# Development of a Residual-Based Anomaly Detection System with Persistence Logic for Marine Diesel Generators

Luis F. Mendoza[1], Edwin Puertas[2], Edwin Paipa-Sanabria[3] and, Juan C Martinez-Santos[4]

[1,2,4] *Universidad Tecnologica de Bolivar, Cartagena, Bolívar, 130001, Colombia*
*luimendoza@utb.edu.co*
*epuerta@utb.edu.co*
*jcmartinezs@utb.edu.co*
[1,3] *Corporación de Ciencia y Tecnología para el Desarrollo de la Industria Naval, Marítima y Fluvial, Cartagena, Bolívar, 130009, Colombia*
*epaipa@cotecmar.com*

## ABSTRACT

Diesel generator engines (DGEs) are critical safety and mission assets for naval platforms, providing continuous electrical power for navigation, communications, habitability, and training operations. In practice, maintenance of auxiliary generation on training ships still relies mainly on time-based tasks and reactive corrective actions, despite the increasing availability of onboard operational data. This paper presents a residual anomaly detection system with persistence logic for the diesel generator engines of a Colombian Navy training ship. The proposed approach targets early detection of abnormal thermal behavior under scarce fault labels through a traceable workflow that combines: (i) data quality gates and preprocessing, including plausibility filtering and multivariate inconsistency treatment; (ii) target and feature definition for nominal regression modeling; (iii) residual monitoring with EWMA smoothing and time varying control limits; and (iv) persistence rules for sustained event declaration. The methodology is organized as an end-to-end workflow aligned with CRISP-DM and a PHM detection-first strategy. Since historical fault labels are limited and not reliably aligned in time, offline evaluation combines predictive assessment on held-out healthy data with Monte Carlo validation under simulated fault scenarios. Results from historical generator monitoring data show that the nominal models provide stable residual baselines for key thermal variables, while the detector can identify simulated abnormal scenarios across different severity levels. This paper provides a traceable workflow for anomaly detection in sparse and irregular shipboard monitoring data, discusses the limitations imposed by manual logs and scarce labels, and outlines future work toward health indexing, operational feedback, and more robust diagnostic support.



https://doi.org/10.36001/IJPHM.2026.v17i1.4762

## 1. INTRODUCTION

Naval platforms demand high reliability and continuity of electrical power to ensure safe operations, crew training readiness, and mission continuity. On a training ship, auxiliary power generation is essential not only for ship services but also for underpinning training cycles and operational safety margins during navigation, maneuvering, and port operations. Diesel generator engines (DGEs) provide this electrical backbone; therefore, detecting incipient abnormal behavior before a functional failure is a key enabler for safe and cost-effective maintenance.

However, when the maritime sector is considered, only 2% of the classed ships operate under a Condition-Based Maintenance (CBM) scheme. This indicates the lack of maturity of this industrial sector within the maintenance analytics context, thus making the implementation of fault diagnosis for marine machinery inconsistent while preventive maintenance is still preferred. If deep learning methodologies are considered, only variational autoencoders have been analysed to perform the anomaly detection task, which are unable to consider sequential patterns. Moreover, there is a lack of available data for marine machinery in academia due to the sensitive and confidential information that can be extracted. It can be considered, therefore, that there is a lack of analysis of fault diagnosis of marine machinery in real scenarios, thus averting some data preparation steps, such as data imputation and steady states identification, which are fundamental when analysing real operational data.(Velasco-Gallego & Lazakis, 2022)

Current shipboard maintenance for auxiliary generation is typically a blend of preventive time-based routines and correc-

tive interventions after alarms, abnormal readings, or outright failure. Such mixed strategies are effective for well-understood wear-out modes. However, they can be inefficient when degradation is intermittent, mode-dependent, or expressed as subtle multivariate changes that do not violate single-variable limits. Condition-based maintenance (CBM) and prognostics and health management (PHM) provide a structured path to leverage monitoring data for early detection, diagnosis, and eventually prognosis (Jardine, Lin, & Banjevic, 2006; Vachtsevanos, Lewis, Roemer, Hess, & Wu, 2006; Lee, Kao, & Yang, 2014). In particular, anomaly detection is appropriate when labeled faults are scarce or incomplete, a common reality for shipboard auxiliary systems, where maintenance records may not align temporally with high-frequency monitoring data and where fault taxonomies are incomplete.

A fault is an unexpected event and input to the system that can occur in any part of the system. Faults can also be classified according to their time behavior, i.e., abrupt fault, incipient fault, and intermittent fault. Any fault in a system causes poor performance and leads the entire system to collapse if it is not timely handled (Ahmad & Mohd-Mokhtar, 2022) A fault is defined as an unpermitted deviation of at least one characteristic property of a variable from an acceptable behavior. Therefore, the fault is a state that may lead to a malfunction or failure of the system. The time dependency of faults can be distinguished, abrupt fault (stepwise), incipient fault (drift-like), intermittent fault. With regard to the process models, the faults can be further classified.(Isermann, 2005)

This paper addresses the gap between available generator monitoring data on a Colombian Navy training ship and actionable, deployable anomaly detection under label scarcity. The work is grounded in a PHM workflow and emphasizes interpretability and operational deployment logic rather than purely offline benchmarking.

The main methodological novelty of this work is not the isolated use of EWMA or persistence rules, which are well established monitoring mechanisms. Rather, the contribution lies in their integration into a traceable residual workflow adapted to sparse, irregular, and manually recorded shipboard generator data. The proposed framework links data quality gates, nominal regression models for each target, residual baseline calibration, EWMA control limits, persistence rules for event declaration, and Monte Carlo evaluation under simulated abnormal scenarios. This integration provides an auditable PHM pipeline focused on detection for cases where dense sensor streams and time-aligned fault labels are not available.

In this study, early detection is defined operationally rather than as real-time fault diagnosis. Because the available records come from hourly manual engineering rounds, one sample corresponds to one recorded operational round. Therefore, early detection refers to identifying a sustained abnormal deviation before it becomes a documented functional failure or maintenance event, and within the temporal resolution allowed by the inspection routine. The objective is not to detect sub-hourly transients, but to provide earlier maintenance awareness when abnormal thermal behavior persists across successive rounds.

## 2. BACKGROUND AND RELATED WORK

PHM frameworks typically distinguish four complementary routes: detection, diagnosis, prognosis, and prescription (Lee et al., 2014). Detection identifies deviations from expected behavior; diagnosis isolates the root cause or fault mode; prognosis estimates future health evolution and remaining useful life; and prescription recommends actions that optimize risk, cost, and availability. CBM operationalizes these routes by scheduling maintenance based on measured condition rather than fixed time intervals (Jardine et al., 2006). For shipboard generator engines, detection is an effective first step because it does not require a complete, labeled fault library and provides actionable early warnings for targeted inspection.

From an operational perspective, anomaly detection sits at the intersection of PHM detection and early diagnostic triage. On naval platforms, generator sets experience frequent operating-point changes driven by ship service loads, maneuvering, and switching policies, which yield strongly multimodal data distributions. In such environments, deterministic limit checks (single-variable thresholds) may fail to capture multivariate "weak signals" of abnormality. At the same time, supervised fault classification is often infeasible due to scarce, delayed, or ambiguous labels. Consequently, anomaly detection methods that learn normal behavior from historical operation and flag statistically significant deviations become a pragmatic route toward CBM adoption, especially when complemented by persistence logic to avoid false alarms during benign transients (Chandola, Banerjee, & Kumar, 2009; Montgomery, 2009). In the present context, this motivation favors a residual-based formulation, since it enables anomalies to be interpreted as departures from expected thermal behavior under given operating conditions, rather than only as abstract deviations in a high-dimensional feature space.

There are three main types of model-based fault detection techniques according to the way of residual generation. They are known as observer-based FD techniques, parity-space-based FD techniques, and parameter-estimation-based FD techniques. Observer-based fault detection techniques correspond to the design of an observer for estimating system output and residual generation. In the parity-space-based approach, residual is generated by eliminating the initial states of dynamic systems and utilizing only system input and measurement data within a finite time window. The prime objective of both techniques is to ensure the robustness of residual against

the process and measurement of unknown inputs. Finally, the parameter-estimation approach is used to detect the slight change/drift in the system parameters by comparing the actual parameters of the nominal process with the estimated parameters (Ahmad & Mohd-Mokhtar, 2022).

Industrial sensor faults are typically categorized into two types: incipient and abrupt faults. Incipient faults, also referred to as successive faults in this study, develop gradually and are difficult to detect early on as sensor values do not deviate significantly from the normal operating range. Conversely, abrupt faults, characterized as short-term faults in this study, occur suddenly and are generally easier to identify. Each type of fault exhibits distinct patterns that are critical for effective detection and analysis. Each fault type represents a unique class, requiring accurate classification of sensor data into predefined categories, including both normal and various faulty conditions. This problem is fundamental to industrial sensor fault diagnosis, where each fault type signifies a distinct class requiring precise identification (Awaisi, Ye, & Sampalli, 2025).

This problem is fundamental to industrial sensor fault diagnosis, where each fault type signifies a distinct class requiring precise identification. Identifying the specific sensor error class helps in performing root cause analysis and facilitates timely error correction. Recently, machine learning (ML) and deep learning (DL) techniques, such as long short-term memory (LSTM), support vector machine (SVM), and deep neural network (DNN), have demonstrated significant potential in fault detection and multi-class classification (Awaisi et al., 2025).

It should be noted that while point anomalies can occur in any data set, collective anomalies can occur only in data sets in which data instances are related. In contrast, occurrence of contextual anomalies depends on the availability of context attributes in the data. A point anomaly or a collective anomaly can also be a contextual anomaly if analyzed with respect to a context. Thus a point anomaly detection problem or collective anomaly detection problem can be transformed to a contextual anomaly detection problem by incorporating the context information.(Chandola et al., 2009)

The quality of statistical analytic can be highly affected by the proportion of missing data, often categorized into the following three types:

- Missing Completely At Random (MCAR): If every measurement in the dataset has the same probability of being missing, the datasets is defined to be missing completely at random. This implies that the causes of the missing data are unrelated to the data. MCAR is an ideal assumption but it rarely occurs in practice.
- Missing At Random (MAR): Suppose only groups of measurements in the datasets have the same probability of being missing, and the observed data define the probability. In that case, we define the dataset to be missing at random. MAR is a more general and realistic assumption than MCAR. Under this assumption, the missingness can be modelled by using the observed data.
- Missing Not At Random (MNAR): This refers to the case when neither MCAR nor MAR holds. When the dataset is MNAR, the fact that the data are missing is systematically related to the unobserved data. It is hard to handle this missing data type because it will require strong assumptions about the patterns of missingness (Xiao et al., 2025).

Recent maritime studies have converged on this formulation of the problem. (Park & Oh, 2023) proposes a predictive maintenance algorithm for ship generator engines that explicitly addresses the scarcity of abnormal data by generating fault-like samples through engine simulations and combining them with collected ship data. Their design choice mirrors a key reality in shipboard PHM: maintenance actions and true fault occurrences are too infrequent (and too heterogeneously documented) to support conventional supervised learning. Although their work leverages simulation to enrich the abnormal class, the underlying objective remains aligned with that of anomaly detection systems: detecting departures from normal operational patterns early enough to enable targeted maintenance planning for generator engines.

