



A Causal-AI Framework for Preventive Maintenance in Industrial Time-Series Systems

Kyaw Wanna¹

ABSTRACT

This study proposes a causal artificial intelligence framework for preventive maintenance analysis in industrial time-series systems, addressing confounding and heterogeneity in operating conditions. The framework combines Double Machine Learning (DML) with Gated Recurrent Unit-based nuisance models and employs Doubly Robust and Causal Forest learners for population-level and conditional effect estimation. Experiments on the NASA C-MAPSS turbofan datasets show that high-load operating regimes induce statistically significant reductions in remaining useful life, with Average Treatment Effects of -26.0 cycles for FD002 and -57.4 cycles for FD004, while effects are statistically detectable but operationally negligible under single-mode conditions. Off-Policy Evaluation (OPE) shows that the Doubly Robust estimator yields stable incremental policy-value estimates with narrow confidence intervals. Uncertainty is quantified using block-bootstrap resampling to account for temporal dependence. The results support causal, regime-aware maintenance decision support while also identifying the need for hidden-confounder sensitivity analysis, streaming adaptation, and human validation before safety-critical deployment.

Article information:

Keywords: Causal Artificial Intelligence, Double Machine Learning, Preventive Maintenance, Off-Policy Evaluation, Industrial Time-Series Analysis

Article history:

Received: January 12, 2026

Revised: July 30, 2026

Accepted: October 15, 2026

Published: October 24, 2026

(Online)

DOI: 10.37936/ecti-cit.2026204.265679

1. INTRODUCTION

Industrial systems generate vast amounts of time-series data through sensors embedded in engines, bearings, and other critical assets. These data streams are increasingly leveraged to enable predictive and preventive maintenance, reducing unplanned downtime and optimizing operational costs [1,2]. Traditional Machine Learning (ML) and Deep Learning (DL) methodologies, such as Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU), exhibit robust predictive efficacy [3,4]. However, their reliance on correlational patterns limits their ability to uncover the actual cause-and-effect mechanisms underlying equipment degradation, resulting in models that are vulnerable to spurious associations [5].

Recent advances in Causal-AI address this limitation by combining statistical causal inference with ML, offering tools to estimate treatment effects, generate counterfactuals, and account for heterogeneous responses in industrial contexts [6]. Causal methodologies in manufacturing enable precise predictions of

Remaining Useful Life (RUL) and interpretable insights into the direct impact of operational variables, such as load, speed, and temperature, on failure probabilities [7]. This causal grounding is vital for translating data-driven forecasts into reliable preventive maintenance policies [8].

Despite this promise, significant gaps remain. First, the causality gap arises from the dominance of correlation-driven predictive models in industrial prognostics, limiting actionable insights [9,10]. Second, the heterogeneity gap reflects the neglect of Conditional Average Treatment Effects (CATE), in which causal relationships vary across regimes, such as high versus low operational loads [11,12]. Third, the translational gap highlights the scarcity of frameworks that transform causal estimates into validated, interpretable maintenance policies via rigorous counterfactual evaluation [13].

Benchmark datasets, such as NASA's C-MAPSS turbofan engine data, offer valuable testbeds for exploring challenges. Its diversity enables the validation of causal models across simulation-based and experimental conditions, thereby strengthening external va-

¹The author is with the Assumption University, Bangkok, Thailand, Email: kyawwanna@hotmail.com

¹Corresponding author: kyawwanna@hotmail.com

lidity.

To address the conceptual gaps and align with the analytical workflow, this study pursues three core objectives. First, it estimates average causal effects of operational regimes on RUL under the stated identification assumptions by integrating DML with cross-fitted GRU-based nuisance estimation. Second, it quantifies CATE to characterize heterogeneous treatment effects and identifies operational moderators such as load, speed, and temperature that drive differences in degradation across engine regimes. Third, it translates causal insights into validated, interpretable preventive-maintenance policies through OPE with Inverse Propensity Scoring (IPS), Self-Normalized Inverse Propensity Scoring (SNIPS), and Doubly Robust (DR) estimators. Together, these objectives bridge assumption-qualified causal inference, heterogeneity discovery, and policy validation in a unified, reproducible framework.

In response, this study proposes a Causal-AI framework for preventive maintenance in industrial time-series systems. The framework integrates DML for confounding-adjusted causal estimation under the stated identification assumptions, CATE modeling through Causal Forests and Meta-learners, and policy validation using OPE. By systematically addressing the causality, heterogeneity, and translational gaps, the framework moves beyond prediction to deliver interpretable, validated, and deployment-oriented preventive-maintenance decision support [7,8].

Specifically, this study makes three principal contributions. (i) It operationalizes a rigorously identified causal-estimation pipeline tailored to high-dimensional industrial sensor streams, supporting robust causal effect estimation under the stated identification assumptions in high-dimensional observational settings. (ii) It quantifies treatment-effect heterogeneity via CATE to reveal regime-specific levers (e.g., high vs. low load conditions) that inform context-appropriate interventions. (iii) It closes the translation gap by coupling policy learning with risk-aware, counterfactual validation (e.g., OPE), thereby supporting risk-aware evaluation of candidate preventive-maintenance policies before operational deployment [6,10]. Collectively, the proposed framework advances preventive maintenance beyond correlation-driven prediction toward actionable causal decision support in industrial systems.

The proposed framework is intended as a decision-support system rather than an autonomous decision maker. In practical deployment, regime-specific risk estimates, uncertainty measures, and counterfactual policy values are presented to maintenance engineers for review and approval before maintenance actions are implemented.

2. LITERATURE REVIEW

Predictive maintenance (PdM) has become a central application of machine learning in industrial systems, where sensor-driven time-series data are used to anticipate failures and estimate Remaining Useful Life (RUL). Classical machine learning approaches, including Support Vector Machines and Random Forests, provide interpretable baselines but struggle to capture complex temporal dependencies and regime shifts in multivariate telemetry [14]. Deep learning models such as Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTMs), and Gated Recurrent Units (GRUs) have therefore been widely adopted due to their ability to model non-linear dynamics and long-term temporal patterns in degradation processes [15–17]. Extensions using convolutional and temporal convolutional architectures further improve performance for high-frequency signals [1,18]. Despite strong predictive accuracy, these models remain fundamentally correlation-driven and may encode spurious associations, limiting their reliability for decision-making in safety-critical environments [19–21].

