Development of a Remaining Useful Life Prediction Model for Marine Diesel Engine Filtration Systems

Joan Suarez[1], Clara Guimaraes[2], Edwin Paipa[3], Juan Carlos Martinez-Santos[4] and Edwin Puerta[5].

[1, 4, 5] *Universidad Tecnologica de Bolivar, Cartagena, Bolívar, 130001, Colombia*
jloaiza@utb.edu.co, jcmartinezs@utb.edu.co, epuerta@utb.edu.co
[2] *Universidade do Estado do Rio de Janeiro, Rio de Janeiro, Rio de Janeiro, 20550-900, Brasil*
paula.clara@graduacao.uerj.br
[1, 2, 3] *Corporación de Ciencia y Tecnología para el Desarrollo de la Industria Naval, Marítima y Fluvial, Cartagena, Bolívar, 130009, Colombia*
jmsuarez@cotecmar.com, cguimaraes@cotecmar.com, epaipa@cotecmar.com

ABSTRACT

Marine diesel propulsion engines are essential to naval platforms, enabling maneuvering, navigation readiness, and training operations. However, maintenance of propulsion consumables—particularly fuel filtration elements—often remains time-based and corrective despite the growing availability of onboard operational records. This paper presents the development and validation of a Remaining Useful Life (RUL) prediction model for the propulsion engine filtration system of the Colombian Navy (ARC) training ship, aiming to estimate time to replacement for cartridge-based filters.

The proposed approach handles imperfect manual operational data and scarce, non-uniform maintenance labels through a Prognostics and Health Management (PHM) workflow guided by the Cross-Industry Standard Process for Data Mining (CRISP DM). It combines physics-informed data quality control using plausibility bounds, outlier mitigation, and time-series reconstruction; expert validation of representative operating cycles using a Delphi protocol; and event logging to align filter-replacement actions with gap-aware approximations. It was trained supervised regression models using an automated machine learning (AutoML) strategy implemented in PyCaret and refined through hyperparameter optimization in Optuna. A Random Forest Regressor model achieved the best performance, reaching a test root mean squared error (RMSE) of 52.92 hours with a coefficient of determination (R^2) of 0.921.

https://doi.org/10.36001/IJPHM.2026.v17i1.4761

1. INTRODUCTION

Marine vessels rely predominantly on diesel engines for propulsion and onboard power generation, making their continuous and efficient operation central to navigational safety, availability, and fuel economy (Upadrashta & Wijaya, 2025; Mwangi et al., 2025). Propulsion plants operate under harsh and highly variable conditions, including load changes, weather-driven profiles, and constrained inspection windows. Under these conditions, maintenance decisions have immediate consequences for operational readiness, cost, and risk. Despite the increasing presence of shipboard monitoring systems, condition assessment in service is still frequently dominated by threshold-based alarms and human interpretation, which can overlook slowly developing degradation trends that remain below alarm limits (Kocak, 2023).

In this context, the central maintenance challenge addressed in this paper is the lack of a condition-informed planning criterion for filtration components operating under variable and uncertain load regimes. Maritime maintenance practices have historically favored preventive strategies based on time or accumulated operating hours. However, field evidence indicates that a substantial fraction of machinery failures are random or non-age-related, limiting the effectiveness of fixed-interval maintenance in preventing unscheduled downtime (Asimakopoulos, 2023). The rapid adoption of Industrial Internet of Things (IIoT) technologies and data analytics has therefore motivated a shift toward predictive maintenance and condition-based maintenance (CBM), aiming to anticipate degradation, reduce unnecessary interventions, and improve spare-parts planning and operational continuity (Sielaff, Lucke, & Wolf, 2024; Kirketerp-Møller, Hyldgaard, Cai, Dodis, & Rytter, 2025). Prognostics and Health Management (PHM) provides a structured framework for this transi-

tion by combining operational data with diagnostic and prognostic methods, where Remaining Useful Life (RUL) estimation is a key decision-support quantity for maintenance optimization (Upadrashta & Wijaya, 2025; Pugalenthi, Park, Hussain, & Raghavan, 2021).

The purpose of this paper is to develop and validate a CRISP-DM-guided PHM pipeline for estimating the RUL of fuel filtration cartridges in a marine diesel propulsion system under realistic shipboard data constraints. The study focuses on the cartridge replacement task of the Colombian Navy training ship, where maintenance decisions are affected by variable operating regimes, imperfect historical records, and the absence of fully curated maintenance labels. Rather than assuming a laboratory-grade condition-monitoring environment or run-to-failure databases, the proposed approach formulates RUL as an event-driven target derived from validated replacement evidence and accumulated operating hours.

The selected application is operationally relevant because filtration components are critical consumables in diesel propulsion systems. Progressive clogging increases flow resistance, can degrade engine performance, increase fuel consumption, and potentially propagate adverse effects to downstream subsystems (Hagmeyer & Zeiler, 2023; Zeiler & Hagmeyer, 2023). Filter health is often inferred from indirect, load-dependent measurements, such as differential pressures, temperatures, and flow-related variables, whose interpretation is complicated by changing operating conditions and noisy signals (Zeiler & Hagmeyer, 2023). Reliable RUL prediction for filtration cartridges can therefore provide a practical lever to move from conservative, time-based replacement toward condition-informed scheduling, balancing reliability with component utilization.

Practical implementation in operational fleets remains challenging. Shipboard data streams often contain missing values, manual transcription errors, sensor noise, and regime-dependent variability that complicate the extraction of health indicators and the training of robust models (Jeon, Noh, et al., 2021). More critically, labeled fault and run-to-failure data are scarce for large marine engines and their subsystems, which constrains supervised learning and limits the direct transfer of laboratory-developed methods to real deployments (Mwangi et al., 2025; Kim, Antariksa, Handayani, Lee, & Lee, 2021; Kirketerp-Møller et al., 2025). Consequently, the main bottlenecks are frequently traceable label construction, defensible data preprocessing, and evaluation protocols that reflect prospective use rather than optimistic, randomly shuffled sampling.

In the present study, the operational time series used for modeling is available as an open dataset of shipboard operational parameters (Llamas Reinoso, Martinez-Santos, & Puertas, 2026); notably, this source does not include maintenance event records. To enable supervised, event-driven labeling, the filter cartridge replacement log was constructed by manually extracting and consolidating onboard maintenance records, and then aligning the identified events with the operational time series. This setting reflects a common constraint in operational PHM: abundant operational measurements may coexist with maintenance actions recorded in heterogeneous, partially structured sources (Kumar & Flores-Cerrillo, 2024).

Under these constraints, this work contributes an end-to-end PHM workflow for RUL estimation of marine diesel fuel-filtration cartridges under imperfect shipboard data conditions. The main methodological contributions are:

- a CRISP-DM-guided PHM workflow that converts heterogeneous operational records and fragmented maintenance evidence into a traceable prognostic dataset;
- an event-driven RUL labeling strategy anchored to validated cartridge-replacement actions and accumulated engine operating hours;
- a physics-informed preprocessing procedure combining plausibility bounds, outlier mitigation, missing-value reconstruction, and expert validation of representative operating cycles;
- a leakage-aware modeling and evaluation protocol based on time-consistent partitions, AutoML benchmarking, hyperparameter optimization, and held-out latest-voyage validation; and
- a practical shipboard case study that identifies data-governance and standardization requirements for condition-informed maintenance planning.

The resulting benchmark of data-driven regressors for filter RUL prediction is intended to be reproducible and applicable to shipboard contexts where labels are scarce, data quality is imperfect, and maintenance planning still depends heavily on time-based routines and expert judgment.

2. BACKGROUND AND RELATED WORK

To position the present study, the authors conducted a targeted search into two thematic blocks: one focused on PHM and machine learning foundations, and another on marine diesel engines, filtration, and RUL validation. The search equations and their intended purpose are summarized in Table 1. The screening template was structured to extract evidence aligned with four guiding questions: (i) which operating conditions influence filter degradation and prediction accuracy, (ii) what machine learning strategies are used for filter RUL estimation in maritime internal combustion engines, (iii) which limitations and research gaps are repeatedly reported, and (iv) what validation metrics and evaluation strategies are used to claim predictive performance.

The search strategy was designed as a targeted background review to support the methodological positioning of the pro-

Table 1. Search equations used to collect candidate literature and their intended purpose.

Block	Scopus query string	Purpose
ML and PHM (G1)	("Machine Learning" OR "Deep Learning" OR "Artificial Intelligence") AND (filter OR separator OR cartridge) AND (RUL OR "Remaining Useful Life" OR prognostics)	Identify ML approaches for filter RUL estimation and feature relevance.
ML and PHM (G2)	("IoT" OR "Internet of Things" OR "Sensor data") AND (filter OR separator) AND ("Remaining Useful Life" OR RUL)	Identify sensing context and data requirements for data-driven prognosis.
ML and PHM (G3)	("CRISP-DM") AND ("Prognostics and Health Management" OR PHM) AND ("Machine Learning" OR "Deep Learning")	Support methodological justification for CRISP-DM-aligned PHM workflows.
ML and PHM (G4)	("CRISP-DM") AND ("Predictive Maintenance" OR "Condition Monitoring") AND ("Industrial IoT" OR "Data Analytics")	Validate CRISP-DM adoption in data-centric engineering projects.
Maritime and filters (G3)	("predictive maintenance") AND ("marine diesel engine" OR "ship engine") AND (filter OR separator)	Analyze how operating environment and duty cycles impact degradation.
Maritime and CRISP-DM (G4)	("CRISP-DM") AND ("machine learning") AND ("marine" OR "ship")	Link ML development structure to maritime PHM applications.
Filter RUL (G4)	("predictive maintenance" OR PHM) AND ("filter degradation") AND ("Remaining Useful Life" OR RUL)	Investigate degradation forecasting and RUL modeling strategies for filters.
Validation focus (G4)	("Remaining Useful Life") AND ("performance evaluation" OR metrics) AND (regression OR forecasting)	Identify evaluation metrics, validation protocols, and reporting practices.

posed PHM/RUL workflow, rather than as a systematic review. Accordingly, potential biases must be acknowledged. Selection bias may arise from prioritizing studies related to PHM, RUL estimation, filtration, and marine diesel engines. Database bias may result from the use of Scopus-indexed literature, while temporal bias may be introduced by emphasizing recent developments in data-driven PHM and maritime predictive maintenance. Language and search-term biases are also possible because the queries were formulated in English and based on specific keyword combinations. Finally, publication bias may occur because successful applications are more likely to be reported than unsuccessful or operationally constrained implementations. These limitations were partially mitigated by using multiple thematic query blocks and linking each query to a defined review purpose, as summarized in Table 1.

Table 1 is therefore retained as a transparency element to document the targeted search logic used to position the study and support the methodological decisions adopted in the proposed PHM/RUL workflow.

2.1. From planned maintenance to PHM in ship engine rooms

Marine propulsion and auxiliary systems operate under duty cycles shaped by weather, mission demands, and port constraints. Historically, engine-room maintenance has relied on planned overhauls and running-hour schedules, with reactive repairs when faults materialize. It can lead to over-maintenance or, conversely, unexpected failures when degradation develops between scheduled interventions (Mwangi et al., 2025; Asimakopoulos, 2023). Because failures in propulsion related systems affect safety, environmental performance, and lifecycle cost, shipboard safety management frameworks emphasize systematic maintenance planning and operational risk control (International Maritime Organization, 2010). At the same time, operational studies show that alarm-based monitoring and human interpretation can miss slowly evolving degradation that remains below fixed thresholds (Kocak, 2023). These pressures have increased interest in PHM and predictive maintenance as means to exploit multivariate operational data for early anomaly detection and decision support in real service (Upadrashta & Wijaya, 2025; Kirketerp-Møller et al., 2025).

