American Journal of Business and Operations Research

♦ ISSN: 2692-2967 Vol: 15, No. 01, 2027 ♦ PP. 50–61

♦ Article DOI: https://doi.org/10.54216/AJBOR.150107



FULL LENGTH ARTICLE

Robust Estimation of Component Cost-Risk Indices in a CODLAG Naval Propulsion Digital Shadow

Shimaa Sami Selim[1,*] **A. S. Tolba**[2] **Samir Aboul Fotouh Saleh**[2] **Mohamed Elhoseny**[3]

[1] Obour University for Science & Technology (OUST), Egypt

[2] Mansoura University, Mansoura, Egypt

[3] College of Computing and Informatics, University of Sharjah, United Arab Emirates

Emails: Shaimaa.selim@oust.edu.eg · astolba@mans.edu.eg · Prof_samir@hotmail.com · Melhoseny@ieee.org

Received: August 12, 2026 Revised: September 14, 2026 Accepted: October 22, 2026 ⋆ Corresponding author

ABSTRACT

Maintenance-economic studies often convert condition indicators into apparent costs without verifying whether the resulting targets are independent of the model inputs. This can produce exceptionally high but operationally misleading accuracy. We formulate component-level estimation for a COmbined Diesel-eLectric And Gas (CODLAG) propulsion plant as a leakage-resistant digital-shadow problem. The primary dataset is the public UCI naval propulsion dataset: 11,934 steady-state observations generated by a calibrated numerical frigate simulator across nine speed settings, 51 compressor-health states, and 26 turbine-health states. Because it contains neither monetary maintenance records nor live vessel telemetry, the outputs are explicitly defined as dimensionless compressor and turbine cost-risk indices rather than currency costs. Health coefficients and every target-derived quantity are excluded from predictors. Models are tuned by three-fold grouped inner validation and assessed by five-fold outer validation in which all observations sharing a degradation pair remain in the same fold. Extra Trees achieved an aggregate cost-risk RMSE of 0.0222 (95% CI half-width 0.0049) and R^2 of 0.9886 (95% CI half-width 0.0049) under nominal interpolation. However, combined noise, missingness, calibration bias, and channel loss reduced R^2 to 0.6335 even after corruption-augmented training, while leave-one-speed-regime-out testing produced mean $R^2 = -0.379$. Scenario-matched 90% split-conformal intervals retained 89.5–90.4% empirical coverage by widening from 0.032 nominally to 0.406 under severe corruption. An external NASA C-MAPSS FD001 benchmark yielded RMSE 16.97 cycles and $R^2 = 0.821$ for remaining-useful-life prediction on 100 unseen engines. The study therefore supports robust condition-risk ranking within the simulated envelope, not direct monetary forecasting or autonomous maintenance scheduling. The principal contribution is a transparent validation protocol that exposes target leakage, simulator determinism, sensor-quality sensitivity, and the boundary between a digital shadow and a deployable digital twin.

Keywords: Naval propulsion ▪ Digital shadow ▪ Condition monitoring ▪ Cost-risk index ▪ Target leakage ▪ Nested grouped validation ▪ Conformal prediction ▪ Sensor degradation

1. INTRODUCTION

Condition-based maintenance (CBM) seeks to align inspection and intervention with observed asset condition rather than a fixed calendar. In maritime propulsion, the potential value is substantial because propulsion failures can combine repair expense, loss of availability, safety exposure, and mission disruption [1, 2, 3]. Yet prediction and maintenance are different

© 2027 The Author(s). Published by ASPG. This article is licensed under CC BY 4.0.

tasks. A prediction model estimates an unobserved condition or future outcome; a maintenance policy additionally selects an action, time, and resource plan under economic and reliability constraints. Recent offshore-wind studies make this distinction explicit by coupling condition assessment with opportunistic scheduling, environmental accessibility, subsystem dependence, and revenue or cost objectives [4, 5, 6].

Digital-twin terminology introduces a second distinction. A digital model has no automated data link; a digital shadow receives one-way updates from a physical or simulated source; and a full digital twin has an operational synchronization and feedback mechanism [7, 8, 9, 10]. The public naval propulsion data used in this work were generated offline by a numerical simulator. They can support the design and testing of an analytical digital shadow, but they cannot demonstrate live synchronization, closed-loop updating, or intervention feedback. Calling the experiment a validated real-time twin would therefore overstate the evidence.

A third distinction concerns the target itself. A naïve formulation may treat products such as compressor decay coefficient $\times$ fuel flow as maintenance costs. These products are neither independently observed expenditures nor expressed in currency, and their factors may also be available to the predictors. A model can consequently reconstruct the formula instead of learning a transportable relationship between measured condition and economic burden. This target-construction problem, compounded by a deterministic simulator grid, explains why R^2 values near 0.9999 do not establish operational validity.

The study asks a narrow and defensible question: can measured propulsion signals estimate normalized component severity, and hence a transparent cost-risk ranking, when target leakage is blocked and sensor defects are introduced? The output is a dimensionless index for prioritization; conversion to monetary cost is deferred until work orders, labor rates, parts, downtime, and failure consequences are available. This framing preserves the useful cost-oriented decision perspective while removing claims unsupported by the data.

1.1 Research Questions and Contributions

The study is organized around four research questions: (RQ1) how much of the apparent near-perfect performance is attributable to target construction and validation design; (RQ2) how accurately can independently withheld compressor and turbine severity be inferred from measured signals; (RQ3) how sensitive is that inference to realistic sensor defects; and (RQ4) does the same leakage-control and validation discipline remain useful on an external engine-degradation benchmark?

The principal contributions are:

- A provenance and target audit that corrects the dataset description, removes currency claims, and formalizes a leakage firewall between measured predictors, latent health coefficients, and derived decision indices.
- A two-stage estimator in which measured signals predict component severity and a transparent post-model mapping converts severity to a cost-risk index. This prevents the model from learning a cost formula whose factors are already present in the input.
- A nested, degradation-state-grouped validation protocol with training-only preprocessing, explicit hyperparameter search, fold-level uncertainty, a random-row diagnostic, and an unseen-speed-regime stress test.
- A realistic data-integrity analysis spanning Gaussian sensor noise, missing values, calibration drift, channel dropout, corruption-augmented training, and scenario-matched split-conformal intervals.
- An external NASA C-MAPSS benchmark and a standards-aligned deployment architecture that clearly separates what is experimentally implemented from the additional functions required for a full digital twin.

These contributions moderate rather than amplify the performance claim. Strong interpolation within the simulator grid is reported together with the failure modes that preclude direct onboard deployment.

2. RELATED WORK AND RESEARCH POSITIONING

2.1 Maritime Condition Monitoring and Digital Representations

Machine-learning research on the same CODLAG simulator has primarily estimated compressor and turbine decay coefficients from steady-state signals. Coraddu et al. evaluated supervised approaches for low-intrusive CBM, and Cipollini et al. studied analysis with limited expert feedback [11, 12]. These studies established the dataset as a condition-estimation benchmark, not a repository of maintenance invoices. More recent maritime research has investigated vessel-machinery prediction [2], component-level predictive-maintenance challenges [13], marine-engine learning pipelines [14], and digital-twin-based comparisons for diesel engines [15]. Their common emphasis is condition, fault, or remaining-life estimation; economic outputs usually require a separate decision model or lifecycle-cost layer.

