Sustainable Anomaly Detection Framework for Autonomous Surface Ship: Adaptive Subsystem-Level Anomaly Detection Algorithm via MLOps

Minji KIM[1], Gwangho YUN[2], Hwasup JANG[3], and Jaecheul PARK[*]

[1,2,3,*] *Korean Register, Seoul, 06178, South Korea*
kimmji@krs.co.kr
ghyun@krs.co.kr
janghs@krs.co.kr
jchpark@krs.co.kr

ABSTRACT

For the stable operation of Maritime Autonomous Surface Ships (MASS), this study proposes a sustainable anomaly detection framework that integrates a subsystem-level Condition-Based Maintenance (CBM) model with an adaptive MLOps pipeline. The main engine is decomposed into 14 functional units, each monitored by a hybrid algorithm that combines an Attention-LSTM-AutoEncoder and an Isolation Forest to detect subtle anomalies. To address model performance degradation caused by gradual data drift in maritime environments, an Autonomous Maintenance Mechanism is developed. This mechanism utilizes state severity (Z-Score) and drift velocity (ΔZ) indicators to algorithmically distinguish between sudden physical faults and gradual sensor drift. Based on this distinction, the MLOps pipeline accumulates confirmed drift in a buffer and selectively retrains and redeploys models using local onboard data once sufficient evidence has been gathered, while bypassing suspected fault conditions to avoid learning anomalous patterns. Experiments on an engine testbed indicate that the proposed system can suppress the Anomaly Rate (AR_t) during data drift and help restore diagnostic reliability, suggesting a practical basis toward self-sustaining condition monitoring for MASS.

1. INTRODUCTION

Ship main engines operate under harsh conditions including high temperatures, pressures, and continuous vibration. Failures can pose severe threats to navigational safety and logistical continuity, making precise condition monitoring and predictive maintenance increasingly important (X. Xu et al., 2021). Traditional Time-Based Maintenance (TBM) does not adequately reflect real-time machinery status, which may lead to unnecessary maintenance or missed early failures (Ahmad & Kamaruddin, 2012). As Maritime Autonomous Surface Ship (MASS) technology advances and seafarer shortages intensify, demand for Condition-Based Maintenance (CBM) solutions is growing (Li et al., 2025; Proulx & Reichard, 2021).

Applying CBM to marine main engines is challenging due to their structural complexity: multiple dynamically interacting subsystems generate inherently multivariate, time-series sensor signals where early anomaly symptoms are difficult to capture using rule-based methods or single AI models (Vanem & Storvik, 2017; Malhotra et al., 2016). Deep learning architectures, particularly Transformer-based attention mechanisms, have shown promise in capturing complex sensor interactions and non-linear temporal patterns at the subsystem level (Liang, Knutsen, Vanem, Zhang, & Æsøy, 2024; J. Xu, Wu, Wang, & Long, 2022).

However, even high-accuracy models can suffer performance degradation from data drift in real operational environments, where the statistical characteristics of sensor data gradually change due to mechanical aging, sensor degradation, and seasonal shifts. Such drift can induce an increase in false alarms and degrade the system's diagnostic reliability. The conventional approach of periodic human intervention for model retraining is difficult to reconcile with the philosophy of autonomous shipping. Therefore, an MLOps framework that monitors data drift and helps maintain model reliability becomes as important as the anomaly detection algorithm itself.

This study proposes a Sustainable Anomaly Detection Framework via Adaptive Subsystem-Level MLOps, consisting of two stages: (1) a subsystem-level hybrid detection model that decomposes the main engine into 14 functional units to improve detection resolution, and (2) an autonomous MLOps pipeline that aims to algorithmically distinguish sudden anoma-

lies requiring immediate crew inspection from gradual drift requiring model updates, using Z-Score and ΔZ indicators. Through testbed experiments encompassing progressive degradation and abnormal scenarios, this study shows that the proposed MLOps-based system can detect subsystem-level anomalies while helping to restore diagnostic reliability by suppressing the Anomaly Rate (AR_t) during data drift events. Overall, this work contributes a technological basis toward improving the self-sustaining reliability of Maritime Autonomous Surface Ships (MASS).

2. RELATED WORK

Traditional ship engine maintenance has evolved from Time-Based Maintenance (TBM) toward Condition-Based Maintenance (CBM) driven by advances in IoT and sensor technology (X. Xu et al., 2021; Ahmad & Kamaruddin, 2012). Deep learning, particularly LSTM and Transformer-based architectures, has become the standard for multivariate time-series anomaly detection in complex machinery (Malhotra et al., 2016; Liang et al., 2024; J. Xu et al., 2022). Hybrid models combining Transformer-AutoEncoder structures with Isolation Forest achieve high precision by robustly evaluating reconstruction error distributions in high-dimensional spaces (Haque & Soliman, 2025). To avoid the "masking effect" in global models, subsystem-level modular approaches decompose facilities into component units for localized detection (H. Xu, Pang, Wang, & Wang, 2023). While AI models perform well in controlled environments, inevitable data drift in dynamic maritime settings degrades performance and inflates false alarm rates (Lu et al., 2019). Emerging MLOps frameworks address this through real-time monitoring and continuous retraining (Kodakandla, 2024; Ashraf et al., 2026); however, existing studies treat fault diagnosis and drift detection separately. This study distinguishes itself by integrating both into a unified adaptive framework using Z-Score and ΔZ indicators.

3. SUBSYSTEM-LEVEL ANOMALY DETECTION MODEL

3.1. Data Acquisition and Feature Engineering

The framework was developed using data from a failure simulation testbed consisting of a MAN B&W 6S46MC-C7 two-stroke engine and a Fuchino CFSR 20.0 dynamometer (up to 7,400 kW). Three sensor systems were used: an Alarm Monitoring System (AMS, 149 signals), a Performance Measurement Indicator (PMI, sampled at 33.3 kS/s/ch), and a Vibration Monitoring System (VMS, 16–32 kHz band). A total of 22 core features directly related to the No. 1 Cylinder System were utilized, summarized in Table 1. High-frequency PMI and VMS data were downsampled to 1 Hz, standard scaled, and segmented using a 60-second sliding window for model input.