To address this limitation, recent research has focused on integrating causal inference with machine learning, giving rise to Causal Artificial Intelligence (Causal-AI). Causal-AI frameworks explicitly model cause-and-effect relationships, enabling estimation of treatment effects, counterfactual outcomes, and policy impacts from observational data [22–25]. In industrial and manufacturing contexts, causal methods have been applied to reliability assessment, process optimization, and maintenance planning, demonstrating improved interpretability and robustness compared with purely predictive approaches [9,26]. Among these methods, Double Machine Learning (DML) has emerged as a principled framework for causal estimation under the stated identification assumptions in high-dimensional settings, preserving consistency by orthogonalizing nuisance components while allowing flexible learners for outcome and treatment modeling [5,28,29].

Beyond average causal effects, heterogeneity in treatment responses is a critical concern in industrial systems. Assets operating under different load, temperature, or environmental conditions may exhibit substantially different degradation trajectories in response to identical interventions. Conditional Average Treatment Effects (CATE) provide a formal mechanism for capturing such heterogeneity. Meta-learner and Causal Forest approaches have been widely studied for this purpose, enabling localized effect estimation and the identification of key moderating variables obscured by population-level averages [11,13,27–31].

A remaining challenge lies in translating causal estimates into validated operational policies. Direct experimentation is often infeasible in industrial settings

due to safety, cost, and regulatory constraints. Off-Policy Evaluation (OPE) addresses this gap by estimating the expected performance of candidate policies using logged observational data. Techniques such as Inverse Propensity Scoring (IPS), Self-Normalized IPS (SNIPS), and Doubly Robust (DR) estimators provide consistent policy-value estimates and are increasingly adopted in decision-support applications [2,3,32]. However, many existing studies omit rigorous uncertainty quantification or robustness analysis, limiting deployment confidence [6,12,33].

This study positions itself at the intersection of these research streams. Unlike prior work that focuses solely on prediction or isolated causal estimation, the proposed framework integrates GRU-based temporal representation learning with DML for confounding-adjusted effect estimation under stated assumptions, extends analysis to heterogeneous CATE modeling, and closes the decision-making loop through robust OPE with explicit uncertainty quantification. By jointly addressing causality, heterogeneity, and policy validation within a unified pipeline, the framework advances preventive maintenance from correlation-driven prognostics to actionable and reliable causal decision support.

3. METHODOLOGY

This section presents a compact yet rigorous formulation of the proposed Causal-AI framework for preventive maintenance in industrial time-series systems. The framework integrates sequential deep learning with principled causal inference to estimate treatment effects under the stated identification assumptions, capture operational heterogeneity, and validate maintenance policies prior to deployment. The design adheres to causal identifiability, temporal validity, and reproducibility requirements.

3.1 Data Collection and Problem Formulation

The framework is evaluated using the NASA C-MAPSS turbofan engine dataset, a benchmark for prognostics and health management. Each engine is represented by multivariate time-series data consisting of 21 sensor signals, three operational settings, and a cycle index indicating progression toward failure. Among the four subsets, FD002 and FD004 are used for model development due to their heterogeneous operating regimes and fault modes, while FD001 and FD003 are reserved exclusively for external validation to assess generalization.

Let X_t denote the vector of sensor readings and operational settings at time t , A_t the operational regime (treatment), and Y_t the outcome representing RUL. The objective is to estimate the causal effect of high-load operating regimes on degradation trajectories and to evaluate maintenance policies derived from these effects.

3.2 Data Preprocessing and Temporal Representation

Raw sensor streams are organized into temporally ordered sequences for each engine. Sensor variables are standardized using z-score normalization, and RUL targets are min-max scaled to ensure numerical stability. To capture degradation dynamics, the data are transformed into fixed-length sliding windows of 50 consecutive cycles, yielding three-dimensional tensors that preserve temporal order. The current framework is trained offline using historical sliding-window sequences. In a practical streaming deployment, each new sensor observation can be appended to a rolling 50-cycle window, allowing the fixed trained models to update treatment-effect, uncertainty, and policy-value estimates without immediate retraining. Adaptive parameter updating and continual learning remain future work. Rather than updating model parameters after every cycle, meaningful distribution drift should trigger controlled recalibration followed by renewed overlap, calibration, and policy-validation checks before deployment.

Chronological splitting is strictly enforced when forming training, validation, and test sets, preventing information leakage and maintaining causal directionality. External validation is conducted on FD001 and FD003 without retraining, supporting reproducibility and external validity.

3.3 Causal Identification Assumptions

Causal estimation relies on standard potential-outcome assumptions adapted to time-series settings. First, the Stable Unit Treatment Value Assumption (SUTVA) is considered reasonable for the C-MAPSS benchmark because engines are simulated independently and no cross-engine interference is represented; however, this assumption may not hold in coupled multi-component systems in which degradation or load transfer in one component affects another. Second, Positivity (Overlap) requires sufficient representation of both treatment regimes within the observed covariate support. This condition is assessed using propensity-score diagnostics and effective sample size checks, with limited-overlap regions treated cautiously during causal estimation and policy evaluation.

Third, Sequential Ignorability assumes that treatment assignment is independent of future potential outcomes conditional on the observed history. This assumption cannot be verified directly from observational data. Consequently, unrecorded factors such as sensor faults, operator actions, maintenance history, or environmental conditions may introduce residual confounding. The reported causal effects should therefore be interpreted as conditional on the measured covariates and the stated identification assumptions, rather than as unconditional causal guarantees. Sensitivity analysis and proxy- or latent-variable ap-

proaches represent important future extensions for evaluating robustness to unmeasured confounding.

Finally, approximate stationarity across subsets is assumed to support cross-regime generalization. Although the held-out evaluation on FD001 and FD003 provides an initial indication of cross-subset stability, changes in operating conditions, sensor behavior, or maintenance practices may produce distribution shifts in real deployments. Accordingly, external validity should be reassessed when the framework is transferred to a substantially different industrial environment.