Digital-twin maintenance reviews stress that synchronization, model updating, and decision feedback are defining operational capabilities [7, 8]. Maritime examples include lifecycle services for the research vessel Gunnerus [16] and fault-diagnosis architectures for autonomous ships [17]. ISO 23247 provides a digital-twin framework, ISO 13374 organizes condition-monitoring data processing, and ISO 55000 links asset information to lifecycle decision making [18, 19, 20]. These standards motivate the proposed architecture, but standards alignment alone does not prove that all twin functions were implemented.

2.2 Prediction versus Maintenance Optimization

Maintenance decisions require more than an accurate point prediction. Cai et al. connected data-driven condition information to CBM optimization [21]. Three recent studies illustrate the same separation in offshore wind: Ye et al. developed a failure-warning model using ARIMA-LASSO and operating-condition-specific EWMA thresholds [4]; Luo et al. proposed an integrated condition-based opportunistic framework balancing component condition and operation-and-maintenance cost [5]; and Ying et al. optimized opportunistic maintenance across dependent subsystems using fault chains [6]. These studies are not direct numerical baselines for a CODLAG simulator, but they provide an important conceptual benchmark: a condition estimator becomes a maintenance method only after actions, constraints, dependencies, and economic consequences are modeled.

2.3 Research Gap

The literature does not justify the broad claim that no AI-enabled cost-aware maritime maintenance research exists. The narrower gap addressed here is methodological: public simulator studies rarely quantify how target provenance, grouped validation, realistic sensor corruption, uncertainty intervals, and digital-twin scope jointly affect a cost-oriented condition index. Table 1 positions this contribution without claiming primacy over all maritime predictive-maintenance research.

Note. CRI = cost-risk index; RUL = remaining useful life.

3. MATERIALS AND METHODS

3.1 System Scope: An Implemented Digital Shadow

Figure 1 separates the validated offline pipeline from functions proposed for future deployment. The implemented system ingests simulator records, applies data-quality and leakage checks, estimates component severity, and returns a CRI with an uncertainty interval and quality flag. It is a digital shadow because information flows from a simulated physical model to an analytical representation. Live vessel synchronization, dynamic state-space filtering, intervention feedback, and economic calibration are shown as future functions and are not counted among the experimental contributions.

3.2 Primary and External Datasets

The primary dataset is the UCI Condition Based Maintenance of Naval Propulsion Plants dataset (DOI 10.24432/C5K31K) [22]. Its 11,934 rows were generated by a numerical frigate simulator representing a CODLAG propulsion plant. The simulator blocks—propeller, hull, gas turbine, gearbox, and controller—were tuned using knowledge of similar propulsion plants, but the released observations are synthetic steady states rather than measurements collected aboard Royal Canadian Navy vessels. The design is a complete grid of nine speeds (3–27 knots in 3-knot increments), 51 compressor coefficients $k_{Mc} \in [0.950, 1.000]$, and 26 turbine coefficients $k_{Mt} \in [0.975, 1.000]$. There are no missing values and no timestamps, vessel identifiers, work orders, failures, or monetary costs.

Sixteen variables are released: lever position, ship speed, gas-turbine shaft torque and rate of revolution, gas-generator speed, starboard and port propeller torque, high-pressure turbine exit temperature, compressor inlet and outlet temperatures, turbine exit pressure, compressor inlet and outlet pressures, exhaust gas pressure, turbine injection control, and fuel flow. Lever position is an exact encoding of speed and

Table 1. Critical positioning against representative related work.

Study	Asset/data	Endpoint	Distinctive contribution	Boundary
Coraddu et al. [11]	CODLAG simulator	Compressor/turbine decay	Supervised condition estimation	No cost labels; simulator grid
Cipollini et al. [12]	CODLAG simulator	Condition indicators	Minimal-feedback CBM analysis	No monetary validation
Jimenez et al. [2]	Vessel machinery	Predictive-maintenance outputs	Maritime PdM framework	Different assets and endpoints
Ye et al. [4]	Offshore wind	Fault-warning signal	ARIMA-LASSO indicators and operating-condition-specific EWMA thresholds	Warning rather than cost estimation
Luo et al. [5]	Offshore wind farm	Condition + maintenance plan	Health-based windows, component priorities, and cost-aware grouping	No CODLAG sensor benchmark
Ying et al. [6]	Wind subsystems	Maintenance schedule	Fault-chain dependence and optimization	Different degradation/cost structure
This study	CODLAG + C-MAPSS	Severity, CRI, and RUL	Leakage firewall, grouped nested CV, corruption and conformal uncertainty	No field costs or live twin

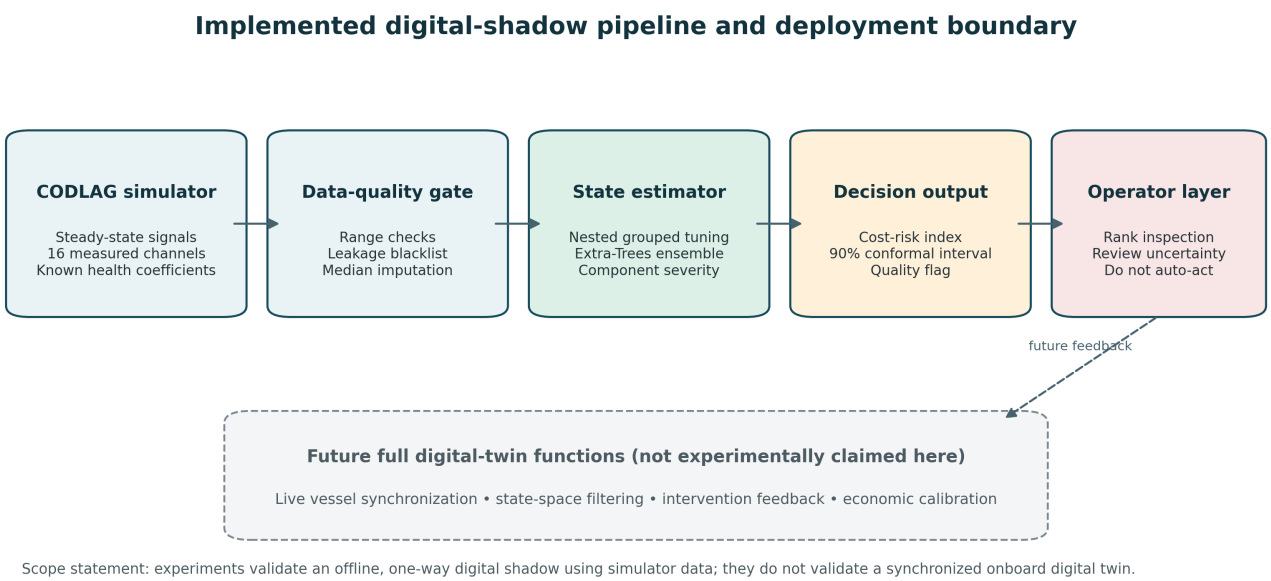


Figure 1. Implemented digital-shadow pipeline and the explicit boundary of a future full digital twin.

Table 2. Dataset provenance and analytical role.