The sensor signals were acquired through a National Instruments CompactDAQ chassis (cDAQ-9189) with dedicated analog input modules; the overall instrumentation and signal flow are illustrated in Figure 1, and the channel configuration is summarized in Table 2. The cylinder-pressure (PMI) channels were sampled at 33.3 kS/s/ch (NI 9253, 24-bit), while temperature, vibration, and crank-angle signals were acquired through the remaining modules and the engine-installed monitoring systems.

Table 1. Summary of Extracted Features by System Category

Category	System Characteristic	Extracted Features
AMS	Thermal Load & Fluid Dynamics	$T_{\text{liner,exh}}$, $T_{\text{liner,man}}$, $T_{\text{exh,out}}$, $T_{\text{cfw,out}}$ (total 4)
PMI	Combustion Efficiency & Pressure	$P_{\max}$, $P_{i,\text{comb}}$, Δ_{crank}, ΔP, $\frac{dP}{d\theta}$, κ_{EVO} (total 6)
VMS	Mechanical Vibration & Wear	$X_{\max}$, σ, X_{rms}, Crest factor, FC, RVF_{CH1}, RVF_{CH2} for CH1 & CH2 (total 12)

The combined instrument-level uncertainty of the acquisition chain is on the order of a few percent of the measured value (e.g., 0–350 bar cylinder-pressure sensors at 61.68–63.03 μA/bar; accelerometers at 100 mV/g $\pm 2\%$). Because the diagnosis logic operates on standardized Z-Score deviations from a per-channel baseline rather than on absolute readings, this uncertainty is largely absorbed by the baseline normalization; it nonetheless sets a lower bound on the smallest detectable condition change and was accounted for when configuring the warning threshold.

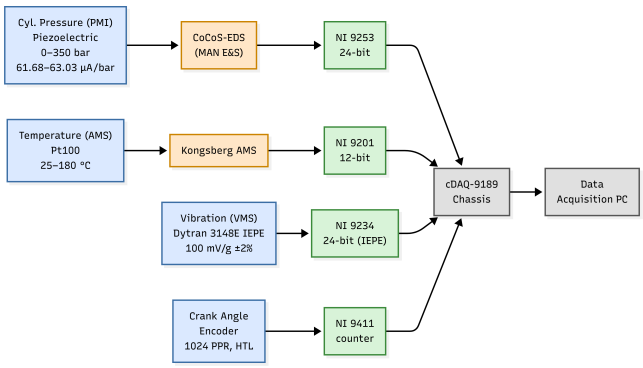

Figure 1. Instrumentation and data acquisition architecture of the engine condition-monitoring system.

3.2. Decomposition into Functional Component Units

Inputting hundreds of multivariate sensor signals into a single global model causes the "masking effect," where subtle incipient faults in specific components are obscured by normal variance of the overall system. To overcome this, the engine is decomposed into 14 functional units (e.g., cylinder, fuel injection, exhaust systems) based on physical domain knowl-

Table 2. Data Acquisition Configuration for the No. 1 Cylinder System (Two-Stroke Engine)

Signal	Sensor / Range	Module (Resolution)
Cyl. pressure (PMI)	Piezoelectric, 0–350 bar, 4–20 mA, 61.68–63.03 μA/bar	NI 9253 (24-bit)
Temperature (AMS)	Pt100, 25–180 °C, 0–10 V	NI 9201 (12-bit)
Vibration (VMS)	Dytran 3148E IEPE, 100 mV/g $\pm$2%	NI 9234 (24-bit, IEPE)
Crank angle	Optical encoder, 1024 PPR, HTL	NI 9411 (counter)

edge, with an independent deep learning model deployed for each unit (Figure 2). This modular architecture maximizes localized detection accuracy, enhances interpretability, and enables lightweight independent retraining via MLOps. This paper validates the framework using the No. 1 Cylinder System model, incorporating 22 multi-modal features across combustion pressure (PMI), temperature (AMS), and mechanical vibration (VMS) domains.

3.3. Attention-LSTM-AutoEncoder and Isolation Forest

The proposed hybrid model combines an Attention-LSTM-based AutoEncoder (Attention-LSTM-AE) with an Isolation Forest, as illustrated in Figure 3.

Step 1: Reconstruction Error Estimation via Attention-LSTM-AE. Given input sequence $X_t = [x_{t-w+1}, \ldots, x_t]$, a Multi-Head Attention layer (4 heads) first extracts inter-sensor correlations, assigning higher weights to fault-critical variables. Subsequent LSTM encoder layers ($64 \rightarrow 32$ nodes) compress data into latent vector h_t, while LSTM decoder layers ($32 \rightarrow 64$ nodes) reconstruct $\hat{X}_t$. The Reconstruction Error (RE) is defined as:

$$RE_t = \frac{1}{D}\|X_t - \hat{X}_t\|^2 \tag{1}$$

RE_t approaches zero for healthy states and spikes sharply upon anomaly occurrence.

Step 2: Isolation Forest Dynamic Anomaly Scoring. Rather than applying a static threshold to RE, the multidimensional RE distribution is fed into an Isolation Forest (IF, 100 trees) that dynamically separates normal load fluctuations from actual faults based on statistical sparsity (Haque & Soliman, 2025; H. Xu et al., 2023). The anomaly score s is:

$$s(x, n) = 2^{-\frac{E(h(x))}{c(n)}} \tag{2}$$

where $c(n)$ normalizes path length. $s \approx 1$ indicates anomaly; $s < 0.5$ indicates normal.

4. DATA DRIFT MONITORING AND AUTONOMOUS MAINTENANCE

While the subsystem-level model in Section 3 detects anomalies under a fixed data distribution, its reliability can degrade over time as the statistical characteristics of sensor signals shift due to mechanical aging, sensor degradation, and seasonal variation. A key difficulty is that such gradual drift and genuine mechanical faults can both manifest as elevated reconstruction errors, so naively retraining on drifted data risks absorbing actual fault patterns into the baseline. To address this, this section presents an autonomous maintenance mechanism that distinguishes the two cases before deciding whether to retrain. We first derive physics-informed engineered features that remain interpretable under varying load (Section 4.1), then introduce two complementary statistical indicators—state severity (Z-Score) and drift velocity (ΔZ)—that characterize both the magnitude and the rate of distribution change (Section 4.2). Based on these indicators, we define a health diagnosis logic and the corresponding MLOps trigger rules (Section 4.3), and finally describe the onboard redeployment pipeline that closes the autonomous maintenance loop (Section 4.4).

