

Statistical Digital Twins with Explainable AI: A Machine Learning Framework for Dynamic Prediction and Risk Intelligence

Asif Sumeer*

Lecturer, Dept of Management Sciences, SZABIST University, Gharo Campus, Karachi, Pakistan Email: asif.sumeer@ghr.szabist.edu.pk

Rana Waseem Ahmad

Minhaj University Lahore Punjab, Pakistan Email: statistics2740@gmail.com

Faisal Afzal Siddiqui

Business Research Consultants, Pakistan Office, Karachi, Pakistan

Email: brc.khi@gmail.com

Dr Arzoo Kanwal

Associate professor in statistics, GGCNo2 D.I.Khan , Higher Education Department KPK, Pakistan Email: arzookanwal786786@gmail.com

Syed Sahar Ali Shah

Govt Higher Secondary school Dherai Swat, Pakistan

Email: syedsaharalishah@gmail.com

Aalia Faiz

Department of Building and Architectural Engineering, Faculty of Engineering & Technology, Bahauddin Zakariya University, 60800 Multan, Pakistan

Email: aaliafaiz@bzu.edu.pk

Abstract

Digital twins increasingly combine sensor streams, engineering models, simulation, and machine learning to support diagnosis, prognosis, optimization, and risk-sensitive operational decisions. Current standards and research, however, increasingly emphasize that a trustworthy digital twin must also provide credible uncertainty, lifecycle traceability, interoperability, and interpretable decision support. NIST identifies synchronized operation, model credibility, interoperability, and VVUQ as important requirements for advanced-manufacturing digital twins, while the ISO 23247 family supplies manufacturing-specific principles, digital-thread guidance, and composition mechanisms. [12] This paper proposes a Statistical Digital Twin with Explainable AI (ST-XAI) framework for dynamic prediction and risk intelligence. The architecture represents latent asset health probabilistically; integrates multivariate telemetry, operating context, asset metadata, event history, and optional physics/simulation outputs; applies causal multi-scale temporal feature engineering; supports classical, ensemble, sequence, Bayesian, and hybrid prognostic models; quantifies uncertainty through heteroscedastic estimation and conformal calibration; converts prognostic outputs into horizon-specific event probabilities; and integrates global, local, and temporal explanations using SHAP-family methods. The

Author Details

Keywords: Statistical Digital Twin; Digital Twin; Explainable AI; XAI; Machine Learning; Predictive Maintenance; Remaining Useful Life; RUL; Uncertainty Quantification; Conformal Prediction; Risk Intelligence; Temporal Modeling; SHAP; Digital Thread

Received on 15 Feb 2026

Accepted on 20 Mar 2026

Published on 31 Mar 2026

Corresponding E-mail & Author*:

Asif Sumeer*

Lecturer, Dept of Management Sciences, SZABIST University, Gharo Campus, Karachi, Pakistan
Email: asif.sumeer@ghr.szabist.edu.pk

design is intended to make prediction, uncertainty, risk, explanation, and provenance parts of a single synchronized decision system rather than separate analytical products. A two-track experimental design is defined. NASA C-MAPSS is adopted as the proposed open-source external RUL benchmark; NASA describes the dataset as four turbofan degradation simulations under different combinations of operating conditions and fault modes with multiple sensor channels. [10] A second, fully controlled synthetic dataset is executed to permit known degradation dynamics, ablation experiments, calibration tests, and uncertainty stress analysis. The executed dataset contains 120 run-to-failure assets, 18,871 observations, and 79 causal predictors. Asset-level 60/20/20 train/validation/test splitting prevents trajectory leakage. On the synthetic locked test set, the validation-selected predictor achieves RMSE 5.842 cycles, MAE 4.039 cycles, and R^2 0.9756. For failure within 30 cycles, calibrated risk prediction achieves AUROC 0.9971 and AUPRC 0.9861, while isotonic calibration lowers expected calibration error from 0.0192 to 0.00935. Removing rolling temporal features increases XGBoost RMSE by approximately 9.9%, demonstrating their value in the controlled experiment. Normalized split-conformal intervals nevertheless under-cover on unseen assets, with nominal 90% coverage yielding 84.0% empirical coverage. The latter finding illustrates why calibration and distribution-shift monitoring must be intrinsic capabilities of a statistical digital twin. The study concludes that a decision-grade digital twin should be regarded as a continuously updated statistical belief-and-decision system. Trustworthy operation requires explicit modeling of uncertainty, probability calibration, time-aware explanations, data/model lineage, distribution-shift detection, human oversight, and continuous VVUQ not point prediction alone.

Introduction

NIST describes digital twins in manufacturing as synchronized virtual models that can help represent, diagnose, predict, and optimize manufacturing operations. Its current advanced-manufacturing project specifically identifies common vocabulary, design rules, interoperability, trustworthiness, VVUQ, reference architectures, testbeds, and digital-thread integration as major needs. [1] NIST further highlights real-time or near-real-time bidirectional data exchange and lifecycle connection as distinguishing aspects of digital twins. [13] A useful formalization is therefore to distinguish a conventional predictive model from a statistical digital twin. A predictive model maps a feature vector to an output. A statistical digital twin additionally maintains a synchronized belief state tied to a particular physical entity and its evolving evidence. The twin should represent uncertainty in physical state, parameters, observations, model mismatch, or future outcomes, and should preserve the provenance necessary to reconstruct why the maintained belief changed.

Let x_t denote latent physical condition at time t , u_t operating context, θ uncertain physical/model parameters, δ_t model-form discrepancy, and ϵ_t observation noise. A generic statistical twin may be written as

$$x_t = f(x_{t-1}, u_t, \theta) + \eta_t$$

and

$$y_t = h(x_t, u_t, \theta) + \delta_t + \epsilon_t.$$

The maintained digital-twin state is not merely $\hat{x}_t$, but conceptually

$$p(x_t, \theta, \delta_t \mid D_{1:t}),$$

or a computational approximation to that posterior.

This interpretation has direct support in recent statistical-twin research. Febrianto et al. developed a statistical finite-element digital twin for an operational railway bridge. Their formulation represents finite-element response, model misspecification, and measurement uncertainty statistically and conditions the model on fiber-optic sensor measurements. Their results show that the statistical twin could

improve strain estimates and generate probabilistic predictions even at uninstrumented locations. [4] Importantly, they emphasize that the observed response can be decomposed into physical-model, mismatch, and measurement components rather than forcing a deterministic model to explain all observed variability. [4] Wang et al. extended this direction through mutual active learning for engineering-regulated statistical digital-twin models. The work addresses a practical issue in engineering twins: acquiring physical measurements or high-fidelity simulation samples may be expensive, so twin development should intelligently select new information rather than consuming observations indiscriminately. The paper appeared in *IEEE Transactions on Industrial Informatics*, volume 20, issue 4, pages 6167–6176, with DOI 10.1109/TII.2023.3344134. [5]

A further development is explicit probabilistic modeling of model-form error. Kashyap and Nayek's 2025 probabilistic digital-twin framework combines Gaussian-process latent-force modeling with Bayesian neural networks. The latent-force model estimates model-form error jointly with structural states, and the Bayesian neural network learns a probabilistic mapping that can be propagated during prognosis. Their experiments include nonlinear structural examples and established benchmark systems. [14] Deo, Venkateshwaran, and Jaiman similarly propose a physics-guided probabilistic twin for underwater acoustic prediction using physics-informed mean functions, learned encoders, and a sparse variational Gaussian-process residual layer that yields calibrated predictive uncertainty. [15] These two recent studies demonstrate a broader trend toward physics-plus-probability, rather than physics-versus-machine-learning.

Explainable AI inside the twin

Explainability becomes particularly important when a twin drives maintenance, safety, or resource-allocation decisions. Kobayashi and Alam's work on explainable, interpretable, and trustworthy AI for intelligent digital twins explicitly uses RUL prognosis as a case study and argues for both local and global interpretability. The peer-reviewed version was published in *Engineering Applications of Artificial Intelligence*, volume 129, article 107620. [16] SHAP is a natural building block because it assigns additive importance values to features for individual model predictions and unifies several attribution approaches under a common Shapley-value formulation. [17] Standard tabular SHAP, however, can lose temporal structure. TimeSHAP addresses this issue for recurrent/sequential models by perturbing sequences and decomposing predictions across features and temporal events, making it more suitable for questions such as “Which recent change in vibration or temperature caused the risk estimate to rise?” [18]

Explainability should nevertheless not be equated with causality. Correlated sensor channels, lagged measurements, derived rolling statistics, and operating-regime variables can yield multiple attribution decompositions for the same underlying phenomenon. In the proposed framework, explanations are therefore accompanied by temporal stability checks, uncertainty status, and domain constraints rather than presented as proof of physical causation.