Dataset	Size	Origin	Inputs	Endpoint	Use limitation
UCI naval propulsion	11,934 steady states	Numerical CODLAG frigate simulator	15 measured predictors	$k_{Mc}, k_{Mt} \rightarrow$ CRI	No field telemetry, time, or money
NASA C-MAPSS FD001	20,631 train cycles; 100 test engines	C-MAPSS turbofan simulator	17 nonconstant predictors	Capped RUL	Aircraft domain; ancillary transfer test

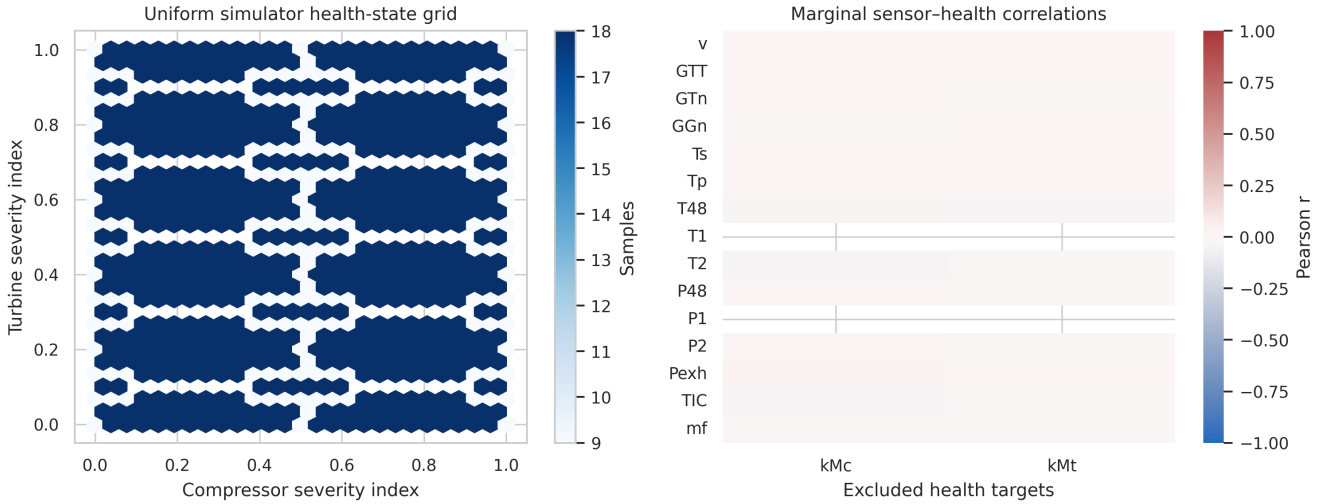


Figure 2. Uniform health-state grid and marginal relationships between measured signals and excluded health targets.

was removed; the remaining 15 measured signals form the predictor vector. Both health coefficients are targets only and never enter the model input.

The external benchmark is NASA C-MAPSS FD001 [23, 24]. It contains 20,631 training-cycle rows from 100 engines and an official test set of 100 unseen engines operating under one condition with one high-pressure-compressor degradation mode. Seventeen nonconstant operating/sensor variables were retained. Remaining useful life was capped at 125 cycles during training, a common stabilization for early healthy cycles, and evaluated at the final observed cycle of each test engine using the official labels. This benchmark does not validate maritime cost; it tests whether the same leakage and hold-out discipline transfers to a distinct propulsion-degradation problem.

Dataset provenance, inputs, endpoints, and interpretation limits are summarized in Table 2; Figure 2 illustrates the CODLAG health-state grid and signal relationships.

3.3 Target Definition: From Unsupported Cost to Cost-Risk

Let k_{Mc} and k_{Mt} denote the simulator's compressor and turbine health coefficients, where 1 indicates the nominal state and lower values indicate stronger degradation. Component severity is normalized to $[0,1]$:

$$z_c = \frac{1-k_{Mc}}{1-0.950}, \qquad z_t = \frac{1-k_{Mt}}{1-0.975}. \tag{1}$$

A transparent aggregate cost-risk index is then calculated after prediction:

$$\begin{aligned} &\mathrm{CRI} = w_c z_c + w_t z_t, \\ &w_c = 0.55, \quad w_t = 0.45, \quad w_c + w_t = 1. \end{aligned} \tag{2}$$

The weights illustrate a slightly higher prioritization of compressor degradation; component results are reported separately so users can substitute validated fleet-specific weights. CRI is dimensionless and monotonic in severity. It is not a price, expected invoice, or lifecycle cost. If fleet records later provide inspection cost c_0, component intervention costs a_j, and downtime cost d, a monetary model could be calibrated as:

$$\mathbb{E}[C \mid x] = c_0 + \sum_j a_j \Pr(z_j \geq \tau_j \mid x) + d\,\mathbb{E}[T_{\mathrm{down}} \mid x]. \tag{3}$$

but none of c_0, a_j, τ_j, d, or downtime outcomes is available in the public dataset. Equation 3 is therefore a future calibration pathway, not an estimated result in this paper.

3.4 Observation and Learning Formulation

For observation i, let $\mathbf{s}_i = [z_{c,i}, z_{t,i}]^{\mathsf{T}}$ denote latent component severity, u_i the operating regime (principally ship speed), and $\mathbf{x}_i \in \mathbb{R}^{15}$ the measured signal vector. The simulator can be represented conceptually as:

$$\mathbf{x}_i = h(\mathbf{s}_i, u_i) + \varepsilon_i, \tag{4}$$

where h is the unknown nonlinear mapping and ε_i represents measurement error. The learning task estimates severity from signals:

$$\hat{\mathbf{s}}_i = f_\theta(\mathbf{x}_i), \qquad \theta^* = \arg\min_\theta \frac{1}{n_{\mathrm{tr}}} \sum_{i \in \mathrm{tr}} \|\mathbf{s}_i - f_\theta(\mathbf{x}_i)\|_2^2. \tag{5}$$

Predicted CRI is computed only after $\hat{\mathbf{s}}_i$ is obtained. This two-stage form makes the target construction auditable and prevents a regressor from receiving the coefficients used to calculate its response.

The model-specific search spaces are summarized in Table 3.

Table 3. Hyperparameter optimization protocol.

Model	Tuned parameters	Search space	Search method
Ridge	α	$\{0.01, 0.1, 1, 10, 100\}$	Exhaustive grid
Random Forest	`max_depth`; `min_samples_leaf`; `max_features`	$\{10, 16, \text{None}\}$; $\{1, 2, 4\}$; $\{0.6, 0.8, 1.0\}$	8 random combinations; 240 trees
Extra Trees	`max_depth`; `min_samples_leaf`; `max_features`	$\{10, 16, \text{None}\}$; $\{1, 2, 4\}$; $\{0.6, 0.8, 1.0\}$	8 random combinations; 240 trees

3.5 Leakage Firewall and Preprocessing

The leakage firewall applies four rules. First, k_{Mc}, k_{Mt}, z_c, z_t, CRI, and every algebraic descendant of these quantities are prohibited from the predictor matrix. Second, lever position is removed because it deterministically duplicates ship speed. Third, imputation, scaling, augmentation, and model tuning are fitted inside each training partition only. Fourth, all nine observations sharing the same (k_{Mc}, k_{Mt}) health pair form one group; a group is never split between training and test. In set notation, if $A(Y)$ denotes the algebraic closure of the response variables, the admissible inputs satisfy:

$$X_{\text{adm}} \cap A(Y) = \varnothing, \qquad G_i = G_j \Longrightarrow \text{fold}(i) = \text{fold}(j). \quad (6)$$

The nominal UCI data contain no missing values. Median imputation is nevertheless included in the pipeline because missingness is introduced during deployment stress tests. Robust scaling is used for Ridge regression; tree ensembles operate on the imputed native scale. No outlier is deleted merely because it is statistically extreme. Values are removed only if they are nonfinite or outside a physically specified range; neither condition occurred in the source file.