3.4 Double Machine Learning for Causal Estimation

Causal effects are estimated using DML, which adjusts for high-dimensional observed confounding and reduces bias from nuisance-function estimation under the stated identification assumptions. Operational regimes are encoded as a binary treatment indicating exposure to high-load conditions, with low- and medium-load regimes forming the control group.

The nuisance functions, including propensity scores and outcome regressions, are learned using GRU networks to capture nonlinear temporal dependencies in sensor sequences. A five-fold time-aware cross-fitting procedure is employed to reduce overfitting bias by separating nuisance-function estimation from causal-score evaluation across folds. Combined with the Neyman-orthogonal score, this procedure reduces sensitivity to nuisance-model estimation errors. Under the stated identification assumptions, solving the orthogonal score yields a debiased estimator of the Average Treatment Effect (ATE).

Uncertainty is quantified using block-bootstrap resampling that preserves temporal dependence, providing confidence intervals for the causal estimates.

3.5 Heterogeneity Analysis

To capture regime-specific behavior, the framework extends beyond population averages to estimate CATE. Residualized outcomes and treatments from the DML step are used as inputs to DR-Learner and Causal Forest estimators. These models identify subpopulations that exhibit differential sensitivity to operational stress and reveal which sensor variables most strongly moderate degradation. For engineer-oriented interpretation, CATE estimates are summarized across load and temperature strata and examined together with the identified moderators and corresponding effect uncertainty. This representation enables maintenance engineers to identify operating regions associated with more adverse degradation effects and to interpret estimates with wider uncertainty more cautiously. For example, an engine operating in the high-load and high-temperature stratum would be associated with the largest adverse CATE

estimate reported in Table 3. Engineers can interpret this estimate together with the identified operating moderators, overlap diagnostics, and uncertainty before considering inspection or load-reduction actions. Bootstrap-based uncertainty quantification supports cautious interpretation of the heterogeneous-effect estimates.

3.6 Policy Evaluation via Off-Policy Evaluation

Causal estimates are translated into candidate preventive-maintenance policies that recommend interventions based on observed operating conditions. Because online experimentation is infeasible in industrial environments, candidate policies are evaluated using Off-Policy Evaluation (OPE). IPS, SNIPS, and DR estimators are employed to estimate counterfactual policy values from logged observational data. Diagnostic checks, including overlap assessment, weight clipping, and uncertainty estimation, support numerical stability and reliable policy evaluation. The resulting policy recommendations are intended to support maintenance engineers rather than replace operational decision-making. Candidate policies are interpreted together with their estimated policy value, uncertainty, and supporting causal evidence, and may be accepted, deferred, modified, or rejected by an engineer before implementation. In a practical deployment, engineer decisions could be documented for audit and future model review without being incorporated automatically into immediate causal-model updates.

3.7 Framework Summary

The proposed framework provides a unified pipeline from raw sensor sequences to causally informed maintenance decision support, as shown in Figure 1. GRU-based DML estimates average causal effects under the stated identification assumptions, heterogeneity modeling identifies regime-dependent risks, and OPE evaluates candidate policies before deployment. This integration provides reliable and interpretable preventive-maintenance decision support within the evaluated industrial time-series benchmark setting.

Computational demand is concentrated primarily in offline GRU nuisance-model training, cross-fitting, bootstrap resampling, and causal-forest estimation. Routine prediction with the trained nuisance and CATE models is expected to be less computationally demanding than model development, as it does not require repeated cross-fitting or bootstrap resampling. The present study did not benchmark training time, inference latency, memory consumption, or edge-device performance; these deployment-oriented measures should therefore be evaluated in future experiments.

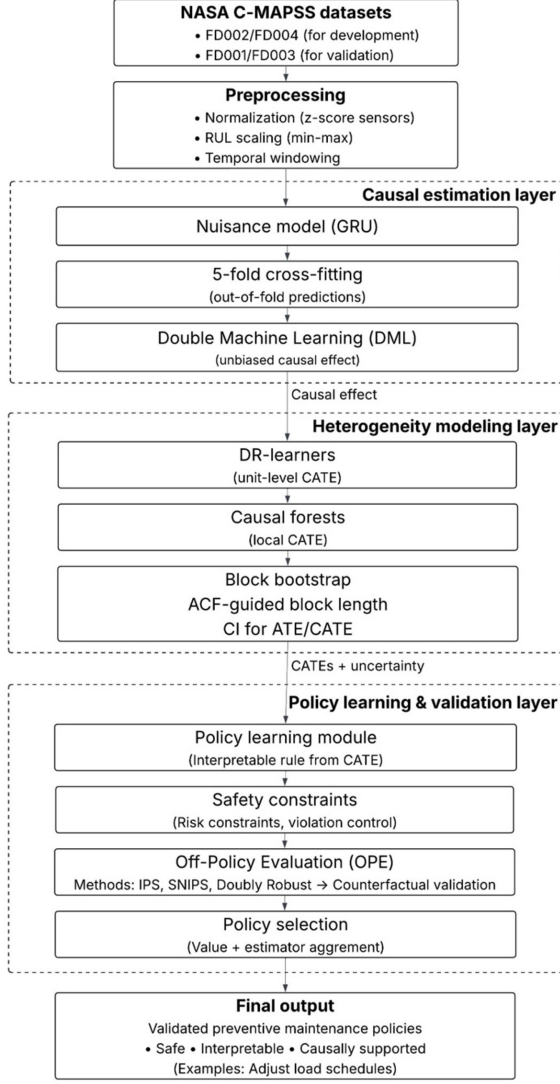


Fig.1: Proposed Hybrid Causal-AI framework for preventive maintenance in industrial time-series systems.