Although often used interchangeably in the literature, it is important to distinguish between Predictive Maintenance (PdM), Prognostics and Health Management (PHM), and the emerging paradigm of Prescriptive Maintenance (PxM). PdM primarily focuses on anticipating maintenance needs or potential failures through condition monitoring, historical data, and predictive models, including anomaly detection and degradation-pattern identification (Dalzochio et al., 2020) ; it is also referred to in the literature as Condition-Based Maintenance (CBM) (Gulati, 2013). PHM, which is the framework adopted in this study, represents a more comprehensive engineering discipline; it extends beyond failure anticipation by explicitly modeling degradation trajectories, estimating Remaining Useful Life (RUL), and integrating this capability into broader lifecycle management, logistics, and operational decision-making (Zio, 2022). Furthermore, recent data-driven maintenance frameworks are moving toward PxM, which builds upon PHM by not only estimating the expected failure time or maintenance

window, but also seeking to recommend specific actions considering operational constraints, logistics, risks, and costs (Ansari, Glawar, & Nemeth, 2019).

Figure 1 summarizes this conceptual progression by illustrating how PdM, PHM, and PxM differ in scope and decision-support capability. In this study, the proposed approach is positioned within the PHM layer because it focuses on degradation modelling and RUL estimation for fuel-filter replacement planning.

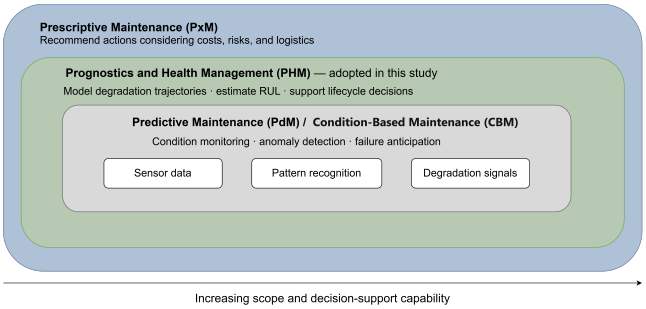


Figure 1. Conceptual relationship between Predictive Maintenance (PdM), Prognostics and Health Management (PHM), and Prescriptive Maintenance (PxM), highlighting their increasing scope and decision-support capability.

Reported failure contributors in propulsion systems commonly involve predictable wear mechanisms and context-dependent, less predictable factors. Practical failure modes often appear in subsystems such as (Kalafatelis et al., 2025):

- Fuel system (e.g., filter blockages, injection-related faults);
- Lubrication system (e.g., insufficient oil supply/quality);
- Intake and exhaust system (e.g., valve obstruction, air filter loading).

This diversity reinforces the need to interpret condition information in relation to the operating context rather than relying solely on absolute thresholds.

2.2. Remaining Useful Life as a prognostic target

Remaining Useful Life is a central PHM target, representing the time remaining before a component reaches a maintenance threshold that requires replacement or restoration. When defined consistently, RUL supports maintenance planning, logistics, and spares provisioning by aligning interventions with actual need and reducing unplanned downtime (Shifat, Yasmin, Hur, & Park, 2021). Beyond operational planning, accurate RUL estimation can inform decisions on component utilization, maintenance policy design, and lifecycle optimization (Quan, Cheng, Guan, Zhang, & Quan, 2025).

RUL estimation methods are commonly grouped into physics-based, data-driven, and hybrid approaches. Physics-based models offer interpretability but require explicit degradation models that are difficult to derive and maintain across changing operating regimes. Data-driven models learn patterns from historical measurements but are sensitive to data representativeness, label quality, and sensor reliability. Hybrid methods combine physical constraints with data-driven learning to improve robustness under uncertainty and incomplete coverage, which is especially relevant when shipboard data are noisy or heterogeneous (Cao, Xiao, Sun, & Gan, 2024; Hagmeyer & Zeiler, 2023).

2.3. Filtration-oriented prognosis in marine engines

Filters (air, oil, fuel) are consumable components that protect engines from harmful particles. Their degradation is typically dominated by clogging (particle buildup), which increases flow resistance and manifests as an increasing pressure drop across the filter medium (Zeiler & Hagmeyer, 2023). In propulsion plants, filtration degradation can therefore influence efficiency, fuel consumption, and downstream stress, making it an operationally meaningful target for prognosis (Kalafatelis et al., 2025; Zeiler & Hagmeyer, 2023). Filtration is also attractive for applied PHM because replacements are more frequent than major failures, enabling event-driven labeling when maintenance actions can be traced reliably. Comparative work on filtration prognosis reports benefits from incorporating physical insight (e.g., plausibility constraints and physics-informed features) into data-driven models, particularly under regime variability and limited sensing (Hagmeyer & Zeiler, 2023; Zeiler & Hagmeyer, 2023). In parallel, industrial case studies emphasize that digitalization and disciplined data-science workflows are often decisive for obtaining stable, actionable predictions from process variables (Florentino & Moura, 2025).

2.4. Shipboard data quality, label scarcity, and methodological implications

A persistent barrier in maritime PHM is data quality. Operational logs may contain gaps, sensor faults, manual transcription errors, and regime-dependent bias; if unaddressed, learned models can overfit artifacts rather than degradation signatures (Jeon et al., 2021). To specifically address these challenges, the recent literature has proposed targeted mitigation strategies. For instance, to tackle sensor faults and poor data quality in harsh environments, (Han, Ellefsen, Li, Holmeset, & Zhang, 2021) demonstrated the use of semi-supervised frameworks, such as LSTM-based Variational Autoencoders, to identify abnormal records and reconstruct latent normal states. To mitigate information gaps and 'data starvation' caused by manual logging, researchers have applied Transfer Learning to adapt knowledge from data-rich vessels (Fan, Sun, Hu, Vladimir, & Mao, 2026), while frameworks like Mar-RUL employ first-order Markov chains for missing data imputation and simulate degradation trajectories when historical failure records are sparse (Velasco-Gallego & Lazakis, 2023). Furthermore, to

overcome regime-dependent bias from load and speed fluctuations, advanced approaches such as Multi-source Adversarial Domain Adaptation (MSDA) have been introduced to dynamically weight data from similar operating domains, effectively separating load variations from actual physical wear (Shi, Wang, & Ding, 2026).

Complementing these data quality solutions, the issue of label scarcity compounds this problem: run-to-failure examples are rare in operational fleets and fault annotations are often incomplete, so many approaches rely on maintenance-event alignment, proxy targets, or expected-behavior baselines (Kim et al., 2021; Kirketerp-Møller et al., 2025). These constraints elevate three requirements for credible RUL reporting: (i) transparent preprocessing with defensible data-quality gates, (ii) traceable label construction anchored to maintenance actions, and (iii) leakage-aware evaluation that preserves temporal structure and approximates prospective deployment (Duan, Vasudevan, et al., 2022; Michalski et al., 2025).

2.5. Model selection under operational variability

Model selection in PHM balances accuracy, robustness across operating modes, and interpretability for engineering adoption, while accounting for the practical constraints summarized in Table 2. Ensemble models are frequently used in applied settings due to their tolerance for heterogeneous features and nonlinearities. At the same time, deep learning architectures can capture temporal dependencies when sufficient data coverage and label quality are available (Michalski et al., 2025; Sun, Ren, et al., 2024). Multi-criteria selection frameworks further emphasize that model choice should account not only for error metrics but also for monitoring requirements, maintainability, and the stability of the end-to-end pipeline under operational drift. In this case study, the vessel's operational profile and the long-standing doctrine governing data recording limited both the quantity and consistency of the available data, making extensive experimentation with data-hungry deep learning models and more complex ensemble schemes impractical. Therefore, a fast-to-apply, low-cost baseline was prioritized as a defensible reference point, while higher-capacity approaches are deferred to future work and to evaluation in a production-like operational environment.

2.6. Open operational data and reproducibility

Open datasets enable reproducibility and comparative baselining in maritime PHM, where proprietary constraints often limit access to realistic operational traces. In this work, the operational time series are drawn from an open dataset of onboard propulsion machinery parameters (Llamas Reinoso et al., 2026). Notably, this public release does not include maintenance event records; event logs required for supervised, event-driven labeling must be derived separately from onboard records, consistent with common practice in operational PHM studies that face fragmented maintenance (Kumar & Flores-Cerrillo, 2024).

Table 2. Key theoretical challenges in shipboard RUL modeling and commonly reported mitigation principles.

Challenge	Implication and mitigation principle
Operating-mode variability	Models must generalize across regimes; prefer mode-aware features and evaluation protocols that preserve temporal structure (Asimakopoulos, 2023; Kirketerp-Møller et al., 2025).
Data quality issues	Missingness and abnormal records can bias learning; apply data-quality gates, reconstruction, and plausibility checks before modeling (Jeon et al., 2021; Duan et al., 2022).
Label scarcity and uncertainty	Limited failures constrain supervised learning; leverage maintenance events, proxy labels, and hybrid constraints to improve robustness (Cao et al., 2024; Hagmeyer & Zeiler, 2023).
Deployment constraints	Practical adoption requires reproducible pipelines and monitoring; include stable preprocessing and maintainability considerations where feasible (Michalski et al., 2025).

3. SYSTEM AND CASE STUDY DESCRIPTION

This section frames the case study from the operational reality of training ship studied and the current limitations in the management of operational and maintenance information onboard. The goal is to make the modeling choices traceable not only to propulsion system, but also to the practical constraints imposed by fragmented records, partial digitization, and information loss over time. Therefore, (i) described the context on the ship and the information-management gap, (ii) limited the scope of the asset to the main propulsion engine and its associated subsystems, and (iii) documented the data sources, the acquisition mode, and (iv) quality constraints that condition the PHM workflow and the construction of RUL-to-replacement labels.

3.1. Training Ship operational context and information management gap

The Ship studied is a brigantine-rigged training vessel and the flagship of the Colombian nation, alternating extended navigation periods with port stays and anchorage. In this profile, propulsion availability is both mission- and safety-relevant: a loss of propulsion capability impacts maneuverability, schedule adherence, training activities, and operational readiness. However, the main constraint addressed in this case study is not the absence of measurements *per se*. Still, the way those measurements and maintenance actions are currently recorded, preserved, and made available for analysis.

To provide a clearer description of the case-study platform,

Figure 2 presents the ARC “Gloria” training ship, while Table 3 summarizes its principal particulars.

Figure 2. ARC “Gloria” training ship considered in the case study.

These characteristics are relevant to contextualize the operational profile of the vessel, its propulsion requirements, and the maintenance environment in which the filtration-system RUL prediction model was developed.

Table 3. Principal particulars of the ARC “Gloria” training ship.

Characteristic	**Value**
Vessel type	Brigantine-rigged training ship
Length overall without bowsprit	64.6 m
Length overall with bowsprit	76.0 m
Beam	10.6 m
Mean draught	4.5 m
Displacement	1,300 t
Total number of sails	23
Autonomy	60 days
Maximum speed under engine	12 knots
Main propulsion engine	MAN 23L 30A four-stroke diesel engine
Engine configuration	Six in-line cylinders
Rated power	1,300 HP / 960 kW

In practice, the ship’s operational and maintenance information is distributed across heterogeneous media and formats: handwritten log sheets from engineering rounds (“minutas”), paper-based maintenance reports and checklists, informal annotations, and isolated spreadsheets created for specific needs.