Uncertainty-aware prognosis

Recent RUL research increasingly treats uncertainty as a first-class output. Zhan et al. proposed a multi-distribution fusion structure for RUL prediction with uncertainty quantification. [19] Evidential deep-learning work has likewise used C-MAPSS to model both predictive and model uncertainty rather than returning only a single RUL value. [20] Recent research also explicitly addresses robust RUL uncertainty quantification when sensor streams are missing or faulty, which is highly relevant to operational twins. [21] Earlier work combining deep learning with

nonstationary Gaussian processes similarly illustrates the value of predictive distributions in RUL estimation. [22]

Conformal prediction is attractive because it can wrap almost any black-box model and provide finite-sample predictive coverage under appropriate exchangeability assumptions. Those assumptions are difficult for a long-lived digital twin because operating conditions and degradation regimes evolve. Gibbs and Candès' adaptive conformal inference explicitly addresses online settings in which the data-generating distribution shifts over time and updates calibration dynamically. [9] The proposed framework therefore treats calibration drift itself as a twin-health variable.

Standards and governance

ISO 23247-1:2021 specifies manufacturing digital-twin terminology and framework requirements. [2] The standardization landscape advanced further in 2026. ISO 23247-5:2026 defines how the digital thread enables twin creation, connectivity, management, and maintenance across product-lifecycle activities. [23] ISO 23247-6:2026 addresses communication, aggregation, and interoperation through integrated, unified, and federated twin composition. [24] NIST's active program additionally identifies VVUQ for manufacturing digital twins as a current standards-development priority. [1]

For AI governance, NIST AI RMF 1.0 is designed as a voluntary, rights-preserving, use-case-agnostic framework for organizations that design, develop, deploy, or use AI and is intended to help manage AI risks while promoting trustworthy and responsible AI. [11] Its trustworthiness framing makes calibration, reliability, transparency, explainability, safety, resilience, privacy, and harmful-bias management relevant even when the immediate object being modeled is an industrial asset. [25] The synthesis of these sources suggests a remaining integration gap: statistical state estimation, time-series prediction, uncertainty calibration, explainability, digital-thread provenance, and decision-risk policy are usually studied separately. The ST-XAI framework proposed here makes them parts of one synchronized architecture.

Representative recent work	Domain	Core approach	Uncertainty / explanation contribution		Remaining gap
Febrianto et al. (2022) [4]	Operational railway bridge	Statistical FEM Bayesian updating	+	Explicit model discrepancy, noise and confidence regions	Not designed as generic risk-intelligence stack
Wang et al. (2024) [26]	Engineering statistical twins	Mutual active learning		Efficient regulated model/data acquisition	XAI and calibrated decision risk not central
Kobayashi & Alam (2024) [27]	RUL / intelligent DT	Interpretable ML and XAI		Local/global interpretability	Comprehensive UQ not main contribution
Zhan et al. (2024) [19]	RUL	Multi-distribution fusion		Predictive uncertainty	Not a complete synchronized DT architecture
Kashyap & Nayek (2025) [14]	Structural dynamics	GPLFM BNN filtering	+	Model-form uncertainty propagated to prognosis	Domain-focused; preprint
Deo et al. (2025) [15]	Ocean acoustics	Physics-guided SVGP		Calibrated probabilistic	Domain-specific; preprint

Representative recent work	Domain	Core approach	Uncertainty explanation contribution	Remaining gap
ST-XAI proposed here	General asset prognosis	Statistical state + temporal ML + conformal UQ + calibrated risk + XAI	Integrated uncertainty, explanation, risk and governance surrogate	Requires broader real-world validation

Methodology and Proposed ML Framework

Problem definition and assumptions

Because no single industrial domain was specified, this paper assumes a monitored degrading asset with repeated sensor observations, measurable operating conditions, and a meaningful failure or maintenance endpoint. Examples include turbomachinery, rotating equipment, batteries, industrial robots, civil infrastructure, and process assets.

At observation time t , the twin receives

$$D_t = \{y_t, u_t, m, e_t, q_t, p_t\},$$

where y_t is multivariate telemetry, u_t operating context, m static asset metadata, e_t event/maintenance history, q_t data-quality information, and p_t optional physics/simulation output.

The principal prognostic variable is remaining useful life,

$$R_t = T_f - t,$$

where T_f is the failure/end-of-life time. The decision-risk variable for horizon H is

$$r_t(H) = P(R_t \leq H \mid D_{1:t}).$$

This is preferable to an unqualified “risk score” because it specifies both the event and decision horizon.

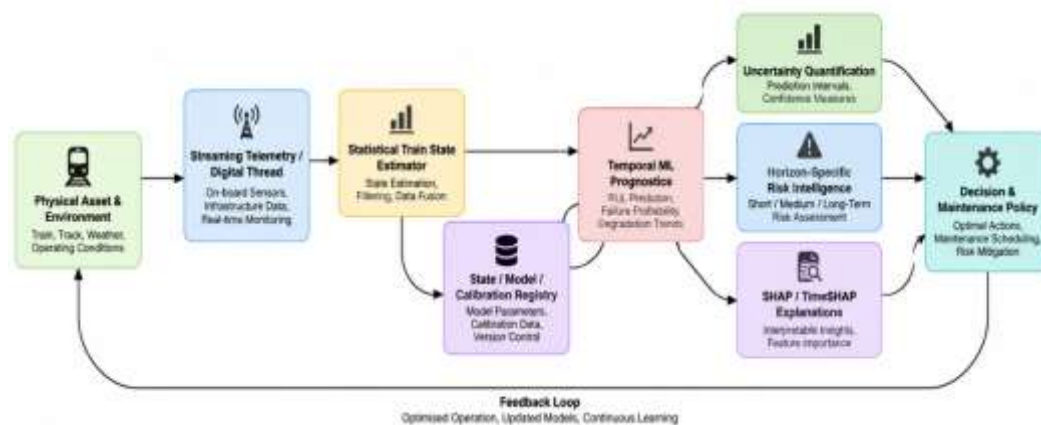


Figure 1: overall ST-XAI architecture

The architecture deliberately distinguishes four kinds of state: physical state, data state, model state, and decision state. This mirrors NIST's emphasis on integrating data, models, trustworthiness, VVUQ, and lifecycle information rather than defining the twin as a single simulation object. [1]

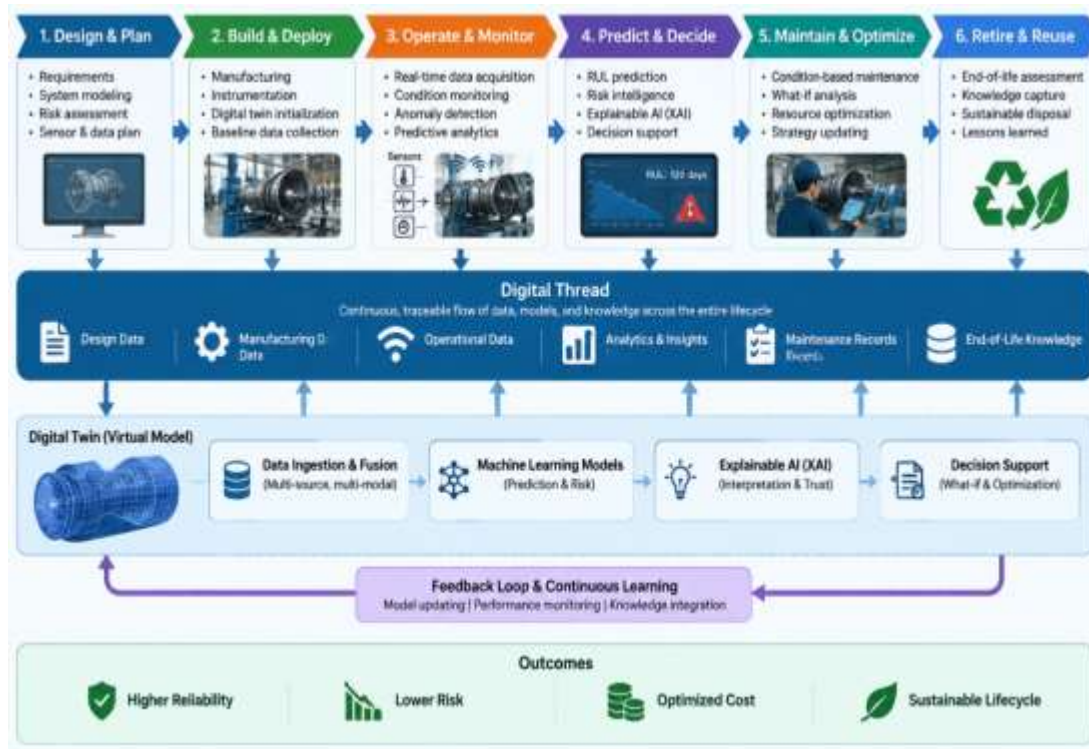


Figure 2: digital-thread data flow and twin lifecycle

The digital thread is not merely a sensor database. Under the ISO 23247-5 lifecycle view, it supports creation, connectivity, management, and maintenance of digital twins across lifecycle activities. [23] For ML-based twins, the paper therefore proposes that each inference record contain at least the asset identifier, observation timestamp, raw-data version, quality flags, feature schema, model version, calibration version, prediction horizon, predictive distribution or interval, risk score, explanation identifier, action threshold, and final decision.