4.1. Domain Knowledge-Based Engineered Features

Three physics-informed features are derived (excluding RPM < 30):

- **Combustion Efficiency** (`F_Pmax_Eff`) $\hookrightarrow P_{\max}/\text{RPM}$
- **Thermal Load Sensitivity** (`F_Temp_Sens`) $\hookrightarrow Temp/\text{RPM}$
- **Vibration per Load** (`F_Vib_Load`) $\hookrightarrow RMS/\text{Fuel Index}$

As shown in Figure 4, each cell represents the absolute correlation between a pair of features (the 22 raw channels of Table 1 together with the three engineered features), with darker cells indicating stronger correlation, and the red boxes mark the rows and columns of the three engineered features. Their consistently low absolute correlation with the raw sensor channels and with one another indicates that they capture largely non-redundant information rather than restating existing measurements.

4.2. Statistical Drift Detection: Z-Score and ΔZ

Two complementary indicators enable autonomous state identification:

State Severity (Z-Score) measures absolute deviation of current batch mean $\mu_{i,t}$ from baseline reference $(\mu_{i,\text{ref}}, \sigma_{i,\text{ref}})$:

$$Z_{i,t} = \frac{\mu_{i,t} - \mu_{i,\text{ref}}}{\sigma_{i,\text{ref}} + \epsilon} \tag{3}$$

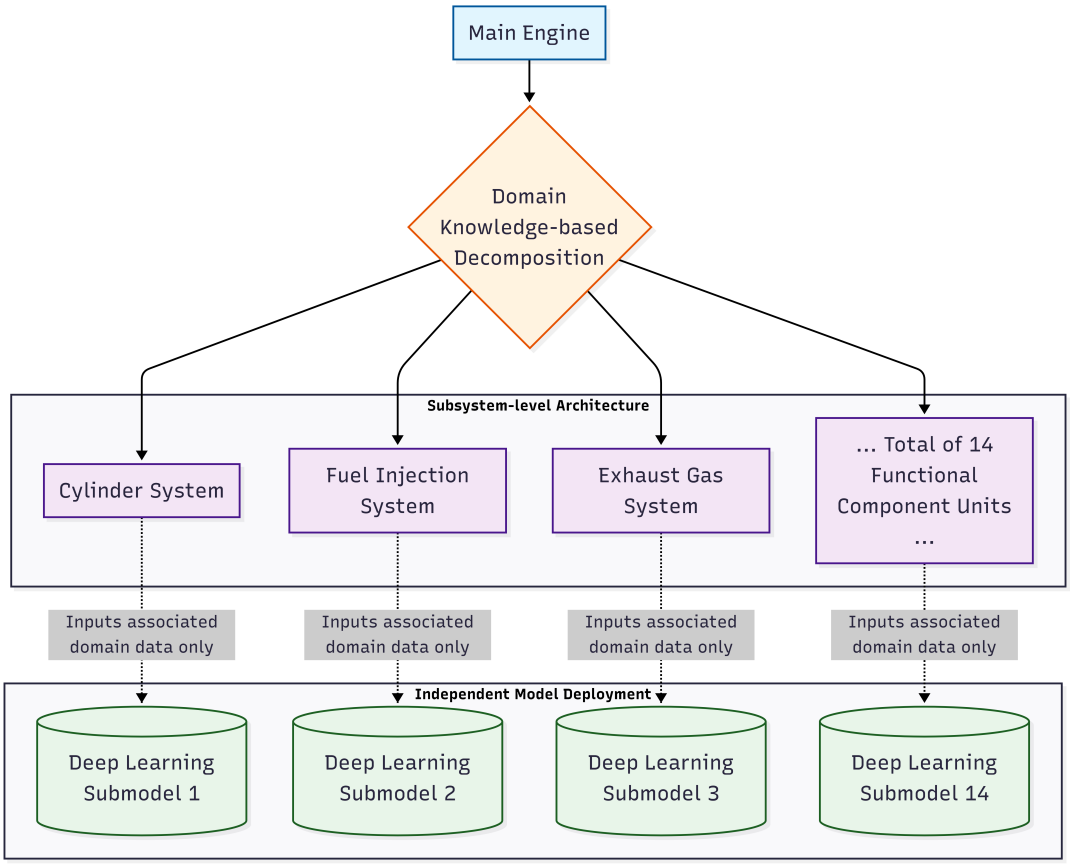

Figure 2. Proposed subsystem-level architecture of the main engine.

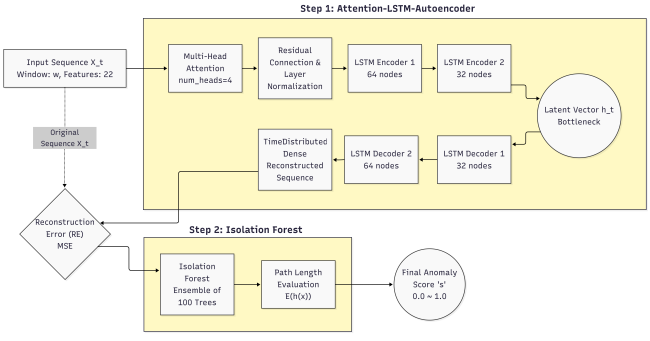

Figure 3. Flowchart of the proposed hybrid model.