These conditions are representative of many in-service maritime environments where formal instrumentation may coexist with predominantly manual operational reporting, and where maintenance history is not fully integrated into a digital asset-management software. Consequently, the present work uses the propulsion system as a realistic testbed to demonstrate a CRISP-DM–guided PHM pipeline that can operate under imperfect, partially digitalized data while explicitly managing uncertainty and incompleteness.

3.2. Main propulsion system and scope of the case study

The main propulsion system of the ARC ’Gloria’ is powered by a MAN 23L 30A four-stroke marine diesel engine, rated at 1,300 HP (960 kW). This reciprocating internal-combustion diesel engine has six in-line cylinders. In its model designation, ’23L’ denotes the piston bore, while ’30A’ refers to the stroke configuration. The engine is fitted with an exhaust-gas turbocharger and intercooler, and is connected to the fuel supply and filtration, lubrication, cooling, intake/exhaust, reduction, shaft-line, and propeller-load subsystems considered in the case study.

The case study is intentionally limited to the ship’s main propulsion engine and the subsystems that most directly condition its reliability and maintainability within routine shipboard practice. From a systemic perspective, the propulsion plant can be viewed as an integrated set composed of: the diesel engine prime mover; fuel supply and filtration; lubricating-oil circuit and filtration; cooling-water circuit; air intake/exhaust path; reduction/shaft line; and the propeller load as an external demand.

Within the Colombian Navy maintenance doctrine, shipboard maintenance is governed by a structured, multi-scope framework that defines responsibilities, allowable interventions, and traceability requirements across five maintenance levels (I–V), ranging from operator routines to overhaul and Original Equipment Manufacturer (OEM)-supported work (Table 4). This layered doctrine shapes what can be executed onboard, what must be escalated, and how actions are recorded, effectively constraining both the availability and the granularity of maintenance evidence (“Doctrina de Material Naval. Tomo III. Mantenimiento”, 2022).

Within this system-of-systems, the present analysis focuses on components that (i) exhibit recurrent interventions, (ii) leave a meaningful operational signature in the recorded parameters, and (iii) can be linked to identifiable replacement actions under the prevailing maintenance philosophy. Figure 3 further highlights an additional operational constraint: maintenance-task definition is implicitly bounded by the subsystem boundary around the propulsion engine, so that the formal maintenance plan typically covers only components attached to the engine, whereas external elements in the fuel supply chain (e.g., upstream tanks and associated hardware) may fall outside scheduled-program coverage and depend on vendor practices that are not guaranteed under variable operating conditions.

Accordingly, special emphasis is placed on the filtration-related maintenance loop (e.g., cartridge changes), since operating conditions and contamination levels influence filter degradation. At the same time, replacement decisions are often driven by time-based routines and operator judgment rather than quantified health indicators. In this context, the proposed RUL model is not intended to override doctrine or

Table 4. Maintenance levels and practical scope defined by the Colombian Navy maintenance doctrine

Level (responsibility / scope)	Key practical activities
I (Operator / daily routine)	Cleaning and visual inspection; check fluid levels (oil/fuel); basic lubrication; detect and report abnormal noise/conditions (CILA: Cleaning–Inspection–Lubrication–Adjustment).
II (Trained technician / simple tasks, no long downtime)	Basic specialized checks; belt tension check/adjustment; top-up of fluids; air filter cleaning; oil and filter change.
III (On-site specialist technician / routine technical work)	Routine calibrations; verification of operating/service parameters; parts replacement.
IV (Workshop team + lead/field engineer / major intervention)	Partial disassembly for maintenance; specialized calibrations; inspection/verification of fits, adjustments, and tolerances; welding and associated inspection.
V (Overhaul + OEM support / highest level)	Full teardown/disassembly; destructive and non-destructive testing (DT/NDT); precision calibration and metrology using specialized instruments and tooling.

expand maintenance scope beyond defined boundaries, but to act as a decision-support tool that increases determinism and consistency when planning a plausible maintenance action within the imposed policy constraints, improving timing and prioritization while preserving traceability and operational feasibility.

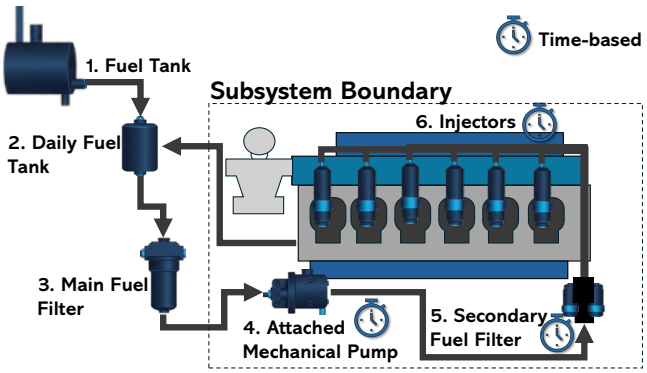


Figure 3. Main propulsion engine and associated subsystem considered in the case study.

3.3. Data sources, acquisition modality, and current maintenance practice

The datasets used in this study were reconstructed from routine onboard practices rather than from an automated, high-frequency sensor architecture. The core sources are:

- **Operation logs from engineering rounds:** periodic readings manually captured by the technical crew during engine-room inspections, typically including variables such as engine angular speed and shaft angular speed, pressures (e.g., lubrication and fuel), and temperatures (e.g., cooling, lubrication, exhaust), plus contextual notes (navigation/port/anchorage condition).
- **Maintenance documentation:** records of preventive and corrective actions (including component replacement notes) that enable identification of *events* such as cartridge changes, inspections, and interventions.
- **Accumulated-operation indicators:** hourmeter/ accumulated-time fields that provide an operational basis to build persistence measures (e.g., hours since last replacement), even when timestamps are incomplete or irregular.

From an operational standpoint, maintenance decisions for propulsion-related components tend to follow time-based routines complemented by corrective actions triggered by observed symptoms, alarms, or performance degradation noted by operators. In the absence of fully digitalized, searchable histories, the practical result is that maintenance planning is often constrained to what can be confirmed in the most recent logs, and long-horizon trend analysis becomes difficult. It creates an opportunity for PHM methods that (i) consolidate existing information, (ii) reconstruct consistent timelines, and (iii) produce interpretable health indicators and RUL estimates aligned with how the crew already operates and documents maintenance.

3.4. Data quality constraints induced by partial digitization

Because the information originates from manual notes and heterogeneous records that are only partially digitized, the resulting datasets exhibit characteristic limitations that directly shape the modeling strategy:

- **Missing values and incomplete rounds:** Some variables are absent due to omitted measurements, operational workload, or partial log completion.
- **Irregular temporal coverage:** record density varies across voyages and operational phases; long gaps can exist, limiting strictly sequence-based approaches and motivating event- and accumulation-based representations (e.g., cycle variables).
- **Transcription and formatting inconsistencies:** manual digitization introduces errors (decimal separators, unit mismatches, swapped fields) requiring plausibility filters, standardization, and outlier handling.
- **Label uncertainty in maintenance events:** replacement actions (e.g., cartridge change) may be recorded with

ambiguous dates, partial descriptions, or without standardized subcomponent identifiers, which complicates the definition of "ground truth" for supervised learning.

- **Information loss across the lifecycle:** missing documents, non-centralized storage, and version drift in spreadsheets lead to irrecoverable historical segments, forcing reconstruction from the surviving evidence rather than from a complete archival baseline.

These constraints motivate a PHM workflow that is resilient to incomplete histories: (i) rigorous data preparation (cleaning, reconstruction of time series, and event extraction), (ii) explicit construction of RUL-to-replacement labels from identifiable replacement events, and (iii) modeling choices that prioritize robustness and interpretability under sparse, human-in-the-loop measurements. In this setting, the prognostic objective is not to replicate laboratory-grade condition monitoring, but to extract actionable value from the information that the ship *already produces*, while highlighting how improved digitalization and standardized recording would immediately increase model reliability and operational impact.

4. METHODOLOGY

CRISP-DM was selected because the main challenge of this study is not limited to algorithm selection, but involves transforming heterogeneous shipboard records into a traceable prognostic dataset suitable for supervised RUL modeling. Unlike purely model-centric workflows, CRISP-DM explicitly begins with business understanding and data understanding (Chapman et al., 2000), which are essential in this case to define the maintenance objective, identify the cartridge-replacement event as the prognostic anchor, assess the limitations of manual records, and determine whether the available evidence can support reliable label construction. This makes CRISP-DM suitable for operational PHM studies where data quality, event traceability, expert knowledge, and evaluation design strongly condition the validity of the final model.

Other frameworks, such as generic machine-learning pipelines or software-oriented MLOps lifecycles, are useful for model implementation and deployment, but they do not provide the same explicit structure for linking maintenance objectives, legacy data assessment, event reconciliation, and model evaluation. In contrast, CRISP-DM offers a problem-to-data-to-model sequence that is consistent with the constraints of this case study. Therefore, it was adopted as a guiding and adapted framework for PHM model development under realistic shipboard conditions. The deployment phase is not claimed as completed; it is addressed only through deployment considerations and future work, since onboard operationalization requires a dedicated monitoring, governance, and MLOps architecture.

In this work, some CRISP-DM stages were mapped to PHM activities tailored to the marine propulsion context: (i) business understanding through the identification of maintenance-management gaps and the definition of the prognostic objective for fuel-filter life; (ii) data understanding through manual and technical documentation validation and the verification of doctrinal philosophy; (iii) data preparation through expert validation of representative operational cycles using a Delphi-based protocol, physics-informed plausibility screening, outlier mitigation, missing-value reconstruction, and event-log consolidation; (iv) modeling through RUL label construction and supervised regression model development; (v) evaluation through model comparison, hyperparameter tuning, and held-out latest-voyage validation; and (vi) deployment, which is reserved for future work once a dedicated monitoring and MLOps architecture is available.

The workflow is structured as shown in Figure 4. Within the Data Preparation phase, is executed over the operational-variable validation through a Delphi expert survey and subsequent preprocessing (outlier removal, missing value imputation, and normalization) and related to the maintenance records, the matching task implied maintenance event reconciliation, including the approximation of out-of-phase events and the computation of RUL labels from hour-meter differences. The resulting matched Dataset is then used in the Modeling phase, where the validated operational series and the maintenance event timeline are integrated into a single, temporally coherent prognostic dataset. This design supports different feasible entry points depending on data availability: when only operational readings are accessible, the preprocessing stream alone can support model development; when maintenance records are also available, the event reconciliation stream enables supervised, event-driven label construction, improving model traceability and methodological rigor.

The deployment-oriented stage (vi) is explicitly reserved as future work, to be addressed once a deployment platform and an MLOps architecture are developed to operationalize the model.

4.1. Business Understanding (BU)

As depicted in Figure 4, the Business Understanding phase establishes the operational and maintenance context that governs all subsequent modeling decisions. It comprises three sequential steps.

Step 1 - First Interview and Field Interview

This step elicits both managerial and operational knowledge to ensure that the data-driven analysis is aligned with the maintenance objectives of the organization. The office-level interview with senior technical personnel was intended to clarify the business motivation behind filter replacement analysis, including availability requirements, maintenance planning criteria, and the operational consequences of delayed or premature cartridge changes.