A complementary line of work emphasizes real-time anomaly detection frameworks tailored to marine machinery. (Velasco-Gallego & Lazakis, 2022) introduces RADIS, which means a real-time anomaly detection intelligent system that couples an LSTM-based variational autoencoder with multi-level Otsu thresholding, and they validate it on sensor parameters from a ship diesel generator. It is closely related to the present study in three ways: (i) it treats the machinery as a multivariate time series rather than isolated channels; (ii) it uses an unsupervised deep generative model to characterize normality and derive anomaly scores; and (iii) it highlights deployment-oriented design choices (thresholding and near-real-time operation) that are essential when moving from offline experimentation to shipboard alerting. However, such approaches generally benefit from denser and more temporally regular data than those available in manual engine-room rounds, where hourly measurements, transcription issues, and discontinuities can limit the reliability of sequence-based modeling.

Beyond merchant shipping, naval logistics contexts also motivate anomaly-oriented fault detection from operating parameters. (Michelena et al., 2023) describes a fault-detection system approach for warship equipment with the explicit purpose of optimizing replacement parts and logistics decisions, comparing one-class techniques that learn normal behavior without requiring expert-crafted fault labels. While their application scope extends to fleet sustainment and provision-

ing, the methodological implication is directly relevant: ship systems benefit from detectors that remain effective when ground truth is limited and when operational variability can dominate the signal. It aligns with the generator-focused objective of the present work: an anomaly detector must remain robust across regimes and produce an alert burden compatible with shipboard maintenance workflows.

Taken together, the literature supports anomaly detection as an appropriate entry point for PHM under scarce labels, but it also shows that methodological choices depend strongly on data quality, temporal continuity, and the level of interpretability required for maintenance use. In this study, these constraints favor a residual-based anomaly detection strategy over a purely multivariate or deep temporal detector as the final decision layer. While multivariate unsupervised models remain useful references for capturing global distributional rarity, the adopted residual-based formulation provides a more direct physical interpretation of deviations at the target-variable level and is better suited to sparse, irregular, and operationally noisy shipboard monitoring data.

## 3. SYSTEM AND CASE STUDY DESCRIPTION

This section describes the operational context and the specific assets considered in the case study, to make the subsequent modeling decisions traceable to the physical system and to the realities of shipboard data collection. We first delimit the scope of the study to a Colombian Navy training ship auxiliary generator engines, then summarize the main characteristics of the generator sets under analysis, and finally document the data-acquisition modality and the associated quality constraints that inform the design of the anomaly-detection workflow.

### 3.1. Colombian Navy training ship context (scope-limited)

The case study corresponds to a Colombian Navy brigantine training ship that performs extended navigation periods interleaved with port stays and anchorage. Within this operating profile, continuity of onboard electrical supply is a safety- and mission-critical requirement, as a result of loss of generation capacity can affect ship service loads, navigation auxiliaries, and operational readiness. The present work intentionally limits its scope to the auxiliary diesel generator engines and to measurements collected during routine engineering rounds, without extending the analysis to other subsystems.

### 3.2. Diesel generator sets under study

Two generator sets are considered, corresponding to the ship's starboard (MG#1) and port (MG#2) units. Both sets are powered by Caterpillar 3406C TA diesel engines (6-cylinder inline, four-stroke, water-cooled) operating at a governed speed of 1800 rpm (60 Hz electrical system), with a 24 V battery starting system. The nominal electrical performance range in service is approximately 275–400 kVA, with nominal voltages of 220–480 V and a standard power factor of 0.8.

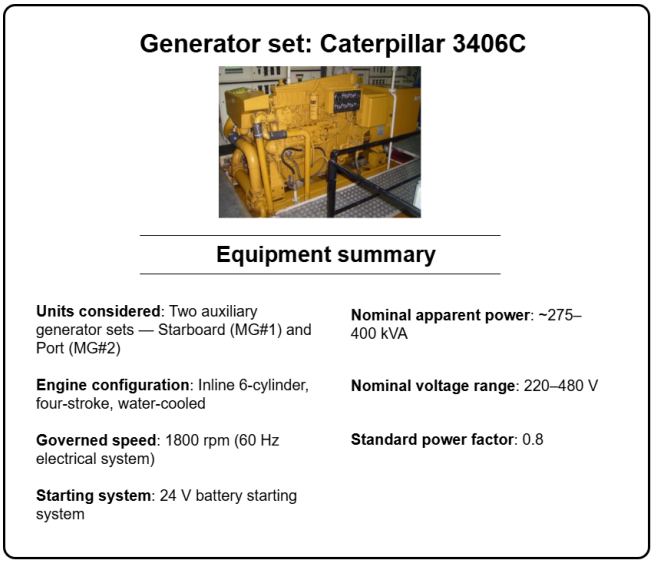


Figure 1. Auxiliary diesel generator sets considered in the case study (MG#1–MG#2)

Operational measurements available for both units include mechanical and thermofluid variables such as RPM, oil pressure, fuel pressure, exhaust gas temperature, cooling-water temperature, lubricating-oil temperature, and intake-air temperature, as well as electrical and load-related variables such as battery voltage, generator current, output power, and frequency. In addition, a categorical operational context variable indicates whether the ship is sailing, in port, or at anchor.

### 3.3. Data acquisition architecture and current maintenance practice

The operational datasets used in this study came from *hourly engine-room inspection rounds* conducted by the ship's technical crew. During each round, the crew entered key diesel-generator operating parameters by hand in engine-room log sheets, locally referred to as "minutas". These records were later digitized into spreadsheets and curated into a structured dataset.

(Llamas Reinoso, Martinez-Santos, & Puertas, 2026) documents the resulting dataset used in this study. This acquisition modality captures generator behavior under authentic shipboard conditions across navigation, anchorage, and port stays. However, it also means that the data were not produced by automated high-frequency sensors, but by human-in-the-loop measurements taken in an operationally demanding environment.

On the other hand, frequency domain filtering methods can be used to more precisely attenuate or eliminate noise in certain frequency bands. In addition to the measurement noise, mea-

surements from individual sensors can have errors due to a broad variety of reasons. These errors can originate from temporary faults or a complete sensor breakdown, inappropriate data handling/preprocessing by the data acquisition system (DAQ), calibration errors, incorrect sensor installation, and inaccurate sensor technology. The methodology required to detect and handle these errors, however, varies based on the impact of the failure on the data. Such errors are quite common in almost all datasets, and therefore, it is recommended to apply these or equivalent filters as the first step of the data cleaning process (Kim, Gupta, & Steen, 2025).

In anomaly detection literature, anomalies in univariate time series or dynamic signals are categorized into three categories: point anomalies, contextual anomalies, and collective anomalies. If a single measurement deviates significantly from the rest of the sensor readings, then a point anomaly is said to have occurred. Contextual anomaly occurs when a measurement is not anomalous in an 'overall sense' but only in a specific context. While control charts can be built to detect point and contextual anomalies, more specialized approaches are needed to detect collective anomalies (Kumar & Flores-Cerrillo, 2024).

The first 3 outlier detection methods are quite simple from an implementation point of view, but the fourth one requires some knowledge about digital filters and an understanding of the physical phenomenon. However, for a more comprehensive outlier detection layer, more advanced methods covering multivariate outlier detection algorithms, so that samples within the normal bounds but deviating from the prominent trends or correlations in the data time series are also detected (Kim et al., 2025).

Digital signals are known to acquire noise during the data recording process. Moving average is one of the simplest and most widely used methods for removing such noise and extracting cleaner signals. This method is effective in smoothing data and reducing random high-frequency noise by averaging over several samples recorded during a moving time frame, so-called window. Depending on how the weight for each sample in a window is defined, these methods become simple moving average, weighted moving average, and exponential moving average. Choosing a window size is very important for the results of data smoothing, but often it is determined empirically. If it is too small for the data, it may still not provide enough smoothing, and if it is too large, it can obscure important data changes (Kim et al., 2025).

Two independent datasets are available, one per generator unit, corresponding to the two auxiliary diesel generator engines installed in the engine room (MG#1 and MG#2). The records cover the ship's operational periods in 2021, 2023, 2024, and 2025, reflecting actual duty cycles and switching policies between generator units (Llamas Reinoso et al., 2026). Both datasets share the same 18-column structure and include timestamp variables (date and time), an operational accumulation indicator (hourmeter), mechanical variables (RPM), fluid pressures (lubricating-oil pressure and fuel supply pressure), thermal variables (exhaust gas temperature, lubricating-oil temperature, cooling-water temperature, and intake/ambient air temperature), and electrical variables (battery voltage, generator current, active power, and frequency). Contextual fields describing the ship's condition (e.g., port, anchorage, origin/destination text fields) are also present, enabling coarse stratification by operating context when needed. Dataset sizes differ due to generator availability, maintenance windows, and variations in log completeness (Table 1).

Table 1. Dataset sizes per generator unit.

| Dataset | Observations | Variables |
|---|---|---|
| MG#1 | 7,405 | 18 |
| MG#2 | 6,126 | 18 |

Considering that the datasets originate from handwritten records later transcribed, they present characteristic quality limitations:

- **Missing or incomplete entries:** Values are absent in multiple variables due to omitted measurements, incomplete rounds, or partial annotations.
- **Irregular temporal coverage:** Some periods exhibit substantially lower record density than others because of ship operational availability, generator activation cycles, and variability in round completion. This discontinuity limits strictly sequence-based modeling and motivates event- and persistence-based detection.
- **Transcription and formatting inconsistencies:** Manual digitization can introduce errors (e.g., misplaced decimal separators, handwriting interpretation), requiring standardization and plausibility checks during data preparation.
- **Manual measurement variability:** Since readings are taken manually in an operational environment, precision and repeatability vary, increasing dispersion in thermal and electrical variables.

These constraints are representative of *in-service maritime operational data* in a large fraction of naval and commercial shipping contexts, where missing values, irregular reporting intervals, data-entry inconsistencies, and measurement uncertainty are routinely reported as recurring challenges for ship performance and machinery analytics (Kim et al., 2025). They also shape how maintenance decisions are made in practice. When monitoring is sparse and labels are incomplete, maintenance organizations tend to rely on time-based routines complemented by corrective actions triggered by alarms or observed performance degradation.

Under these constraints, anomaly detection is an appropriate first PHM route because measurements are intermittent and

time-synchronized fault confirmation is limited. The system learns expected healthy behavior from available operational records and flags *sustained* departures that are more likely to warrant targeted inspection, troubleshooting, and maintenance planning than isolated spikes.

## 4. METHODOLOGY

This section presents the end-to-end methodology used to develop the proposed anomaly detection system, linking the operational problem to the data-driven design choices adopted in the study. The approach is structured as a traceable workflow aligned with CRISP-DM and with a PHM detection-first strategy, moving from data ingestion and preparation to nominal modeling and residual-based monitoring. The following subsections define the objective and framing, describe the development stages (A–F), and specify the procedures, assumptions, and parameters used to enable reproducible implementation and evaluation under scarce fault labels.

### 4.1. Objective and methodological framing

The methodological objective of this work is to develop an anomaly detection system for the auxiliary diesel generator engines under study using measurements from hourly engineering rounds. Given the absence of labeled fault events, the system is formulated as *nominal modeling* of key thermal responses, followed by residual-based monitoring to detect sustained deviations that may indicate abnormal operation or incipient degradation.

Another class of machine fault diagnostic approaches is the model-based approaches. These approaches utilize physics specific, explicit mathematical model of the monitored machine. Finally, the residuals are evaluated to arrive at fault detection, isolation and identification. Model-based approaches can be more effective than other model-free approaches if a correct and accurate model is built (Jardine et al., 2006).

