anomaly detection results remains an important direction for future work. Providing clearer explanations for detected anomalies could support mission operators in diagnosing system behaviour and responding more effectively to unexpected events.

## 6.6 Final Remarks

This thesis presented a systematic investigation of anomaly detection in spacecraft telemetry using the ESA Anomaly Detection Benchmark dataset. By combining exploratory data analysis, reproducible experimentation through the TimeEval framework, and comparative evaluation of forecasting-based and reconstruction-based algorithms, the work addressed the objectives outlined in Chapter 1.

The results demonstrate that successful anomaly detection in spacecraft telemetry requires not only robust machine learning algorithms but also a careful understanding of the structural and temporal characteristics of the monitored signals.

The objectives defined in Chapter 1 have therefore been addressed through the combined analysis of telemetry structure, reproducible experimental evaluation, and comparative assessment of forecasting-based and reconstruction-based anomaly detection models.

Overall, the findings of this thesis emphasise the importance of integrating domain-aware data analysis with systematic algorithm evaluation. Such an approach provides a stronger foundation for developing reliable anomaly detection systems capable of supporting future spacecraft monitoring and mission operations.

# References


[1] K. Kotowski *et al.*, "European Space Agency Benchmark for Anomaly Detection in Satellite Telemetry," Aug. 17, 2025, *arXiv*: arXiv:2406.17826. doi: 10.48550/arXiv.2406.17826.

[2] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 1–58, Jul. 2009, doi: 10.1145/1541880.1541882.

[3] 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*, London United Kingdom: ACM, Jul. 2018, pp. 387–395. doi: 10.1145/3219819.3219845.

[4] Z. Zamanzadeh Darban, G. I. Webb, S. Pan, C. Aggarwal, and M. Salehi, "Deep Learning for Time Series Anomaly Detection: A Survey," *ACM Comput. Surv.*, vol. 57, no. 1, pp. 1–42, Jan. 2025, doi: 10.1145/3691338.

[5] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[6] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," Dec. 10, 2022, *arXiv*: arXiv:1312.6114. doi: 10.48550/arXiv.1312.6114.

[7] A. Enttsel *et al.*, "Anomaly Detection-Reconstruction Trade-Off in Autoencoder-Based Compression," in *2024 32nd European Signal Processing Conference (EUSIPCO)*, Lyon, France: IEEE, Aug. 2024, pp. 1957–1961. doi: 10.23919/EUSIPCO63174.2024.10715213.

[8] G. G. González, S. M. Tagliafico, A. Fernández, G. G. Sena, J. Acuña, and P. Casas, "One Model to Find Them All Deep Learning for Multivariate Time-Series Anomaly Detection in Mobile Network Data," *IEEE Trans. Netw. Serv. Manage.*, vol. 21, no. 2, pp. 1601–1616, Apr. 2024, doi: 10.1109/TNSM.2023.3340146.

[9] A. van den Oord *et al.*, "WaveNet: A Generative Model for Raw Audio," Sep. 19, 2016, *arXiv*: arXiv:1609.03499. doi: 10.48550/arXiv.1609.03499.

[10] P. Malhotra, A. Ramakrishnan, G. Anand, L. Vig, P. Agarwal, and G. Shroff, "LSTM-based Encoder-Decoder for Multi-sensor Anomaly Detection," Jul. 11, 2016, *arXiv*: arXiv:1607.00148. doi: 10.48550/arXiv.1607.00148.

[11] KP Labs, "Few-Shot Anomaly Detection in Satellite Telemetry," European Space Agency (ESA), Noordwijk, Netherlands, Contract Number: 4000141301, 2024. [Online]. Available: https://activities.esa.int/4000141301