3.6 Sensor-Corruption Model and Robust Training

To avoid treating clean simulator data as operationally representative, test signals are transformed by a corruption operator:

$$\begin{aligned} \mathbf{x}' &= \mathbf{M} \odot (\mathbf{x} + \mathbf{b} + \eta) + (\mathbf{1} - \mathbf{M}) \odot \text{NA}, \\ \eta_j &\sim \mathcal{N}(0, \rho^2 \sigma_j^2). \end{aligned} \quad (7)$$

where $\mathbf{M}$ is a Bernoulli availability mask, $\mathbf{b}$ is a calibration offset, ρ controls noise relative to the training standard deviation, and $\odot$ denotes elementwise multiplication. Eight scenarios are evaluated: nominal; 1% and 3% Gaussian noise; 5% and 15% missing completely at random (MCAR); 2% range-based bias applied to T48, T2, and P2; complete dropout of T48 and P2; and a combined severe scenario with 3% noise, 15% MCAR, and 2% bias. These are controlled sensitivity scenarios, not estimates of naval sensor failure rates.

For corruption-augmented training, each clean training row is paired with one independently corrupted copy using 2% noise, 10% MCAR, and 1% bias in the same three channels. Targets are unchanged. Augmentation is performed after the outer split, so no test information is injected.

3.7 Candidate Models and Hyperparameter Optimization

Three models were selected to cover interpretable linear regularization and nonlinear ensembles: Ridge regression [25], Random Forest [26], and Extremely Randomized Trees (Extra Trees) [27]. Every implemented candidate is reported, and no untested method is included in the workflow. Hyperparameters are optimized in a three-fold inner GroupKFold using negative root-mean-square error averaged across the two normalized severity outputs. Ridge uses exhaustive grid search. Each tree ensemble uses eight draws from a prespecified 27-combination grid. The modal winning Extra-Trees configuration across outer folds used 240 trees, `max_features`=0.8, unrestricted depth, `min_samples_leaf`=1, and `random_state`=20260903.

The nested assessment and transfer workflow is shown in Figure 3.

3.8 Validation, Uncertainty, and Metrics

Primary performance is estimated by five outer folds grouped by the 1,326 unique degradation pairs. Hyperparameter selection uses three grouped folds nested inside each outer training set. The reported 95% confidence intervals are t intervals over the five outer-fold scores; they quantify fold-to-fold variability, not all sources of epistemic uncertainty. A random 80/20 row split is retained only as an inflation diagnostic. A leave-one-speed-regime-out analysis trains on eight speeds and tests on the ninth to measure operating-envelope transfer.

For each outer fold, 20% of training groups are reserved for split-conformal calibration [28]. Absolute CRI residuals define the finite-sample quantile:

$$\hat{q}_{0.90} = \text{Quantile}_{\lceil (n_{\text{cal}}+1)0.90 \rceil / n_{\text{cal}}} \left\{ \left| \text{CRI}_i - \widehat{\text{CRI}}_i \right| : i \in \text{cal} \right\}, \quad (8)$$

$$I_{0.90}(x) = \left[\max\left(0, \widehat{\text{CRI}}(x) - \hat{q}_{0.90}\right), \min\left(1, \widehat{\text{CRI}}(x) + \hat{q}_{0.90}\right) \right]. \quad (9)$$

Because corruption changes the residual distribution, calibration and test data receive the same named corruption when scenario-matched coverage is assessed. This is a sensitivity calculation; exchangeability may fail under real temporal drift.

Point metrics are mean absolute error (MAE), root-mean-square error (RMSE), and the coefficient of determination R^2:

$$\begin{aligned} \text{MAE} &= n^{-1} \sum_i |y_i - \hat{y}_i|, \\ \text{RMSE} &= \sqrt{n^{-1} \sum_i (y_i - \hat{y}_i)^2}, \\ R^2 &= 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}. \end{aligned} \quad (10)$$

Algorithm 1: Leakage-resistant grouped estimation. *Inputs:* Measured signals X; health targets S; health-pair groups G; candidate models M.

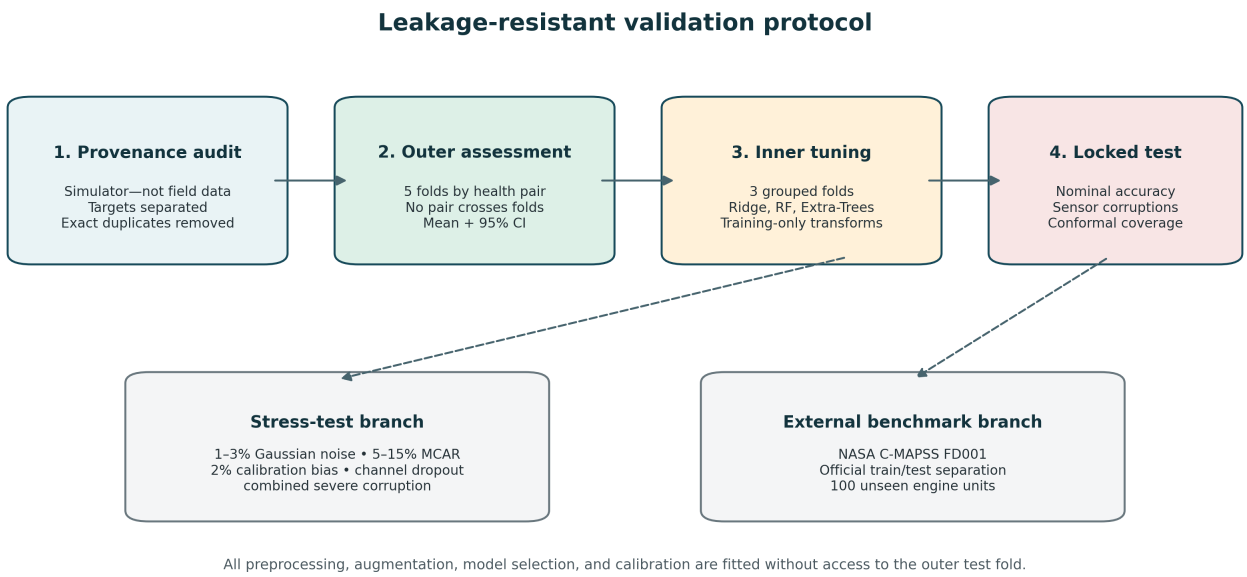


Figure 3. Nested grouped assessment, sensor-corruption branch, and external benchmark protocol.

1. Create severity targets from k_{Mc} and k_{Mt}; construct CRI only for evaluation and decision output.
2. Delete k_{Mc}, k_{Mt}, every target-derived field, and deterministic duplicate lever position from X.
3. For each of five outer grouped folds, reserve one set of health pairs as the untouched test fold.
4. Within the remaining groups, optimize each candidate by three-fold grouped validation; fit imputation and scaling inside each inner training fold.
5. Refit the selected configuration on the outer training data; predict component severity and CRI on the untouched health pairs.
6. Aggregate fold metrics and compute t-based 95% confidence intervals.