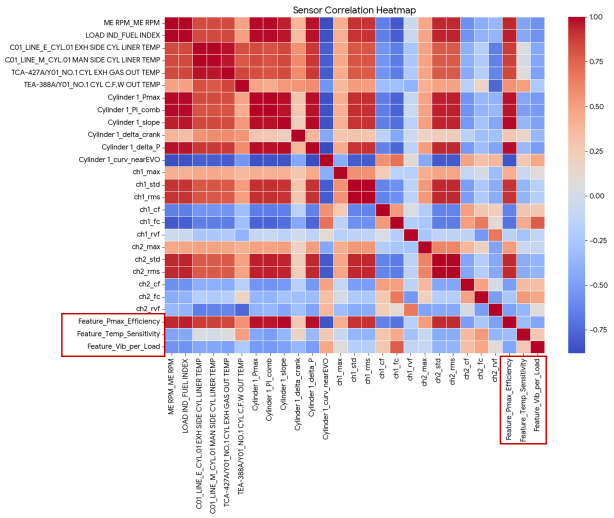

Figure 4. Sensor correlation heatmap of the 22 raw features (Table 1) together with the three engineered features (red boxes: `Feature_Pmax_Efficiency`, `Feature_Temp_Sensitivity`, and `Feature_Vib_per_Load`). The engineered features show consistently low absolute correlation with the raw channels and with each other, indicating that they provide largely non-redundant information.

Rate of Change (ΔZ) evaluates acceleration of severity relative to the previous batch:

$$\Delta Z_{i,t} = Z_{i,t} - Z_{i,t-1} \tag{4}$$

Their combination distinguishes sudden faults from gradual drift and underpins the four diagnostic states defined in Section 4.3. A high Z accompanied by a high ΔZ reflects a rapidly developing mechanical fault (*Potential Fault*), which must bypass retraining. In contrast, an elevated Z with a low ΔZ indicates an environmental or aging-induced distribution shift rather than a sudden fault; depending on its severity, this is diagnosed as *Gradual Aging* (warning-level deviation) or *Accelerated Aging* (critical-level deviation), both of which are treated as drift candidates that are accumulated for retraining under the rules in Section 4.3.

4.3. Health Diagnosis Logic and MLOps Trigger Rules

For N_{total} features, critical and warning counts are aggregated as follows:

$$N_{\text{critical}} = \sum_{i=1}^{N_{\text{total}}} \mathbb{I}(|Z_{i,t}| > 3.0) \tag{5}$$

$$N_{\text{warning}} = \sum_{i=1}^{N_{\text{total}}} \mathbb{I}(|Z_{i,t}| > 1.5) \tag{6}$$

The full rule set is summarized in Table 3. The system distinguishes four states—Stable, Potential Fault, Gradual Aging, and Accelerated Aging—and triggers differentiated Airflow pipeline actions accordingly. Critically, a breach characterized by high volatility (Potential Fault) bypasses retraining and issues a physical inspection alarm, as it indicates a sudden mechanical failure. Conversely, sustained breaches with a low rate of change indicate environmental or aging-induced drift; such batches are accumulated in the retraining buffer (subject to the AR_t gate described below) rather than triggering immediate retraining.

A dual-monitoring network integrates engine health diagnosis with the monitoring of model-data inconsistency, represented by the "Anomaly Rate (AR_t)". Since ground truth labels are unavailable in real-time maritime operations, AR_t is defined as the proportion of input samples within a monitoring batch that the model classifies as anomalous based on the reconstruction error distribution. It is calculated as follows:

$$AR_t = \frac{1}{N_{\text{batch}}} \sum_{i=1}^{N_{\text{batch}}} \mathbb{I}(s_i > 0.5) \tag{7}$$

where s_i is the anomaly score of the i-th sample derived from the Isolation Forest, N_{batch} is the total number of samples in a monitoring batch, and $\mathbb{I}(\cdot)$ is the indicator function that outputs 1 if the condition is met and 0 otherwise. A batch is counted as a drift sample when it is diagnosed as a drift-related state (Gradual or Accelerated Aging) *and* its model–data inconsistency is elevated ($AR_t \geq 7.0\%$); batches that are stable, or whose drift is not accompanied by elevated AR_t, are not counted. Faults (Potential Fault) are excluded from the buffer entirely. Over a sliding window of recent batches, retraining is triggered when the buffer holds at least 20 batches and the proportion of counted drift samples reaches 50% (the 20/50 Strategic Sample Rule). In this framework, AR_t serves as a label-free supporting gate confirming that a diagnosed drift is materially degrading model performance, rather than as a standalone diagnostic metric.

Both Gradual and Accelerated Aging are characterized by a low rate of change ($|\Delta Z_{i,t}| < 0.2$), which distinguishes environmental or aging-induced drift from sudden mechanical failures (the latter exhibiting high $|\Delta Z|$); Accelerated Aging additionally involves critical-level severity ($|Z_{i,t}| \geq 3.0$). The 20/50 Strategic Sample Rule is evaluated over a sliding window of recent batches: retraining is triggered only when at least 20 batches have accumulated in the buffer and at least 50% of them are counted as drift samples (i.e., GA/AA with $AR_t \geq 7\%$). Requiring a sustained majority of drift samples, rather than a single excursion, prevents transient fluctuations from triggering unnecessary retraining. Even when retraining proceeds, the subsequent validation gate guards against catastrophic forgetting of past fault patterns before any model is redeployed. The decision flowchart for this autonomous maintenance cycle is shown in Figure 5.

4.4. Onboard Redeployment Pipeline

Upon retraining trigger, the MLOps framework retrieves recent buffered data to learn the "new normal." Only subsystem models corresponding to drifted features are selectively retrained for computational efficiency. Updated models are versioned in the MLflow Model Registry, validated against catastrophic forgetting of past fault patterns, and hot-swap deployed to seamlessly replace obsolete models, completing a fully closed-loop autonomous maintenance cycle.