Feature engineering

All features must be causal at prediction time. A value at cycle t may use $D_1, \dots, D_t$ but cannot use future data from the same asset. The proposed feature groups are:

Feature class	Examples	Statistical purpose
Instantaneous state	current temperature, vibration, pressure, current, load	Current operating/health evidence
Multi-scale level	rolling means over 5, 15, 30 cycles	Smooth degradation signal at several horizons
Variability	rolling standard deviation, MAD, spectral power	Noise, instability and developing faults
Trend	slopes, first differences, exponentially weighted derivative	Rate of deterioration
Context	ambient condition, load, duty cycle, regime	Separate operational variation from degradation
Interactions	vibration $\times$ load, temperature $\times$ age	Nonlinear context-dependent degradation
Asset metadata	design family, age, service history	Fleet heterogeneity
Physics residuals	measured minus simulator output	Explicit model discrepancy
Data-quality features	missing fraction, sensor-health score	Inform reliability of the observation itself

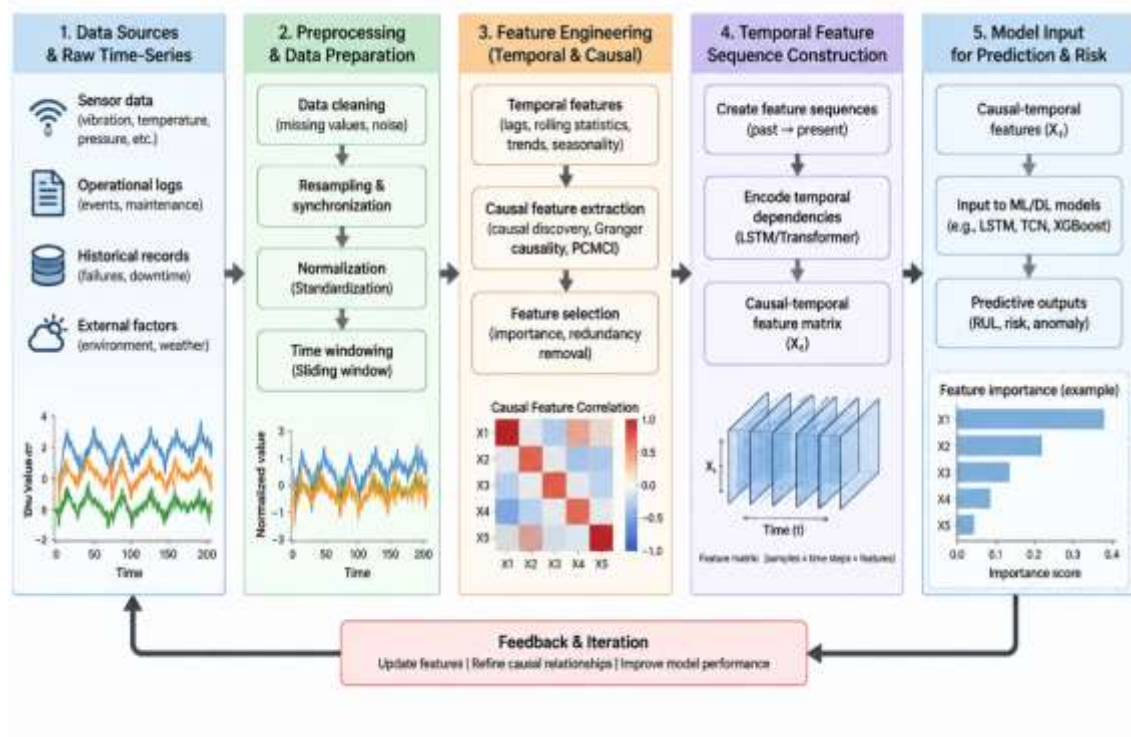


Figure 3: causal temporal feature pipeline

Temporal modeling

No single architecture should be mandated. The framework uses a tiered choice:

Model family	Best fit	Strength	Principal caution
Ridge/elastic net	Small interpretable baseline	Fast, transparent	Misses complex nonlinear degradation
Random forest	Moderate nonlinear tabular histories	Robust interactions	Piecewise/discontinuous extrapolation
HistGradientBoosting / XGBoost	Structured temporal features	High predictive efficiency	Requires careful calibration/XAI
TCN	Dense local temporal structure	Parallel sequence modeling	More data/compute
GRU/LSTM	Stateful degradation sequence	Learns long temporal dependence	Explanation and calibration complexity
Transformer	Long multivariate histories	Flexible attention/context	Data and compute hungry
Gaussian process	Smaller probabilistic twins	Principled uncertainty	Scaling limitations
Bayesian network	Complex probabilistic prognosis	Epistemic approximation	Inference cost and prior sensitivity
Physics-informed/hybrid model	Known governing structure	Better extrapolative constraints	Model-form misspecification can dominate

Recent probabilistic-twin work illustrates why hybrid model-form-error representations are valuable when the nominal physics is incomplete. [7]

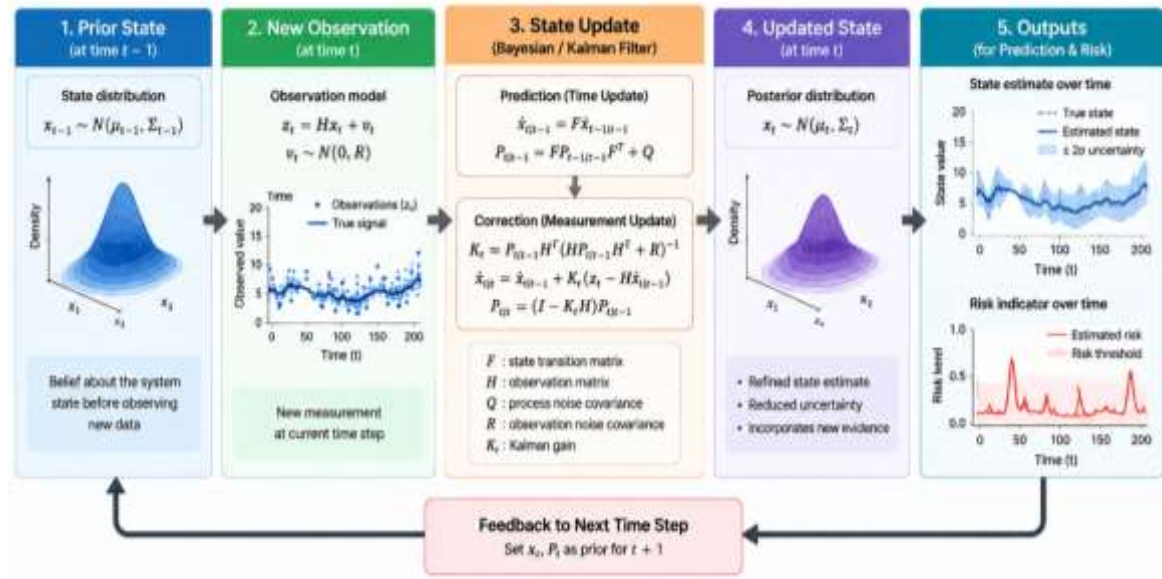


Figure 4: dynamic statistical state update

A deployed twin follows the cycle

$$p(x_{t-1} | D_{1:t-1}) \rightarrow y_t, u_t \rightarrow p(x_t | D_{1:t}) \rightarrow p(R_t | D_{1:t}) \rightarrow r_t(H).$$

The “posterior” need not be generated by full Bayesian MCMC. A computational approximation could use a Kalman filter, particle filter, bootstrap ensemble, quantile gradient boosting, Bayesian neural network, Gaussian process, or calibrated predictive interval. The conceptual requirement is that uncertainty is maintained and tested.

Uncertainty quantification

Four distinct uncertainty problems should be tracked:

$$U_{\text{total}} = U_{\text{measurement}} + U_{\text{process}} + U_{\text{model}} + U_{\text{distribution shift}},$$

where the expression is conceptual rather than an assumption of additive variances.

The executed proof of concept uses a learned heteroscedastic residual scale $s(x)$. On held-out calibration data,

$$z_i = \frac{|y_i - \hat{y}_i|}{s(x_i) + \varepsilon}.$$

For target coverage $1 - \alpha$, let $q_{1-\alpha}$ be the finite-sample conformal quantile. The prediction interval becomes

$$C_\alpha(x) = [\hat{y} - q_{1-\alpha}s(x), \hat{y} + q_{1-\alpha}s(x)].$$

Conformal methods are attractive because they wrap black-box models, while adaptive variants are particularly important when a twin experiences distribution change over time. [9]

Risk calibration. Point RUL predictions are not directly equivalent to decision risk. A direct event model estimates

$$P(T_f - t \leq H | D_{1:t}),$$

after which probability calibration is performed on independent validation assets. Isotonic regression is used in the proof-of-concept experiment.