Output: Fold-level predictions, metrics, tuned parameters, and leakage audit.

Algorithm 2: Corruption-robust prediction and conformal intervals. *Inputs:* Outer training/calibration/test groups; fitted estimator; corruption scenario q.

1. Generate a corrupted copy of outer training signals using 2% noise, 10% MCAR, and 1% bias; append it to clean training data.
2. Fit the robust estimator without accessing calibration or test targets.
3. Apply the same named deployment scenario q separately to calibration and test signals.
4. Compute calibration absolute residuals and the finite-sample 0.90 conformal quantile.
5. Return clipped CRI point estimates, 90% intervals, and a sensor-quality scenario flag.

Output: Scenario-specific RMSE, R^2, empirical coverage, and interval width.

3.9 Experimental Platform and Reproducibility

Analyses were executed on an x86-64 Linux environment using Python 3.12.13, NumPy 2.3.5, pandas 2.2.3, SciPy 1.17.0, scikit-learn 1.8.0, Matplotlib 3.10.8, and seaborn 0.13.2. Parallel tree fitting used all available logical cores. The UCI data were produced by the original authors' numerical CODLAG simulator; this study did not rerun or modify that proprietary simulation software. The NASA benchmark was generated with NASA's Commercial Modular Aero-Propulsion System Simulation (C-MAPSS). Fixed seed 20260903 controls splitting, hyperparameter sampling, corruption, and ensemble randomness.

4. RESULTS

4.1 Provenance and Leakage Audit

The audit established three constraints. First, the primary observations are simulator outputs, not real-time ship sensor records. Second, products of decay coefficients and fuel flow are nonmonetary algebraic proxies. Third, their factors and close deterministic correlates make near-perfect reconstruction unsurprising. Reproducing the algebraic target while supplying its components yielded mean $R^2 = 0.9999988$ and RMSE = 0.000540. Even after removing the health coefficients but retaining simulator signals, the same product target remained near perfect ($R^2 = 0.9999980$), demonstrating that the deterministic grid and target formula—not merely model quality—drive the result.

Under the corrected formulation, a random row split yielded CRI $R^2 = 0.9946$ and RMSE = 0.0155. When entire degradation pairs were withheld, performance decreased to $R^2 = 0.9886$ and RMSE = 0.0222. The difference is not catastrophic, but it confirms that row-wise splitting is optimistic because the same health pair appears at multiple speeds across training and test.

The contrast between algebraic-proxy reconstruction and leakage-free grouped estimation is illustrated in Figure 4.

4.2 Nested Grouped Model Comparison

Extra Trees achieved the lowest grouped error for the aggregate CRI and the two component severities (Table 4). The 95% intervals are deliberately retained rather than presenting

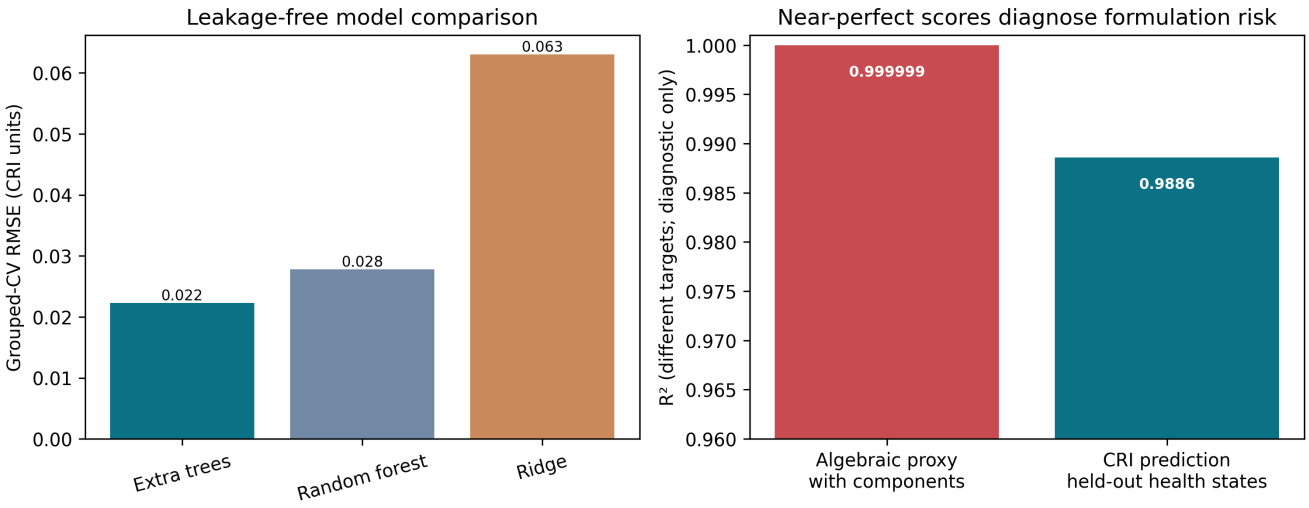


Figure 4. Leakage-free grouped model comparison and diagnostic contrast with algebraic proxy reconstruction.

Table 4. Five-fold nested health-state-grouped performance.

Model	CRI MAE (±95% CI)	CRI RMSE (±95% CI)	CRI R^2 (±95% CI)	Compressor R^2	Turbine R^2
Extra Trees	0.00819 ± 0.00170	0.02225 ± 0.00486	0.98858 ± 0.00487	0.99616	0.98981
Random Forest	0.00973 ± 0.00172	0.02780 ± 0.00535	0.98226 ± 0.00618	0.99414	0.98446
Ridge	0.04936 ± 0.00032	0.06301 ± 0.00073	0.91064 ± 0.00191	0.79163	0.91074

a single favorable split. Ridge performed materially worse, indicating that orthogonalization or linear regularization alone does not capture the nonlinear simulator response. Random Forest was competitive but less accurate than Extra Trees. These results support Extra Trees as the point estimator for the subsequent robustness analysis; they do not imply superiority on unobserved vessels.

Fold-level scores underlying these summaries are provided in Appendix Table A1.

Note. Confidence intervals are t intervals across five outer folds. Lower MAE/RMSE and higher R^2 are better.

4.3 Sensitivity to Sensor-Data Defects

The clean simulator result is not robust to realistic data defects (Table 5 and Figure 5). With clean-only training, 15% MCAR reduced R^2 to 0.702, two-channel dropout to 0.673, and combined severe corruption to 0.286. Corruption-augmented training slightly increased nominal RMSE from 0.0222 to 0.0239 but reduced severe-case RMSE by 28.3%, from 0.1781 to 0.1276. The improvement is meaningful, yet an RMSE of 0.1276 on a $[0, 1]$ index is too large for autonomous decisions. Robust augmentation should therefore be viewed as damage limitation, not a substitute for sensor-quality controls.

Note. Improvement is relative RMSE reduction from clean-only to corruption-augmented training; negative means worse.