Any technical system consists of inputs, outputs, and state variables. To improve analysis of these systems several approaches use observers or structural equations to calculate residual values. Without loss of generality, the underlying assumption is here that the system and its residuals were designed in such a way that during the system's normal operation the residual values are zero (Diedrich & Niggemann, 2022).

This formulation was prioritized over a purely multivariate anomaly detector as the final decision layer as the objective was not only to score statistical rarity, but also to generate interpretable and persistent alerts under operationally realistic data constraints. In model-based fault detection, residuals represent the deviation between the observed behavior and the behavior expected under normal operating conditions, which gives them direct physical meaning at the monitored variable level (Isermann, 2005; Diedrich & Niggemann, 2022). This property is especially valuable in maintenance-oriented settings, where alert review benefits from knowing whether the deviation is associated with exhaust, cooling water, or lubricating oil thermal behavior rather than only with a global anomaly score. In addition, hourly engineering rounds yield sparse and irregularly sampled observations, a setting that makes sequence-centered deep architectures harder to justify as the primary decision mechanism without stronger temporal continuity assumptions (Li & Marlin, 2020). For this reason, multivariate unsupervised detectors were considered useful as supporting tools for data cleaning and distributional screening, whereas the final anomaly logic was based on residuals and persistence rules to preserve interpretability, operational traceability, and robustness against short-lived transients.

To ensure continuity from operational need to modeling decisions, we adopt CRISP-DM as a transversal data methodology (business understanding $\rightarrow$ data understanding $\rightarrow$ data preparation $\rightarrow$ modeling $\rightarrow$ evaluation), aligned with the PHM "detection-first" route as an entry point to condition-based maintenance decision support (Kumar & Flores-Cerrillo, 2024).

### 4.2. Development stages

The methodology is organized into six stages (A–F), reflecting the actual development sequence: reading and structuring the data; cleaning and preprocessing; defining targets and features; training nominal models; constructing the anomaly scoring and event logic; and performing offline evaluation under scarce fault labels.

Figure 2 summarizes the end-to-end offline workflow.

### 4.3. Stage A: Data ingestion and structuring (CRISP-DM: data understanding)

Stage A converts the digitized engineering-round logs into a consistent analytical dataset that can be processed reproducibly in the subsequent stages. The inputs are provided as spreadsheet tables (one per generator and operational period) that include numeric operational readings (pressures, temperatures, electrical variables, governed variables) and contextual/text fields (navigation/port descriptors).

**A1. Reading and schema harmonization.** All tables are ingested using a unified schema: column names are standardized, units are checked for consistency, and a single variable dictionary is enforced across MG#1 and MG#2. This step ensures that downstream cleaning rules and model pipelines apply identically to both generator datasets.

**A2. Type conversion and parsing.** The digitization process may cause numerical columns to be imported as text, including temperatures, voltages, currents, power, and fre-

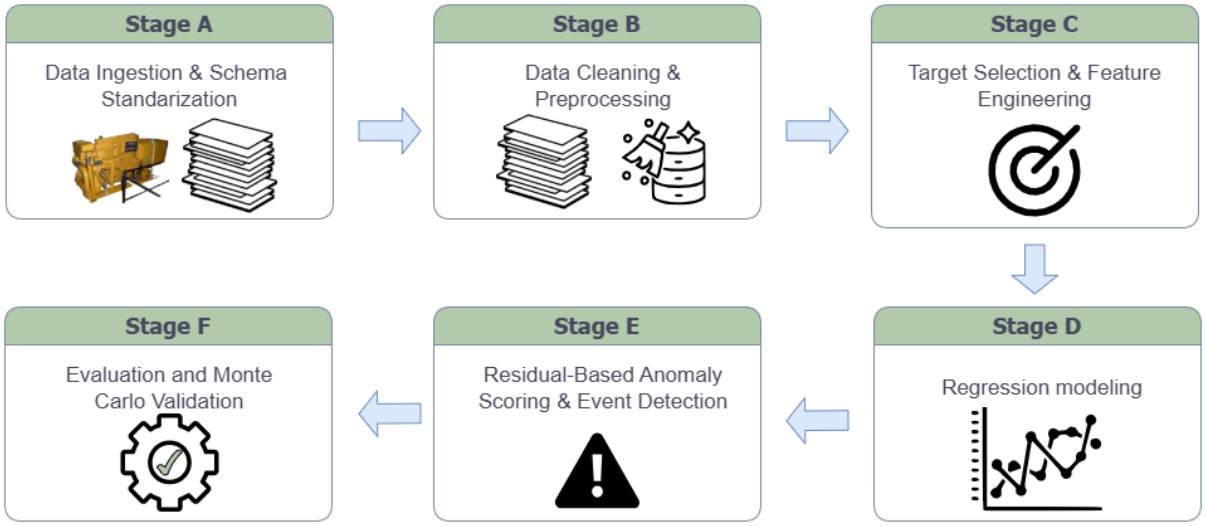


Figure 2. Traceable workflow used to build the anomaly detection system for diesel generators

quency. Stage A explicitly converts all operational channels to numerical format and preserves contextual fields only for auditing, traceability, and chronological ordering during cleaning.

**A3. Temporal discontinuity and decision on time variables.** Although each record contains date/time information, the available coverage does not support treatment as a continuous time series. The logs correspond to operational periods in 2021 and 2023–2025, with a complete year (2022) missing, and additional gaps emerge after removing empty log rows in Stage B. Moreover, the observations come from hourly engineering rounds rather than uninterrupted sensor streams, so temporal spacing is only locally regular and not sufficiently continuous across the full dataset to justify sequence-based modeling assumptions. Consequently, the data are treated as *operational snapshots for modeling purposes*, where each record represents the observed condition of the generator under a given operating state rather than a point within a fully continuous trajectory. For this reason, date and time are not used as model inputs. Similarly, the hourmeter, while conceptually related to accumulated wear, exhibits severe discontinuities under this coverage and is therefore excluded from the feature space in subsequent modeling stages. Time-related fields are nevertheless retained for (i) ordering records during cleaning, (ii) preserving traceability of observations, and (iii) reporting and analyzing detected events in the Results section. This choice avoids imposing artificial temporal continuity on irregular records while remaining consistent with the residual- and persistence-based detection logic adopted later in the workflow.

When the data is collected at non-uniform or irregular sampling intervals, data synchronization needs to be applied to ensure the accuracy of data analysis. Techniques such as dynamic temporal warping (DTW) can be used to align temporally shifted data sequences, but other methods, such as data resampling, are also commonly used, depending on the nature of the non-uniformity. It utilizes low-sampling-rate sensor measurements collected from manufacturing equipment indicators and treats them as feature-level condition indicators. These indicators provide coarse yet operationally meaningful snapshots of system behavior. Incoming observations are initially processed through operating state classification, which distinguishes idle-state data from operating state data. Idle-state observations occur during periods of stable system behavior with minimal load or steady operating conditions, where sensor measurements show relatively low variance. This classification step does not involve anomaly detection; rather, it determines how each observation will be processed subsequently (Song & Kim, 2026).

Missing data are defined as two types: incomplete data and complete missing data. Incomplete data refer to situations where there are parts of the sensor monitoring data that are missing, usually due to sensor failures or other limitations that result in the loss of certain data segments, but still result in partially valid data. In contrast, complete missing data refer to a situation where the sensor fails to monitor any data, leaving the relevant portion completely empty. This is usually caused by equipment failures, communication breakdowns, or other serious problems that prevent access to any valid information (Teng, Yi, & Wang, 2025).

The omission method removes samples with missing values from further analysis. Although simple to apply, this strategy can substantially reduce the effective sample size. It can also break the continuity of the available records, making temporal information harder to analyze (Zhang & Thorburn, 2022).

At the end of Stage A, the output is a standardized dataset for each generator. Each dataset has consistent column types and a reproducible schema, which enables traceable cleaning in Stage B and feature–target specification in Stage C.

### 4.4. Stage B: Cleaning and preprocessing (CRISP-DM: data preparation)

Stage B transforms the standardized tables into an *operationally valid* dataset suitable for nominal modeling. The cleaning strategy is conservative and explicitly traceable, reflecting two constraints of shipboard manual logs: (i) they contain digitization/transcription artifacts and (ii) they do not form a continuous, uniformly sampled time series.

The assimilated dataset was further processed through a comprehensive outlier detection layer, comprising several submethods for detecting different types of outliers in the data time series. These methods focused on 4 types of outliers:

(a) Obvious outliers (samples violating min/max physical limits of individual data variables);

(b) Repeated values;

(c) Drop-outs (missing values); and

(d) Spikes (sudden jumps in time series) (Kim et al., 2025).

**B1. Scope of cleaning: operational variables only.** Cleaning and preprocessing focus on the **operational variables** (mechanical, thermal, and electrical measurements), as these channels encode the generator's physical state and are required for nominal modeling and residual monitoring. Contextual text fields (e.g., navigation origin and destination, port and anchorage descriptors) are retained for descriptive analysis and later event interpretation. However, these fields are not part of the numerical cleaning process and are not used as direct model inputs. In addition, Stage B acknowledges that some channels are governed (e.g., RPM and frequency) and exhibit low variability, while electrical load proxies (kW and amperes) are highly redundant; these considerations guide feature selection in Stage C but do not prevent using the governed variables as *monitoring* variables during cleaning and plausibility checks.

**B2. Manual correction of evident digitization errors.** The first and most impactful step is a targeted manual correction of values that are incompatible with the equipment's physics or the local operating context. The procedure follows three explicit criteria:

1. **Statistical plausibility:** assess the frequency and distribution of valid values for each variable to identify entries that are clearly outside the typical support.
2. **Operational ranges:** use engineering plausibility ranges defined by ship maintenance personnel, reviewed against manufacturer-consistent operating expectations and the observed operating support of the data, as guardrails for plausibility.
3. **Local temporal coherence:** verify the suspected entry against immediately preceding and following records to decide whether a corrected value can be inferred without introducing speculative assumptions.

This step addresses typical errors such as repeated digits (e.g., RPM recorded as 18101) or misplaced decimals (e.g., exhaust temperature recorded as 23°C). When a correction cannot be inferred reliably (e.g., kW and amperes recorded as zero while thermal and mechanical channels indicate active operation), the record is removed to avoid injecting contradictions into the learned nominal behavior.

The first step of data processing is data cleaning. This is an important step since data, especially event data, which is usually entered manually, always contains errors. Data cleaning ensures, or at least increases the chance, that clean (error-free) data are used for further analysis and modelling. Data errors are caused by many factors including the human factor mentioned above. For condition monitoring data, data errors may be caused by sensor faults. Sometimes it requires manual examination of data. Graphical tools would be very helpful to finding and removing data errors (Jardine et al., 2006).

**B3. Missingness handling and controlled interpolation.** Moreover, ship performance datasets are known to be highly unbalanced, leading to the formation of such sparse regions (Kim et al., 2025).

After manual correction, fully empty rows are removed (common in handwritten logs where dates may be written continuously, but measurements are absent). This removal introduces additional *interruptions* in the time axis, reinforcing the decision in Stage A to avoid sequence-based modeling assumptions.

This phenomenon is characterized by scattered and intermittent gaps in the dataset, which are of extremely short duration. Sporadic missing is commonly caused by transient issues such as fiber optic transmission failures. These brief gaps are not confined to specific segments but are dispersed throughout the entire dataset (Xiao et al., 2025).