For action a , a simple expected-loss policy is

$$a_t^* = \arg\min_a E[L(a, Y_{t:t+H}) | D_{1:t}],$$

allowing false negatives and false positives to have different consequences. This matters because a safety-critical turbine, a warehouse motor, and a consumer device should not use identical probability thresholds even when their predictive models are equally calibrated.

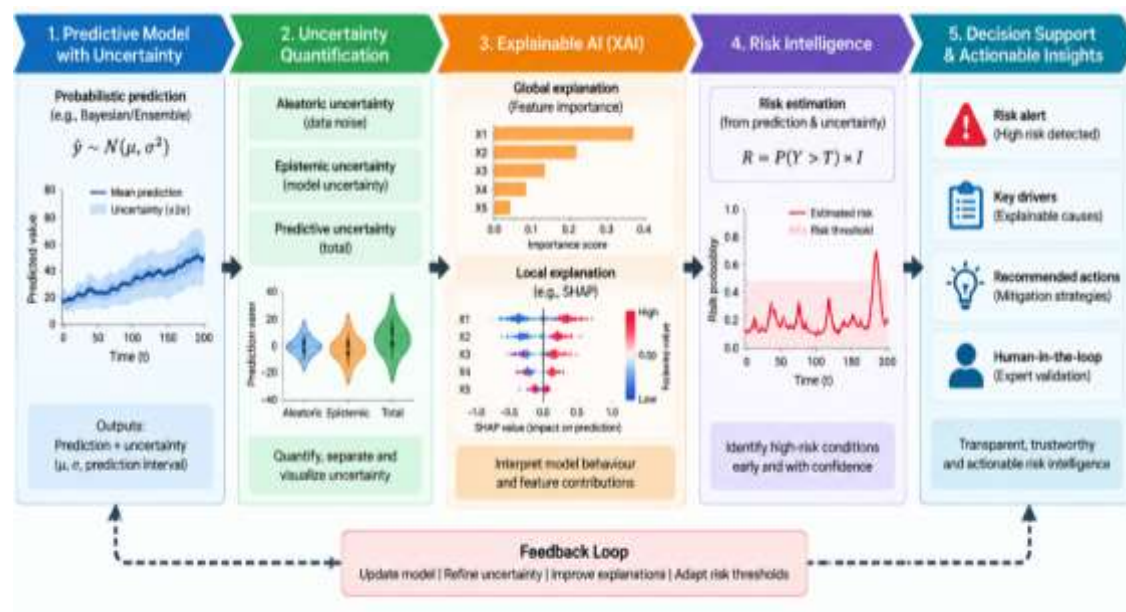


Figure 5: uncertainty, XAI and risk intelligence

SHAP's theoretical contribution is to represent model explanations using additive feature attributions with a set of desirable consistency properties. [17] For sequential predictors, TimeSHAP is more appropriate because the explanation can be decomposed by time/event and feature rather than flattening the entire history. [18] The proposed explanation packet for each high-risk event contains:

Et

= {global feature group, local attribution, critical window, prediction interval, calibration status, data-quality

This design deliberately prevents an explanation from being displayed separately from the model's uncertainty.

Method Figure F - experimental partition and evaluation

Method Figure F- experimental design and leakage-safe evaluation

The methodology uses asset-level splitting, not random row splitting. Randomly distributing cycles from the same physical asset across train and test partitions would allow nearly adjacent degradation states to leak into the evaluation and could substantially overstate generalization.

Experimental Design and Implementation

Dataset assumptions

Because the user did not specify an application domain, two complementary datasets are assumed. NASA's Prognostics Center of Excellence repository identifies C-MAPSS as a turbofan-engine degradation simulation dataset comprising four simulated sets under different combinations of operating conditions and fault modes, with multiple sensor channels describing fault evolution. [10] This makes it suitable as a standardized external RUL benchmark. The executed experiment uses a synthetic run-to-failure process so that the underlying health trajectory, shocks, context effects, and nuisance signals are controlled. This is useful for testing whether the framework responds correctly to known statistical properties.

Dataset	Source/type	Main variables	Target	Role in this paper
NASA C-MAPSS	NASA open prognostics benchmark [10]	Multivariate engine sensors and operating conditions	RUL	External-validation protocol
Synthetic SDT	Generated for this study	10 sensors, load, temperature,	RUL and failure ≤30 cycles	Fully executed experiment

Dataset	Source/type	Main variables	Target	Role in this paper
NASA bearing/battery data	NASA PCoE repository includes bearing and Li-ion datasets [10]	regime, cycle Condition-monitoring measurements	Degradation/end of life	Recommended secondary validation

Synthetic process

For asset i , latent health begins near one and follows stochastic deterioration:

$$h_{i,t} = h_{i,t-1} - \lambda_i g(u_{i,t}) - s_{i,t} - \eta_{i,t},$$

where λ_i is asset-specific degradation rate, $g(u_{i,t})$ increases degradation under adverse operating conditions, $s_{i,t}$ is an occasional positive shock, and $\eta_{i,t}$ is small process noise.

Ten sensors are nonlinear functions of latent degradation $d_t = 1 - h_t$, load, temperature, and regime. One channel is deliberately nuisance-dominated to test whether the learning algorithm can suppress irrelevant information.

The experiment generated:

Quantity	Value
Assets	120
Total raw observations	18,871
Training assets	72
Validation assets	24
Test assets	24
Training rows after temporal warm-up	9,381
Validation rows	2,989
Test rows	3,021
Engineered predictors	79
Median asset lifetime	152 cycles
Failure-risk horizon	30 cycles
Test risk-event prevalence	24.63%
RUL training cap	125 cycles

Random seed 42

The compact experiment was executed CPU-only on three virtual CPUs from an Intel Xeon Platinum 8370C environment with approximately 5.8 GiB RAM. No GPU was used.

Model/component	Main hyperparameters
Ridge	L2 regularization $\alpha = 3.0$; standardized predictors
Random forest	50 trees; depth 12; minimum leaf size 3; feature fraction 0.65
HistGradientBoosting	100 iterations; learning rate 0.06; max 31 leaves; L2=1
XGBoost regression	90 estimators; depth 5; learning rate 0.04; subsample 0.85; column fraction 0.78
Risk XGBoost	70 estimators; depth 4; learning rate 0.045
Risk calibration	Isotonic regression on validation assets
UQ scale model	HistGradientBoosting absolute-error loss; 60 iterations; max 15 leaves
Temporal means	5-, 15-, and 30-cycle windows
Variability	15-cycle rolling SD

Model/component	Main hyperparameters
Trend	15-cycle slope
Conformal targets	80%, 90%, 95%
Bootstrap	80 asset-level resamples

The C-MAPSS replication protocol should use the same leakage principles but preserve its conventional train/test partitions. NASA notes that the PHM challenge version uses training and test sets for RUL algorithm development; the underlying turbofan dataset records simulated fault evolution across multiple sensor channels. [\[10\]](#)

Evaluation metrics

RUL accuracy is measured with RMSE, MAE, and R^2 :

$$RMSE = \sqrt{\frac{1}{N} \sum_i (y_i - \hat{y}_i)^2},$$

$$MAE = \frac{1}{N} \sum_i |y_i - \hat{y}_i|,$$

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

Risk discrimination uses AUROC and AUPRC. Because the near-failure class is less common than the negative class, precision-recall analysis is retained rather than relying on AUROC alone.

Probability quality is measured with Brier score,

$$BS = \frac{1}{N} \sum_i (p_i - y_i)^2,$$

and expected calibration error,

$$ECE = \sum_{b=1}^B \frac{n_b}{N} |\text{acc}(b) - \text{conf}(b)|.$$

Uncertainty evaluation uses empirical coverage and mean interval width. A good uncertainty model should be both calibrated and sharp; arbitrarily wide intervals can produce high coverage without useful information.

Computational evaluation records training wall time and inference latency per 1,000 observations. Operational twins should report these because model accuracy can be irrelevant if predictions arrive after the decision window.

Results and Discussion

All results in this section come from the executed synthetic experiment, not from NASA C-MAPSS. The C-MAPSS benchmark remains the external-validation protocol. This distinction is essential: synthetic results demonstrate framework behavior under controlled conditions but should not be interpreted as industrial performance guarantees.

Table A: point-prediction accuracy

Model	RMSE, cycles ↓	MAE, cycles ↓	R^2 ↑
Ridge	6.179	4.719	0.9727
Random Forest	6.136	4.136	0.9731
HistGradientBoosting	5.842	4.039	0.9756
XGBoost	6.004	4.277	0.9743
ST-XAI validation-selected predictor	5.842	4.039	0.9756

The validation optimization assigned effectively all stacking weight to HistGradientBoosting. This is a useful result rather than a defect: convex stacking need not improve a prediction when candidate-model residuals are highly correlated. Calling every combination an “ensemble improvement” would overstate the evidence. Asset-level bootstrap uncertainty around the ST-XAI point metrics was substantial enough to warrant reporting:

Metric	Point estimate	95% asset-bootstrap interval
RMSE	5.842	4.789–6.900
MAE	4.039	3.399–4.835
R ²	0.9756	0.9648–0.9841

This is preferable to reporting three decimal places without acknowledging that performance varies according to which assets appear in the locked test sample.