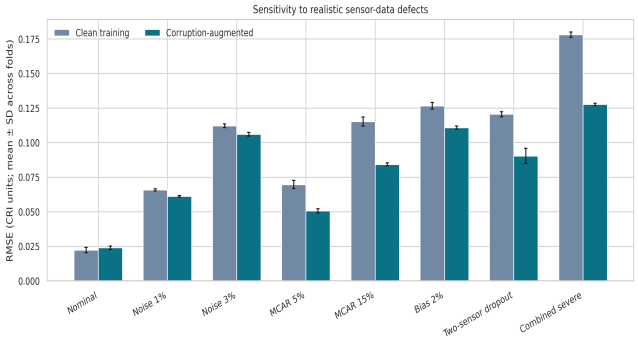


Figure 5. CRI error under realistic sensor-noise, missingness, bias, and dropout scenarios.

4.4 Prediction Uncertainty

Scenario-matched split-conformal intervals attained mean empirical coverage close to the nominal 90% target: 89.8% under nominal data, 90.3% under 3% noise, 89.5% under 15% MCAR, and 90.4% under combined severe corruption. Coverage alone is not evidence of usefulness. Mean interval width increased from 0.032 nominally to 0.338 under 3% noise, 0.263 under 15% MCAR, and 0.406 under the severe combination. The widening is an appropriate warning that a point estimate becomes less actionable as data quality deteriorates.

Table 5. Sensor-corruption sensitivity of the selected Extra-Trees estimator.

Test condition	Clean-train RMSE	Clean-train R^2	Augmented RMSE	Augmented R^2	RMSE improvement
Nominal	0.0222	0.9888	0.0239	0.9871	−7.7%
Noise 1%	0.0657	0.9029	0.0612	0.9158	+6.9%
Noise 3%	0.1121	0.7172	0.1061	0.7468	+5.4%
MCAR 5%	0.0697	0.8906	0.0507	0.9422	+27.3%
MCAR 15%	0.1151	0.7016	0.0842	0.8403	+26.8%
Bias 2%	0.1265	0.6396	0.1108	0.7239	+12.5%
Two-sensor dropout	0.1205	0.6731	0.0904	0.8156	+25.0%
Combined severe	0.1781	0.2865	0.1276	0.6335	+28.3%

Table 6. Scenario-matched 90% split-conformal interval performance.

Scenario	Empirical coverage (mean ± SD)	Mean interval width
Nominal	0.898 ± 0.010	0.032
Noise 3%	0.903 ± 0.008	0.338
MCAR 15%	0.895 ± 0.018	0.263
Combined severe	0.904 ± 0.008	0.406

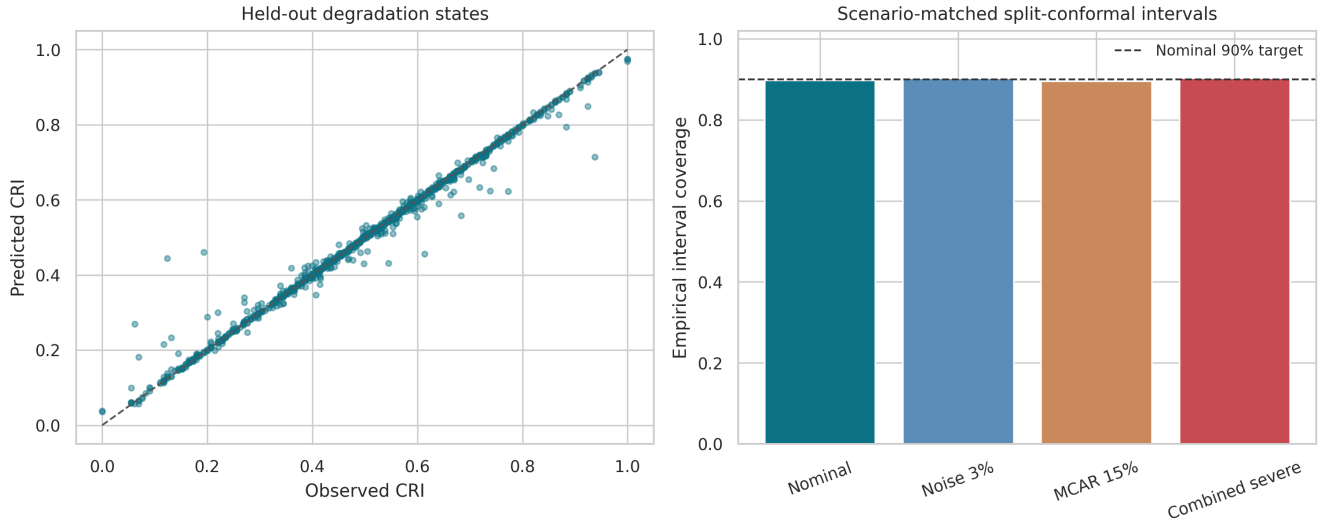


Figure 6. Held-out degradation-state predictions and scenario-matched conformal coverage.

Coverage and interval-width results are reported in Table 6, with observed-versus-predicted agreement shown in Figure 6.

4.5 Unseen Operating-Regime Transfer

The leave-one-speed-regime-out test is the strongest warning against overgeneralization. Across the nine held-out speeds, mean MAE was 0.200, mean RMSE was 0.242, and mean $R^2 = -0.379$. At 27 knots, RMSE reached 0.381 and $R^2 = -2.268$. The estimator therefore interpolates well among withheld health states when familiar speed regimes remain in training, but it does not extrapolate reliably to a completely unseen speed. A deployed system must either cover the intended operating envelope during calibration or use an explicit physics-informed adaptation mechanism.

Speed-specific errors are detailed in Table 7.

Table 7. Leave-one-speed-regime-out generalization.

Held-out speed (knots)	CRI MAE	CRI RMSE	CRI R^2
3	0.166	0.205	0.055
6	0.166	0.201	0.092
9	0.215	0.265	−0.578
12	0.183	0.223	−0.120
15	0.179	0.218	−0.070
18	0.180	0.220	−0.085
21	0.193	0.236	−0.250
24	0.188	0.230	−0.191
27	0.327	0.381	−2.268

4.6 External C-MAPSS Benchmark

On 100 official C-MAPSS FD001 test engines, Extra Trees achieved MAE 12.01 cycles, RMSE 16.97 cycles, and $R^2 = 0.821$; Random Forest achieved RMSE 17.28 and $R^2 = 0.814$; Ridge achieved RMSE 20.60 and $R^2 = 0.736$. The external result is intentionally less spectacular than the CODLAG interpolation score and is more consistent with a difficult fleet holdout. It supports the portability of the validation workflow, not the portability of a maritime cost function.

Model-level external benchmark metrics are summarized in Table 8.

Table 8. External validation on NASA C-MAPSS FD001 (100 unseen engines).

Model	MAE (cycles)	RMSE (cycles)	R^2
Extra Trees	12.01	16.97	0.821
Random Forest	12.24	17.28	0.814
Ridge	16.37	20.60	0.736

5. DISCUSSION

5.1 What High Nominal Accuracy Does—and Does Not Show

The nominal R^2 remains high because the primary benchmark is a dense, deterministic simulator grid. The grouped protocol prevents identical health pairs from crossing folds, but adjacent grid states remain highly regular and the sensors encode them strongly. The result demonstrates interpolation across unseen health pairs within represented operating regimes. It does not demonstrate performance under ship-to-ship variability, sea state, sensor aging, maintenance history, controller

changes, or adversarial conditions.