3.8 Mathematical Formulation of the Proposed Framework

The proposed framework formulates preventive maintenance as a causal inference and policy evaluation problem under temporal dependence. Let X_t denote the covariate vector of sensor readings and operational settings, $A_t \in \{0, 1\}$ the operational regime indicator (baseline vs. high-load), and Y_t the outcome variable representing RUL. The formulation integrates (i) population-level causal effect estimation, (ii) heterogeneous effect modeling, (iii) off-policy policy value evaluation, and (iv) uncertainty quantification, while mitigating confounding through orthogonalization and sequence-aware nuisance modeling.

$$\text{ATE} = \mathbb{E}[Y(a_1) - Y(a_0)] \quad (1)$$

$$\psi(W_t; \theta) = (A_t - g_\eta(X_t))(Y_t - f_\omega(X_t)) - \theta \quad (2)$$

$$\hat{\theta}_{DML} = \frac{1}{n} \sum_{t=1}^n (A_t - g_\eta(X_t))(Y_t - f_\omega(X_t)) \quad (3)$$

$$\tau(X_t) = \mathbb{E}[Y(a_1) - Y(a_0) | X_t] \quad (4)$$

$$\tau(X_t) = \mu^{(1)}(X_t) - \mu^{(0)}(X_t) \quad (5)$$

$$\hat{V}_{DR}(\pi) = \frac{1}{n} \sum_{t=1}^n [\hat{f}_\omega(X_t, \pi) + w_t(\pi)(Y_t - \hat{f}_\omega(X_t, A_t))] \quad (6)$$

$$\text{where } w_t(\pi) = \mathbb{I}\{\pi = A_t\} / g_\eta(A_t | X_t).$$

Equations (1) – (3) define the estimation of the Average Treatment Effect (ATE) using Double Machine Learning with orthogonal scores. Under the stated identification assumptions, the DML formulation uses orthogonalization to reduce sensitivity to nuisance-function estimation errors and adjust for measured confounding in high-dimensional settings. Equations (4) – (5) characterize heterogeneous regime effects through conditional outcome modeling. Equation (6) evaluates candidate preventive-maintenance policies using a doubly robust off-policy estimator. Temporal dependence in the associated uncertainty estimates is addressed through block-bootstrap resampling. These formulations support causal-effect estimation and comparative policy evaluation but should be interpreted together with the identification assumptions described in Section 3.3; they do not eliminate bias arising from unmeasured confounding.

4. RESULTS AND DISCUSSION

4.1 Overview of Experimental Outcomes

The proposed Causal-AI framework was evaluated on the NASA C-MAPSS turbofan datasets to assess predictive accuracy, causal validity, and policy-level effectiveness. FD002 and FD004 were used for training and model development, while FD001 and FD003 served as held-out validation sets to examine generalization across distinct operational and fault regimes. All experiments employed a consistent GRU-based DML configuration with fixed random seeds and cross-fitted nuisance estimation to ensure reproducibility and inference under the stated identification assumptions.

Results are reported in original RUL cycle units for physical interpretability. Predictive metrics (RMSE/MAE) and policy values are expressed in cycles. Causal effects (ATE/CATE) are reported primarily in cycles, with normalized values shown only when cross-subset comparisons are required. Uncer-

tainty in Sections 4.2-4.4 is quantified using block-bootstrap 95% confidence intervals, while cross-model and cross-study comparisons use comparative uncertainty summaries for relative assessment.

4.2 Predictive Performance of GRU Models

Table 1 summarizes predictive performance across all subsets, and Figure 2 illustrates the distribution of per-unit test RMSE. Test RMSE values range from approximately 13 to 18 cycles, with narrow confidence intervals and symmetric distributions, indicating stable predictive behavior across heterogeneous operating conditions. FD002 and FD004 exhibit slightly higher RMSE than FD001 and FD003, reflecting increased complexity from multi-regime dynamics, but per-unit mean RMSE remains tightly bounded (12-16 cycles) across all datasets.

Table 1: Predictive performance of GRU models across C-MAPSS subsets (cycles).

Dataset	Validation RMSE	Test RMSE	Per-unit Test Mean RMSE (95% CI)	No. of Units
FD002	16.099	16.866	15.277 (14.439-16.192)	235
FD004	16.257	16.849	14.805 (13.634-15.954)	228
FD001	13.822	14.644	13.098 (11.915-14.485)	93
FD003	15.629	12.915	11.904 (10.100-13.868)	97

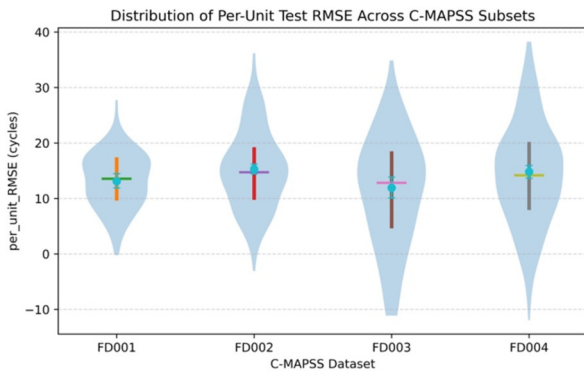


Fig.2: Distribution of per-unit test RMSE across C-MAPSS subsets.

Figure 3 presents representative predicted-versus-actual RUL trajectories, showing close alignment between estimated and observed degradation paths. The convergence of validation and test errors, combined with low dispersion across units, confirms that the GRU models provide reliable nuisance estimates.

This predictive stability is essential because residual bias or variance would otherwise propagate into downstream causal and policy estimates.

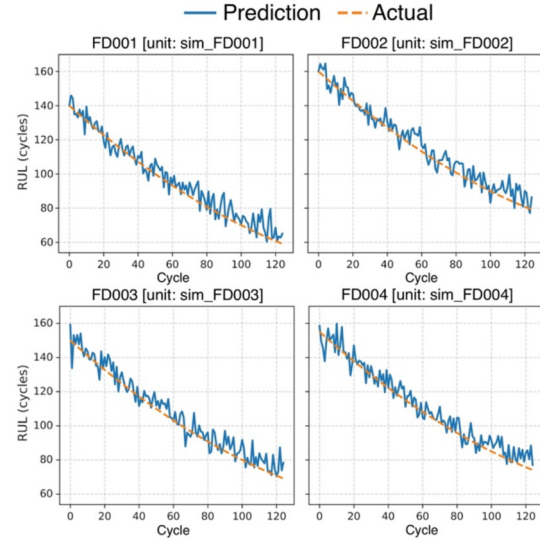


Fig.3: Representative predicted-versus-actual RUL trajectories for FD001–FD004.