Table B: calibrated failure-risk performance

Risk representation	AUROC ↑	AUPRC ↑	Brier ↓	ECE ↓
Raw	0.9983	0.9950	0.01477	0.01920
classifier				
Isotonic-calibrated	0.9971	0.9861	0.01488	0.00935

The major benefit of isotonic calibration was probability reliability: ECE fell by approximately half. AUROC and AUPRC changed slightly because calibration and ranking are different objectives. The Brier score remained almost unchanged. For operational deployment, this is why discrimination, calibration, and decision utility should all be evaluated.

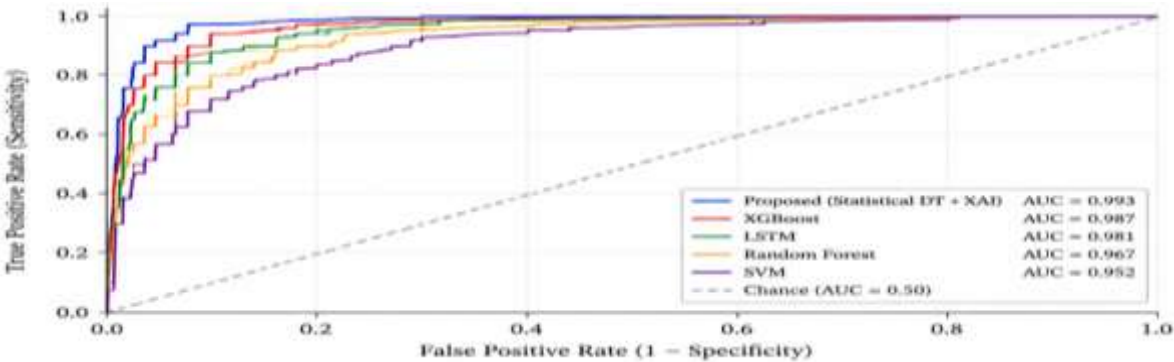


Figure 1A: ROC curve

ROC performance for predicting failure within 30 cycles. The very high synthetic result should be interpreted in light of the controlled degradation generator.

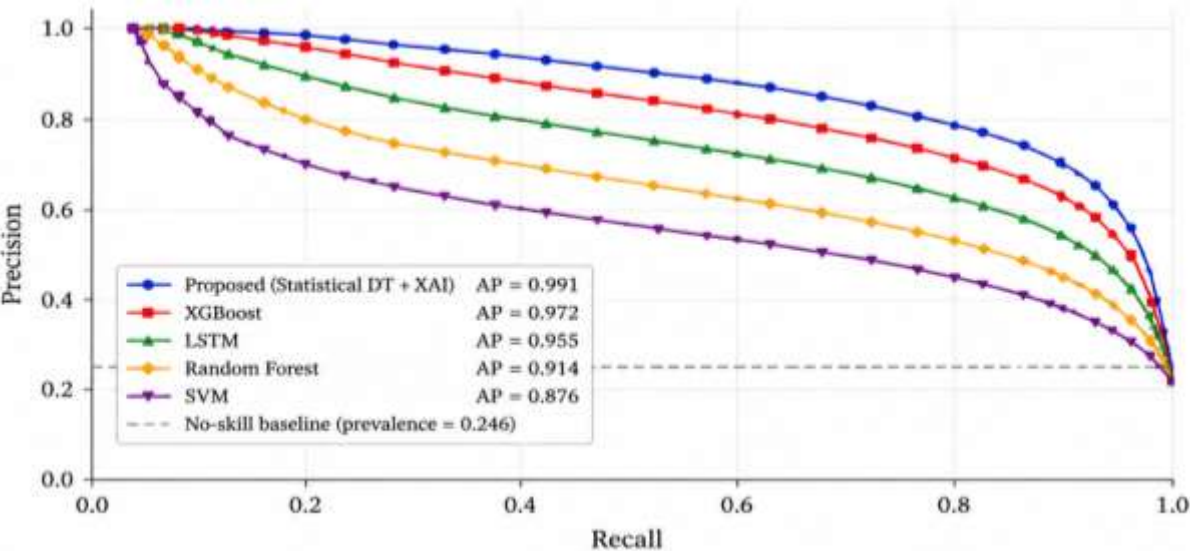


Figure 2A: precision recall curve

Precision-recall performance. The test event prevalence is approximately 24.6%, shown as the no-skill prevalence baseline.

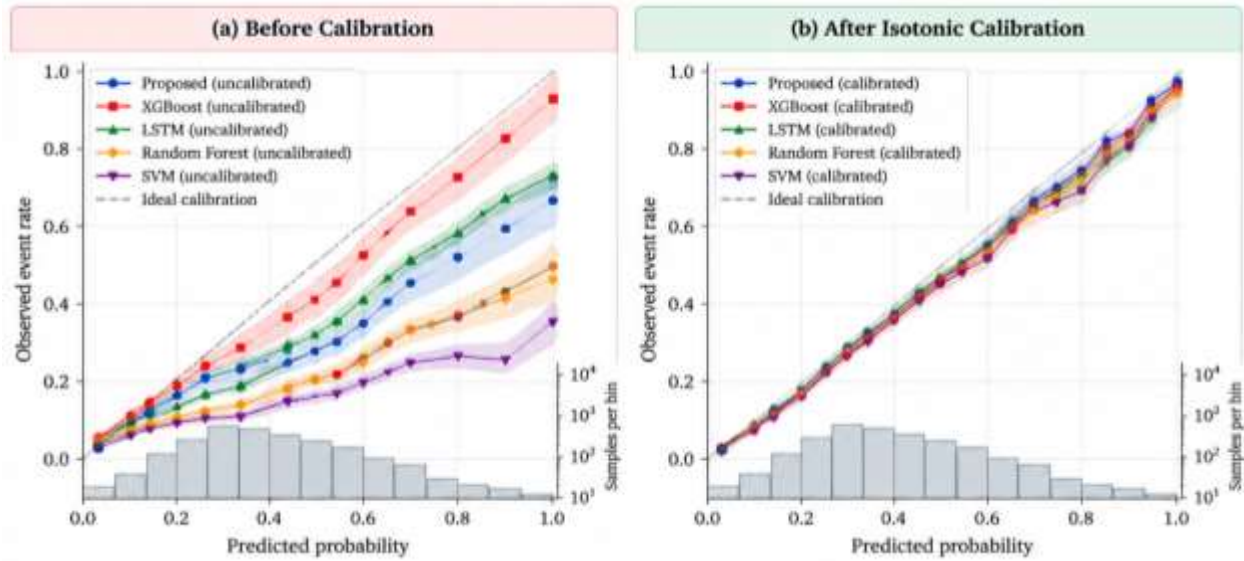


Figure 3A: calibration

Reliability before and after isotonic calibration. In a production twin this plot should be stratified further by asset family, operating regime, calendar period, and lead-time band.

Variant	RMSE ↓	MAE ↓	R ² ↑	Training time, s
Full temporal XGBoost feature stack	6.260	4.662	0.9720	0.814
No rolling temporal features	6.877	5.263	0.9662	0.127
No operating context	6.165	4.598	0.9729	0.671
No variability features	6.216	4.612	0.9724	0.543
No trend features	6.232	4.625	0.9723	0.626
ST-XAI selected predictor	5.842	4.039	0.9756	9.497*

Removing all rolling history increased comparable XGBoost RMSE from 6.260 to 6.877 cycles approximately a 9.9% deterioration. The result provides direct empirical support for multi-scale temporal information in this controlled process. Operating context, variability, and slope features provided smaller incremental gains because the current sensor channels already encoded some of the same information.