When the data is collected at non-uniform or irregular sampling intervals, data synchronization needs to be applied to ensure the accuracy of data analysis. Techniques such as dynamic temporal warping (DTW) can be used to align temporally shifted data sequences, but other methods, such as data resampling, are also commonly used, depending on the nature of the non-uniformity. Therefore, the application of

robust data processing techniques, such as statistical filtering and regular sensor calibration, is essential to improve the signal-to-noise ratio and overall trustworthiness of in-service data (Kim et al., 2025).

Additionally, domain knowledge, such as the ship's logs, sensor operation methods, and weather data, can be utilised to infer the missingness mechanism. Handling such missing values is an essential preprocessing step to improve the quality of data, increase the reliability of data-driven decision-making, and maximise the performance of predictive models, beyond simply filling in the data (Kim et al., 2025).

Understanding the missing mechanism is crucial as it reveals the pattern or structure of the missing values. Additionally, the Missing Rate (MR), calculated as the proportion of missing values to the total number of data entries, is a key metric (Xiao et al., 2025).

Missing data are a pervasive issue in data-driven modelling. The absence of data could cause bias in the statistical analysis, leading to invalid conclusions. Moreover, the lost data makes many data modelling techniques ineffective because they resume complete information for all the variables included. Hence, efficient ways of handling the missing data are urgently needed. Common known methods to deal with missing values range from data omission to sophisticated imputation algorithms (Xiao et al., 2025).

For internal gaps within numeric operational variables, Stage B applies linear interpolation **only** for short gaps of at most two consecutive missing samples. Larger gaps are not interpolated to prevent synthetic trajectories from dominating the statistical structure of the data set.

**B4. Hard plausibility filtering using physical operating ranges.** Next, a hard filter is applied using engineering plausibility ranges for the Caterpillar 3406C TA generator set. These ranges were defined from maintenance-personnel recommendations, cross-checked against manufacturer-consistent operating expectations and the observed support of the operational datasets. They act as a first-order consistency check to remove values that remain physically implausible after manual corrections. Table 2 reports the ranges used.

**B5. Multivariate inconsistency treatment via LOF (as a cleaning tool).** PCA is a well-known outlier detection algorithm, however, it is not proven effective for all types of data errors and outliers. Thus, for a more robust approach, specific filters/detection methods should be used as the first step of the data cleaning process and advanced methods like PCA should be used to detect more complex outliers (Kim et al., 2025).

Even if a value lies inside the hard physical limits, it can be inconsistent in the multivariate operating space (e.g., atypical combinations of thermal and load variables). To treat these cases, Stage B uses the Local Outlier Factor (LOF) as a *data-cleaning* mechanism (not as a final anomaly detector). LOF is configured as shown in Table 3. Observations flagged as LOF outliers are temporarily set to missing and then reconstructed using the same controlled interpolation rule (gap $\leq$ 2) described in B3.

Table 2. Normal operating ranges used as physical plausibility limits for Stage B preprocessing (Caterpillar 3406C TA generator set).

| Variable | Min | Max |
|---|---|---|
| Angular Speed (rpm) | 1795 | 1810 |
| Oil pressure | 40 psi | 88 psi |
| Fuel pressure | 25 psi | 45 psi |
| Exhaust temperature | 220°C | 360°C |
| Lube-oil temperature | 40°C | 105°C |
| Cooling-water temperature | 35°C | 95°C |
| Air temperature | N/A | 90°C |
| Battery voltage | 24 V | 26 V |
| Current (amperes) | N/A | 426 A |
| Active power (kW) | N/A | 260 kW |
| Frequency | 59.5 Hz | 60.5 Hz |

Table 3. LOF hyperparameters used in Stage B for multivariate inconsistency treatment (cleaning).

| Parameter | Value |
|---|---|
| Number of neighbors | 20 |
| Contamination | 0.03 |
| Distance metric | Euclidean (Minkowski, $p = 2$) |

**B6. Traceability log of the cleaning process.** To guarantee reproducibility and enable auditing, all transformations are recorded as a stepwise traceability log. Table 4 summarizes, for each generator, the number of rows before/after each step and the magnitude of corrections applied.

Table 4 shows that data cleaning had a strong effect on both datasets, especially on MG#2. This effect should be interpreted with care. In MG#2, the reduction in usable rows is due mostly to the large number of records with date information present but most monitored variables left blank. In other words, many rows existed structurally in the logs, but they did not contain enough engineering information to support correction or modeling. As a result, the impact observed in Table 4 reflects not only preprocessing decisions, but also the high proportion of null and near-empty records in the original digitized source.

This introduces an important limitation. The final datasets depend on the cleaning and correction rules applied during pre-

Table 4. Traceability log of Stage B cleaning process for MG#1 and MG#2 (rows before/after and rows affected per step).

| Step | Rows before | Rows after | Rows affected |
|---|---|---|---|
| Initial state (format standardization) | 7405 | 7405 | 0 |
| Manual value corrections | 7405 | 5404 | 2001 |
| Automatic interpolation (gap $\leq 2$) | 5404 | 5404 | 8 |
| Physical range filtering | 5404 | 5367 | 37 |
| LOF-based inconsistency treatment | 5367 | 5367 | 115 |
| **MG#2** | | | |
| Initial state (format standardization) | 6126 | 6126 | 0 |
| Manual value corrections | 6126 | 3042 | 3084 |
| Automatic interpolation (gap $\leq 2$) | 3042 | 3042 | 0 |
| Physical range filtering | 3042 | 3014 | 26 |
| LOF-based inconsistency treatment | 3014 | 3014 | 34 |

processing. For MG#2, they are also constrained by the limited raw information available in the original logs. Therefore, the cleaned datasets should be interpreted as operationally usable inputs for nominal modeling, not as neutral replicas of the raw records.

At the end of Stage B, the cleaned datasets for MG#1 and MG#2 provide the basis for Stage C. In that stage, targets and features are defined for nominal regression modeling, and in Stage E, residual-based anomaly scores and persistence rules are applied.

### 4.5. Stage C: Target definition and feature specification (CRISP-DM: data understanding → preparation)

Stage C formalizes the learning problem as three supervised regression tasks for nominal modeling, using as targets the exhaust temperature, cooling water temperature, and lubricating oil temperature. These variables were selected because they are physically meaningful indicators of generator thermal condition and are expected to respond to changes in load, combustion quality, and heat-transfer performance.

The variable selection also deserves careful consideration. Including unnecessary variables can make the data noisier and reduce the effectiveness of the fault detection model. A generic guide is to include only those variables that can assist in early fault detection; a variable that does not show any change in behavior under the influence of process faults of interest should be excluded. Data pre-processing includes, amongst others, identification and removal of outliers, noise reduction, transformation of variables, and extraction of features. The overall objective of this step is to increase the 'information content' of training dataset so that the PM model's ability to distinguish between normal and faulty operations is bolstered (Kumar & Flores-Cerrillo, 2024).

After cleaning and harmonizing the numeric operational variables, the final implementation adopted a *common reduced input schema* for the three targets. Rather than maximizing the number of candidate predictors, the objective was to preserve variables that are physically interpretable, routinely available, and sufficiently informative for expected thermal behavior, while excluding channels that could introduce redundancy, leakage, or limited additional value. In the final feature design, the retained predictors were: oil pressure, fuel pressure, intake-air temperature, battery voltage, and active power (`kW`).

This decision reflects three main considerations. First, only one explicit load proxy was retained: `kW`. Although current (`Amperes`) is also associated with load, both variables are strongly related, so keeping both would add redundancy without clear benefit. Second, governed channels such as engine speed (`RPM`) and electrical frequency were excluded because they exhibit limited variability under the operating regime of the generator and therefore contribute little discriminative information for regression. Third, the remaining thermal variables were not used as predictors of each other in the final schema. Although they are correlated, including one thermal target to predict another could blur the interpretation of the residuals and weaken the causal meaning of the downstream anomaly detector. For this reason, the final regression models estimate each thermal target from a compact set of non-target operational inputs rather than from the other thermal responses.

Table 5. Final target set and common input schema used for the nominal regression modeling.

| Target | Input schema used in the final implementation |
|---|---|
| Exhaust temp. | Oil pressure, Fuel pressure, Air temperature, Battery voltage, Power (`kW`) |
| Cooling-water temp. | Oil pressure, Fuel pressure, Air temperature, Battery voltage, Power (`kW`) |
| Lubricating-oil temp. | Oil pressure, Fuel pressure, Air temperature, Battery voltage, Power (`kW`) |

The three target models used the same predictor subset, preserving a consistent inference schema from nominal modeling to residual monitoring. Table 5 summarizes the final feature specification used in the implemented pipeline.

### 4.6. Stage D: Nominal regression modeling (CRISP-DM: modeling)

Stage D trains supervised regression models to estimate the expected thermal response of generators under healthy operating conditions. Let $\mathbf{x}_t \in \mathbb{R}^m$ denote the operational input vector for record $t$, and let $y_{k,t}$ be the observed value of the thermal target $k$, where $k \in \{\text{Exhaust}, \text{Water}, \text{Oil}\}$. For each target, a predictor $\hat{f}_k(\cdot)$ is learned to map the current operating condition to the expected healthy value, as expressed in Eq. (1):

$$\hat{y}_{k,t} = \hat{f}_k(\mathbf{x}_t). \tag{1}$$

This formulation is consistent with a residual-based anomaly detection strategy where the regression model is not intended to classify faults directly, but to provide a reference estimate of nominal thermal behavior against which deviations can later be monitored.

**D1. Model portfolio and target-specific input schema.** To avoid committing a priori to a single model family, a portfolio of regression algorithms was evaluated for each target variable. The candidate set included linear baselines and nonlinear learners capable of capturing interactions between load and thermal response, including linear regression, regularized linear models, support vector regression, $k$-nearest neighbors, random forests, and gradient-boosting-based regressors. Each target was modeled independently using a target-specific input schema derived from the cleaned operational variables and excluding the target itself and other variables deemed unsuitable as direct predictors. This design preserves interpretability and avoids leakage from variables that could encode the response too directly.

Parameter-estimation methods infer model parameters from measured input–output data and use deviations from the estimated model as diagnostic evidence. These methods can be applied to a wide range of engineering systems because they provide information about system dynamics without requiring direct measurement of all internal process variables. They are especially useful for component fault detection and can also support sensor and actuator fault detection. Their effectiveness has been demonstrated in industrial processes and automated control systems (Ahmad & Mohd-Mokhtar, 2022).

**D2. Common pipeline and data partitioning.** All candidate regressors were trained within a common scikit-learn pipeline composed of: (i) feature scaling through StandardScaler, followed by (ii) the regression model. Although scaling is not required by all learners, its inclusion ensures comparability across model families and is particularly important for scale-sensitive methods such as SVR, KNN, and penalized linear models.

Regression analysis supports hypothesis testing, confidence intervals for coefficients, and interpretable inference. Its main strengths are simplicity, parameter interpretability, availability of closed-form estimators in the linear-Gaussian case, and well-established diagnostic tools. However, its limitations include reliance on linearity, independence, and homoscedasticity assumptions, sensitivity to extrapolation beyond the observed range, and possible misspecification when degradation dynamics are nonlinear or regime-switching (Jarosz-Kozyro & Baranowski, 2025).

For each target, the healthy dataset was partitioned into three subsets: a training subset for model fitting, a calibration subset for residual-baseline estimation, and an internal test subset for offline predictive assessment. To avoid temporal data leakage, cleaned records were ordered chronologically within each generator before partitioning. The split was performed by contiguous record blocks rather than by random mixing: earlier records were used for model fitting, a subsequent block was used for residual-baseline calibration, and the final held-out block was reserved for internal testing. The test block was not used to train regressors, estimate residual baseline parameters, tune EWMA settings, or define persistence rules.