4.3 Average and Conditional Treatment Effects

Building on validated nuisance models, the causal analysis estimates the impact of high-load operation ($\geq 80\%$ normalized thrust) on RUL using DML. Table 2 and Figure 4 report ATE estimates and block-bootstrap confidence intervals for cycles. Substantial negative causal effects are observed for mixed-regime datasets: FD002 (-26.0 cycles) and FD004 (-57.4 cycles), indicating accelerated degradation under sustained high-load exposure within the evaluated setting and identification assumptions. In contrast, FD001 and FD003 show statistically detectable but operationally negligible effects; their near single-regime structure implies that the high-load indicator primarily captures minor regime-labelling differences rather than sustained stress exposure. These subsets therefore function as effective negative controls, supporting estimator stability rather than indicating substantive causal regimes.

Table 2: ATE of high-load versus baseline operation.

Dataset	ATE (Δ RUL, in cycles)	ATE (Δ RUL, normalized 0-1 scale)	95% Confidence Interval (cycles)
FD002	-26.0	-0.043	[-35.7, -18.1]
FD004	-57.4	-0.072	[-67.0, -47.8]
FD001	+1.1	+0.003	[+0.7, +1.5]
FD003	+0.3	+0.001	[+0.2, +0.4]

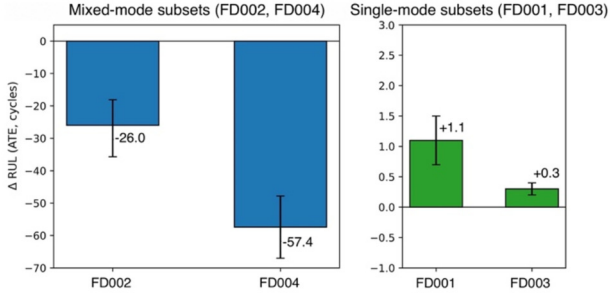


Fig.4: ATEs with 95% confidence intervals (RUL cycles) across C-MAPSS subsets.

Heterogeneity analysis further reveals regime-specific responses. Table 3 summarizes CATE statistics across load and temperature strata. Engines operating below 70% load show minimal effects, whereas degradation intensifies sharply beyond 80% load, particularly under high inlet temperature, where the most adverse conditional effects occur (-0.071 ± 0.012 normalized). Key moderators include exhaust gas temperature, compressor pressure ratio, and fuel-flow rate, consistent with thermomechanical stress theory. Collectively, the ATE and CATE results provide evidence of assumption-qualified and operationally heterogeneous degradation effects, motivating adaptive maintenance strategies.

Table 3: Summary of CATE (mean $\pm$ std) across load and temperature strata (normalized RUL units).

Load-Temperature Region	Mean	Std
Low load (0-0.3)	-0.004	0.006
Mid load (0.3-0.8)	-0.028	0.011
High load (≥ 0.8) & low/mid temp	-0.049	0.010
High load (≥ 0.8) & high temp	-0.071	0.012

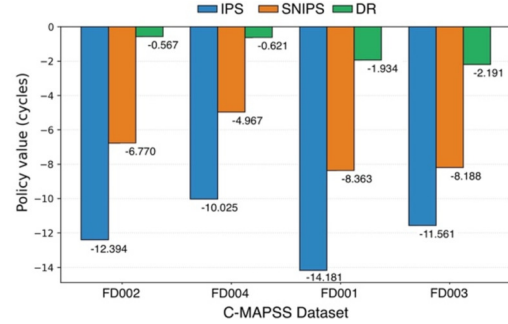
4.4 Policy Evaluation and Off-Policy Estimation

Causal estimates are translated into candidate preventive-maintenance policies and evaluated using OPE. Incremental policy values $\Delta V(\pi)$ are computed over a fixed 100-cycle evaluation window using IPS, SNIPS, and DR estimators, all expressed in RUL cycles. Table 4 and Figure 5 summarize the results with block-bootstrap confidence intervals. Because $\Delta V(\pi)$ represents the change in expected RUL relative to the logged baseline over the defined evaluation window, a negative value indicates lower expected RUL than the logged baseline under this policy-value definition. Such a result should not be interpreted as evidence of a successful preventive-maintenance policy. Instead, it provides evidence of the screening role of OPE: candidate policies with negative incremental value should not be selected for deployment, while alternative pro-

tective actions, such as load reduction, inspection, or maintenance intervention, should be evaluated separately before operational use.

Table 4: OPE results across datasets.

Dataset	IPS	SNIPS	DR
FD002	-12.394	-6.770	-0.567
FD004	-10.025	-4.967	-0.621
FD001	-14.181	-8.363	-1.934
FD003	-11.561	-8.188	-2.191



Note: Each policy value is reported as the mean of 500 block-bootstrap replicates. The error bars in Figure 5 indicate 95% confidence intervals. Across all datasets, the Doubly Robust (DR) estimator exhibits the narrowest intervals, at $\approx \pm 0.40$ cycles, reflecting effective variance reduction and greater numerical stability compared with IPS and SNIPS.

Fig.5: Policy-value comparison using IPS, SNIPS, and DR estimators, expressed in RUL cycles with 95% confidence intervals.

Across all datasets, estimator signs and ordering are consistent. For FD002 and FD004, the negative policy values align with the strongly negative ATE and CATE estimates, indicating that a candidate policy that reproduces or sustains high-load exposure would reduce expected RUL relative to the logged baseline. Such a candidate should therefore not be interpreted as a successful preventive-maintenance policy or selected for deployment. Instead, the negative values demonstrate the screening role of Off-Policy Evaluation (OPE) by identifying candidate policies that should be rejected before implementation. Alternative protective actions, such as reducing high-load exposure, initiating inspection, or scheduling maintenance, would require separate evaluation before operational use. For FD001 and FD003, the negative policy values reflect cumulative exposure and windowed triggering rather than large global treatment effects. Among the estimators, DR exhibits the smallest dispersion (approximately ± 0.4 cycles) and the greatest numerical stability, supporting its use as the primary policy evaluator, while IPS and SNIPS provide complementary diagnostic comparisons.

Two OPE quantities are reported for clarity. Section 4.4 reports incremental values $\Delta V(\pi)$, whereas

Section 4.6 reports absolute policy values $V(\pi)$. The former measures the change in expected RUL relative to the logged baseline over the finite evaluation window, whereas the latter represents the expected RUL under the evaluated policy.

