Academic Journal of Applied Sciences
ISSN: 3105-0549 (Print), ISSN: 3105-0557 (Online) | Vol. 3, No. 1, 2026

Multi-Horizon Machine Learning for Bridge Deterioration Prediction and Risk-Based Maintenance Decision-Making

-- Coherence-Constrained Forecasting, Trajectory-Level Conformal Uncertainty, and Budget-Constrained Scheduling

Gabriel Stanton, Elena Moretti, Maya Whitaker *

Department of Computer Science, University of Arizona, Tucson, AZ, USA

Abstract: Bridge management agencies must decide not only which structures to treat, but when to treat them over a multi-year programme. Most machine-learning pipelines for bridge condition, however, forecast a single fixed horizon, report only marginal uncertainty, and feed a top-k selection rule that ignores treatment cost and timing. This paper proposes an integrated multi-horizon framework. (i) A horizon-conditioned learner (HCM) is trained on horizon-stacked data with a monotone constraint on the horizon variable, which guarantees coherent deterioration trajectories and, as a by-product, a valid discrete-time hazard and remaining-useful-life estimate. (ii) Predictive uncertainty is certified at the trajectory level by a locally adaptive max-norm conformal band, and calibrated first-passage risks are kept monotone by a coherence-preserving isotonic projection. (iii) A multi-horizon scheduling policy (MHSP) allocates a per-period budget by maximising discounted, consequence-weighted risk averted, solved as a linear-programming relaxation of a multiple-choice knapsack. On a 11,963-bridge NBI-schema longitudinal benchmark with explicit intervention counterfactuals, the HCM eliminates trajectory incoherence entirely (0.0% of bridges, versus 32.2–53.5% for independent per-horizon models and 43.8% for a recursive model) while also being the most accurate learner at every horizon (ROC-AUC 0.952 to 0.903 for h = 2 to 10 years). Per-horizon 90% conformal bands attain only 73.6% joint trajectory coverage; the proposed adaptive max-norm band reaches 89.9% joint coverage while being 9.9% narrower than a Bonferroni band. At a 15% programme budget, MHSP averts 51.9% of consequence-weighted Poor bridge-years, against 33.1% for a single-horizon risk×consequence index, 22.8% for condition-first and 10.6% for age-first triage, reaching 79.9% of a perfect-foresight upper bound. Code and benchmark are fully seeded and reproducible.

Keywords: Bridge management systems, multi-horizon forecasting, gradient boosting, conformal prediction, monotone constraints, probability calibration, remaining useful life, maintenance scheduling, risk-based decision making.

1. Introduction

The United States maintains more than 617,000 highway bridges. The 2021 ASCE Infrastructure Report Card graded the stock a "C", noting that 42% of bridges are at least 50 years old, that roughly 7.5% are in Poor condition, and that the repair backlog exceeds $125 billion [1]. Agencies cannot inspect, repair or replace every deficient structure, so scarce funds must be steered to the interventions that avert the most harm. Since the 1970s, biennial inspections have populated the National Bridge Inventory (NBI), in which trained inspectors assign 0–9 condition ratings to a bridge's deck, superstructure and substructure [2]. These ratings drive national performance reporting and, through systems such as Pontis, network-level resource allocation [3].

Classical decision support rests on Markov-chain and hazard-based deterioration models [4–9] embedded in life-cycle reliability formulations [10–12]. A growing body of work instead applies machine learning (ML) to NBI data [13–17], and the recent risk-based framework of Zhang and Qiu [18] makes an important step forward: it couples gradient-boosted deterioration forecasts with split conformal intervals, isotonic calibration and SHAP attribution, and then multiplies a calibrated, uncertainty-inflated risk by a failure-consequence score to obtain a Bridge Maintenance Priority Index (BMPI) that clearly dominates condition-first and age-first triage. Their results establish convincingly that prediction and prioritisation should be treated as one probabilistic problem rather than two.

Building directly on that position, we observe that a deployed bridge programme is intrinsically multi-horizon. An agency does not ask only "will this bridge be Poor in ten years?"; it asks "in which two-year cycle will it cross the Poor threshold, how confident is that whole trajectory, and given an annual budget, in which cycle should it be treated?" Three technical gaps follow.