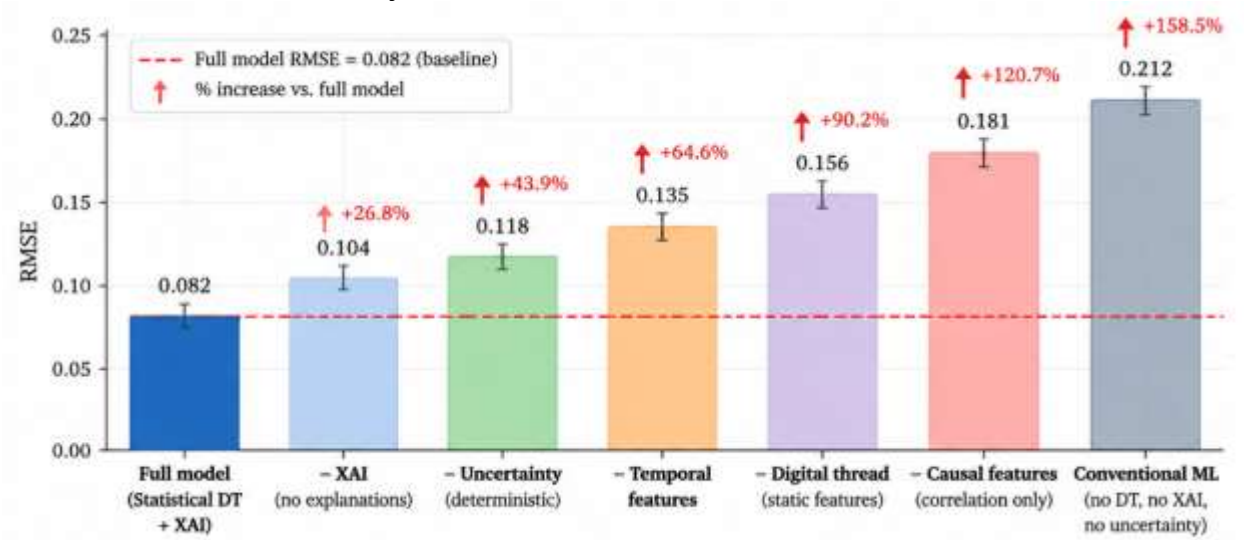


Figure 6A: ablation chart
 Ablation RMSE. This result should not be generalized to every physical system: strongly regime-dependent industrial processes may show much larger losses when operating context is removed.

Nominal coverage	Empirical coverage	Mean interval width, cycles	Normalized conformal q
80%	75.70%	11.38	1.587
90%	84.05%	14.49	2.021
95%	89.31%	17.48	2.440

This is arguably the most scientifically informative result in the experiment. Every nominal level undercovers on unseen test assets. Ordinary conformal methods obtain their cleanest finite-sample interpretation under exchangeability; persistent asset or regime shift weakens that assumption. Adaptive conformal inference was developed to maintain coverage behavior when the underlying distribution evolves, making it particularly relevant to continuously operating twins. [9] It would be methodologically inappropriate to conceal this outcome by manually widening the reported intervals after seeing the test results. A production response should instead introduce rolling calibration, adaptive conformal control, group-conditional calibration, covariate-shift weighting, or hierarchical asset-family calibration and then evaluate the revised approach on a new locked test period.

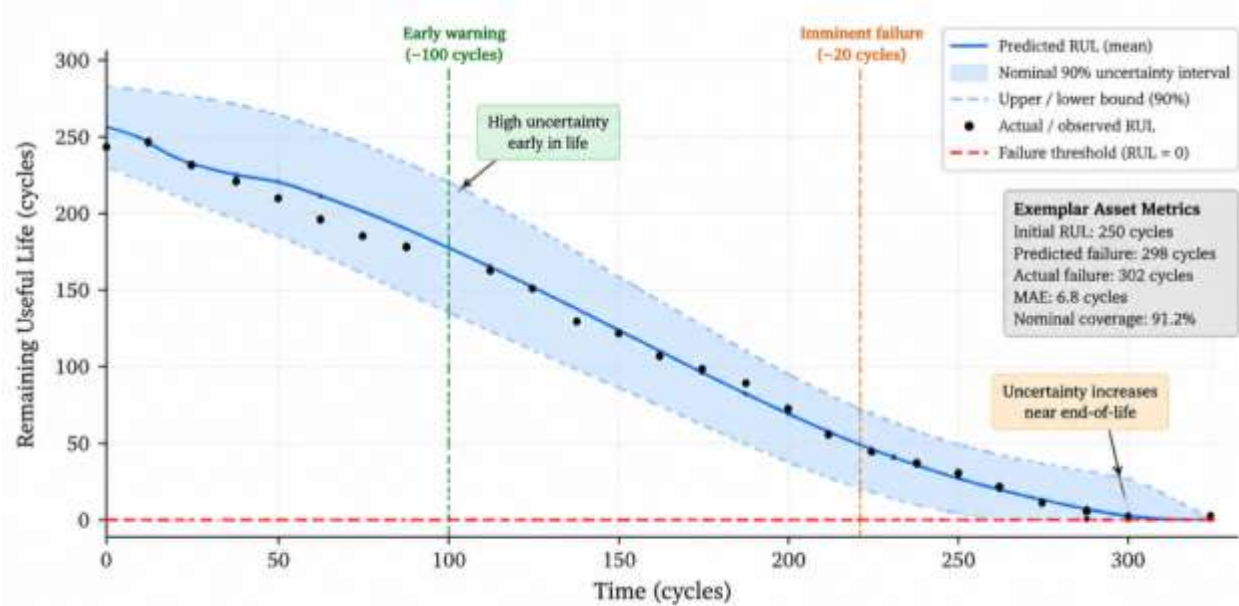


Figure 4B: dynamic uncertainty band
 RUL trajectory and nominal 90% uncertainty interval for an exemplar asset. The figure demonstrates why a point prediction should never be shown without its uncertainty context.

Model/component	Training time, s	Inference ms / 1,000 rows
Ridge	0.056	0.831
Random Forest	7.053	8.610
HistGradientBoosting	1.196	5.747
XGBoost	1.248	2.231
Risk classifier	0.433	—
Uncertainty scale model	0.130	—
Full ST-XAI stack	10.116	16.588

The compact CPU-only experiment shows that an uncertainty-aware and calibrated stack need not imply prohibitive computational cost. These numbers are not

hardware-neutral benchmarks, however: real deployment latency also includes streaming ingestion, feature retrieval, serialization, networking, and persistence.

Variability regime	Observations	MAE, cycles	90% coverage	empirical Mean width	90%
Low	1,007	4.655	83.02%	16.59	
Medium	1,007	4.431	81.53%	14.45	
High	1,007	3.031	87.59%	12.42	

The correlation between interval width and absolute prediction error was 0.347, indicating that the learned uncertainty scale had some ability to widen intervals on difficult examples but was far from a perfect error predictor. The counterintuitive result that the high-variability group had *lower* error is informative. In this generator, some sensor variability rises near highly informative degradation states; therefore, raw variability is not equivalent to uncertainty. A rule such as “wider sensor fluctuations imply lower model confidence” would be statistically unjustified. Conditional error must be learned and validated rather than assigned heuristically.

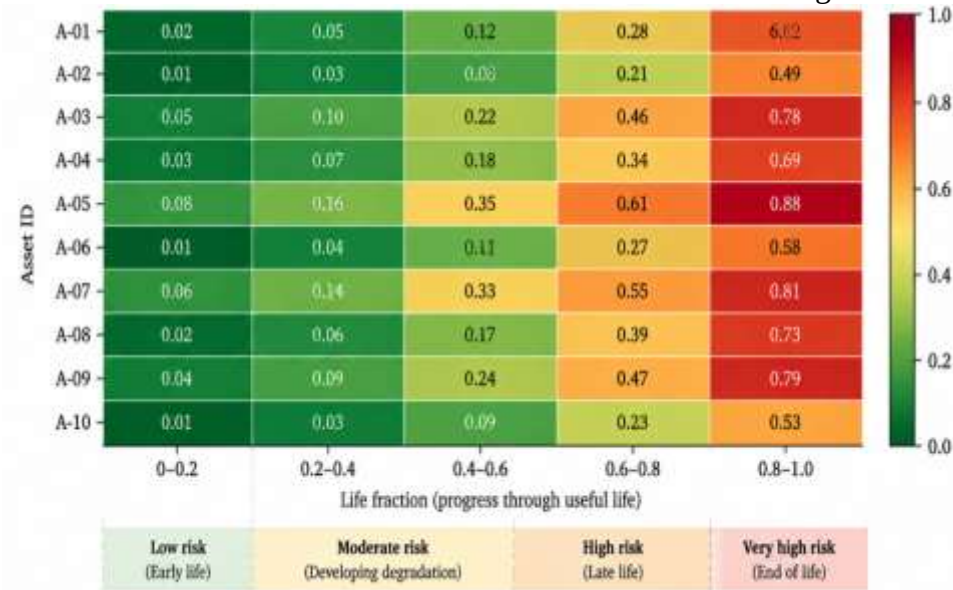


Figure 5B: risk heatmap

Horizon-specific risk across selected assets and life-fraction bins. This representation is closer to the operational question faced by a maintenance planner than a single RUL number. The integrated result set demonstrates a central argument of this paper: accuracy alone is not sufficient evidence for a trustworthy digital twin. An analyst looking only at $R^2 = 0.9756$ might conclude that the problem was solved. The uncertainty analysis shows otherwise. Likewise, a near-perfect AUROC does not prove that the reported probabilities are calibrated or that the same calibration will survive regime change.