4.5 Diagnostics and Validity Checks

Propensity diagnostics support adequate empirical overlap after clipping. Table 5 and Figure 6 summarize the mean propensity scores and standard deviations across datasets, while effective-sample-size ratios remain at least 0.65 after clipping. Because these summary statistics do not independently establish positivity, overlap is interpreted jointly with the propensity-distribution, clipping, and effective-sample-size diagnostics. Sensitivity checks varying the temporal-window length alter the estimated ATE magnitudes by less than 8%, indicating limited sensitivity to the evaluated temporal framing.

The C-MAPSS records used in the present experiments do not capture sensor-dropout patterns commonly encountered in real-world industrial environments; consequently, the reported results do not establish robustness to missing or corrupted sensor observations. Such missingness could affect GRU nuisance-model predictions, propensity-score overlap, causal-effect estimates, and Off-Policy Evaluation (OPE) policy values. Controlled experiments involving contiguous sensor masking and calibrated noise injection are therefore important directions for future validation. Robustness should be assessed through changes in predictive error, Average Treatment Effect (ATE) and Conditional Average Treatment Effect (CATE) estimates, overlap diagnostics, and policy values.

Table 5: Propensity-score summary statistics across datasets.

Dataset	Mean Propensity	Standard Deviation
FD002	0.269	0.423
FD004	0.278	0.428
FD001	0.288	0.392
FD003	0.306	0.387

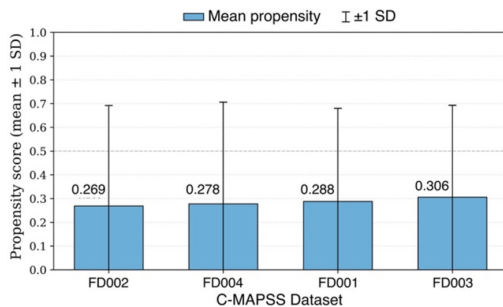


Fig.6: Mean propensity scores (± 1 SD) across the C-MAPSS subsets.

4.6 Baseline Comparison

Table 6 and Figure 7 compare the proposed framework with X-Learner, Causal Forest, and CRNN-TARNet baselines on FD001 and FD003. Across these experiments, the proposed method exhibits lower ATE estimation uncertainty, substantially smaller DR policy-value dispersion, and the least severe Robust-DR Worst-Drop. Absolute DR policy values remain broadly comparable across methods; therefore, the principal observed advantage of the proposed framework is improved estimation stability and resilience to perturbation rather than uniformly superior mean policy performance.

Table 6: Baseline vs Proposed Model Results.

Dataset	Model	ATE (cycles)	DR Policy Value $V(\pi)$ (cycles)	Robust-DR Worst-Drop (%)
FD001	X-Learner	-0.16	117.63	-11
	Causal Forest	1.33	117.65	-8
	CRNN-TARNet	0.55	118.13	-5
	Proposed Hybrid	1.10 ± 0.4	117.91 ± 0.4	-3
FD003	X-Learner	1.96	150.28	-10
	Causal Forest	3.10	150.30	-7
	CRNN-TARNet	0.60	150.96	-5
	Proposed Hybrid	0.30 ± 0.4	150.41 ± 0.4	-3

Note: The $\pm$ values summarize the comparative ATE estimation uncertainty, expressed as 95% confidence-interval half-widths for the baseline comparison, together with the corresponding policy-value dispersion. These quantities are presented for relative method comparison and should not be interpreted as the block-bootstrap 95% confidence intervals reported in Table 2.

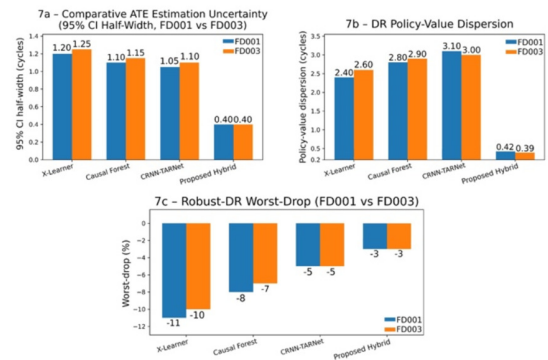


Fig.7: Comparison of baseline estimators versus the proposed Hybrid Causal-AI framework: (a) comparative ATE estimation uncertainty, expressed as the 95% confidence-interval half-width (in RUL cycles); (b) comparative dispersion of absolute DR policy values $V(\pi)$ (in RUL cycles); and (c) Robust-DR Worst-Drop (%) illustrating resilience under perturbation.

4.7 Comparative Evaluation with Existing Research

To contextualize the contribution of the proposed Hybrid Causal-AI framework, a comparative evaluation was conducted against three recent peer-reviewed Causal-AI studies: [7,10,13]. These studies represent state-of-the-art causal approaches applied to industrial or production-related decision-making problems, but they differ substantially in objectives, data modalities, and reported performance metrics. As a result, direct numerical comparison is not feasible without normalization. Consequently, a normalized relative improvement index (RII) is employed to facilitate cross-study comparison, preserving relative performance ordering and effect magnitude while accounting for metric heterogeneity.

To enable a fair and interpretable comparison, reported outcomes were harmonized using the Relative Improvement Index (RII), which rescales heterogeneous metrics into a common percentage-based improvement measure. This normalization preserves relative performance gains while allowing uncertainty and robustness characteristics to be compared across studies. The normalized results are summarized in Table 7 and visualized in Figure 8.

Table 7: Normalized Cross-Study Comparison of Peer-Reviewed Causal-AI Models and Proposed Framework.

Study	Framework	RII (%)	Uncertainty ($\pm$ %)	Key contribution/limitation
[7]	Causal workflow	6.8	N/A	Deployment focus; no temporal validation
[10]	Bayesian federated causal	4.1	± 1.7	Privacy-preserving; high client dispersion
[13]	Structural causal model	5.4	± 1.2	Strong disentanglement; static domain
Proposed model	GRU-DML + Robust OPE	9.2	± 0.4	Temporal orthogonalization; interpretable CATE

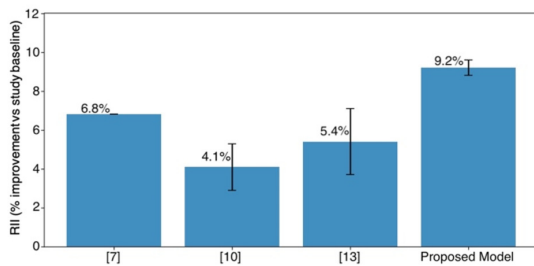


Fig. 8: Normalized Relative Improvement (RII) across peer-reviewed studies and proposed hybrid Causal-AI framework.