The leakage audit also shows why accuracy should not be the paper's principal contribution. Algebraic proxies can be reconstructed almost exactly even when they have no economic semantics. The stronger contribution is the method for detecting and preventing that error: an explicit target lineage, predictor blacklist, grouped split, stress tests, and uncertainty-aware output. A value near $R^2 = 0.9999$ is therefore treated as a diagnostic warning rather than a headline result.

5.2 Relation to Maintenance Decision Making

CRI can rank observations for inspection or analyst review, but it does not choose a maintenance action. The offshore-wind literature demonstrates the missing decision layer: condition-based maintenance windows, cost and resource constraints, component dependencies, and fault chains alter the optimal schedule [5, 6]. A naval implementation would likewise require mission availability, redundancy, dock capacity, spare-parts lead times, labor, safety constraints, and the consequence of delayed intervention. Equation 3 sketches how field costs could enter, but optimizing those decisions is beyond the present evidence.

A practical operator interface should therefore display component severity, aggregate CRI, conformal interval, sensor-quality indicators, and an out-of-distribution warning. Automatic work-order generation should be disabled when intervals are wide, channels are missing, or the operating regime lies outside training support. Human review is not merely a deployment preference; it follows directly from the severe-corruption and unseen-speed results.

5.3 Digital-Shadow Fidelity and the Path to a Full Twin

The implemented pipeline satisfies only the analytical portion of a digital-twin architecture. A full twin would require identity-resolved physical assets, timestamped bidirectional data exchange, an online state estimator, model-version governance, synchronized configuration, and feedback from inspections and interventions. Fidelity should be evaluated with state-estimation residuals, synchronization latency, drift statistics, and decision outcomes. None can be measured in the present static file. Figure 1 therefore labels the implemented artifact as a digital shadow and places closed-loop functions outside the validated boundary.

ISO 13374 can structure acquisition, manipulation, state detection, health assessment, prognosis, and advisory functions [19]. ISO 23247 can guide entity models and service interfaces [18], while ISO 55000 can govern how condition evidence enters asset-management decisions [20]. These standards improve architectural coherence, but real compliance and effectiveness require deployment-specific verification.

5.4 Robustness, Uncertainty, and Model Governance

Corruption augmentation trades a small nominal penalty for improved degraded-input performance. The benefit is largest for missing values and channel dropout because the model has seen masked patterns during training. Calibration bias remains difficult because systematic shift is not removed by median imputation; reference-sensor checks or online recalibration are needed. Similarly, conformal intervals preserve coverage only by becoming wide. In operations, interval width should be part of the abstention rule, not hidden behind a nominal coverage percentage.

Model governance should record the training envelope, dataset lineage, target definition, hyperparameters, software versions, sensor mapping, and calibration date. Monitoring should track missingness, univariate range violations, multivariate distance, prediction-interval width, and post-maintenance residuals. A model should be retrained only after data-quality review and should be revalidated by asset or voyage groups rather than random rows.

6. LIMITATIONS AND FUTURE RESEARCH

This study has six material limitations. First, the primary dataset is numerical and deterministic; it contains no sea-state variability, measurement chronology, fleet heterogeneity, or maintenance events. Second, CRI is dimensionless and cannot be interpreted as money, savings, or return on investment. Third, the grid consists of steady states rather than degradation trajectories, so temporal forecasting, remaining-life dynamics, and causal intervention effects cannot be inferred. Fourth, the external C-MAPSS benchmark concerns aircraft turbofans and RUL, not naval maintenance cost; it validates workflow transfer only. Fifth, scenario-matched conformal coverage assumes exchangeability within each corruption scenario and may fail under abrupt or sequential distribution shift. Sixth, the study implements no live synchronization, state-space filter, feedback control, or maintenance optimizer.

Future work should proceed in stages. A prospective dataset should join timestamped onboard signals to asset identity, operating context, alarms, inspections, replaced parts, labor hours, logistics, downtime, and verified expenditure. Cost categories must be defined before modeling and recorded independently of predictors. Validation should hold out whole vessels and later time blocks. A hierarchical or physics-informed state estimator can then separate asset-specific baselines from shared degradation. Monetary calibration should be evaluated with decision-curve analysis and cost-sensitive loss, followed by a constrained policy that considers mission readiness and safety. Finally, an onboard pilot should measure synchronization latency, drift, abstention frequency, false-alarm burden, lead time, and realized maintenance outcomes.

The unseen-speed failure also suggests a targeted experimental-design priority: collect calibration observations at the edges and transitions of the intended operating envelope. Domain adaptation should not be accepted solely because it restores a high aggregate R^2; it must improve vessel- and regime-held-out performance without degrading uncertainty calibration.

7. CONCLUSION

This study frames the task as robust, uncertainty-aware estimation of component cost-risk indices from simulated condition signals rather than near-perfect monetary maintenance-cost forecasting. The public CODLAG data are described as numerical-simulator outputs. Health targets and their descendants are excluded from predictors, models are tuned and tested by degradation-state groups, and performance is reported under sensor defects and an external engine benchmark. Extra Trees provides strong nominal interpolation (CRI

$R^2 = 0.9886$), but performance deteriorates under severe corruption ($R^2 = 0.6335$ after robust augmentation) and fails under unseen speed regimes (mean $R^2 = -0.379$). Conformal intervals appropriately widen as data quality worsens.

Accordingly, the framework is suitable as a research-grade digital shadow and a basis for condition-risk ranking within a represented simulator envelope. It is not yet a validated real-time digital twin, a monetary cost estimator, or an autonomous maintenance policy. This narrower conclusion is scientifically stronger because it aligns the claim with the available evidence and makes the remaining path to field validation explicit.

DATA AVAILABILITY

The primary dataset is publicly available from the UCI Machine Learning Repository as "Condition Based Maintenance of Naval Propulsion Plants," DOI: 10.24432/C5K31K. The external benchmark is publicly available from the NASA Prognostics Center of Excellence repository as the "Turbofan Engine Degradation Simulation Data Set" (C-MAPSS). The study uses only these public simulator datasets; no confidential naval operational records were accessed.

REFERENCES

[1] R. Ahmad and S. Kamaruddin, "An overview of time-based and condition-based maintenance in industrial applications," *Computers & Industrial Engineering*, vol. 63, no. 1, pp. 135–149, 2012.

[2] V. J. Jimenez, N. Bouhmala, and A. H. Gausdal, "Developing a predictive maintenance model for vessel machinery," *Journal of Ocean Engineering and Science*, vol. 5, no. 4, pp. 358–386, 2020.

[3] D. Kimera and F. N. Nangolo, "Maintenance optimization for marine mechanical systems," *Proceedings of the Institution of Mechanical Engineers, Part M: Journal of Engineering for the Maritime Environment*, vol. 234, no. 2, pp. 446–462, 2020.

[4] H. Ye, W. Zhu, H. Li, W. Ji, C. G. Soares, and J. Wang, "Failure warning for offshore wind turbines based on Autoregressive models," *Ocean Engineering*, vol. 332, p. 121448, 2025.

[5] J. Luo, X. Luo, X. Ma, Y. Zan, and X. Bai, "An integrated condition-based opportunistic maintenance framework for offshore wind farms," *Reliability Engineering & System Safety*, vol. 256, p. 110701, 2025.