Thus, Stage D serves a dual role: it produces the predictive model itself and also preserves a target-specific healthy reference needed later for the statistical calibration of the residual detector in Stage E. Model performance on the internal test subset was assessed using MSE, RMSE, and $R^2$:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2, \tag{2}$$

$$\text{RMSE} = \sqrt{\text{MSE}}, \tag{3}$$

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}. \tag{4}$$

These metrics quantify how well the model reproduces normal behavior, which is a prerequisite for residual-based monitoring. If the predictor is unstable or systematically biased under healthy conditions, the residual will mix modeling error with potential fault effects and become less informative for anomaly detection.

**D3. Model selection, residual diagnostics, and exported artifacts.** For each target, the retained nominal model was

selected from the best-performing candidates according to internal-test RMSE, $R^2$, and residual stability. RMSE was used as the primary predictive screening metric, while $R^2$ and residual dispersion were used as complementary criteria to verify that the selected model provided a suitable nominal baseline for residual monitoring. In addition to standard predictive metrics, residual behavior was inspected to verify that the healthy prediction error remained sufficiently centered and bounded to support downstream statistical monitoring. This diagnostic step is essential because, in a residual-based framework, an isolated large prediction error cannot be interpreted directly as a fault; rather, anomaly inference depends on whether the observed residuals remain compatible with the residual distribution expected under nominal operation.

To make the model-selection process transparent, Table 6 summarizes the leading nominal regression candidates for each monitored target. The table reports internal-test RMSE, $R^2$, and test residual standard deviation, together with the final model retained for residual generation. The selection was performed independently for each thermal target because the three monitored responses exhibited different predictive difficulty and residual behavior. The label RandomForest (baseline) denotes the default Random Forest configuration used as an internal reference, whereas RandomForest (100 trees) and RandomForest (200 trees) denote explicitly parameterized alternatives.

Table 6 shows that the selected nominal regressors were tree-based models for the three monitored thermal targets. For exhaust temperature, RandomForest with 100 trees was retained because it provided a strong overall screening balance, with internal-test performance practically equivalent to the 200-tree variant while preserving a compact and stable residual baseline. For cooling-water temperature and lubricating-oil temperature, RandomForest with 200 trees was retained because it achieved the best or near-best predictive performance among the leading candidates, with the lowest internal-test RMSE and stable residual dispersion. These results make explicit that the residual-based detector was not built on an unspecified or generic regression model; instead, each monitored target used a selected nominal model whose predictive adequacy and residual behavior were verified before Stage E.

The final exported artifacts for each target included the fitted regression pipeline, the input schema, the internal performance metrics, and the calibration subset later used to estimate the healthy residual baseline for Stage E. Figure 3 compares the residual distributions obtained from the calibration and internal test subsets for the three thermal targets. The close overlap between calibration and test residuals, together with their concentration around zero, supports the assumption that the healthy residual baseline is sufficiently stable to serve as the reference for subsequent EWMA-based anomaly monitoring.

### 4.7. Stage E: Residual-based anomaly scoring and persistent event detection (CRISP-DM: modeling → evaluation)

Stage E transforms the regression outputs into an anomaly detection signal by monitoring the residuals relative to the healthy baseline characterized in Stage D. As shown by the residual diagnostics in the calibration and internal test subsets, the selected models produce prediction errors that remain reasonably centered and stable under normal operation. Based on this result, Stage E assumes that healthy residuals should stay within a statistically consistent band, while abnormal conditions are expected to induce deviations that are not only larger in magnitude, but also *persistent* over successive records.

Under low-sampling-rate monitoring, permanent damage or degradation in manufacturing equipment typically manifests as small but systematic deviations from stable idle-state behavior. As a result, deviations induced by persistent degradation become more consistently represented across time and across heterogeneous sensor features, even when their absolute magnitudes differ. The resulting deviation sequence therefore provides a stable and interpretable basis for alarm decision-making under low-frequency monitoring. This sequence is used as the input to the persistence-based alarm decision mechanism described in the next subsection, where sustained deviation behavior is explicitly considered (Song & Kim, 2026).

Residual evaluation refers to the process of extracting fault information from residuals to differentiate between fault and disturbance. Hence, a model-based fault detection system consists of two subsystems: residual generation, residual evaluation, and threshold computation. In case of fault occurrence, an alarm is generated to notify the operator, or some control action is taken to compensate the effect of fault for the smooth operation of an entire system. The process of modifying the control action according to fault nature is called fault-tolerant control (FTC). Fault detection system should be robust against all undesired inputs such as process and measurement disturbance. In a parity-space-based approach, a residual is generated by eliminating the effect of initial states of a dynamic system and utilizing only the system input and measurement data within a finite time window. The inconsistency arises in the residual in case of abnormal behavior evolving in the system dynamics. The core of the parameter-estimation technique is based on system identification by utilizing the system's measured input and output data. In this technique, system parameters of a practical system are identified either offline or online under the normal operating condition while assuming that the fault is reflected in the system's physical parameters. Any discrepancy in process parameters

Table 6. Predictive screening and selected nominal regression model by monitored target. RMSE and $R^2$ were computed on the internal test subset.

| Target | Candidate model | Test RMSE | $R^2$ | Test residual std. | Selected |
|---|---|---|---|---|---|
| Exhaust temp. | RandomForest (100 trees) | 9.088 | 0.806 | 9.088 | Yes |
| Exhaust temp. | RandomForest (200 trees) | 9.097 | 0.806 | 9.098 | No |
| Exhaust temp. | XGBRegressor (baseline) | 9.269 | 0.798 | 9.270 | No |
| Cooling-water temp. | RandomForest (200 trees) | 1.343 | 0.820 | 1.343 | Yes |
| Cooling-water temp. | RandomForest (100 trees) | 1.343 | 0.820 | 1.344 | No |
| Cooling-water temp. | RandomForest (baseline) | 1.352 | 0.818 | 1.352 | No |
| Lubricating-oil temp. | RandomForest (200 trees) | 1.728 | 0.765 | 1.728 | Yes |
| Lubricating-oil temp. | RandomForest (100 trees) | 1.734 | 0.763 | 1.734 | No |
| Lubricating-oil temp. | RandomForest (baseline) | 1.796 | 0.746 | 1.795 | No |

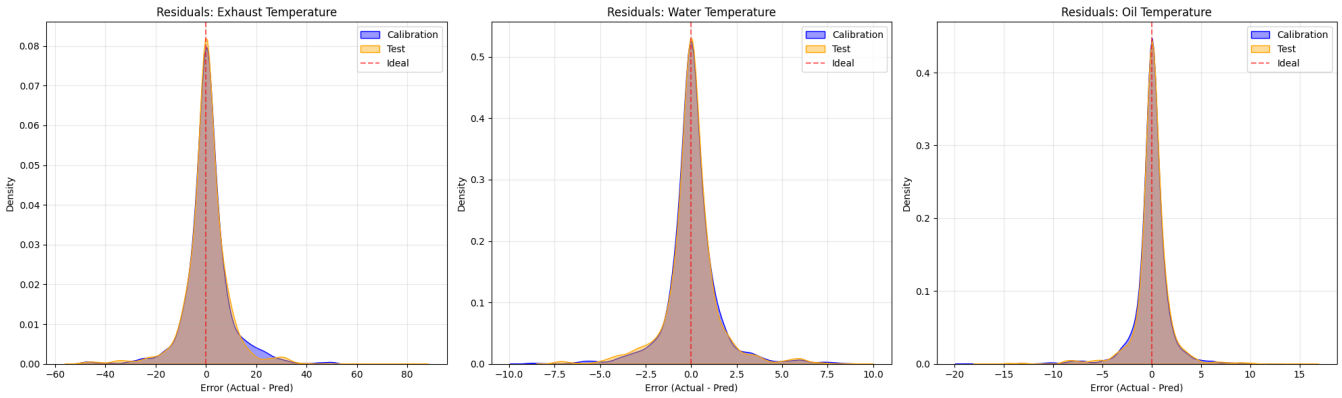


Figure 3. Residual distributions of the selected nominal models in the calibration and internal test subsets for exhaust, cooling-water, and lubricating-oil temperature.

indicates the change/fault in the system (Ahmad & Mohd-Mokhtar, 2022).

**E1. Residual definition.** The robustness of the fault detection method is checked by the measure of sensitivity to noise, disturbance, and uncertainty. Likewise, robustness and the performance of the fault detection method are determined in terms of FAR and MDR. Based on the above survey, the following points are highlighted: An observer generates the residual for fault detection, and it is synthesized to zero in fault-free cases. Observer-based and parity-space-based FD methods produce the alike residual in terms of residual characteristics. However, the observer-based method shows more robustness to uncertainty as compared to the parity-space-based residual generator (Ahmad & Mohd-Mokhtar, 2022).

For each thermal target $k$ at record $t$, the residual is defined as the difference between the observed value and the value expected under nominal operation, as given in Eq. (5):

$$r_{k,t} = y_{k,t} - \hat{y}_{k,t}. \tag{5}$$

This residual is the basic anomaly-sensitive quantity monitored in Stage E. However, a nonzero residual is not interpreted directly as a fault, since part of it corresponds to the intrinsic prediction error of the regression model. Instead, anomaly inference is based on whether the residual sequence becomes statistically inconsistent with the healthy residual baseline characterized from the calibration subset in Stage D.

**E2. Noise-robust monitoring via EWMA.** Manual rounds and operational variability may introduce isolated residual fluctuations. To reduce this effect, the residual sequence is smoothed using an exponentially weighted moving average (EWMA). For each target $k$, the monitoring statistic is updated as

$$z_{k,t} = \lambda r_{k,t} + (1-\lambda) z_{k,t-1}, \qquad z_{k,0} = \mu_{k,0}, \qquad \lambda \in (0,1]. \tag{6}$$

Here, $\mu_{k,0}$ is the healthy residual center estimated from the calibration subset. In the implemented detector, the EWMA statistic is initialized at this healthy center rather than at the first observed residual. This avoids artificial start-up bias and keeps the initial monitoring behavior consistent with the nominal residual reference.

The smoothing parameter $\lambda$ in Eq. (6) controls detector memory: smaller values produce stronger smoothing and reduce sensitivity to isolated spikes, whereas larger values react faster to abrupt shifts. In this study, $\lambda$ was not treated as a universal constant, but as a detector parameter selected through a limited offline sensitivity search. Alternative smoothing factors were compared by examining their effect on event detection, false-alarm burden, and detection latency under simulated abnormal scenarios. Lower $\lambda$ values reduced reaction to isolated residual spikes but tended to delay sustained alarm activation, while higher values increased responsiveness at the cost of greater sensitivity to local residual variability. Based on this trade-off, $\lambda = 0.2$ was retained as the reference smoothing factor for the benchmark.

Failures in capturing and transmitting data, obtained from sensors, normally occur in one capture and are normalized in the next. The reason is that the disturbances related to these problems are mostly transient, caused by imbalances that lead to momentary instabilities. Thus, when interference occurs in a single measurement, generating a single incorrect value, there is a high probability that these data will be misinterpreted. When considering subsequences of data, if any value of the subsequence is discrepant from the rest of the data, the subsequence is interpreted as a failure (collective outlier), otherwise, as an event (contextual outlier). When a measurement affected by a fault is analyzed as a time series (slice), the sub-sequence corresponding to that fault will be different from the normal trend of the entire time series. However, this will only be true if the deviation from normality is large enough to influence the measurement of the entire segment, which occurs in persistent failures. In the case of transient failures, a single abnormal data point, the influence on series measurement is less significant. Thus, when analyzed in a data segment, the disturbance caused is mitigated due to the presence of several other normal values (Costa, Nassar, & Dantas, 2022).