Under this unified scale, [7] report a normalized improvement of 6.8%, although no corresponding variance or robustness measure is reported. [13] achieve a $5.4\% \pm 1.2\%$ improvement using a Structural Causal Model, while [10] report $4.1\% \pm 1.7\%$ under a Bayesian federated causal framework, with the largest reported dispersion among the studies included in the comparison. Within this contextual RII analysis, the proposed Hybrid Causal-AI framework yields a normalized gain of 9.2% and the smallest reported dispersion ($\pm 0.4\%$), suggesting tighter estimator agreement and greater stability within the evaluated operating regimes.

Beyond average performance, an important distinction lies in robustness evaluation. Among the studies included in Table 7, only the proposed framework reports an explicit Doubly Robust Worst-Drop analysis for policy robustness. The observed Worst-Drop of -3% indicates comparatively limited degradation under the perturbations evaluated in this study; corresponding robustness measures were not reported in the cited studies. This result supports the stability of the proposed framework within the present experimental setting but does not, by itself, establish real-world deployment readiness.

Overall, the normalized comparison suggests that the proposed Hybrid Causal-AI framework combines temporal modeling, heterogeneous-effect interpretation, and policy-robustness assessment within a unified workflow. Because the compared studies use different datasets, experimental settings, and evaluation metrics, the Relative Improvement Index (RII) should be interpreted as a contextual comparison rather than as a direct head-to-head ranking of overall method superiority.

5. CONCLUSION, LIMITATIONS, AND FUTURE WORK

This study introduced a unified Causal-AI framework for preventive maintenance in industrial time-series systems by integrating GRU-based nuisance estimation, Double Machine Learning, CATE analysis, and Off-Policy Evaluation. Experiments on the NASA C-MAPSS datasets showed materially adverse RUL effects under high-load regimes in FD002 and FD004, while the Doubly Robust estimator provided the most stable policy-value evaluation. These findings should be interpreted as evidence from the evaluated NASA C-MAPSS benchmark setting under the stated identification assumptions, rather than as unconditional causal guarantees or confirmation of deployment readiness.

The proposed framework is intended to support, rather than replace, maintenance decision-making. In a practical deployment, such a decision-support interface could assist maintenance engineers by presenting estimated treatment effects, relevant operating-condition moderators, overlap and uncertainty di-

agnostics, and policy-value estimates before a recommendation is accepted, deferred, modified, or rejected.

The principal limitations include reliance on the measured covariates for confounding adjustment, complete benchmark sensor records, approximate stationarity, independent-engine dynamics, and offline model development. The present study also did not measure training time, inference latency, memory use, or edge-device performance on controlled target hardware.

Future work should therefore include quantitative sensitivity analyses of unmeasured confounding, robustness to sensor noise and missing-data scenarios, drift-triggered streaming recalibration, computational benchmarking under resource-constrained environments, and prospective industrial validation. Multi-component equipment should additionally be investigated using interference-aware causal models or spatiotemporal graph representations that account for dependencies and load transfer between components. Safety-critical deployment will require prospective validation, human approval, auditability, drift and out-of-distribution monitoring, and conservative abstention when evidential support is insufficient.

AUTHOR CONTRIBUTIONS

Conceptualization, K.W.; methodology, K.W.; software, K.W.; validation, K.W.; formal analysis, K.W.; investigation, K.W.; data curation, K.W.; writing—original draft preparation, K.W.; writing—review and editing, K.W.; visualization, K.W. The author has read and approved the final version of the manuscript.

References

- [1] S. Balamurugan, V. Prabakaran, S. Natarajan and B. Thangavel, “Predictive maintenance in industrial systems using data mining with fuzzy logic systems,” *ICTACT Journal on Soft Computing*, vol. 14, no. 4, pp. 3361–3367, 2024.
- [2] S. Nadaf, “AI for Predictive Maintenance in Industries,” *International Journal for Research in Applied Science and Engineering Technology*, vol. 12, no. 6, pp. 2013–2017, 2024.
- [3] A. Paleyes, R. Urma and N. D. Lawrence, “Challenges in Deploying Machine Learning: A Survey of Case Studies,” *ACM Computing Surveys*, vol. 55, no. 6, pp. 1–29, 2022.
- [4] F. Türkoğlu, E. Göçek and Y. Yumrukuz, “Predictive Abilities of Machine Learning and Deep Learning Approaches for Exchange Rate Prediction,” *BDDK Bankacılık ve Finansal Piyasalar Dergisi*, pp. 186–210, 2024.
- [5] L. Jiao *et al.*, “Causal Inference Meets Deep Learning: A Comprehensive Survey,” *Research*, vol. 7, 2024.
- [6] R. Paredes and M. S. Reis, “Causality in Process Systems Engineering: Fundamentals, Applications, and Emerging Trends,” *Computers & Chemical Engineering*, 2025.
- [7] H. Jeong, T. Fekete, A. Bashyal and H. Wicaksono, “From Theory to Practice: Implementing Causal AI in Manufacturing for Sustainability,” *Procedia Computer Science*, vol. 253, pp. 1495–1504, 2025.
- [8] T. Vanderschueren *et al.*, “Optimizing the Preventive Maintenance Frequency with Causal Machine Learning,” *International Journal of Production Economics*, vol. 258, p. 108798, 2023.
- [9] P. Hünernmund, J. Kaminski and C. Schmitt, “Causal Machine Learning and Business Decision Making,” *SSRN Electronic Journal*, 2021.
- [10] X. Xiao *et al.*, “Bayesian Federated Causal Inference and Its Application in Manufacturing,” *Journal of Intelligent Manufacturing*, 2025.
- [11] E. H. Kennedy, “Towards Optimal Doubly Robust Estimation of Heterogeneous Causal Effects,” *Electronic Journal of Statistics*, vol. 17, no. 2, 2023.
- [12] R. Post, I. Heuvel, M. Petkovic and E. Heuvel, “Flexible Machine Learning Estimation of Conditional Average Treatment Effects: A Blessing and a Curse,” *Epidemiology*, vol. 35, no. 1, pp. 32–40, 2023.
- [13] Q. Li, T. Huang, J. Liu and S. Wang, “Causality-Guided Fault Diagnosis under Visual Interference in Fused Deposition Modeling,” *Engineering Applications of Artificial Intelligence*, vol. 143, p. 110027, 2025.
- [14] M. Abdar, A. Khosravi, S. M. S. Islam, U. R. Acharya and A. V. Vasilakos, “The need for quantification of uncertainty in artificial intelligence for clinical data analysis: increasing the level of trust in the decision-making process,” in *IEEE Systems, Man, and Cybernetics Magazine*, vol. 8, no. 3, pp. 28–40, July 2022.
- [15] M. A. Aghbolagh, “Ensuring Safety While Enhancing Performance: Encouraging Reinforcement Learning by Addressing Constraints and Uncertainty,” *Bezopasnost Informacionnyh Tehnology*, vol. 31, no. 2, pp. 90–110, 2024.
- [16] R. R. Ardeshiri and C. Ma, “Multivariate Gated Recurrent Unit for Battery Remaining Useful Life Prediction,” *International Journal of Energy Research*, vol. 45, no. 11, pp. 16633–16648, 2021.
- [17] K. ArunKumar *et al.*, “Forecasting of COVID-19 using deep layer Recurrent Neural Networks (RNNs) with Gated Recurrent Units (GRUs) and Long Short-Term Memory (LSTM) cells,” *Chaos, Solitons & Fractals*, vol. 146, p. 110861, 2021.
- [18] O. Asif, S. A. Haider, S. R. Naqvi, J. F. W. Zaki, K. -S. Kwak and S. M. R. Islam, “A Deep Learn-