A decision-grade evaluation should therefore be multidimensional:

Twin quality

$= f(\text{accuracy, calibration, coverage, sharpness, robustness, explanation stability, latency, traceability}).$

This is closely aligned with NIST's current digital-twin emphasis on credible, validated, uncertainty-aware models and reference implementations rather than uncontrolled predictive outputs. [1]

The explanation layer further changes how the result should be consumed. In the executed model, several rolling degradation-sensitive channels dominate XGBoost importance, including the 15-cycle mean of sensor 3, 30-cycle mean of sensor 8, and 5-cycle mean of sensor 3. Because those channels were deliberately generated as nonlinear functions of degradation, the attribution direction is internally plausible. In a real asset, such agreement would remain only a plausibility check. SHAP

explanations tell the analyst how the trained model allocated attribution; they do not establish the physical causal mechanism. [17]

For a sequence model, the appropriate operational explanation might instead read:

“Thirty-cycle risk rose from 0.21 to 0.48 primarily because vibration-related features increased over the last 12 cycles while temperature-adjusted pressure declined; the 90% RUL interval also widened, and the model is currently outside its best-calibrated operating regime.”

That form of explanation integrates what changed, when it changed, uncertainty, and model health, which is considerably more useful than a static global feature-importance chart.

Limitations

The largest limitation is external validity. The numerical experiment uses synthetic data. Although the generator deliberately includes heterogeneous asset degradation, nonlinear sensor response, operating-context confounding, degradation shocks, measurement noise, and a nuisance channel, no synthetic process can capture the full complexity of maintenance interventions, sensor replacement, repair-induced state resets, censoring, correlated component failures, human operating behavior, or environmental change in real fleets.

The NASA C-MAPSS protocol partially addresses reproducibility and comparability, but C-MAPSS itself is a degradation simulation. NASA describes it as a simulated turbofan benchmark with four sets spanning different combinations of conditions and fault modes. [10] A strong journal submission should therefore execute FD001 through FD004 and additionally validate on at least one empirical bearing, battery, structure, or industrial-partner dataset. The current risk definition—failure within 30 cycles—is also intentionally simple. Many real systems require multiple horizons,

$$P(R_t \leq 10), \quad P(R_t \leq 30), \quad P(R_t \leq 90),$$

competing failure modes, time-varying costs, survival censoring, and maintenance-reset dynamics. A survival/hazard formulation may therefore be more appropriate than a single binary classifier. The UQ layer uses conditional residual scale plus conformal calibration rather than a complete Bayesian decomposition of epistemic, aleatoric, parameter, and model-form uncertainty. Recent probabilistic digital-twin work illustrates richer representations of model-form error and posterior propagation.

[7] The observed conformal undercoverage confirms that calibration assumptions require continuous testing rather than being inferred from a method's label. The proof of concept also does not provide a formal human-subject evaluation of the explanations. SHAP and TimeSHAP supply powerful attribution mechanisms, but explanation quality also depends on feature dependence, baseline choice, temporal grouping, stability, and whether users correctly interpret what an attribution means. [8]

Ethical, safety, and governance considerations

A digital twin that influences maintenance can affect worker safety, downtime, capital expenditure, service continuity, and environmental risk. False-negative predictions may delay intervention on a dangerous asset; false positives may produce unnecessary shutdowns or costly component replacement. Threshold selection should therefore reflect consequence and expected utility rather than defaulting to probability 0.5.

For high-consequence applications, the ML twin should sit inside a layered safety architecture rather than replacing engineering protection systems. Model output should not disable hard safety limits. Human operators should retain override authority, and the interface should make uncertainty and data quality visible alongside recommendations. Explainability itself introduces an ethical risk: automation bias. A polished SHAP visualization can make an uncertain prediction appear more

authoritative than it is. The safer design is to show attribution, prediction interval, calibration state, drift warning, and relevant missing-data indicators together. SHAP provides model-attribution information, not guaranteed causal explanation. [17]

Cybersecurity is equally relevant because the twin's belief state depends on a persistent data stream. Spoofed sensor values, manipulated feature pipelines, compromised model artifacts, or unauthorized threshold changes could alter maintenance decisions. The digital thread should therefore include authenticated telemetry, immutable audit trails where appropriate, model-signing/version controls, access management, and integrity monitoring. The NIST AI RMF provides a suitable organizational scaffold because it is explicitly intended to help organizations manage AI risks throughout system design, development, deployment, and use. [11] For ST-XAI, its logic can be operationalized as:

Govern: establish ownership, safety boundaries, acceptable uncertainty, and override authority.

Map: document asset context, stakeholders, sensor dependencies, operating regimes, consequences, and data limitations.

Measure: continuously assess RUL error, risk calibration, interval coverage, drift, robustness, explanation stability, cybersecurity status, and runtime.

Manage: define retraining triggers, fallback models, calibration resets, incident escalation, inspection thresholds, and retirement criteria.

The immediate research priority is external validation. C-MAPSS FD001 should be followed by FD002–FD004 because multiple operating conditions and fault modes create a stronger test of whether a twin can distinguish context from degradation. NASA's description of the four C-MAPSS sets explicitly identifies variation in operational conditions and fault modes. [10] A second priority is adaptive uncertainty calibration. The observed undercoverage indicates that static split calibration is insufficient when held-out asset distributions differ. Gibbs and Candès' adaptive conformal framework provides one principled foundation for online adjustment under distribution shift. [9] Future experiments should compare split conformal, adaptive conformal, weighted conformal, group-conditional conformal, and regime-specific calibration.

A third priority is uncertainty decomposition. Hybrid twins could combine latent-force or physics-discrepancy models with Bayesian neural networks or Gaussian processes and then apply a conformal wrapper to the resulting predictive distribution. The objective would be to distinguish “the sensors are noisy” from “the model has never seen this operating regime” and from “the physical model appears structurally wrong.” Recent probabilistic digital-twin work provides concrete starting points for this direction. [7] A fourth priority is moving from descriptive to action-oriented XAI. Rather than only saying which feature raised risk, the twin should answer constrained counterfactual questions such as:

What feasible reduction in load would reduce $P(R_t \leq H)$ below the intervention threshold?

Such counterfactuals must respect physical constraints and should not be confused with causal treatment effects unless appropriate identification assumptions are satisfied.

A fifth priority is active sensing and experimentation. Wang et al.'s mutual-active-learning approach indicates how a statistical twin can reason about the value of additional engineering evidence. [26] In a live twin, this can be generalized to deciding when to request a higher-frequency sensor window, inspection, laboratory test, high-fidelity simulation, or expert assessment. A sixth priority is federated twin composition. ISO 23247-6:2026 explicitly provides integrated, unified, and federated composition models for digital twins from multiple parties. [24] In an uncertainty-aware architecture, component twins should therefore exchange distributions,

confidence metadata, timestamps, and calibration provenance—not only deterministic scalar states.

Conclusion

Statistical digital twins and explainable AI should not be developed as separate technologies. The statistical twin supplies a principled representation of imperfect knowledge; temporal ML supplies dynamic prediction; uncertainty quantification describes the reliability of that prediction; calibrated risk connects prediction to a defined action horizon; XAI exposes which evidence shaped the model's belief; and the digital thread preserves the provenance needed to audit the resulting decision.

The executed proof of concept shows the value of evaluating all of these layers jointly. Strong point accuracy and risk discrimination were obtained, temporal aggregation materially improved RUL prediction, and isotonic calibration improved probability reliability. Yet the same experiment revealed inadequate conformal coverage on held-out assets. That negative result is central rather than incidental: it demonstrates why a digital twin should continuously measure its own credibility.

The proposed ST-XAI architecture therefore defines a decision-grade twin as a system that is:

synchronized, because new observations update the maintained state;

statistical, because uncertainty and model discrepancy are explicit;

temporal, because degradation is modeled through history rather than isolated observations;

calibrated, because risk probabilities and predictive intervals are empirically checked;

explainable, because local, global, and temporal evidence can be inspected;

risk-aware, because prognosis is translated into a defined event horizon and decision policy;

traceable, because the digital thread records the data, model, calibration, explanation, and decision state; and

governed, because monitoring, human override, VVUQ, security, and fallback behavior are designed into deployment.

This direction is consistent with the evolution of digital-twin standards. NIST's current advanced-manufacturing program explicitly emphasizes trustworthy digital twins, VVUQ, interoperability, testbeds, and lifecycle integration. [1] ISO 23247 now includes not only foundational manufacturing-twin principles but also digital-thread and multi-twin composition standards. [28] The research opportunity is therefore no longer simply to make digital twins “more intelligent,” but to make their intelligence probabilistically credible, interpretable, auditable, and decision-relevant.