5. EXPERIMENTS AND RESULTS

5.1. Setup and Preprocessing

Experiments used streaming data from the engine testbed, selecting 22 core features for the No. 1 Cylinder System (4 AMS, 6 PMI, 12 VMS). Engine load was sequentially controlled from idle to 25%, 50%, 75%, 85%, and 100%, with a 20-minute stabilization and 5-minute collection phase per stage. A sliding window of size 60 with stride 1 generated 41 sequences per 100-tick block (input shape: $(N_{\text{batch}}, 60, 22)$), split 6:2:2 for training, validation, and testing. An Apache Airflow DAG ran continuous inference with batch size $= 100$,

Table 3. Health Diagnosis Logic and MLOps Pipeline Operations

Diagnosis	**Rule Set (Statistical Thresholds)**	**Action**	**Buffer / Pipeline**
Stable	No CRITICAL ($\|Z\| < 3.0$), WARNING count $< 50\%$, AND $\|\Delta Z\| \leq 0.2$	Normal Operation	Recorded, not counted as drift
Potential Fault	≥ 1 CRITICAL sensor AND $\|\Delta Z\| > 0.5$ (high-velocity spike)	Issue Crew Alarm; isolate fault data	Excluded from buffer
Gradual Aging	No CRITICAL, WARNING count $\geq 50\%$, AND $\|\Delta Z\| < 0.2$ (low-velocity shift)	Monitor / accumulate	Counted as drift if $AR_t \geq 7\%$
Accelerated Aging	≥ 1 CRITICAL sensor AND $\|\Delta Z\| < 0.2$ (high-severity, low-velocity drift)	Monitor / accumulate	Counted as drift if $AR_t \geq 7\%$

Retraining is triggered by the 20/50 Strategic Sample Rule: ≥ 20 batches buffered over a sliding window with $\geq 50\%$ counted as drift samples.

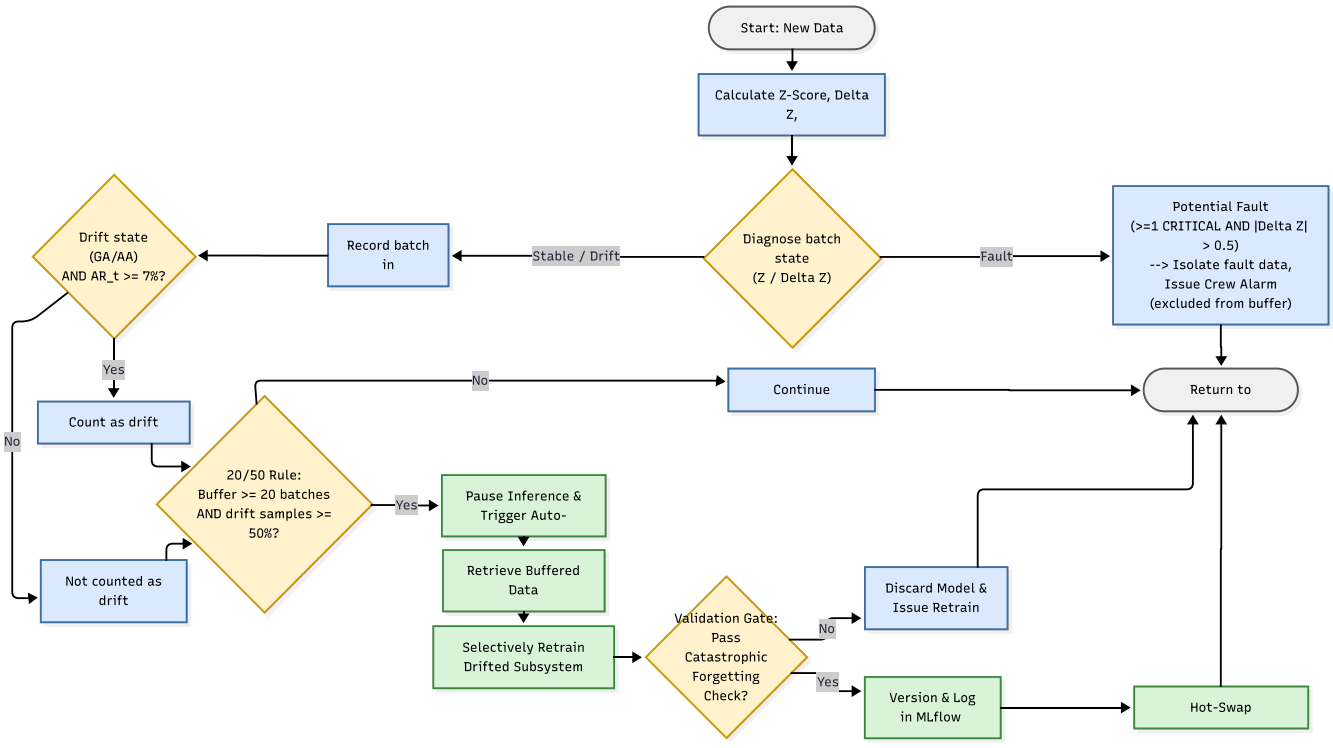

Figure 5. Decision flowchart of the proposed MLOps pipeline.

step size $= 80$, logging Z-Score and ΔZ in real time on the MLflow dashboard (Figure 6).

Figure 6. Real-time MLflow logging records of Z-Score and system health index.

5.2. Preprocessing Effectiveness

Table 4 compares anomaly detection performance across data configurations. Combining proposed engineered features with low-load filtering achieved 98.5% accuracy and a 1.8% false alarm rate, substantially outperforming raw sensor data baselines.

5.3. Fault Discrimination and Drift Detection

As illustrated in Figure 7, the framework successfully distinguished between two distinct conditions by evaluating both severity and volatility:

- **Sudden Anomaly:** During a fuel injection valve failure, the Z-Score of `Feature_Pmax_Efficiency` spiked significantly (> 3.0) with a high rate of change ($\Delta Z > 0.5$). Following the logic in Table 3, the system issued a "Potential Fault" alarm for immediate physical inspection and correctly bypassed the retraining pipeline to avoid learning the fault pattern.
- **Gradual Drift:** During periods of minor environmental fluctuation, multiple features exceeded the warning threshold (1.5) but maintained a low rate of change ($\Delta Z < 0.2$). This state was correctly diagnosed as "Gradual Aging" rather than a mechanical failure, and the corresponding batches were accumulated in the drift buffer for retraining once the 20/50 rule was satisfied.

5.4. Autonomous Adaptation During Data Drift

Figure 8 demonstrates the autonomous recovery process in three phases. During Drift Accumulation, the Diagnostic Reliability (representing the model's ability to correctly represent the operational state) declined from 97.5% to ∼85%. This decline was driven by increasing reconstruction errors as feature Z-Scores progressively breached critical thresholds (> 3.0) with $\Delta Z < 0.05$. Upon satisfying the MLOps Rule Set conditions, the system autonomously triggered the Airflow retraining pipeline without human intervention. Following the redeployment of the retrained model (v2.0), the Z-Scores reset to stable levels near 0 and the "Diagnostic Reliability" (accuracy) recovered from approximately 85% to over 97%, establishing a new baseline for the shifted data distribution. Because ground-truth labels are unavailable in real operation, this recovery is monitored online through the label-free AR_t, which returns to a low level after redeployment.