Statistical Process Control (SPC) is the application of statistical tools to monitor and improve process quality [1]. Nowadays, various controls are used not only in industrial production, but also in many other industries. Statistical quality control is an important tool widely used in the service delivery field to monitor the entire operation. The Exponentially Weighted Moving Average (EWMA) gained further refinement allowing simultaneous monitoring of process mean and variability (Ferezagia, Kesa, Sellyra, Anggara, & Lee, 2025).

**E3. Time-varying control limits and pointwise exceedance.** For each target, the healthy residual baseline is represented by two parameters: the center $\mu_{k,0}$ and the residual dispersion $\sigma_{k,0}$, both estimated from the calibration subset reserved in Stage D. Given the EWMA recursion in Eq. (6), the corresponding standard deviation of the EWMA statistic at time $t$ is computed as shown in Eq. (7):

$$\sigma_{k,t}^{\mathrm{EWMA}} = \sigma_{k,0}\sqrt{\frac{\lambda}{2-\lambda}\,[1-(1-\lambda)^{2t}]}, \tag{7}$$

when time-varying limits are enabled. This formulation narrows the limits during the initial steps and lets them converge gradually to the steady-state EWMA variance. Accordingly, the upper and lower control limits are defined by Eq. (8):

$$\mathrm{UCL}_{k,t} = \mu_{k,0} + L\,\sigma_{k,t}^{\mathrm{EWMA}}, \qquad \mathrm{LCL}_{k,t} = \mu_{k,0} - L\,\sigma_{k,t}^{\mathrm{EWMA}}, \tag{8}$$

The multiplier $L$ was selected together with $\lambda$ during the same offline sensitivity search. Smaller $L$ values make the detector more sensitive to residual deviations, but also increase the probability of false alarms under local variability in the healthy residual baseline. Larger $L$ values reduce false alarms, but may miss weaker or slowly developing deviations and increase detection latency. The adopted value $L = 3.0$ was therefore retained as a conservative control-limit setting, consistent with the objective of limiting nuisance alarms in manually recorded shipboard data. A pointwise exceedance indicator is then defined by Eq. (9):

$$a_{k,t} = \mathbb{I}(z_{k,t} > \mathrm{UCL}_{k,t} \quad \vee \quad z_{k,t} < \mathrm{LCL}_{k,t})\,. \tag{9}$$

Thus, the detector does not operate directly on raw residual magnitude alone, but on the smoothed residual statistic in Eq. (6), evaluated against the time-consistent control limits in Eq. (8). Figure 4 illustrates this process for exhaust temperature under a simulated injector-wear scenario, showing the observed and predicted signals, the EWMA trajectory, the control limits, and the resulting alarm behavior.

**E4. Persistence-based event declaration.** Anomaly scoring and event declaration are carried out independently for each monitored target. To reduce false alarms caused by isolated limit crossings, the detector requires a minimum number of consecutive exceedances before declaring a sustained alarm. Let $c_{k,t}$ denote the running count of consecutive violations for target $k$, as defined in Eq. (10):

$$c_{k,t} = \begin{cases} c_{k,t-1} + 1, & \text{if } a_{k,t} = 1, \\ 0, & \text{if } a_{k,t} = 0. \end{cases} \tag{10}$$

A sustained alarm is declared only when this counter reaches a predefined minimum run length $d$, as shown in Eq. (11):

$$\mathrm{Alarm}_{k,t} = \mathbb{I}\,(c_{k,t} \geq d)\,. \tag{11}$$

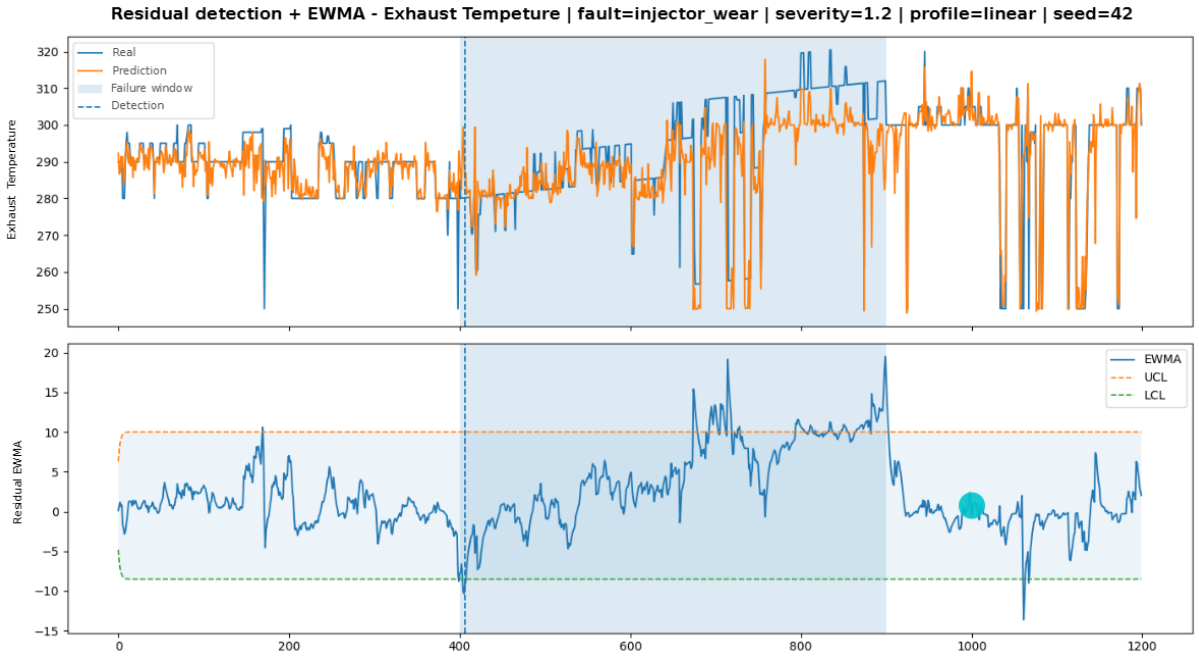


Figure 4. Residual-based anomaly monitoring for exhaust temperature under simulated injector wear, showing observed and predicted signals, EWMA evolution, time-varying control limits, and persistent alarm activation.

The persistence length $d$ was also included in the offline sensitivity search. Shorter persistence requirements increase sensitivity and may reduce detection delay, but they also allow isolated threshold crossings to become alarms. Longer persistence requirements reduce nuisance alarms, but they delay event declaration and may reduce sensitivity to short abnormal intervals. The retained value $d = 3$ was selected as a conservative compromise: it requires sustained evidence across consecutive records while keeping the alarm logic compatible with the hourly manual-round sampling scheme. This rule is consistent with the operational rationale of the detector: a single large residual may still be attributable to transient noise or model imprecision, whereas repeated out-of-limit EWMA values, as identified through Eqs. (9) and (10), provide stronger evidence that the residual sequence is no longer compatible with healthy behavior. Consequently, event declaration is based on persistence of statistically abnormal residual behavior rather than on isolated prediction errors.

Overall, the residuals and detector parameters were validated as part of the complete monitoring configuration rather than as isolated quantities. Residual suitability was first checked through calibration and internal-test residual diagnostics, verifying that healthy residuals remained centered and sufficiently stable to define a nominal baseline. The EWMA parameters $\lambda$ and $L$, together with the persistence length $d$, were then examined through offline sensitivity runs under simulated abnormal scenarios. The retained configuration, $\lambda = 0.2$, $L = 3.0$, and $d = 3$, provided a conservative balance between sensitivity, false-alarm control, and detection latency for the manually recorded shipboard dataset. These values should therefore be interpreted as benchmark-tuned reference settings, not as universally optimal constants for other vessels, fleets, or sampling conditions.

To make the parameter-selection process explicit, Table 7 reports the offline sensitivity analysis conducted for the three monitored thermal targets. The sweep varied the EWMA smoothing factor $\lambda$, the control-limit multiplier $L$, and the persistence length $d$. Each configuration was evaluated using the same Monte Carlo evaluation design, target-specific simulated abnormal scenarios, nominal regression models, residual-based monitoring logic, and metric definitions used in the detector benchmark. The objective was not to identify a universal optimum for all metrics, but to examine how detector behavior changed in terms of event detection, latency, false alarms, and point-wise overlap with the simulated abnormal interval.

Table 7 shows that the EWMA and persistence parameters mainly affect latency, false-alarm burden, and point-wise abnormal interval coverage. The effect is most visible for exhaust temperature, where event detection ranged from 0.900 to 0.950 depending on the configuration. For cooling-water and lubricating-oil temperature, event detection remained equal to 1.000 across the evaluated settings, so the parameter choice mainly affected how early the events were detected, how many alarms appeared outside the injected abnormal interval, and how much of the abnormal interval was covered by the sustained alarm state.

Table 7. Offline sensitivity analysis of EWMA and persistence parameters across monitored thermal targets. Det. denotes event detection rate, Lat. detection latency in samples, and FA false alarms.

| Target | $\lambda$ | $L$ | $d$ | Det. | Lat. | FA pre | FA post | Prec. | Rec. | F1 |
|---|---|---|---|---|---|---|---|---|---|---|
| Exhaust temp. | 0.20 | 3.0 | 3 | 0.950 | 169.447 | 5.800 | 5.200 | 0.685 | 0.242 | 0.346 |
| Exhaust temp. | 0.20 | 3.2 | 3 | 0.925 | 173.973 | 4.575 | 4.250 | 0.679 | 0.207 | 0.306 |
| Exhaust temp. | 0.20 | 3.2 | 4 | 0.900 | 177.861 | 3.075 | 3.225 | 0.700 | 0.186 | 0.281 |
| Exhaust temp. | 0.15 | 3.0 | 3 | 0.950 | 161.658 | 7.825 | 7.425 | 0.723 | 0.309 | 0.421 |
| Exhaust temp. | 0.15 | 3.2 | 4 | 0.950 | 177.658 | 4.825 | 5.000 | 0.726 | 0.260 | 0.369 |
| Cooling-water temp. | 0.20 | 3.0 | 3 | 1.000 | 57.125 | 11.275 | 19.350 | 0.876 | 0.836 | 0.853 |
| Cooling-water temp. | 0.20 | 3.2 | 3 | 1.000 | 60.700 | 8.650 | 17.725 | 0.890 | 0.825 | 0.855 |
| Cooling-water temp. | 0.20 | 3.2 | 4 | 1.000 | 62.625 | 6.775 | 16.425 | 0.904 | 0.818 | 0.857 |
| Cooling-water temp. | 0.15 | 3.0 | 3 | 1.000 | 50.900 | 17.250 | 26.750 | 0.845 | 0.857 | 0.849 |
| Cooling-water temp. | 0.15 | 3.2 | 4 | 1.000 | 56.800 | 11.775 | 23.575 | 0.875 | 0.840 | 0.855 |
| Lubricating-oil temp. | 0.20 | 3.0 | 3 | 1.000 | 126.650 | 9.850 | 6.000 | 0.862 | 0.530 | 0.646 |
| Lubricating-oil temp. | 0.20 | 3.2 | 3 | 1.000 | 143.825 | 8.275 | 5.025 | 0.869 | 0.497 | 0.620 |
| Lubricating-oil temp. | 0.20 | 3.2 | 4 | 1.000 | 151.200 | 6.350 | 4.400 | 0.889 | 0.480 | 0.608 |
| Lubricating-oil temp. | 0.15 | 3.0 | 3 | 1.000 | 120.250 | 13.025 | 9.500 | 0.856 | 0.587 | 0.687 |
| Lubricating-oil temp. | 0.15 | 3.2 | 4 | 1.000 | 133.275 | 9.400 | 7.550 | 0.882 | 0.552 | 0.668 |

For exhaust temperature, the more aggressive configuration $\lambda = 0.15$, $L = 3.0$, and $d = 3$ achieved lower latency and higher F1-score than the retained configuration, but it also increased false alarms before and after the injected abnormal interval. Increasing $L$ to 3.2 and increasing $d$ to 4 reduced the false-alarm burden, but delayed detection and reduced recall. A similar pattern was observed for lubricating-oil temperature: the aggressive configuration improved recall and F1-score, but at the cost of more false alarms. For cooling-water temperature, all configurations performed strongly and the differences in F1-score were small; however, stricter settings reduced false alarms while slightly increasing latency.