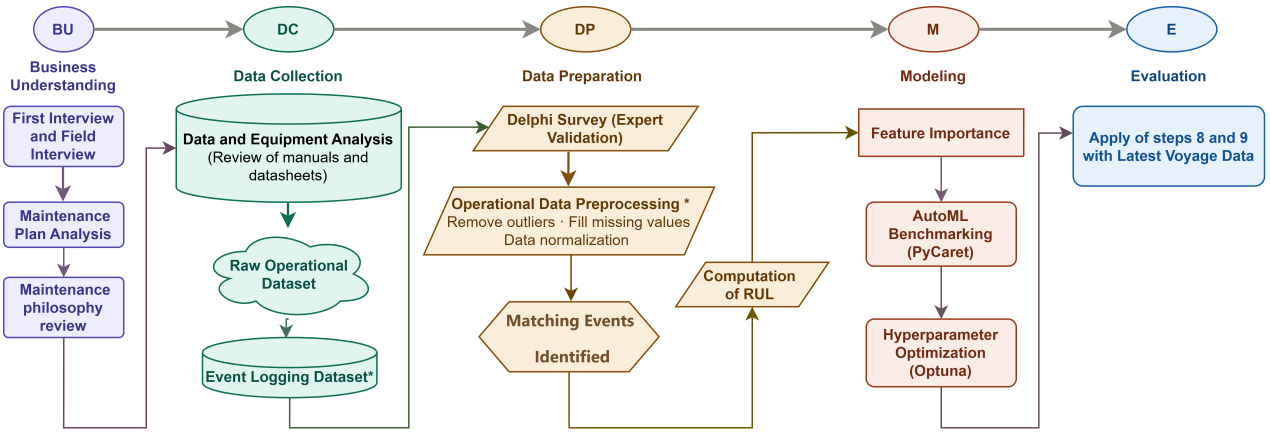


Figure 4. CRISP-DM-guided pre-deployment PHM workflow for RUL prediction of fuel filtration cartridges. The implemented phases are Business Understanding (BU), Data Collection/Data Understanding (DC/DU), Data Preparation (DP), Modeling (M), and Evaluation (E). Deployment is not claimed as completed and is addressed separately as future operationalization.

Complementarily, the field interview with engine-room personnel characterizes how filter cartridge replacements are executed and recorded in practice. It includes clarifying local conventions, such as how a "change" is logged, which subsystem is referenced, and how shifts and port operations affect record timing, as well as verifying maintenance templates or check formats used onboard. The output of this step is an agreed interpretation of what constitutes a valid filter-change event, which business and operational objectives motivate its analysis, and which contextual cues are required to disambiguate records, such as operating regime or voyage context.

Step 2 - Maintenance Plan Analysis

In this step reviews the maintenance plan and the declared maintenance philosophy to identify the primary replacement event used as the prognostic anchor (cartridge/filter change) and to characterize expected timing gaps between prescribed interventions and recorded execution. This stage reconciles planned periodicity with field practice, explicitly documenting typical delays, asynchronous logging patterns, and window effects (e.g., tasks performed during port stays but logged later).

Step 3 - Maintenance Philosophy Review

At this stage, consensus was also reached on how to handle temporal offsets introduced by manual recordkeeping and the absence of ERP-based processing. In particular, the maintenance doctrine required formal traceability only for maintenance actions at level 3 or higher, which can result in incomplete or delayed timestamps for routine consumable replacements. To preserve chronological consistency, event alignment was defined relative to the engine-hour meter: when a filter-change record lacked an exact operational timestamp, the event was matched to the last available engine-hour reading before the record, or to the closest engine-hour reading observed when the engine was restarted, and operation resumed.

4.2. Data Collection (DC)

The Data Collection phase converts maintenance knowledge into a formally defined prognostic target by structuring, validating, and temporally aligning cartridge replacement events. As depicted in Figure 4, this phase covers Steps 4, 5, and 6, governing maintenance semantics, planning assumptions, and event-to-data alignment. The objective is to ensure that the RUL variable reflects observed maintenance behavior rather than nominal schedules.

Step 4 - Data and Equipment Analysis

In this step a functional review of the propulsion and auxiliary subsystems was performed using manuals and datasheets, verifying that the measured variables contained in the operational data set are coherent.

The case study focuses on the main propulsion engine. Functionally, the propulsion plant integrates: (i) an air-based starting system (30 bar) with redundant compressors, (ii) gear-based synchronization, (iii) an exhaust-driven turbocharging line coupled with an intercooler, (iv) a fuel injection and filtration chain including duplex pre-filtration and cartridge-based filtration stages, (v) a gear-type lubrication circuit that also supplies the turbocharger, and (vi) a seawater-cooled thermal management architecture split into high-temperature (HT) and low-temperature (LT) circuits (including oil and charge-air cooling). These functional subsystems motivate both the selection of operational variables used in this work (rotational speeds, thermal and pressure variables, and filtration-related indicators).

Step 5 - Raw Operational Data

The operational time series used in this study are obtained from

an open Zenodo release for ARC *Gloria* (Llamas Reinoso et al., 2026). Because these data originate from hourly manual rounds rather than automated high-frequency sensing, Step 2 emphasizes engineering plausibility bounds and cross parameter expectations to identify transcription inconsistencies, delayed logging, and regime transitions that could distort temporal interpretation. Representative checks include rotational consistency between `Engine_RPM` and `Shaft_RPM`, coherence between propeller pitch and load indicators, thermodynamic compatibility between exhaust gas temperatures and engine speed, and pressure–temperature consistency within lubrication and cooling circuits.

Step 6 - Event Logging Dataset

In this step, the maintenance-event documentation was reviewed and consolidated to identify the key labels required for subsequent event-based analysis. Unlike the operational records collected during engineering rounds, this stage focused only on formats related to maintenance activities, failure reports, inspections, interventions, and component replacement records. These documents were manually examined to determine how relevant events were described, classified, and recorded in practice, with particular attention to labels associated with filter cartridge changes, failure symptoms, corrective actions, affected subsystems, dates, accumulated operating hours, and maintenance observations.

The identification of these labels was supported by validation with onboard personnel and experienced staff familiar with the maintenance formats, ensuring that the extracted event categories reflected the actual terminology and recording conventions used in the vessel. As a result, this step produced a structured event-logging dataset that preserves the traceability of maintenance and failure records while providing the basis for defining valid replacement events, linking interventions to operational history, and preparing the event labels required for later data integration and modeling stages.

4.3. Data Preparation (DP)

The Data Preparation phase transforms the raw inputs from both collection streams into a temporally coherent, expert-validated dataset ready for modeling. The activities executed were: The Delphi survey to assess the veracity of significative operational data, operational-variable preprocessing and maintenance event reconciliation and RUL label construction as shown in Figure 4.

Step 7 - Delphi Survey

In this step it formalizes operational validation through a structured Delphi survey involving five engine-room domain experts. The Delphi survey technique was adopted because it provides a structured approach for collecting and integrating expert judgement when direct experimental validation is limited or when specialized operational knowledge is required (Hasson, Keeney, & McKenna, 2000). In this study, the Delphi-based validation was used to assess whether the extracted operating cycles were plausible representations of real shipboard engine-room conditions, due to the output of Step 5 is a technically constrained raw dataset that has been validated against the asset's functional behavior and features and is therefore suitable for structured expert review. Particularly, this is done because of the operational data were manually recorded and the interpretation of ENGINE ANGULAR SPEED (EAS) profiles and continuous operating periods required domain-specific knowledge from personnel familiar with the vessel and its propulsion system.

The size of the expert panel was carefully defined based on strict criteria to ensure reliable operational validation. Instead of aiming for a statistically large sample, the selection focused on specialized, system-related expertise. To be included, experts needed hands-on navigation experience, familiarity with engine-room routines, and practical knowledge of the vessel's main propulsion engine. Because of the vessel's unique propulsion system, the pool of personnel meeting all these requirements was naturally limited. Therefore, a panel of five selected experts was considered appropriate for this Delphi-based validation, as the goal was not statistical representativeness but the evaluation of operational feasibility by domain specialists. To minimize individual bias, expert responses were aggregated and analyzed using acceptance rate, median rating, interquartile range, and Kendall's coefficient of concordance. Twelve representative operational cycles were evaluated according to two ordinal criteria: (i) plausibility of the EAS profile and (ii) plausibility of continuous operating time. Experts assigned Likert-type ratings to each cycle under both evaluation dimensions.

Let $x_{ij}^{(d)}$ denote the ordinal rating assigned by expert j to cycle i under dimension d, where $i = 1, \ldots, m$, $j = 1, \ldots, n$, $m = 12$, $n = 5$, and $d \in \{\mathrm{EAS}, \mathrm{TIME}\}$. Validation rigor was quantified through four complementary indicators: acceptance rate, median rating, interquartile range (IQR), and Kendall's coefficient of concordance.

The acceptance rate measures the proportion of ratings that reach or exceed the minimum acceptable category c_{acc}:

$$AR^{(d)} = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbf{1}\left(x_{ij}^{(d)} \geq c_{\mathrm{acc}}\right), \tag{1}$$

where $\mathbf{1}(\cdot)$ is the indicator function.

Cycle-wise central tendency was summarized using the median expert rating:

$$\tilde{x}_i^{(d)} = \mathrm{median}\left(x_{i1}^{(d)}, x_{i2}^{(d)}, \ldots, x_{in}^{(d)}\right). \tag{2}$$

Because survey responses were ordinal, robust non-parametric

statistics were used to quantify dispersion. The interquartile range was computed as

$$IQR_i^{(d)} = Q_{3,i}^{(d)} - Q_{1,i}^{(d)}, \tag{3}$$

where $Q_{1,i}^{(d)}$ and $Q_{3,i}^{(d)}$ denote the first and third quartiles of the expert ratings for cycle i.

Inter-expert agreement across the n experts was quantified using Kendall's coefficient of concordance. Let $r_{ij}^{(d)}$ denote the rank assigned by expert j to cycle i. The total rank for cycle i is

$$R_i^{(d)} = \sum_{j=1}^{n} r_{ij}^{(d)}. \tag{4}$$

Let

$$\bar{R}^{(d)} = \frac{1}{m} \sum_{i=1}^{m} R_i^{(d)} \tag{5}$$

be the mean rank sum. The dispersion of rank sums is then

$$S^{(d)} = \sum_{i=1}^{m} \left(R_i^{(d)} - \bar{R}^{(d)} \right)^2. \tag{6}$$

Kendall's coefficient of concordance is defined as

$$W^{(d)} = \frac{12 S^{(d)}}{n^2 (m^3 - m)}. \tag{7}$$

Across the twelve operational cycles, acceptance was high for both evaluation dimensions (EAS plausibility: 86.7%; operating-time plausibility: 90.0%). Median ratings remained consistently within the agreement region of the ordinal scale, while dispersion was typically limited to $IQR \approx 1$, indicating limited disagreement among experts. Kendall's coefficients showed moderate but nontrivial concordance ($W = 0.237$ for EAS plausibility and $W = 0.176$ for operating-time plausibility), which is consistent with heterogeneous yet convergent expert judgment in real operational environments.

Collectively, the high acceptance rates shown in Figure 5, low ordinal dispersion, and nonzero concordance indicate that the extracted cycles constitute operationally credible abstractions of real ship duty profiles. This validation step strengthens the business and data understanding stages by constraining subsequent processing to expert-supported operating regimes, thereby reducing the risk of propagating atypical or weakly representative patterns into segmentation, temporal alignment, and downstream event-matching stages.