[6] F. Ying, L. Huang, M. Song, and Y. Fu, "Optimization of an opportunistic maintenance strategy for wind turbines based on subsystem fault chains," *Ocean Engineering*, vol. 339, p. 122106, 2025.

[7] I. Errandonea, S. Beltrán, and S. Arrizabalaga, "Digital twin for maintenance: A literature review," *Computers in Industry*, vol. 123, p. 103316, 2020.

[8] F. Tao, B. Xiao, Q. Qi, J. Cheng, and P. Ji, "Digital twin modeling," *Journal of Manufacturing Systems*, vol. 64, pp. 372–389, 2022.

[9] B. Schleich, N. Anwer, L. Mathieu, and S. Wartzack, "Shaping the digital twin for design and production engineering," *CIRP Annals*, vol. 66, no. 1, pp. 141–144, 2017.

[10] B. R. Barricelli, E. Casiraghi, and D. Fogli, "A survey on digital twin: Definitions, characteristics, applications, and design implications," *IEEE Access*, vol. 7, pp. 167 653–167 671, 2019.

[11] A. Coraddu, L. Oneto, A. Ghio, S. Savio, D. Anguita, and M. Figari, "Machine learning approaches for improving condition-based maintenance of naval propulsion plants," *Proceedings of the Institution of Mechanical Engineers, Part M*, vol. 230, no. 1, pp. 136–153, 2016.

[12] F. Cipollini, L. Oneto, A. Coraddu, A. J. Murphy, and D. Anguita, "Condition-based maintenance of naval propulsion systems: Data analysis with minimal feedback," *Reliability Engineering & System Safety*, vol. 177, pp. 12–23, 2018.

[13] A. S. Kalafatelis, N. Nomikos, A. Karditsa, and P. Trakadas, "Towards predictive maintenance in the maritime industry: A component-based overview," *Journal of Marine Science and Engineering*, vol. 13, p. 425, 2025.

[14] F. Maione, P. Lino, G. Maione, and G. Giannino, "A machine learning framework for condition-based maintenance of marine diesel engines: A case study," *Algorithms*, vol. 17, no. 9, p. 411, 2024.

[15] M. Acanfora, L. Mocerino, G. Giannino, and U. Campora, "A comparison between digital-twin-based methodologies for predictive maintenance of marine diesel engine," *Proceedings of SPEEDAM*, pp. 1254–1259, 2024.

[16] H. Zhang, G. Li, L. I. Hatledal, Y. Chu, A. Ellefsen, and P. Han, "A digital twin of the research vessel gunnerus for lifecycle services: Outlining key technologies," *IEEE Robotics & Automation Magazine*, vol. 30, no. 3, pp. 6–19, 2023.

[17] A. Hasan, T. Asfihani, O. Osen, and R. T. Bye, "Leveraging digital twins for fault diagnosis in autonomous ships," *Ocean Engineering*, vol. 292, p. 116546, 2024.

[18] International Organization for Standardization, "ISO 23247-1:2021, Automation systems and integration—Digital twin framework for manufacturing—Part 1: Overview and general principles," International Organization for Standardization, Tech. Rep., 2021.

[19] International Organization for Standardization, "ISO 13374-1:2003, Condition monitoring and diagnostics of machines—Data processing, communication and presentation—Part 1: General guidelines," International Organization for Standardization, Tech. Rep., 2003.

[20] International Organization for Standardization, "ISO 55000:2014, Asset management—Overview, principles and terminology," International Organization for Standardization, Tech. Rep., 2014.

[21] Y. Cai, R. H. Teunter, and B. de Jonge, "A data-driven approach for condition-based maintenance optimization," *European Journal of Operational Research*, vol. 311, no. 3, pp. 1036–1049, 2023.

[22] A. Coraddu, L. Oneto, A. Ghio, S. Savio, D. Anguita, and M. Figari, "Condition Based Maintenance of Naval Propulsion Plants," UCI Machine Learning Repository, 2014.

[23] A. Saxena and K. Goebel, "Turbofan Engine Degradation Simulation Data Set," NASA Ames Prognostics Data Repository, Moffett Field, CA, 2008.

[24] A. Saxena, K. Goebel, D. Simon, and N. Eklund, "Damage propagation modeling for aircraft engine run-to-failure simulation," in *Proceedings of the 1st International Conference on Prognostics and Health Management*, 2008.

[25] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[26] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[27] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 63, pp. 3–42, 2006.

[28] J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman, "Distribution-free predictive inference for regression," *Journal of the American Statistical Association*, vol. 113, no. 523, pp. 1094–1111, 2018.

[29] Y. Lei, N. Li, L. Guo, N. Li, T. Yan, and J. Lin, "Machinery health prognostics: A systematic review from data acquisition to rul prediction," *Mechanical Systems and Signal Processing*, vol. 104, pp. 799–834, 2018.

[30] A. A. Murtaza, A. Saher, M. H. Zafar, S. K. R. Moosavi, M. F. Aftab, and F. Sanfilippo, "Paradigm shift for predictive maintenance and condition monitoring from industry 4.0 to industry 5.0," *Results in Engineering*, vol. 22, p. 102935, 2024.

APPENDICES

A. FOLD-LEVEL PRIMARY RESULTS

Table A1. Complete fold-level results used to compute Table 4.

Outer fold	Model	CRI MAE	CRI RMSE	CRI R^2
1	Extra Trees	0.01034	0.02755	0.98288
2	Extra Trees	0.00854	0.02234	0.98873
3	Extra Trees	0.00779	0.02339	0.98774
4	Extra Trees	0.00668	0.01669	0.99376
5	Extra Trees	0.00759	0.02126	0.98979
1	Random Forest	0.01196	0.02955	0.98032
2	Random Forest	0.00997	0.02939	0.98049
3	Random Forest	0.00908	0.02827	0.98210
4	Random Forest	0.00828	0.02036	0.99072
5	Random Forest	0.00938	0.03142	0.97770
1	Ridge	0.04948	0.06325	0.90981
2	Ridge	0.04923	0.06296	0.91050
3	Ridge	0.04921	0.06262	0.91215
4	Ridge	0.04976	0.06387	0.90861
5	Ridge	0.04912	0.06237	0.91216

B. CORRUPTION DEFINITIONS AND REPRODUCIBILITY CHECKS

The corruption scenarios and augmentation settings are specified in Appendix Table B1.

Table B1. Exact sensor-corruption scenarios.

Scenario	Operation	Parameters/channels
Noise 1% / 3%	Independent $\mathcal{N}(0, \rho^2\sigma_j^2)$ noise	$\rho = 0.01$ or 0.03
MCAR 5% / 15%	Independent feature-cell masking	$p = 0.05$ or 0.15
Bias 2%	Add 0.02 of training range	T48, T2, P2
Two-sensor dropout	Mask full channels	T48 and P2
Combined severe	3% noise + 15% MCAR + 2% bias	All signals; bias on T48, T2, P2
Training augmentation	2% noise + 10% MCAR + 1% bias	One corrupted copy per training row

Reproducibility checks included: fixed random seed; no duplicate health pair across outer folds; preprocessing fit only on training partitions; no target-derived predictor; fold-level metrics retained; all mentioned models present in the reported results; and source-data provenance checked against the UCI and NASA repository descriptions.