5.5. Sensitivity Analysis of Decision Thresholds

The sensitivity of the two decision thresholds was examined as summarized in Fig. 9. Both thresholds were evaluated on the testbed data. For the Z-Score threshold (Fig. 9(a)), the performance metrics vary only gradually over the $Z = 2.0$–3.0 range, suggesting that the selected value $|Z| = 3.0$ lies within a relatively stable region rather than a point of abrupt change. This value is consistent with the 3σ rule, which covers approximately 99.7% of a normal distribution, and is in line with the observation that the average Z-Score for normal operation (≈ 1.6) remains well below the critical threshold, whereas that of faulty operation (≈ 3.8) exceeds it. Although a normal-operation mean of ≈ 1.6 may momentarily place individual channels above the warning level ($|Z| > 1.5$), the Stable state requires fewer than 50% of channels to exceed this level and none to reach the critical level ($|Z| > 3.0$), so such fluctuations do not trigger a false diagnosis.

The ΔZ threshold separates sudden faults (high rate of change) from gradual wear-induced drift (low rate of change). As shown in Fig. 9(b), a small ΔZ threshold routes most sudden faults to inspection but tends to misclassify gradual drift, whereas a large threshold shows the opposite tendency; over the evaluated range, the balanced score is highest near $\Delta Z = 0.2$, which guided the selection of this value. While these results indicate a reasonable operating region for both thresholds, the analysis is based on the current testbed and further validation across a broader set of in-service conditions remains useful as future work.

6. LIMITATIONS AND FUTURE WORK

Several limitations should be acknowledged when interpreting the present results. First, all data were collected from a controlled engine testbed, and the fault and drift scenarios were induced through simulated or manually configured conditions rather than observed during actual voyages. While the testbed reproduces realistic load profiles and fault signatures, it cannot fully capture the long-term, compound effects of real maritime operation, such as multi-seasonal environmental variation, cumulative mechanical wear, and the gradual sensor degradation that develops over a vessel's service life. The framework's behavior under such prolonged, continuous operating conditions therefore remains to be validated.

Table 4. Comparison of anomaly detection performance across feature configurations.

Feature Set	Low-load Filter	Accuracy (%)	F1-Score	False Alarm Rate (%)
Raw Sensor Data	×	72.4	0.68	18.4
Raw Sensor Data	✓	84.2	0.81	9.2
Engineered Features	×	81.5	0.79	12.1
Engineered Features (Proposed)	✓	**98.5**	**0.98**	**1.8**

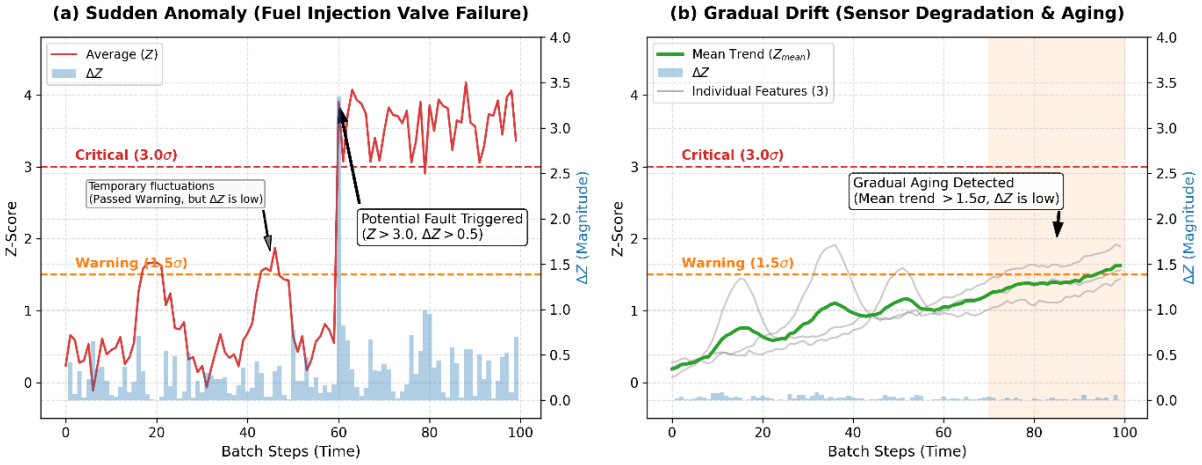

Figure 7. System responses: Sudden anomaly (left) vs. Gradual drift (right).

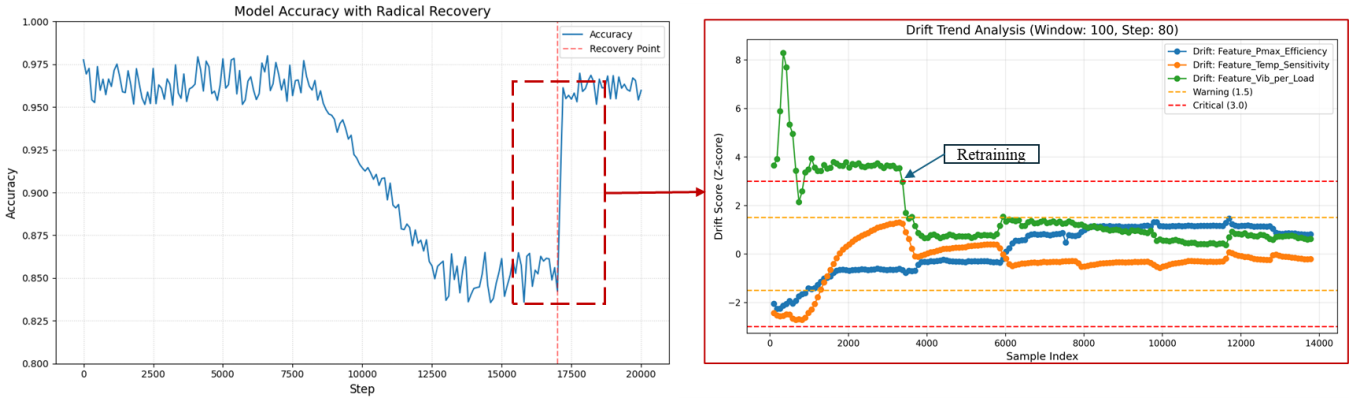

Figure 8. Autonomous adaptation of the MLOps pipeline during data drift: recovery of Diagnostic Reliability (left) and Z-score reset (right).