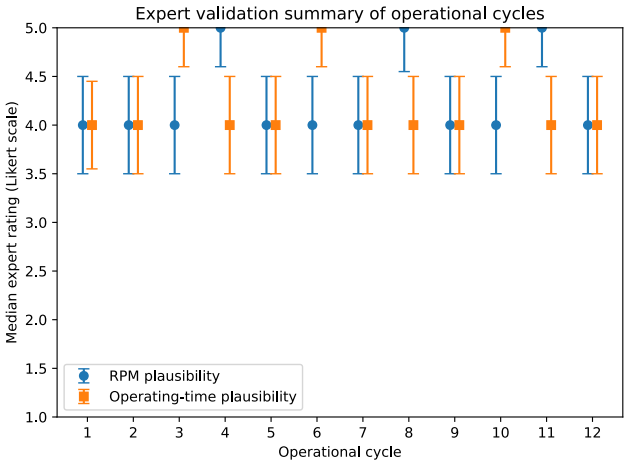


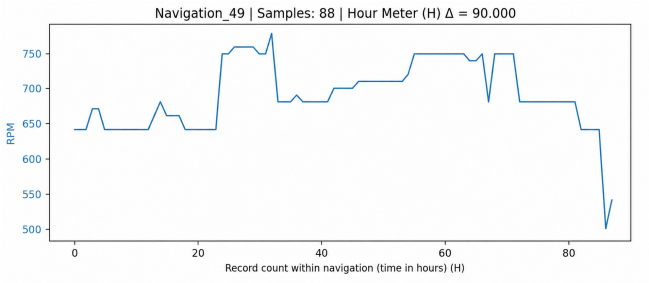
Figure 5. Expert validation summary of representative operational cycles. Markers denote median expert ratings and vertical whiskers represent the interquartile range (IQR) across the five experts. Two evaluation dimensions are shown: EAS-profile plausibility and continuous operating-time plausibility.

Figure 6 illustrates two representative operational cycles from the validated set. (a) Navigation_49 (88 samples, Hour Meter $\Delta = 90$ h) represents a short voyage with a clear load transition around record 25 and a sharp drop near the end, while (b) Navigation_28 (358 samples, Hour Meter $\Delta = 362$ h) captures a longer navigation period with higher variability and multiple regime changes. Both profiles were evaluated by the expert panel and received acceptance ratings within the agreement region, confirming their representativeness of real shipboard operating conditions.

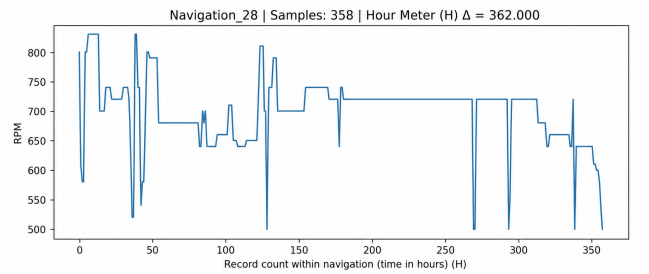
Step 8 - Operational Data Pre-processing

This stage it performs deterministic preprocessing of the validated hourly operational parameters. First, outliers are removed using engineering bounds and cross-parameter consistency checks derived from Steps 3 and 4, rather than relying solely on statistical deviation criteria. This approach mitigates physically implausible spikes caused by transcription errors or logging inconsistencies. Second, missing or invalid entries are reconstructed by localized interpolation across temporally adjacent valid records, preserving chronological continuity while avoiding artificial smoothing of degradation trajectories.

Localized interpolation was preferred over multivariate imputation methods such as MICE because the missing values occurred within hourly manual operational rounds where chronological continuity and physical interpretability were more critical than maximizing multivariate statistical reconstruction. MICE would require assumptions about the missing-data mechanism and stable cross-variable relationships that are difficult to verify under changing ship operating regimes, manual transcription errors, and limited degradation cycles. In contrast, localized interpolation uses temporally adjacent valid



(a) Navigation_49 (88 samples, Hour Meter $\Delta = 90$ h).



(b) Navigation_28 (358 samples, Hour Meter $\Delta = 362$ h).

Figure 6. Representative EAS (RPM) profiles of two validated operational cycles evaluated in the Delphi survey.

records within the same operational context, reducing the risk of introducing synthetic degradation patterns or cross-regime artifacts. Therefore, imputation was applied conservatively and only when the surrounding records supported a physically plausible reconstruction.

Finally, the quantitative operational parameters are normalized to reduce scale dominance during model training. In contrast, text fields are removed, because they are not consistently structured for modeling at the hourly logging resolution. Given the manual acquisition frequency (hourly), normalization is applied to preserve relative regime variation while avoiding suppression of operational dynamics.

The output of this stage is the processed operational dataset, as indicated in the Data Preparation phase of Figure 4. As part of this preprocessing stage, an exploratory data assessment was conducted to characterize the analytical suitability of the operational records before model training. This assessment included verification of temporal coverage, missing-value patterns, physically implausible observations, outlier behavior, and the distribution of the constructed RUL labels across degradation cycles. The review confirmed that the available records covered the 2021–2025 period with hourly manual logging, but with non-uniform density across voyages and operating contexts. Missingness was not uniformly distributed across variables: some continuous thermodynamic and pressure variables showed sufficient coverage for learning, whereas several binary state indicators, alarm fields, and operational-mode descriptors were sparsely populated and therefore unsuitable as stable predictors. Outlier analysis was performed using engineering plausibility limits and cross-parameter consistency checks rather than purely statistical thresholds, in order to distinguish physically possible operating variation from transcription or formatting errors. The RUL labels were also examined at the cycle level to verify that each labeled record was bounded by a validated cartridge-replacement event and that the resulting remaining-life values were consistent with accumulated operating hours rather than calendar-time irregularities.

To quantify the effect of preprocessing, Table 5 summarizes the transition from the raw operational log to the final integrated modeling dataset. The comparison is reported at both row and cell level because the source data limitations were related not only to the number of available records, but also to the completeness of the hourly manual rounds.

Table 5. Quantified impact of operational data preprocessing and integration.

Item	Raw log	Final dataset
Records	10,227	6,915
Columns	44	48
Cell-level entries	449,988	331,920
Missing entries	136,589	15,886
Missingness	30.4%	4.8%

As shown in Table 5, the final integrated dataset retained 6,915 labeled engine-operation records from the 10,227 raw operational records, corresponding to a reduction of 3,312 records, or 32.4% of the initial rows. At the cell level, the number of missing entries decreased from 136,589 in the raw operational log to 15,886 in the final integrated dataset. This reduced the overall missingness from 30.4% to 4.8%, equivalent to a reduction of 120,703 missing entries and 25.6 percentage points. The increase from 44 raw columns to 48 final columns reflects the addition of engineered fields required for temporal alignment, degradation-cycle identification, cartridge-change tagging, and RUL-label construction.

Table 6 summarizes the key properties of the final integrated dataset, while Table 8 lists the retained operational predictors grouped by propulsion subsystem. Variables excluded during screening were either redundant with retained predictors (confirmed by Pearson correlation checks) or insufficiently stable across operating regimes to contribute reliable predictive signal.

The exploratory assessment also explains the reduction from 48 original variables to 34 retained predictors reported in Table 6. This means that 14 variables were not retained as model predictors after coverage, redundancy, and feature-screening checks. A subset of the original variables — primarily binary state indicators (e.g., on/off flags, alarm states, and operational mode fields) — exhibited sparse coverage across the dataset:

Table 6. Summary of the operational dataset used for RUL model training.

Property	Value
Total records	6,915
Degradation cycles	17
Original variables	48
Retained predictors	34
Target variable	RUL (hours to next replacement)
Logging frequency	Hourly (manual rounds)
Time span	2021–2025

these variables recorded values only occasionally and for short intervals, resulting in a high proportion of missing entries relative to the total number of records. Under hourly manual logging, such sparsity indicates that these fields were not systematically captured during engineering rounds, making them unreliable as continuous predictors for a regression model. Variables for which the ratio of valid entries to total records was insufficient to support stable learning were therefore excluded during the screening stage, prioritizing predictors with consistent coverage across the full observation window and across multiple degradation cycles.

Step 9 - Matching Events Identified

In this stage, it constructs the maintenance event log by extracting and validating filter-change records and explicitly tagging confirmed cartridge replacement events. For the article, only the columns that add modeling value are retained: the logged date (timestamp proxy), the raw subject text, the event type, the coded event label, and the associated subsystem; digitization traceability fields are omitted (Table 7). Because the operational dataset is sampled on an hourly grid while maintenance actions may be recorded at irregular times, temporal reconciliation is applied whenever an event is *out-of-phase* with the operational sampling. This reconciliation uses gap-aware approximations to place events onto the nearest consistent boundary while preserving chronological order within each operating context. The output of this stage is a validated event timeline that can be aligned with the preprocessed operational series produced in step 8.

The validated operational series and the validated maintenance event timeline (event reconciliation stream) are integrated into a single, temporally coherent prognostic dataset.

A total of 18 real cartridge-change events were extracted from maintenance records. These events were reconciled with operational data using a temporal proximity matching strategy based on hour-meter alignment. For each annotated replacement date, the closest operational record in terms of accumulated hour-meter reading was identified, producing a structured mapping between maintenance events and operational state. The resulting event–hour-meter pairs define discrete degradation cycles anchored to physically meaningful replacement boundaries.

Subsequently, operational records were assigned to degradation cycles using hour-meter segmentation. A total of 6,915 engine-operation rows were successfully labeled with a cycle identifier, resulting in 17 effective degradation cycles after overlapping or duplicated annotations were reconciled. This step ensures chronological continuity within each cycle and prevents cross-cycle leakage during label construction.

RUL values were then computed exclusively from hour-meter differences relative to the next validated replacement event. By defining RUL as the remaining accumulated operational hours until the next cartridge change, truncated at zero, the target variable becomes independent of calendar irregularities and robust to variations in operational intensity. The resulting RUL distribution spans from 286 to 876 hours, reflecting the observed variability in filter service life under real operating conditions.

The final integrated dataset, therefore, consists of 6,915 temporally ordered operational records, each associated with (i) a validated degradation cycle, (ii) a physically aligned replacement boundary, and (iii) a forward-consistent RUL label.

This integration step is critical for the subsequent modeling stage, as it ensures that the learning algorithm is exposed to degradation trajectories that are physically grounded, temporally coherent, and validated both statistically (preprocessing stream) and operationally (event reconciliation stream). By constructing RUL purely from hour-meter progression within validated cycles, the model is trained on a consistent degradation signal rather than administrative timestamps.

4.4. Modeling (M)

Model development follows an AutoML-first strategy to establish a robust regression baseline under heterogeneous ship operation and nonstationary duty cycles. As illustrated in Figure 4, the workflow is organized into three stages—Step 10 (Feature Importance), Step 11 (AutoML Benchmarking), and Step 12 (Hyperparameter Optimization)—to ensure (i) event-anchored RUL labels, (ii) leakage-aware, time consistent evaluation, and (iii) reproducible model selection and refinement.

For the modeling stage, was regarded a variant of the mainstream RUL prediction literature. RUL is conventionally defined as the time remaining until component failure, and models are trained on run-to-failure datasets in which degradation trajectories are observed from a healthy state to complete breakdown (Zhang et al., 2022; Pan et al., 2025). Under this paradigm, the prognostic target is defined by a failure event, and take a practical departure from this convention. The present work departs from this convention in a practically motivated way: rather than waiting for the filtration cartridge to reach a failure state, the RUL target is defined as the remaining accumulated operating hours until the next scheduled cartridge replacement, as expressed in Eq. 8. This formulation reflects

Table 7. Cleaned maintenance-event log schema used in Step 9 for event identification and labeling (examples). Digitization traceability fields (document/page) are omitted.