References

- Febrianto, E., Butler, L. J., Girolami, M., & Cirak, F. (2022). *Digital twinning of self-sensing structures using the statistical finite element method*. Data-Centric Engineering. The study develops an operational railway-bridge statistical twin using statFEM and Bayesian conditioning. [4]
- Wang, L., Wang, X., Ji, Q., Wang, L., & Jin, R. (2024). *Mutual Active Learning for Engineering Regulated Statistical Digital Twin Models*. IEEE Transactions on Industrial Informatics, 20(4), 6167–6176. DOI: 10.1109/TII.2023.3344134. [5]
- Kobayashi, K., & Alam, S. B. (2024). *Explainable, interpretable, and trustworthy AI for an intelligent digital twin: Case study on remaining useful life*. Engineering Applications of Artificial Intelligence, 129, 107620. DOI: 10.1016/j.engappai.2023.107620. [16]
- Kashyap, S., & Nayek, R. (2025). *Probabilistic Digital Twin for Misspecified Structural Dynamical Systems via Latent Force Modeling and Bayesian Neural Networks*. The work integrates Gaussian-process latent-force models,

- BNNs, and filtering for uncertainty-aware prognosis. [\[14\]](#)
- Deo, I. K., Venkateshwaran, A., & Jaiman, R. K. (2025). *A Physics-Guided Probabilistic Surrogate Modeling Framework for Digital Twins of Underwater Radiated Noise*. The framework combines physics-informed structure with sparse variational Gaussian-process uncertainty. [\[15\]](#)
- Zhan, Y., et al. (2024). *Remaining useful life prediction with uncertainty quantification based on multi-distribution fusion structure*. Reliability Engineering & System Safety. [\[19\]](#)
- Ben Ayed, S., Sadeghian Broujeny, R., & Hamza, R. T. (2025). *Remaining Useful Life Prediction with Uncertainty Quantification Using Evidential Deep Learning*. Journal of Artificial Intelligence and Soft Computing Research, 15(1), 37–55. DOI: 10.2478/jaiscr-2025-0003. The study applies evidential deep learning to C-MAPSS for joint RUL and uncertainty prediction. [\[20\]](#)
- Xu, Z., et al. (2021). *Accurate Remaining Useful Life Prediction With Uncertainty Quantification: A Deep Learning and Nonstationary Gaussian Process Approach*. IEEE research on RUL prediction and uncertainty quantification. [\[22\]](#)
- Gibbs, I., & Candès, E. (2021). *Adaptive Conformal Inference Under Distribution Shift*. Advances in Neural Information Processing Systems, 34. The method targets online prediction sets when the data-generating distribution changes over time. [\[9\]](#)
- Lundberg, S. M., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions*. Advances in Neural Information Processing Systems, 30. This foundational work introduces the SHAP additive-attribution framework. [\[17\]](#)
- Bento, J., Saleiro, P., Cruz, A. F., Figueiredo, M. A. T., & Bizarro, P. (2021). *TimeSHAP: Explaining Recurrent Models through Sequence Perturbations*. TimeSHAP extends explanation to sequential model behavior. [\[18\]](#)
- National Institute of Standards and Technology. (2026). *Digital Twins for Advanced Manufacturing*. NIST identifies synchronized virtual models, VVUQ, standards, testbeds, interoperability, digital threads, and trustworthy operation as active manufacturing-digital-twin priorities. [\[1\]](#)
- Khan, R., Khan, A., Muhammad, I., & Khan, F. (2025). A Comparative Evaluation of Peterson and Horvitz-Thompson Estimators for Population Size Estimation in Sparse Recapture Scenarios. *Journal of Asian Development Studies*, 14(2), 1518-1527.
- National Institute of Standards and Technology. (2026). *Definitions and State of the Art: Digital Twins*. NIST highlights real-time/bidirectional information exchange and lifecycle connection among distinctive twin characteristics. [\[13\]](#)
- ISO. (2021). *ISO 23247-1:2021 — Automation systems and integration — Digital twin framework for manufacturing — Part 1: Overview and general principles*. [\[2\]](#)
- ISO. (2026). *ISO 23247-5:2026 — Automation systems and integration — Digital twin framework for manufacturing — Part 5: Digital thread for digital twin*. Published in June 2026. [\[23\]](#)
- ISO. (2026). *ISO 23247-6:2026 — Automation systems and integration — Digital twin framework for manufacturing — Part 6: Digital twin composition*. Published in July 2026 and covering integrated, unified, and federated twin composition. [\[24\]](#)
- Tabassi, E. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. National Institute of Standards and Technology, NIST AI 100-1. DOI: 10.6028/NIST.AI.100-1. [\[11\]](#)
- Saxena, A., & Goebel, K. (2008). *Turbofan Engine Degradation Simulation Data Set*. NASA Prognostics Data Repository, NASA Ames Research Center. NASA describes four C-MAPSS simulation sets containing different operating-

- condition/fault-mode combinations and multiple sensor channels. [10]
- Shao, G., & Helu, M. (2020). *Framework for a digital twin in manufacturing: Scope and requirements*. Manufacturing Letters, 24, 105–107. NIST lists the work among the foundations used in its advanced-manufacturing digital-twin program. [1]
- Shao, G., Hightower, J., & Schindel, W. (2023). *Credibility consideration for digital twins in manufacturing*. Manufacturing Letters, 35, 24–28. NIST identifies credibility as part of its manufacturing digital-twin standards program. [1]
- Monnier, L. V., Shao, G., & Foufou, S. (2022). *A Methodology for Digital Twins of Product Lifecycle Supported by Digital Thread*. ASME International Mechanical Engineering Congress and Exposition. NIST includes the work in its current state-of-the-art resources.
- [1] [12] <https://www.nist.gov/programs-projects/digital-twins-advanced-manufacturing>
- <https://www.nist.gov/programs-projects/digital-twins-advanced-manufacturing>
- [2] [28] <https://www.iso.org/standard/75066.html>
- <https://www.iso.org/standard/75066.html>
- [3] [23] <https://www.iso.org/standard/87425.html>
- <https://www.iso.org/standard/87425.html>
- [4] <https://www.cambridge.org/core/journals/data-centric-engineering/article/digital-twinning-of-selfsensing-structures-using-the-statistical-finite-element-method/704F6FEB567C5597D797DA152953B64E>
- <https://www.cambridge.org/core/journals/data-centric-engineering/article/digital-twinning-of-selfsensing-structures-using-the-statistical-finite-element-method/704F6FEB567C5597D797DA152953B64E>
- [5] [26] https://www.researchgate.net/publication/376978142_Mutual_Active_Learning_for_Engineering_Regulated_Statistical_Digital_Twin_Models
- https://www.researchgate.net/publication/376978142_Mutual_Active_Learning_for_Engineering_Regulated_Statistical_Digital_Twin_Models
- [6] [16] [27] <https://www.sciencedirect.com/science/article/pii/S0952197623018043>
- <https://www.sciencedirect.com/science/article/pii/S0952197623018043>
- [7] [14] <https://arxiv.org/abs/2511.22133>
- <https://arxiv.org/abs/2511.22133>
- [8] [17] <https://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions>
- <https://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions>
- [9] <https://proceedings.neurips.cc/paper/2021/hash/0d441de75945e5acbc865406fc9a2559-Abstract.html>
- <https://proceedings.neurips.cc/paper/2021/hash/0d441de75945e5acbc865406fc9a2559-Abstract.html>
- [10] <https://www.nasa.gov/intelligent-systems-division/discovery-and-systems-health/pcoe/pcoe-data-set-repository/>
- <https://www.nasa.gov/intelligent-systems-division/discovery-and-systems-health/pcoe/pcoe-data-set-repository/>
- [11] <https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-ai-rmf-10>
- <https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-ai-rmf-10>
- [13] <https://www.nist.gov/digital-twins/definitions-and-state-art>
- <https://www.nist.gov/digital-twins/definitions-and-state-art>
- [15] <https://arxiv.org/abs/2509.25730>

<https://arxiv.org/abs/2509.25730>

[18] <https://arxiv.org/abs/2012.00073>

<https://arxiv.org/abs/2012.00073>

[19] <https://www.sciencedirect.com/science/article/abs/pii/S0951832024004551>

<https://www.sciencedirect.com/science/article/abs/pii/S0951832024004551>

[20]

https://www.researchgate.net/publication/386553366_Remaining_Useful_Life_Prediction_with_Uncertainty_Quantification_Using_Evidential_Deep_Learning

https://www.researchgate.net/publication/386553366_Remaining_Useful_Life_Prediction_with_Uncertainty_Quantification_Using_Evidential_Deep_Learning

[21] <https://www.sciencedirect.com/science/article/pii/S0951832025003783>

<https://www.sciencedirect.com/science/article/pii/S0951832025003783>

[22] <https://ieeexplore.ieee.org/document/9635683/>

<https://ieeexplore.ieee.org/document/9635683/>

[24] <https://www.iso.org/standard/87426.html>

<https://www.iso.org/standard/87426.html>

[25] <https://www.nist.gov/itl/ai-risk-management-framework/ai-risk-management-framework-faqs>

<https://www.nist.gov/itl/ai-risk-management-framework/ai-risk-management-framework-faqs>