Based on these results, the retained setting $\lambda = 0.2$, $L = 3.0$, and $d = 3$ was selected as a common benchmark configuration rather than as the numerically best point for every individual metric. It preserves high event detection, competitive latency, and acceptable F1-score while avoiding the higher false-alarm burden of the more aggressive $\lambda = 0.15$, $L = 3.0$, $d = 3$ alternative. This choice also keeps the detector configuration consistent across the three monitored targets, which supports fair comparison in the subsequent baseline benchmark.

### 4.8. Stage F: Evaluation and Monte Carlo validation under scarce labels

A key limitation throughout this study is the lack of fault labels that are both reliable and temporally aligned with the historical records. This limitation prevents the proposed anomaly detection system from being evaluated solely through a standard supervised comparison against real failure cases. In addition, because the records come from hourly manual engineering rounds, detection latency is measured in samples, where one sample corresponds to one recorded operational round. The evaluation therefore interprets early detection as the ability to identify sustained abnormal deviations within the temporal resolution of the available inspection records, rather than as real-time diagnosis from high-frequency sensor streams.

The offline evaluation combines two components: predictive assessment of the nominal models on healthy held-out data, and residual-detector assessment through simulated fault scenarios.

The first part is addressed in Stage D, where the regression models are evaluated on internal test data. This step is necessary, but it is not enough to describe the detector as a whole. The final detection decision depends not only on regression accuracy, but also on how residuals evolve over time, how the EWMA reacts, and how the persistence logic behaves. For that reason, the second part of the evaluation examines the full detection pipeline through repeated Monte Carlo runs. In each run, a healthy operational window is sampled, a fault pattern is injected in a controlled way, and the detector response is recorded.

**F1. Motivation for Monte Carlo-based evaluation.** Monte Carlo-based evaluation was used because the detector is sequential and the available historical records do not include fully reliable, time-aligned fault labels. The result of a single replay can depend on the healthy window sampled, the local residual pattern in that window, the injected degradation profile, and the random seed used during simulation. Repeating the experiment across several seeds and scenario instances therefore makes the analysis less dependent on any single run and provides a more stable characterization of detector behavior.

During Monte Carlo replay, the nominal model, residual baseline, and detector parameters were fixed before fault injection. Simulated abnormal windows were therefore used only for detector evaluation, not for fitting the regression models, estimating the healthy residual distribution, or calibrating the alarm thresholds.

In this work, Monte Carlo-based evaluation is not presented as a substitute for real-fault validation. It is used as an offline procedure to examine sensitivity, detection delay, and false alarms under controlled conditions. Real-fault validation was not performed because the available maintenance records do not provide sufficiently reliable temporal alignment with the hourly engineering-round measurements. As a result, confirmed fault onset times could not be matched consistently with the monitoring records used by the detector.

It is also important to clarify the intended meaning of representativeness in this evaluation. The simulated scenarios were not designed to reproduce the full physical degradation mechanisms of marine diesel generators, nor to claim equivalence with real fault events. Instead, they were designed to reproduce observable data-level signatures that are relevant for residual-based thermal monitoring: sustained positive or negative deviations, progressive drifts, step-like changes, gain-type deviations, and short spikes. These signatures correspond to generic manifestations of abnormal behavior that a detector would observe in the monitored thermal variables, independently of the exact root cause. Therefore, the representativeness of the scenarios is limited to the residual and signal-behavior level, not to full physics-based fault replication.

**F2. Scenario definition and fault injection.** The simulation process starts from healthy operational windows extracted from the cleaned dataset. Controlled perturbations are then added to represent departures from normal behavior while preserving the variability of the original window. Each scenario is defined by a fault type, a start point, an end point, a severity level, and a degradation profile. In the implemented benchmark, severity was treated as a controlled perturbation multiplier rather than as a real fault class. For each monitored target, the detector was evaluated under severity levels of 0.6, 0.8, 1.0, 1.2, and 1.3, covering weak to stronger abnormal deviations. Progressive profiles, such as linear and piecewise growth, were included because many thermal deviations are more likely to develop over several observations than to appear as abrupt step changes.

Constant faults represent cases in which the sensor output remains fixed, regardless of the actual measured value. In this condition, the sensor is effectively stuck at a particular value. Spike faults, in contrast, introduce sudden and large deviations in the sensor readings. These deviations are temporary, short-lived, and may represent sensor glitches or external disturbances (Awaisi et al., 2025).

Gain is a multiplicative sensor fault. It refers to a condition where the sensor output is consistently higher or lower than the expected value by a certain factor, indicating a systematic error in the sensor response. In mathematical terms, a gain fault can be represented by multiplying the healthy sensor value by a constant factor that reflects the deviation from the expected output (Awaisi et al., 2025).

For the exhaust-temperature case, injector-wear scenarios were used as representative examples of thermally relevant abnormal behavior. These simulated faults do not attempt to reproduce the full physical complexity of real injector degradation or combustion deterioration processes. Their role is more limited: they introduce controlled perturbations with interpretable direction and magnitude so that the detector can be tested in a repeatable way.

The selected perturbation profiles were chosen to cover different abnormal-behavior patterns that are relevant for the proposed detector. Linear and piecewise profiles emulate slowly developing deviations, which are representative of incipient thermal deterioration or gradual efficiency loss. Step-like and gain-type perturbations emulate systematic sensor or process shifts, while spike-like perturbations represent short-duration disturbances or measurement artifacts. Consequently, the scenarios provide a controlled and repeatable way to test whether the proposed detector reacts to different classes of abnormal signal behavior. However, they should be interpreted as detector stress cases rather than as validated physical replicas of specific diesel-engine failure modes.

**F3. Experimental protocol.** For each target and fault scenario, the detector is executed repeatedly over healthy windows sampled with different random seeds. Each run follows the same sequence: sample a healthy window, estimate the expected target trajectory with the selected nominal regression model, compute residuals, monitor them with EWMA, and generate event-level and state-level alarms through persistence rules. Repetition across seeds captures variability in both the operating context and replay selection.

The experiments were organized by fault severity, with multiple repetitions for each severity level. This makes it possible to report not only an overall average, but also how detection rate, latency, and false alarms change as the deviation becomes stronger. This matters in scarce-label settings, since detector usefulness depends not only on the detection of severe deviations, but also on its response to weaker ones.

**F4. Evaluation metrics and summary of results.** The evaluation uses two complementary groups of metrics. The first group contains point-wise classification measures: precision, recall, and F1-score. These metrics compare the detector

Table 8. Detector-level baseline comparison across monitored thermal targets under simulated abnormal scenarios.

| Target | Detector | Det. | Lat. | FA pre | FA post | Prec. | Rec. | F1 |
|---|---|---|---|---|---|---|---|---|
| Exhaust temp. | EWMA + persistence | 0.950 | 169.447 | 5.800 | 5.200 | 0.685 | 0.242 | 0.346 |
| Exhaust temp. | Raw residual threshold | 0.150 | 234.000 | 0.050 | 0.275 | 0.150 | 0.002 | 0.004 |
| Exhaust temp. | Sliding z-score | 0.075 | 277.000 | 0.450 | 0.300 | 0.075 | 0.001 | 0.002 |
| Exhaust temp. | LOF on residuals | 0.150 | 247.500 | 0.175 | 0.900 | 0.106 | 0.002 | 0.004 |
| Exhaust temp. | Isolation Forest residuals | 0.600 | 155.333 | 0.575 | 1.350 | 0.448 | 0.036 | 0.064 |
| Cooling-water temp. | EWMA + persistence | 1.000 | 57.125 | 11.275 | 19.350 | 0.876 | 0.836 | 0.853 |
| Cooling-water temp. | Raw residual threshold | 1.000 | 153.075 | 0.300 | 0.625 | 0.963 | 0.537 | 0.679 |
| Cooling-water temp. | Sliding z-score | 0.225 | 149.333 | 0.325 | 2.975 | 0.080 | 0.003 | 0.006 |
| Cooling-water temp. | LOF on residuals | 0.775 | 93.516 | 0.825 | 5.275 | 0.727 | 0.554 | 0.623 |
| Cooling-water temp. | Isolation Forest residuals | 1.000 | 88.000 | 1.000 | 6.175 | 0.938 | 0.728 | 0.812 |
| Lubricating-oil temp. | EWMA + persistence | 1.000 | 126.650 | 9.850 | 6.000 | 0.862 | 0.530 | 0.646 |
| Lubricating-oil temp. | Raw residual threshold | 0.500 | 330.800 | 0.675 | 0.150 | 0.436 | 0.016 | 0.031 |
| Lubricating-oil temp. | Sliding z-score | 0.225 | 250.222 | 0.475 | 1.375 | 0.126 | 0.003 | 0.007 |
| Lubricating-oil temp. | LOF on residuals | 0.500 | 217.700 | 0.700 | 0.675 | 0.454 | 0.127 | 0.173 |
| Lubricating-oil temp. | Isolation Forest residuals | 0.850 | 193.029 | 1.350 | 1.025 | 0.762 | 0.290 | 0.388 |

state with the simulated abnormal interval after fault injection. The second group contains event-oriented indicators: detection rate, detection latency, false alarms before fault onset, and false alarms after the fault window. These measures are needed because point-wise agreement alone does not fully describe detector behavior. A detector may overlap with the abnormal interval and still be of limited operational use if it detects too late or produces too many unnecessary alarms.

Since the implemented detector separates event triggering from sustained alarm-state labeling, both views are useful. Event metrics describe whether the abnormal condition is detected and how long detection takes, whereas point-wise metrics describe how closely the alarm state overlaps with the simulated abnormal interval once the detector has been activated. Detection latency is reported in samples. Because the monitoring records are based on hourly manual rounds, one sample corresponds to one recorded operational round. Thus, latency values should be interpreted as delays in recorded inspection rounds rather than as delays in continuous sensor time. The system is therefore intended to provide early warning at the inspection-record level, not to capture sub-hourly transients or replace real-time sensor-based protection systems.

To assess the proposed detector against simpler alternatives, a baseline comparison was conducted across the three monitored thermal targets. The values in Table 8 are averages over the simulated severity levels evaluated for each target. Therefore, the table is not intended to show monotonic behavior with severity, but to compare overall detector behavior.

The proposed detector was compared with four residual baselines: a fixed threshold on raw residuals, a sliding z-score detector, LOF on residuals, and Isolation Forest on residuals. For each target, all detectors used the same nominal regression model, simulated windows, perturbation settings, residual stream, and persistence length. Therefore, the comparison isolates the effect of the detector logic rather than differences in regression modeling, scenario generation, or replay selection.