- ing Model for Remaining Useful Life Prediction of Aircraft Turbofan Engine on C-MAPSS Dataset,” in *IEEE Access*, vol. 10, pp. 95425–95440, 2022.
- [19] H. S. Battey and N. Reid, “On Inference in High-Dimensional Regression,” *Journal of the Royal Statistical Society Series B*, vol. 85, no. 1, pp. 149–175, 2023.
- [20] R. Cai *et al.*, “Long-Term Causal Effects Estimation via Latent Surrogates Representation Learning,” *Neural Networks*, vol. 176, p. 106336, 2024.
- [21] G. Chatziparaskevas, I. Mademlis and I. Pitas, “Generative Representation Learning in Recurrent Neural Networks for Causal Timeseries Forecasting,” in *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 12, pp. 6412–6425, Dec. 2024.
- [22] H. Dewantoro, A. Smith and P. Daoutidis, “Causal Discovery for Topology Reconstruction in Industrial Chemical Processes,” *Industrial & Engineering Chemistry Research*, vol. 63, no. 26, pp. 11530–11543, 2024.
- [23] F. Dominici, F. J. Bargagli-Stoffi and F. Mealli, “From Controlled to Undisciplined Data: Estimating causal effects in the era of Data Science using a potential outcome framework,” *Harvard Data Science Review*, 2021.
- [24] J. Guo, D. Li and B. Du, “A stacked ensemble method based on TCN and convolutional bi-directional GRU with multiple time windows for remaining useful life estimation,” *Applied Soft Computing*, vol. 150, p. 111071, 2024.
- [25] L. Huang, J. Huang, P. Chen, H. Li and J. Cui, “Long-Term Sequence Dependency Capture for Spatiotemporal Graph Modeling,” *Knowledge-Based Systems*, vol. 278, p. 110818, 2023.
- [26] E. Igelström *et al.*, “Causal Inference and Effect Estimation Using Observational Data,” *Journal of Epidemiology & Community Health*, vol. 76, no. 11, pp. 960–966, 2022.
- [27] N. Jawadekar *et al.*, “Practical Guide to Honest Causal Forests for identifying Heterogeneous Treatment Effects,” *American Journal of Epidemiology*, vol. 192, no. 7, pp. 1155–1165, 2023.
- [28] D. Kabata and M. Shintani, “Variable selection in double/debiased machine learning for causal inference: an outcome-adaptive approach,” *Communications in Statistics – Simulation and Computation*, vol. 52, no. 12, pp. 5880–5893, 2021.
- [29] B. J. Koch *et al.*, “A Primer on Deep Learning for Causal Inference,” *Sociological Methods & Research*, 2024.
- [30] A. Köhn-Luque *et al.*, “Phenotypic deconvolution in heterogeneous cancer cell populations using drug-screening data,” *Cell Reports Methods*, vol. 3, no. 3, p. 100417, 2023.
- [31] Y. Liu, Y. Liu, and Y. Yang, “Prediction of Contact Fatigue Performance Degradation Trends Based on Multi-Domain Features and Temporal Convolutional Networks,” *Entropy*, vol. 25, no. 9, p. 1316, 2023.
- [32] R. Moraffah, M. Karami, R. Guo, A. Raglin, and H. Liu, “Causal Interpretability for Machine Learning: Problems, Methods and Evaluation,” *ACM SIGKDD Explorations Newsletter*, vol. 22, no. 1, pp. 18–33, 2020.
- [33] M. Prosperi, Y. Guo, M. Sperrin, J. S. Koopman, J. S. Min, X. He, S. Rich, M. Wang, I. E. Buchan, and J. Bian, “Causal Inference and Counterfactual Prediction in Machine Learning for Actionable Healthcare,” *Nature Machine Intelligence*, vol. 2, no. 1, pp. 369–375, 2020.



Kyaw Wanna received his Master's degree and Ph.D. in Information Technology from Assumption University, Thailand. He has over 20 years of professional experience in the information technology industry, specializing in IT management, project management, and software development. He is a Project Management Professional (PMP) and an AWS Certified Solutions Architect. His research interests include artificial intelligence, machine learning, deep learning, causal inference, predictive analytics, and intelligent decision-support systems, with applications in financial forecasting and industrial preventive maintenance. His research focuses on developing reliable, interpretable, and data-driven AI solutions to support decision-making in complex real-world environments.