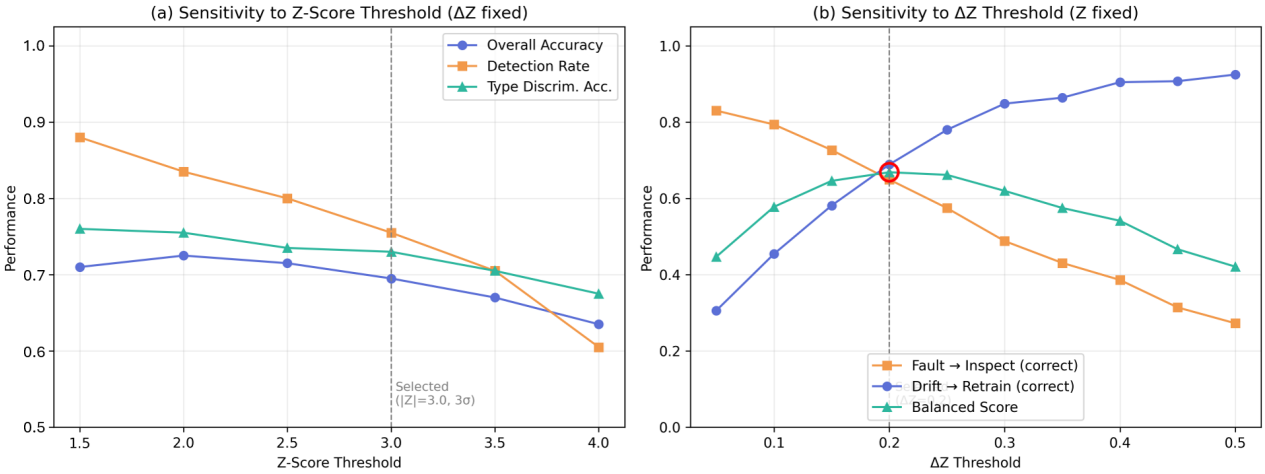

Figure 9. Sensitivity analysis of the decision thresholds. (a) Sensitivity to the Z-Score threshold, computed from the testbed data (22 features, window size 60). The overall accuracy and the type discrimination accuracy vary gradually over the $Z = 2.0$–3.0 region, and the selected value $|Z| = 3.0$ corresponds to the conventional 3σ criterion. (b) Sensitivity to the ΔZ threshold, computed from the testbed data. A small ΔZ threshold routes most sudden faults to inspection but tends to misclassify gradual drift, whereas a large threshold shows the opposite tendency, with the balanced score highest near $\Delta Z = 0.2$.

Second, the statistical thresholds used in the diagnosis logic—most notably the critical and warning Z-Score levels, the ΔZ velocity bounds, the AR_t trigger point, and the 20/50 sample rule—were configured based on engineering judgment and the characteristics of the present testbed data. Although the sensitivity analysis in Section 5.5 indicated that the selected values lie within a stable operating region for this testbed, their transferability across different engine types, vessel classes, or sensor configurations has not yet been established and requires broader validation.

To address these limitations, future work will proceed along three directions. First, we plan to validate and recalibrate the framework using long-term operational data from in-service vessels, enabling assessment under genuine drift and degradation rather than induced scenarios. Second, building on the present sensitivity analysis, we will investigate data-driven or adaptive threshold-setting methods (e.g., distribution-based calibration of the baseline reference) so that the diagnosis logic can generalize across heterogeneous platforms with reduced manual tuning. Finally, we intend to extend the framework to additional shipboard equipment, such as propulsion shafting and generators, and to investigate model lightweighting for edge deployment in communication-constrained maritime environments.

7. CONCLUSION

This study proposed and evaluated an MLOps-based adaptive anomaly detection framework for Maritime Autonomous Surface Ship (MASS) main engines. The key contributions are summarized as follows:

- **Hybrid Architecture for Diagnostic Robustness:** The Attention-LSTM-AE and Isolation Forest combination helps suppress false alarms during normal load fluctuations while isolating physical defects, such as exhaust valve leaks.
- **State Discrimination via Z-Score and ΔZ:** The framework introduces a velocity-based logic that aims to decouple sudden anomalies requiring immediate inspection from gradual drift requiring model updates, so that suspected mechanical failures are less likely to be absorbed into the baseline by the autonomous retraining process.
- **Autonomous Closed-Loop Maintenance:** The Apache Airflow and MLflow orchestrated pipeline uses the Anomaly Rate (AR_t)—a label-free proxy for model–data inconsistency—to admit drifted batches into the retraining buffer, retrains selectively using local buffer data, and redeploys updated models via hot-swapping. In the labeled testbed experiments, this AR_t-driven cycle restored the diagnostic reliability (accuracy), which had declined from approximately 97.5% to 85% during drift accumulation and recovered to over 97% after autonomous redeployment, confirming that the label-free maintenance loop tracks true model performance.

Beyond the specific algorithm, this study contributes to the PHM research community and the maritime industry in three respects. First, it addresses a gap in maritime PHM, where fault diagnosis and data-drift handling have often been studied in isolation, by integrating them into a single lifecycle that links detection, diagnosis, and model maintenance. Second, the velocity-based distinction between sudden faults and gradual drift offers a transferable design principle for safety-

critical PHM applications in general, where indiscriminate retraining on shifting data can erode the very fault patterns a model is meant to recognize. Third, by demonstrating an Airflow- and MLflow-orchestrated pipeline on a marine main-engine testbed, this work provides a reference case for deploying self-maintaining CBM toward the operational reliability requirements of MASS, an area of growing importance as crewless operation and digital classification advance.

Ultimately, this framework provides a strong baseline for sustainable CBM in autonomous maritime systems. While the current evaluation is based on engine testbed data, the transition to real-world, in-service vessel operations—supported by further empirical validation and framework expansion—will be the next critical step toward ensuring the long-term operational reliability of MASS.