Date logged	Raw subject	Event type	Coded event	Subsystem
22/03/2024	Filters Racor 2020	Preventive maintenance	Fuel primary filter replacement	Fuel system
12/04/2024	Corrective maintenance	Failure	Cracked piping	Seawater open cooling system
24/06/2024	Filters Racor 2020	Preventive maintenance	Fuel primary filter replacement	Fuel system
17/07/2024	Fuel secondary filters	Preventive maintenance	Fuel secondary filter replacement	Fuel system
26/08/2024	Maintenance	Preventive maintenance	Oil filter replacement	Lubrication system

the actual maintenance philosophy onboard the ARC Gloria, where cartridges are replaced preventively based on operational experience and time-based routines, not in response to catastrophic filter failure. Consequently, the model does not learn to anticipate breakdown but rather to estimate how much operational life remains before a maintenance action is due — a condition-based scheduling objective that supports maintenance planning for maintenance planning without requiring run-to-failure observations, which are neither available nor desirable in operational naval contexts.

Step 10 - Feature Importance

At this stage, a preliminary assessment of variable importance is performed to support interpretability and screen for non-informative or unstable covariates before training. Variable relevance was assessed using two complementary approaches. First, a Random Forest impurity-based criterion (Gini/impurity importance) was computed by aggregating the reduction in node impurity (MSE for regression) attributable to each feature across all trees and normalizing the result to a $[0, 1]$ scale. It provides a fast, model-internal relevance signal but can favor high-variance or high-cardinality variables. To mitigate this bias and to measure predictive dependence directly, permutation importance was also applied by randomly shuffling each feature and quantifying the degradation in model performance (e.g., RMSE/MAE) relative to the unshuffled baseline. Using both methods reduced reliance on a single criterion and improved robustness against the limitations of each. Based on this combined screening, 34 out of the 48 operational variables were retained. This selection was further supported by Pearson correlation checks to remove redundant covariates, as shown Table 8. For example, highly correlated EAS-derived indicators and start-up cylinder temperature channels were consolidated, reducing six temperature signals to two averaged temperature features.

The top-ranked variables are dominated by cooling and hydraulic system indicators: Clutch Pressure (0.195), Freshwater Outlet Temperature (0.085), Freshwater Pressure (0.073), and Seawater Pressure (0.071). This pattern is physically consistent with the load-dependent nature of filter degradation, as thermal and hydraulic states are tightly coupled to engine

Table 8. Representative operational predictors retained for RUL modeling, grouped by subsystem. Importance weights are shown in parentheses.

Subsystem	Predictors
Rotational	EAS (0.005), Shaft Angular Speed (0.009), Propeller Pitch (0.005)
Fuel system	Fuel Pressure (0.039), Oil Pressure Before Filter (0.004), Oil Pressure After Filter (0.009)
Lubrication	Lubricating Oil Inlet Temperature (0.038), Lubricating Oil Outlet Temperature (0.026), Gearbox Oil Pressure (0.027)
Cooling	Freshwater Inlet Temperature (0.051), Freshwater Outlet Temperature (0.085), Freshwater Pressure (0.073), Seawater Pressure (0.071), Seawater Outlet Temperature (0.026), Air Cooler Inlet Temperature (0.054), Air Cooler Outlet Temperature (0.018)
Mechanical	Clutch Pressure (0.195), Support Bearing 2 Temperature (0.062), Thrust Bearing Temperature (0.059), Servo Piston Pressure (0.019), Starting Air Pressure (0.001)
Combustion	Boost Air Pressure (0.055), Exhaust Gas Temperature (0.021), Cylinder Temp. Group G1 (1,5,6) (0.015), Cylinder Temp. Group G2 (2,3,4) (0.011)
Target	**RUL_Hours (remaining hours to cartridge replacement)**

demand and, consequently, to fuel consumption rates and contamination accumulation in the cartridge. Bearing temperatures (Support Bearing 2: 0.062; Thrust Bearing: 0.059) and turbocharging indicators (Boost Air Pressure: 0.055; Air Cooler Inlet Temperature: 0.054) further reinforce this interpretation. Notably, the variables most directly associated with the filtration subsystem—Oil Pressure Before Filter (0.004) and Oil Pressure After Filter (0.009)—rank among the lowest, suggesting that under the available manual logging resolution, the model relies on the broader propulsion plant context as a indicator for degradation state rather than on direct filter measurements.

Finally, the dataset is partitioned using a time-consistent split to reflect prospective use: training and validation are drawn from earlier periods, and evaluation is performed on tempo-

rally subsequent observations to reduce leakage and optimistic bias.

Step 11 - AutoML Benchmarking

This stage formalizes the prognostic target by combining the validated operational series obtained from the Data Preparation phase with the validated maintenance-event timeline obtained from the Data Collection phase. In this work, two distinct uses of the RUL definition are considered. First, RUL is constructed retrospectively as an event-driven label from validated cartridge-replacement events. Second, this constructed label is used as the target variable in a supervised regression problem, where the model estimates remaining cartridge life from the operational state of the propulsion system.

(i) Preprocessing/event alignment (event-driven RUL label). During the event reconciliation stream, RUL is not estimated by the model. Instead, it is assigned retrospectively to each operational record using the engine hour-meter and the next validated cartridge-replacement event as the reference horizon. This formulation is necessary because the available records do not correspond to conventional run-to-failure trajectories; rather, they reflect preventive or condition-informed replacement actions recorded during real shipboard operation.

For each validated degradation cycle, let H_c^- denote the hour-meter value at the previous cartridge replacement, H_c^+ the hour-meter value at the next validated cartridge replacement, and H_i the hour-meter value of an operational record i located within that cycle. The RUL label assigned to record i is defined as:

$$\mathrm{RUL}_i = \max\big(0,\ H_c^+ - H_i\big)\,, \tag{8}$$

where H_c^+ represents the replacement horizon and H_i represents the accumulated operating time associated with the operational record being labeled. The operator $\max(0,\cdot)$ prevents negative values for records located at or beyond the replacement boundary and preserves the physical interpretation of RUL as a non-negative remaining-time quantity.

Equivalently, if the observed service duration of cartridge cycle c is defined as $L_c = H_c^+ - H_c^-$, and the accumulated cartridge age at record i is $a_i = H_i - H_c^-$, then the same label can be written as:

$$\mathrm{RUL}_i = \max(0,\ L_c - a_i)\,. \tag{9}$$

For example, consider a cartridge that was installed at $H_c^- = 12{,}000$ h and replaced at $H_c^+ = 12{,}620$ h. The observed service duration of that cartridge cycle is therefore $L_c = 620$ h. If an operational record was collected at $H_i = 12{,}250$ h, the accumulated cartridge age at that record is $a_i = 250$ h, and the corresponding RUL label is:

$$\mathrm{RUL}_i = 12{,}620 - 12{,}250 = 370\ \mathrm{h}.$$

Thus, the operational feature vector measured at $H_i = 12{,}250$ h is paired with a continuous target value of 370 remaining operating hours until the next validated cartridge replacement. This procedure anchors the target variable to real maintenance evidence while avoiding dependence on calendar time, which may be irregular due to port stays, delayed logging, or non-uniform vessel operation.

(ii) Model training and testing (state-to-RUL mapping). For learning, the same event-driven horizon definition is retained, but the task is reformulated as predicting the previously constructed RUL label from the operational state rather than computing it directly from the hour-meter index. The resulting state-to-RUL mapping is expressed in Eq. 10:

$$\widehat{\mathrm{RUL}}(i) = f_\theta\big(\mathbf{x}_i\big), \tag{10}$$

where $f_\theta(\cdot)$ denotes the trained predictive model with hyperparameters θ, and $\mathbf{x}_i$ represents the operational feature vector associated with record i, including EAS and related thermodynamic, pressure, cooling, lubrication, and load-dependent indicators. In this stage, RUL_i is the event-driven target constructed during data preparation, whereas $\widehat{\mathrm{RUL}}(i)$ is the model-based estimate of the remaining cartridge life inferred from operating conditions.

The problem was formulated as a regression task because the operational quantity of interest is the remaining accumulated operating time until cartridge replacement, expressed in hours. This continuous estimate is directly useful for maintenance-window planning, spare-parts coordination, and intervention prioritization. A classification formulation based on RUL categories would require defining discrete thresholds, such as short, medium, or long remaining life. Unless these thresholds are explicitly linked to onboard maintenance policies, voyage duration, spare-part logistics, or risk tolerance, they may become arbitrary and may discard relevant temporal information. For instance, two records with substantially different remaining lives could be assigned to the same class, even though they imply different planning decisions.

In addition, the available dataset contains a limited number of validated degradation cycles. Discretizing the RUL range into categories could introduce class imbalance and reduce the amount of information available for learning. For these reasons, regression was adopted as the primary formulation in this stage. Nevertheless, categorical RUL bands remain a relevant future extension for operational decision-support interfaces, particularly if alert levels or maintenance-priority classes are later defined together with onboard maintenance personnel.

In this step implements model benchmarking and selection through PyCaret as an automated comparison layer across multiple regression families under a standardized preprocessing and cross-validation protocol.

PyCaret was used as a low-code AutoML environment to standardize model setup, preprocessing, training, cross-validation, and comparison across candidate regression algorithms (Ali, 2020). This enabled a reproducible benchmarking stage under the same data partitioning and evaluation criteria. Since the target variable is RUL expressed in accumulated operating hours, the error metrics were reported in time units. RMSE was selected as the primary selection criterion because it is directly interpretable in hours and penalizes large forecasting errors. MAE was used as a complementary metric because it provides the average absolute prediction error in hours and is less sensitive to extreme deviations. MSE was also reported as the squared-error counterpart of RMSE, with units of h^2, while R^2 was included as a dimensionless goodness-of-fit indicator.

Candidate models and their comparative performance are summarized in Table 9. Based on this screening, the Random Forest regressor was selected as the best-performing family. After identifying Random Forest, Optuna was used to refine its hyperparameters through an efficient and structured search over the model-configuration space (Akiba, Sano, Yanase, Ohta, & Koyama, 2019). This optimization was conducted under the same time-consistent evaluation protocol to preserve consistency with the benchmarking stage and reduce the risk of optimistic performance estimates caused by temporal leakage.

Table 9. Regression performance summary for candidate RUL models.

Model	RMSE (h)	R^2	MAE (h)	MSE (h^2)
Random Forest Regressor	66.94	0.87	34.63	4550.23
Light Gradient Boosting Machine	72.55	0.85	43.76	5331.48
Gradient Boosting Regressor	103.64	0.70	79.18	10749.84
Lasso Regression	170.99	0.18	136.62	29261.49
Ridge Regression	177.03	0.12	138.43	31452.87
Elastic Net	172.69	0.16	139.34	29862.94

Step 12 - Hyperparameter Optimization (Optuna)

After the AutoML benchmarking stage, the Random Forest Regressor was selected as the best-performing model family and was subsequently refined using Optuna. Hyperparameter optimization was therefore applied only to the selected Random Forest model, while the remaining candidate regressors were retained as benchmark baselines under the common PyCaret setup. The search space used in the Optuna optimization is summarized in Table 10.

Table 10. Optuna search space used for Random Forest optimization.