Table 8 shows that the proposed detector achieved the highest F1-score for the three monitored thermal targets. For exhaust temperature, all detectors showed limited point-wise recall, reflecting the difficulty of this target under the simulated injector-wear scenarios. However, the proposed detector achieved the highest detection rate and F1-score, whereas the fixed threshold, sliding z-score, and LOF baselines detected only a small fraction of the abnormal replays. Isolation Forest reduced latency for some detected exhaust events, but its recall and F1-score remained substantially lower than those of the proposed detector.

For the cooling-water target, the proposed detector achieved the highest F1-score and the shortest detection latency. Isolation Forest also performed competitively, but with lower recall and F1-score. The fixed residual threshold obtained high precision and a perfect detection rate, but it reacted later and covered a smaller portion of the abnormal interval. For the lubricating-oil target, the proposed detector again provided the best overall balance, achieving perfect event-level detection and the highest F1-score, while the simpler baselines either detected fewer events or produced much lower recall.

The comparison also shows the expected trade-off of the proposed detector. EWMA smoothing and persistence improve sustained abnormal-pattern coverage, but they can increase the number of alarms before or after the injected abnormal interval, especially for cooling-water temperature. Therefore, the proposed method should be interpreted as a sensitivity-oriented early-warning layer rather than a highly selective diagnostic classifier. This result supports the retained residual-based EWMA–persistence detector logic, while also motivating future work on adaptive thresholding or target-specific

tuning to reduce false-alarm burden.

## 5. CONCLUSIONS AND FUTURE WORK

This study developed and evaluated an offline residual-based anomaly-detection workflow for the auxiliary diesel generator engines of the study vessel using hourly engineering-round data.

The results show that nominal regression models can provide usable residual baselines for key thermal variables despite low-frequency and irregular records. Residual monitoring with EWMA smoothing, control limits, and persistence rules enables detection of sustained abnormal patterns under simulated fault scenarios without labeled historical failures. The study also confirms that data quality is central to the workflow, since missing values, transcription inconsistencies, and temporal discontinuities strongly affect the final modeling dataset.

These findings suggest that residual-based detection can provide a practical structure for anomaly monitoring in shipboard contexts where continuous sensing and fault labels are limited. More broadly, the workflow offers a traceable basis for organizing detection logic, preprocessing decisions, and offline evaluation in low-frequency maritime monitoring environments.

The study is limited by the use of manually recorded operational rounds and by the absence of time-aligned fault labels for direct validation on real failure events. The resulting baseline is also sensitive to preprocessing decisions, especially when the source datasets contain a high proportion of incomplete records.

Future work should focus on linking detected events with maintenance findings to improve alarm interpretation and support operational labeling. A second line of work is to extend the current workflow toward a composite health index that integrates residual behavior across variables and operating regimes.

## ACKNOWLEDGMENT

We gratefully acknowledge COTECMAR for providing the space and resources required to conduct this research. We also thank the Technological University of Bolivar (UTB) for granting access to its facilities and computational services used to run the models, which were essential to this work. In addition, we recognize the support of Minciencias through Convocatoria 950 (2024), which funded the scholarship resources that enabled this research.

## REFERENCES

Ahmad, M., & Mohd-Mokhtar, R. (2022). A survey on model-based fault detection techniques for linear time-invariant systems with numerical analysis. *Journal of Science and Technology*.

Awaisi, K. S., Ye, Q., & Sampalli, S. (2025). Set: A shared-encoder transformer scheme for multi-sensor, multi-class fault classification in industrial iot. *IEEE Transactions on Machine Learning in Communications and Networking*.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, *41*(3), 1–58. Retrieved from https://doi.org/10.1145/1541880.1541882 doi: 10.1145/1541880.1541882

Costa, F. S., Nassar, S. M., & Dantas, M. A. (2022). Focuser: A fog online context-aware up-to-date sensor ranking method. *Journal of Sensor and Actuator Networks*.

Diedrich, A., & Niggemann, O. (2022). On residual-based diagnosis of physical systems. *Engineering Applications of Artificial Intelligence*, *109*, 104636. doi: 10.1016/j.engappai.2021.104636

Ferezagia, D. V., Kesa, D. D., Sellyra, E. C., Anggara, D., & Lee, C. W. (2025). Control chart approach to monitor and improve production processes: Maximum exponentially weighted moving average using auxiliary variable (av) and multiple measurement (me). *Statistics, Optimization and Information Computing*.

Isermann, R. (2005). Model-based fault-detection and diagnosis – status and applications. *Annual Reviews in Control*, *29*(1), 71–85. doi: 10.1016/j.arcontrol.2004.12.002

Jardine, A. K. S., Lin, D., & Banjevic, D. (2006). A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mechanical Systems and Signal Processing*, *20*(7), 1483–1510. Retrieved from https://doi.org/10.1016/j.ymssp.2005.09.012 doi: 10.1016/j.ymssp.2005.09.012

Jarosz-Kozyro, A., & Baranowski, J. (2025). Recent advances in data-driven methods for degradation modeling across applications. *Processes*.

Kim, Y., Gupta, P., & Steen, S. (2025). A comprehensive review of data processing for ship performance analysis. *Applied Ocean Research*, *162*, 104737. Retrieved from https://doi.org/10.1016/j.apor.2025.104737 doi: 10.1016/j.apor.2025.104737

Kumar, A., & Flores-Cerrillo, J. (2024). *Machine learning in python for process and equipment condition monitoring, and predictive maintenance*. Online: MLforPSE. Retrieved from https://mlforpse.com/

Lee, J., Kao, H.-A., & Yang, S. (2014). Service innovation and smart analytics for industry 4.0 and big data environment. *Procedia CIRP*, *16*, 3–8. Retrieved from https://doi.org/10.1016/j.procir.2014.02.001 doi: 10.1016/j.procir.2014.02.001

Li, S. C.-X., & Marlin, B. M. (2020). Learning from irregularly-sampled time series: A missing data per-

spective. In *Proceedings of the 37th international conference on machine learning* (Vol. 119, pp. 5937–5946). PMLR.

Llamas Reinoso, J., Martinez-Santos, J. C., & Puertas, E. (2026, January). *Main machinery operational data of the training ship arc gloria (2021–2025).* Zenodo. Retrieved from https://doi.org/10.5281/zenodo.18307281 doi: 10.5281/zenodo.18307281

Michelena, Á., López, V. C., López, F. L., Arce, E., Mendoza García, J., Suárez-García, A., ... Quintián, H. (2023). A fault-detection system approach for the optimization of warship equipment replacement parts based on operation parameters. *Sensors*, *23*(7), 3389. Retrieved from https://doi.org/10.3390/s23073389 doi: 10.3390/s23073389

Montgomery, D. C. (2009). *Introduction to statistical quality control* (6th ed.). Hoboken, NJ: Wiley.

Park, J., & Oh, J. (2023). A machine learning based predictive maintenance algorithm for ship generator engines using engine simulations and collected ship data. *Energy*, *285*, 129269. Retrieved from https://doi.org/10.1016/j.energy.2023.129269 doi: 10.1016/j.energy.2023.129269

Song, J., & Kim, H. S. (2026). Persistence-based absolute relative error for alarm-centric monitoring under low-frequency manufacturing. *Mathematics*.

Teng, Z., Yi, X., & Wang, B. (2025). Dynamic relative advantage-driven multi-fault synergistic diagnosis method for motors under imbalanced missing data rates. *Journal of Dynamics, Monitoring and Diagnostics*.

Vachtsevanos, G., Lewis, F., Roemer, M., Hess, A., & Wu, B. (2006). *Intelligent fault diagnosis and prognosis for engineering systems*. Hoboken, NJ: Wiley.

Velasco-Gallego, C., & Lazakis, I. (2022). RADIS: A real-time anomaly detection intelligent system for fault diagnosis of marine machinery. *Expert Systems with Applications*, *204*, 117634. Retrieved from https://doi.org/10.1016/j.eswa.2022.117634 doi: 10.1016/j.eswa.2022.117634

Xiao, H., Cao, R., Chen, Z., Hong, C., Wang, J., Yao, M., ... Luo, T. (2025). Handling missing data in large-scale tbm datasets: Methods, strategies, and applications.

Zhang, Y., & Thorburn, P. J. (2022). Handling missing data in near real-time environmental monitoring: A system and a review of selected methods. *Future Generation Computer Systems*.

## BIOGRAPHIES

**Luis F. Mendoza Cardona** is a Mechatronics Engineer. He is currently pursuing the M.S. degree in Engineering with an emphasis on Computing at the Universidad Tecnologica de Bolivar in Cartagena, Bolivar, Colombia (since early 2025). He has worked as a reliability engineer at COTECMAR under the Integrated Logistic Support (ILS) framework, focusing on reliability-centered practices and maintenance decision support. His research interests include reliability engineering, maintenance, and data analytics, with a particular focus on Prognostics and Health Management (PHM) approaches.

**Edwin Puertas** is an Artificial Intelligence software architect and Natural Language Processing (NLP) researcher with 20 years of experience spanning academia and industry. He is currently an Associate Professor at the Technological University of Bolívar, where he also serves as Head of the Master's and Doctoral Programs in Engineering. He is an active member of the Artificial Intelligence Standards Committee. His work focuses on bridging academic research and real-world adoption by leading innovative AI projects and designing scalable, production-grade systems that apply advanced machine learning to complex challenges across multiple sectors. His research interests include NLP, human–computer interaction, and natural language understanding, as well as AI-enabled software architectures for data-intensive applications. He has taught courses in AI, NLP, Big Data, and Software Engineering and has led initiatives that connect academic research with industry needs to foster technology transfer and practical impact. Multiple scholarships and awards have supported his contributions. He is a Senior Member of the IEEE and regularly participates in international conferences and workshops, contributing to global standards and best practices in Artificial Intelligence.

**Edwin Paipa-Sanabria** is a Naval Engineer and a Navy officer with professional experience in integrated logistics support (ILS) frameworks and naval sustainment. He has worked on ILS-related initiatives at COTECMAR, contributing to the structuring of lifecycle support strategies for naval assets. He currently serves as a Division Head, leading innovation projects to strengthen operational readiness and maintenance capabilities. His most recent work is focused on Maintenance 5.0 initiatives, aligning digital transformation and human-centered technologies with next-generation maintenance practices in maritime and defense contexts.

**Juan C.Martínez-Santos** received the B.Sc. degree in Electronic Engineering (2001) and the M.Sc. degree in Electrical Power (2004) from the Universidad Industrial de Santander, Bucaramanga, Colombia, and the

Ph.D. degree in Computer Engineering from Northeastern University, Boston, MA, USA (2013). He was a Fulbright Scholar — DNP — Colciencias (2007). He has been with the Technological University of Bolívar since 2004 and has been an Associate Professor since 2007. His research focuses on computer architecture and organization, with applications in hardware support for computer security in multicore and multiprocessor architectures, as well as advanced digital design techniques, including hardware description languages, intellectual property (IP) modules, programmable systems, embedded systems, and hardware/software co-design. Since 2012, he has led an interdisciplinary group with an ICT backbone. He currently teaches in the Faculty of Engineering. He is responsible for courses in Computer Architecture and Assembly (Computer Engineering), Microprocessors (Electrical and Electronics Engineering), Microcontrollers (Mechatronics Engineering), and Advanced Digital Design Techniques (graduate-level, electronics and computer track).