ACKNOWLEDGMENT

This work was supported by the Technology Innovation Program (RS-2024-00458756, Development of Engine Equipment for Ships Integrating Intelligent Autonomous Maintenance System) funded by the Ministry of Trade, Industry & Energy (MOTIE, Korea).

NOMENCLATURE

AI & Mathematical Symbols

t	Current time step
w	Sliding window size
D	Number of input features
$X_t, \hat{X}_t$	Input / Reconstructed sequence at time t
RE_t	Reconstruction Error at time t
s_i	Anomaly score of the i-th sample
$c(n)$	Avg. path length of BST (normalization)
AR_t	Anomaly Rate at time t
N_{batch}	Number of samples in a monitoring batch

Health Monitoring & Statistical Indicators

$Z_{i,t}$	State severity (Z-Score) for feature i at time t
$\Delta Z_{i,t}$	Rate of change for feature i at time t
$\mu_{\text{ref}}, \sigma_{\text{ref}}$	Baseline reference mean and std. deviation
$N_{\text{crit}}, N_{\text{warn}}$	Features exceeding critical / warning thresholds

Abbreviations

MLOps	Machine Learning Operations
LSTM	Long Short-Term Memory
AE	AutoEncoder
IF	Isolation Forest
AMS/PMI/VMS	Alarm Monitoring / Performance Measuring / Vibration Monitoring System

REFERENCES

Ahmad, R., & Kamaruddin, S. (2012). An overview of time-based and condition-based maintenance in industrial application. *Computers & Industrial Engineering*, *63*(1), 135–149. doi: 10.1016/j.cie.2012.02.002

Ashraf, W. M., Ansar, T., Ahmed, F., Hussain, J., Abbas, M. M., & Dua, V. (2026). From drift to adaptation to the failed ml model: Transfer learning in industrial mlops. *arXiv preprint arXiv:2602.00957*.

Haque, A., & Soliman, H. (2025). A transformer-based autoencoder with isolation forest and xgboost for malfunction and intrusion detection. *Future Internet*, *17*(4), 164. doi: 10.3390/fi17040164

Kodakandla, N. (2024). Data drift detection and mitigation: A comprehensive mlops approach for real-time systems. *International Journal of Science and Research Archive*, *12*(1), 3127–3139.

Li, H., Meng, X., Liu, J., Zhang, W., Zhou, X., & Yang, X. (2025). A framework of predictive maintenance for maritime autonomous surface ships considering component degradation. *Journal of Marine Science and Engineering*, *13*(3), 512. doi: 10.3390/jmse13030512

Liang, Q., Knutsen, K. E., Vanem, E., Zhang, H., & Æsøy, V. (2024). Unsupervised anomaly detection in marine diesel engines using transformer neural networks and residual analysis. *International Journal of Prognostics and Health Management*, *15*(1), 1–18.

Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., & Zhang, G. (2019). Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering*, *31*(12), 2346–2363. doi: 10.1109/TKDE.2018.2876857

Malhotra, P., Ramakrishnan, A., Anand, G., Goyal, L., Lodha, P., & Singh, P. (2016). Lstm-based encoder-decoder for multi-sensor anomaly detection. *arXiv preprint arXiv:1607.00148*.

Proulx, C., & Reichard, K. (2021). Automated condition-based maintenance (cbm) using artificial intelligence (ai) and machine learning (ml) for unmanned systems. *Annual Conference of the PHM Society*, *13*(1).

Vanem, E., & Storvik, G. O. (2017). Anomaly detection using dynamical linear models and sequential testing on a marine engine system. *International Journal of Prognostics and Health Management*, *8*(1).

Xu, H., Pang, G., Wang, Y., & Wang, Y. (2023). Deep isolation forest for anomaly detection. *IEEE Transactions on Knowledge and Data Engineering*, *35*(12), 12591–12604. doi: 10.1109/TKDE.2023.3270299

Xu, J., Wu, H., Wang, J., & Long, M. (2022). Anomaly transformer: Time series anomaly detection with association discrepancy. In *International conference on learning representations (iclr)*.

Xu, X., Yan, X., Yang, K., Zhao, J., Sheng, C., & Yuan,

C. (2021). Review of condition monitoring and fault diagnosis for marine power systems. *Transportation Safety and Environment*, *3*(2), 85–102. doi: 10.1093/tse/tdab004

BIOGRAPHIES

Minji KIM received her B.S. degree from Korea Maritime & Ocean University, Busan, South Korea, in 2017, and her M.S. degree from Inha University, Incheon, South Korea, in 2020. After holding roles at Hanwha Ocean and the Research Institute of Medium & Small Shipbuilding (RIMS), she joined the Korean Register (KR) as a Senior Researcher at the AI Convergence Center. Her research focuses on AI-based maritime systems, object detection, and Machine Learning Operations (MLOps). She is a recipient of outstanding paper and poster awards from the Journal of Computational Design and Engineering (JCDE) and the Society of CAD/CAM Engineers (CDE).

Gwangho YUN received the B.S. (2021) and M.S. (2023) degrees in Naval Architecture and Ocean Engineering from Pusan National University, Busan, South Korea. He is currently a Researcher at the Korean Register (KR). His research interests include autonomous surface vessels and condition-based maintenance (CBM).

Hwasup JANG received his B.S., M.S., and Ph.D. degrees in Civil Engineering from Wonkwang University, Iksan, South Korea (2003–2009). Following his tenure as a Senior Researcher at the Korea Institute of Civil Engineering and Building Technology (KICT), he joined the Korean Register (KR) in 2010. He currently serves as the Executive Director (Head) of the AI Convergence Center at KR. His research interests include autonomous ships, AI-CAE, and AI strategy.

Jaecheul PARK received his Ph.D. in Ship Engine System Engineering from Mokpo National Maritime University in 2012. A former ROK Navy Combat Officer with 14 years of R&D experience at the Korean Register (KR), he is currently a Part Leader at KR's AI Convergence Center. He leads the development of Condition-Based Maintenance (CBM) technology, focusing on AI-driven fault diagnosis to establish global "Digital Classification" services.