Hyperparameter	Search space
`n_estimators`	Integer: 200–1500
`max_depth`	Categorical: {10, 20, 30, `None`}
`min_samples_split`	Integer: 2–20
`min_samples_leaf`	Integer: 1–10
`max_features`	Categorical: {`sqrt`, `log2`}
`bootstrap`	Categorical: {`True`, `False`}
`n_jobs`	Fixed: -1
`random_state`	Fixed: 42

After this optimization stage, the Random Forest model achieved a validation RMSE of 42.93 h with $R^2 = 0.942$, improving upon the pre-tuning baseline reported in Table 9. The selected configuration included `n_estimators = 300`, `max_depth = 20`, `min_samples_split = 2`, `min_samples_leaf = 1`, and `bootstrap = True`. This configuration favored a sufficiently expressive ensemble, with enough trees and tree depth to capture nonlinear degradation-related patterns while preserving the robustness associated with Random Forest averaging.

The Optuna optimization results showed that `min_samples_leaf` and `n_estimators` were the most influential hyperparameters in the final Random Forest configuration, with relative importance values of 0.39 and 0.34, respectively. This suggests that model performance was mainly affected by the granularity allowed at the terminal nodes and by the number of trees included in the ensemble. These findings are consistent with the need to balance local sensitivity to operating-state variations with generalization across heterogeneous ship operating regimes.

4.5. Evaluation (E)

The Evaluation phase assesses the finalized model on a held-out test segment designed to approximate prospective operation, emulating a deployment workflow in which new operational batches are ingested, processed with fixed rules, and scored without reusing future information.

Step 13 - Apply of steps 8 and 9 with latest-voyage data

In this stage evaluates the finalized artifact on a held-out test segment designed to approximate prospective operation. As indicated in Figure 4, this stage uses the most recent voyage data as an external validation slice, applies the same preprocessing logic (Steps 8 and 9) to ensure feature and label consistency, and then computes out-of-sample errors. This design emulates a deployment workflow in which new operational batches are ingested, processed with fixed rules, and scored without reusing future information. The prediction error plot in Figure 7 summarizes the agreement between estimated and observed RUL under this validation setting.

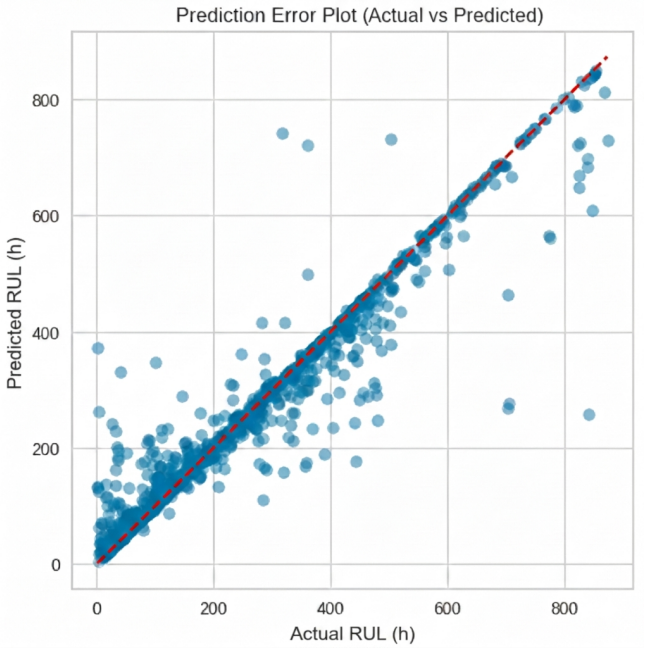


Figure 7. Prediction error plot comparing estimated versus observed RUL on the evaluation set.

5. RESULTS

The finalized model was assessed on the held-out "latest voyage" set (Step 13), yielding an RMSE of 52.92 h with R^2 of 0.921. This error scale supports the model as a defensible baseline for time-to-replacement estimation under operational variability, is suitable for window-based planning and prioritization, and motivates future work on richer condition indicators and production-like evaluation.

Figure 8 summarizes the empirical distribution of prediction errors, defined as $e = \hat{y} - y$, on the held-out evaluation set. The histogram provides a nonparametric estimate of the residual density $\hat{f}_e(\cdot)$, while the overlaid kernel density estimate offers a smooth approximation of the same quantity. The mass concentration near $e \approx 0$ indicates that most predictions are closely centered around the ground true, whereas the visibly heavier tails and sparse extreme values suggest non-negligible probability of large deviations, consistent with operational regimes that are harder to model. The near-coincidence of mean and median around zero is indicative of limited global bias (i.e., $E[e] \approx 0$), while the spread—summarized by the marked $\pm 1\sigma$ band—quantifies the typical dispersion of errors and complements RMSE as a scale-sensitive measure. Overall, the residual shape supports the interpretation of the reported RMSE as driven primarily by a narrow central error mode with occasional large-magnitude errors, motivating future work on uncertainty-aware prediction intervals and feature enrichment to mitigate tail risk (Figure 8).

The concentration of probability mass around zero indicates

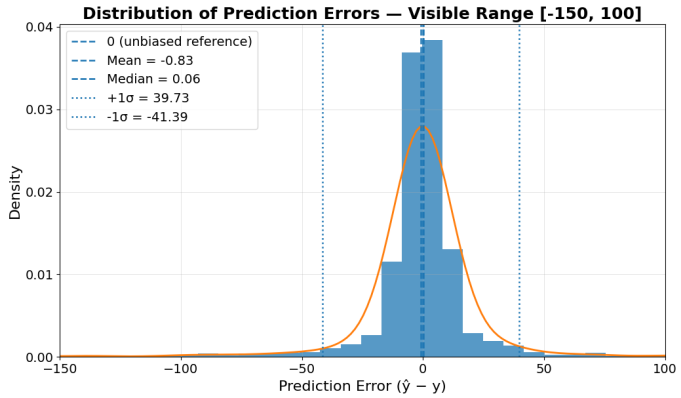


Figure 8. Distribution of prediction errors $e = \hat{y} - y$ on the held-out "latest voyage" test set.

that the model is largely unbiased in expectation, i.e., $E[e] \approx 0$, while the dispersion captured by the $\pm 1\sigma$ band reflects the typical magnitude of prediction deviations. In this context, the reported RMSE of 52.92 h can be interpreted as the root second moment of the residual distribution, $RMSE = \sqrt{E[e^2]}$, indicating that the overall error scale is primarily driven by the central spread of the distribution with additional contribution from the heavier tails observed.

6. DEPLOYMENT CONSIDERATIONS

Although the following consumption approach is proposed, packaging the trained model with (i) an explicit feature list (schema contract), (ii) tuned hyperparameters, and (iii) training-time evaluation metrics, and then validating incoming operational batches against the schema, applying the same preprocessing pipeline, and scoring to produce an RUL estimate, its current scope is limited to manual, offline testing. At this stage, the procedure is intended to support controlled experiments and analyst-driven validation. At the same time, a dedicated platform is developed to continuously monitor, process, and archive operational data for systematic retraining. The required automation (data ingestion, quality monitoring, labeling support, retraining triggers, and governance) is therefore explicitly deferred and described in the future-work roadmap.

7. CONCLUSIONS

This work presented a CRISP-DM-guided and PHM-oriented pipeline to estimate the Remaining Useful Life (RUL) of a critical shipboard asset on the training ship under realistic operational constraints. The central problem addressed in this study is the lack of a predefined maintenance-planning criterion for machinery operating under variable-load conditions, where degradation is not governed by constant operating regimes but by dynamic and uncertain usage patterns. In this context, maintenance decisions based exclusively on fixed-time routines or accumulated experience may be insufficient

to anticipate failure-related events or to optimize intervention timing.

The fragmented, manual, and largely analog nature of the current shipboard information-management framework constituted a key practical consideration in the development of the study. Operational and maintenance information was distributed across handwritten logs, isolated spreadsheets, and non-standard reports, resulting in limited traceability, heterogeneous formats, and reduced analytics readiness. These conditions guided the methodological actions adopted, including manual reconstruction, data cleaning, event extraction, and reconciliation of maintenance evidence. Consequently, the study demonstrates that PHM-oriented knowledge can be progressively extracted from legacy shipboard records when they are systematically interpreted, validated, and structured.

Within the scope of the main propulsion engine and its associated components, with emphasis on the filter replacement process as a measurable and operationally actionable target, the study showed that it is feasible to construct prognostic labels from maintenance events and to learn data-driven mappings between operating conditions and time-to-replacement. A key research contribution lies in the identification and formulation of an event-driven RUL labeling strategy, where the remaining useful life is retrospectively constructed from validated replacement or intervention records and accumulated operating hours. This approach provides a supervised learning basis for anticipating maintenance-relevant events, supporting a transition from purely time-based replacement practices toward condition-informed prognostic reasoning.

The proposed modeling approach was intentionally limited to Machine Learning methods rather than Deep Learning architectures. This decision was consistent with the available data conditions, including the limited volume of historical records, the manual origin of the data, and the need for interpretability in a maintenance-management context. Under these constraints, the use of supervised regression models allowed the study to prioritize traceability, explainability, and practical applicability over model complexity. In addition, expert validation through a Delphi-based process strengthened the interpretation of maintenance events, variable relevance, and operational assumptions, reducing the risk of constructing labels or modeling relationships that were inconsistent with onboard practice.

Although the resulting model provides a methodological basis for RUL estimation, its current state should be understood as a research and decision-support prototype rather than as an immediately deployable operational system. The absence of standardized digital event logging, continuous data capture, robust integration with maintenance-management systems, and real-time validation limits the possibility of onboard deployment at this stage. Nevertheless, the study identifies the specific organizational and technical conditions required to move toward future implementation, including unified data models, structured maintenance-event coding, standardized inspection templates, consistent sensor and engineering-round records, and stronger data governance practices.

The potential impact of this approach is significant at operational, tactical, and organizational levels. At the operational level, RUL estimation could help technical crews anticipate maintenance needs, reduce uncertainty in filter replacement decisions, and improve the use of available maintenance windows. At the tactical level, it could support maintenance planning, spare-parts management, workload prioritization, and coordination between voyage schedules and intervention opportunities. At the organizational level, the pipeline provides evidence that digital transformation initiatives in naval maintenance should not be limited to digitizing existing formats, but should aim to produce reliable, standardized, and model-ready data for future PHM applications.

Consequently, the main contribution of this work is threefold: (i) a replicable end-to-end methodology for developing RUL estimation from legacy shipboard maintenance records under realistic data constraints; (ii) an event-driven RUL labeling strategy that enables supervised learning for anticipating failure-related or maintenance-relevant events; and (iii) a clear identification of the digitalization, standardization, and governance improvements required to increase data reliability, improve future model training, and support the long-term adoption of data-driven maintenance in naval contexts. In this sense, the work contributes not only to the development of a predictive model, but also to the methodological foundation needed to evolve from manual and experience-based maintenance management toward a more traceable, condition-aware, and PHM-enabled maintenance framework.

7.1. Future Work

Future work should advance from the current research prototype toward a controlled validation and monitoring framework for eventual operational use. A first priority is the implementation of a data-validation layer capable of checking feature schema consistency, units, plausible ranges, missing values, and event-label integrity before model inference. This is essential to reduce the risk of unreliable predictions in a shipboard context where data are still partially manual and heterogeneous.

A second direction is to define model lifecycle criteria within an MLOps-oriented framework, including measurable indicators of data drift, performance degradation, and conditions for model review, recalibration, or retraining. Given the constraints of naval operation, such a framework should consider intermittent connectivity, limited onboard computational capacity, and the impracticality of continuous retraining. Therefore, future deployment studies should evaluate the trade-off between predictive accuracy, model complexity, inference cost,

and maintainability.

From a modeling perspective, subsequent experiments should compare the current Machine Learning baseline with alternatives capable of incorporating temporal behavior. Gradient boosting models combined with lag-based feature engineering represent a practical next step, as they may capture part of the degradation dynamics while preserving efficiency and interpretability. As data volume, quality, and continuity improve, sequential architectures such as LSTM or Temporal Convolutional Networks could also be assessed to determine whether their additional complexity provides operationally meaningful gains.

Finally, future work should strengthen the data foundation required for PHM adoption by promoting standardized maintenance-event logging, consistent asset coding, and better integration between operational records and maintenance evidence. These improvements would increase the reliability of RUL labels, support larger and higher-quality training datasets, and enable more robust condition-aware maintenance planning at operational, tactical, and organizational levels.

ACKNOWLEDGMENT

The authors thank COTECMAR for providing the space and resources required to carry out this research. We also acknowledge 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 made this research possible. The authors further extend their gratitude to the personnel of the Colombian Navy (ARC) for facilitating the visits to the training ship, supporting the interview processes, and providing access to valuable operational knowledge. The authors also sincerely thank the experts who participated in the Delphi methodology, whose contribution was a fundamental pillar of this research.

NOMENCLATURE

AR	acceptance rate
BU	Business Understanding (CRISP-DM)
D	Deployment (CRISP-DM)
d	evaluation dimension
DP	Data Preparation (CRISP-DM)
DU	Data Understanding (CRISP-DM)
E	Evaluation (CRISP-DM)
H_c	next replacement horizon
$h(t)$	accumulated operating age
$\mathbf{1}(\cdot)$	indicator function
IQR	interquartile range
m	number of evaluated cycles
M	Modeling (CRISP-DM)
n	number of experts
$Q_{1,i}^{(d)}$	first quartile of ratings for cycle i in dimension d
$Q_{3,i}^{(d)}$	third quartile of ratings for cycle i in dimension d
$r_{ij}^{(d)}$	rank assigned by expert j to cycle i in dimension d
$R_i^{(d)}$	sum of ranks for cycle i in dimension d
R^2	coefficient of determination
$RMSE$	root mean square error
EAS	engine angular speed
RUL	Remaining Useful Life
$S^{(d)}$	dispersion of rank sums in dimension d
W	Kendall's coefficient of concordance
$x_{ij}^{(d)}$	ordinal rating by expert j to cycle i in dimension d
$\hat{x}$	feature vector (input to the model)
$\tilde{x}_i^{(d)}$	median rating of cycle i in dimension d
$\bar{R}^{(d)}$	mean rank sum in dimension d
c_{acc}	minimum acceptable rating category

REFERENCES

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th acm sigkdd international conference on knowledge discovery and data mining* (pp. 2623–2631). doi: 10.1145/3292500.3330701

Ali, M. (2020). *Pycaret: An open source, low-code machine learning library in python.* `https://www.pycaret.org`. (Software)

Ansari, F., Glawar, R., & Nemeth, T. (2019). Prima: A prescriptive maintenance model for cyber-physical production systems. *International Journal of Computer Integrated Manufacturing*, *32*(4–5), 482–503. doi: 10.1080/0951192X.2019.1571236

Asimakopoulos, I. (2023). Data-driven condition monitoring of two-stroke marine diesel engines. *Ships and Offshore Structures*. doi: 10.1080/17445302.2023.2237302

Cao, H., Xiao, W., Sun, J., & Gan, M.-G. (2024). A hybrid data- and model-driven learning framework for remain-

ing useful life prediction. *Engineering Applications of Artificial Intelligence*. doi: 10.1016/j.engappai.2024.108557

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). Crisp-dm 1.0: Step-by-step data mining guide [Computer software manual].

Dalzochio, J., Kunst, R., Pignaton, E., Binotto, A., Sanyal, S., Favilla, J., & Barbosa, J. (2020). Machine learning and reasoning for predictive maintenance in industry 4.0: Current status and challenges. *Computers in Industry*, *123*, 103298. doi: 10.1016/j.compind.2020.103298

Doctrina de material naval. tomo iii. mantenimiento (Segunda edición ed.) [Computer software manual]. (2022). Bogotá, D.C., Colombia. (Doctrina Naval – Nivel Operacional (DOC. MAT. NAV.), ARC OP4-3.3)

Duan, X., Vasudevan, A., et al. (2022). A data scientific approach towards predictive maintenance applications. In *Advances in transdisciplinary engineering.* doi: 10.3233/ATDE220148

Fan, A., Sun, S., Hu, Z., Vladimir, N., & Mao, W. (2026). Enhancing transfer learning strategies for ship fuel consumption prediction under data scarcity. *Ocean Engineering*, *351*, 124398. doi: 10.1016/j.oceaneng.2026.124398

Florentino, R. B., & Moura, L. G. L. (2025). The estimation of the remaining useful life of ceramic plates used in iron ore filtration through a reliability model and machine learning methods applied to industrial process variables of a pims. *Applied Sciences*.

Gulati, R. (2013). *Maintenance and reliability best practices* (3rd ed.). New York, NY: Industrial Press.

Hagmeyer, S., & Zeiler, P. (2023). A comparative study on methods for fusing data-driven and physics-based approaches for rul prediction in filtration processes. *IEEE Access*. doi: 10.1109/ACCESS.2023.3265722

Han, P., Ellefsen, A. L., Li, G., Holmeset, F. T., & Zhang, H. (2021). Fault detection with lstm-based variational autoencoder for maritime components. *IEEE Sensors Journal*, *21*(19), 21903–21912. doi: 10.1109/JSEN.2021.3105226

Hasson, F., Keeney, S., & McKenna, H. (2000). Research guidelines for the delphi survey technique. *Journal of Advanced Nursing*, *32*(4), 1008–1015. doi: 10.1046/j.1365-2648.2000.t01-1-01567.x

International Maritime Organization. (2010). *International Safety Management (ISM) Code and Guidelines on Implementation of the ISM Code.* (IMO, London)

Jeon, M., Noh, Y., et al. (2021). Data gap analysis of ship and maritime data using meta learning. *Applied Soft Computing*. doi: 10.1016/j.asoc.2020.107048

Kalafatelis, A. S., Nomikos, N., Giannopoulos, A., Alexandridis, G., Karditsa, A., & Trakadas, P. (2025). Towards predictive maintenance in the maritime industry: A component-based overview. *Journal of Marine Science and Engineering*.

Kim, D., Antariksa, G., Handayani, M. P., Lee, S., & Lee, J. (2021). Explainable anomaly detection framework for maritime main engine sensor data. *Sensors*, *5200*.

Kirketerp-Møller, T., Hyldgaard, M. W., Cai, J., Dodis, A.-I., & Rytter, N. G. M. (2025). Data-driven predictive maintenance for two-stroke marine diesel engines using machine learning and mlops. *Journal of Ocean Engineering and Science*. doi: 10.1016/j.joes.2025.11.011

Kocak, G. (2023). Condition monitoring and fault diagnosis of a marine diesel engine with machine learning techniques. *Pomorstvo*. doi: 10.31217/p.37.1.4

Kumar, A., & Flores-Cerrillo, J. (2024). *Machine learning in Python for process and equipment condition monitoring, and predictive maintenance: From data to process insights* (1st ed.). MLforPSE. Retrieved 2026-03-01, from https://leanpub.com/ML-Python-for-PM-PdM (First published January 2024)

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

Michalski, M. A. D. C., et al. (2025). A multi-criteria framework for selecting machine learning techniques in predictive maintenance. *IEEE Access*. doi: 10.1109/ACCESS.2025.3604754

Mwangi, M. G. W., Park, M. H., Chun, K. W., Noh, J. H., Kim, C., & Lee, W. J. (2025). Hybrid ai-driven condition monitoring and rul forecasting for multi-fault diagnosis in two-stroke marine diesel engines. *Journal of Ocean Engineering and Science*.

Pan, Y., Kang, S., Kong, L., Wu, J., Yang, Y., & Zuo, H. (2025). Remaining useful life prediction methods of equipment components based on deep learning for sustainable manufacturing: a literature review. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, *39*, e4. doi: 10.1017/S0890060424000271

Pugalenthi, K., Park, H., Hussain, S., & Raghavan, N. (2021). Hybrid particle filter trained neural network for prognosis of lithium-ion batteries. *IEEE Access*, 135132–135143.

Quan, R., Cheng, G., Guan, X., Zhang, G., & Quan, J. (2025). A ho-bigru-transformer based pemfc degradation prediction method under different current conditions. *Renewable Energy*, *124132*.

Shi, Z., Wang, Z., & Ding, H. (2026). Multi-source data fusion for marine four-stroke diesel engine fault diagnosis: An adaptive weight transfer learning framework. *Energy Conversion and Management*, *353*, 121194. doi: 10.1016/j.enconman.2026.121194

Shifat, T. A., Yasmin, R., Hur, J.-W., & Park, H. (2021). A data driven rul estimation framework of electric motor using

machine learning. *Energies*. doi: 10.3390/en14113156

Sielaff, L., Lucke, D., & Wolf, Y. (2024). A reference model for predictive maintenance model development. In *Procedia cirp*.

Sun, J., Ren, H., et al. (2024). Fusion of multi-layer attention mechanisms and cnn-lstm for time-series prognostics in marine applications. *Journal of Marine Science and Engineering*. doi: 10.3390/jmse12060990

Upadrashta, D., & Wijaya, T. (2025). Ai/ml based anomaly detection and fault diagnosis of turbocharged marine diesel engines: Experimental study on engine of an operational vessel. *Information*. doi: 10.3390/info17010016

Velasco-Gallego, C., & Lazakis, I. (2023). Mar-rul: A remaining useful life prediction approach for fault prognostics of marine machinery. *Applied Ocean Research*, *140*, 103734. doi: 10.1016/j.apor.2023.103734

Zeiler, P., & Hagmeyer, S. (2023). A comparative study on methods for fusing data-driven and physics-based models for hybrid remaining useful life prediction of air filters. *IEEE Access*, 35737–35753.

Zhang, J., Jiang, Y., Wu, S., Li, X., Luo, H., & Yin, S. (2022). Prediction of remaining useful life based on bidirectional gated recurrent unit with temporal self-attention mechanism. *Reliability Engineering & System Safety*, *221*, 108297. doi: 10.1016/j.ress.2021.108297

Zio, E. (2022). Prognostics and health management (phm): Where are we and where do we (need to) go in theory and practice. *Reliability Engineering & System Safety*, *218*, 108119. doi: 10.1016/j.ress.2021.108119

BIOGRAPHIES

Joan Suarez Mechanical Engineer, graduated from the Technological University of Bolívar in 2023. He has developed software prototypes for COTECMAR and web-based interfaces for engineering decision support. His work includes optimization libraries for freight-transport routing (e.g., genetic algorithms), and discrete-event simulation for decision-making under uncertainty using pseudo-random number generation and RAM-based analysis. He currently works at COTECMAR as a Reliability Engineer. He is pursuing the M.S. degree in Engineering with an emphasis on Computing at the Universidad Tecnologica de Bolivar, Cartagena, Bolívar, Colombia (since early 2024).

Clara Olimpia Guimarães de Paula is a student of Naval Construction Technology at the Rio de Janeiro State University (UERJ) in Rio de Janeiro, Brazil. She is currently completing a professional internship and conducting applied research at the Science and Technology Corporation for the Development of the Naval, Maritime, and Riverine Industry (COTECMAR) in Cartagena, Colombia. Her involvement has contributed to the research area focusing on the intersection of predictive maintenance and Industry 4.0 within the maritime sector, where her current work involves developing Machine Learning models, with a special focus on marine diesel engines and machinery.

Edwin Paipa 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 Carlos Martinez-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).

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 Commit-

tee. 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, natural language understanding, and 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.