First, forecasting. A bank of independent per-horizon models—the natural extension of single-horizon practice—produces trajectories that are frequently physically impossible: predicted condition that improves with time under a do-nothing scenario, or a cumulative probability of having entered the Poor state that decreases as the horizon lengthens. Such incoherence is not a cosmetic defect; it invalidates the implied hazard function and therefore any remaining-useful-life (RUL) estimate derived from it.

Second, uncertainty. Split conformal prediction certifies marginal coverage at one horizon. A planner, however, consumes an entire trajectory, and the probability that all horizons are simultaneously covered is far below the nominal level. Trajectory-level (simultaneous) validity has not been reported for bridge deterioration.

Third, decisions. Top-k selection by a priority index implicitly assumes equal treatment cost and immediate action. Real programmes face per-period budgets, heterogeneous

rehabilitation costs and a timing choice: treating a bridge early wastes remaining service life, treating it late concedes Poor bridge-years.

This paper contributes:

1) A horizon-conditioned model (HCM): a single learner trained on horizon-stacked data with a monotone constraint on the horizon variable. It provably yields coherent trajectories, answers horizons absent from the training grid, and induces a valid discrete-time survival function.

2) Trajectory-level conformal bands and coherence-preserving calibration: a locally adaptive max-norm nonconformity score gives simultaneous coverage over the whole 10-year path, and a cumulative-maximum projection restores monotonicity destroyed by per-horizon isotonic regression at no measurable calibration cost.

3) A multi-horizon scheduling policy (MHSP): the programme is formulated as a multiple-choice knapsack over (bridge, treatment period) pairs with per-period budgets, solved by an LP relaxation with greedy repair, and evaluated against ground-truth intervention counterfactuals rather than against predicted quantities.

Because multi-year, agency-curated NBI extracts are rarely redistributable, and because policy evaluation additionally requires counterfactual outcomes that no observational archive contains, all experiments use a seeded NBI-schema longitudinal benchmark whose schema follows the FHWA coding guide [2] and whose deterioration mechanisms and aggregate statistics are calibrated to published figures [1], [7], [10], [14]. The pipeline transfers unchanged to operational inventory data.

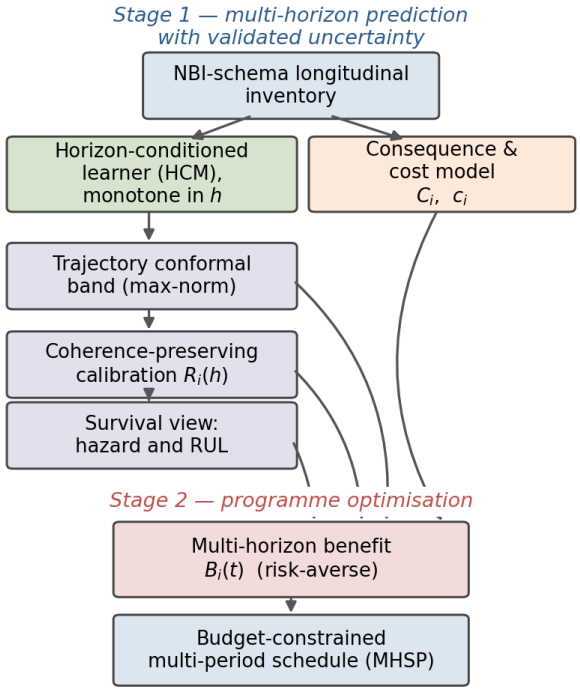

Figure 1. Overview of the proposed framework. Stage 1 produces coherent multi-horizon forecasts equipped with trajectory-level conformal bands, coherence-preserving calibrated risks and a survival/RUL view; Stage 2 turns them into a budget-constrained multi-period repair schedule.

2. Related Work

2.1. Deterioration Modelling and Life-Cycle Management

Markov-chain models dominate classical bridge management, representing condition as discrete states with transition probabilities estimated from historical ratings [4], [6], [7]. Morcous examined the state-independence and stationarity assumptions underlying such models for bridge decks [7], and later work extended them to railway elements and to corrosion-initiation with non-linear propagation [8], [9]. Hazard- and duration-based models relax the discrete-state view; Mauch and Madanat fitted semiparametric hazard-rate models of reinforced-concrete deck deterioration [5]. These models are interpretable and well suited to network simulation, but they typically use few covariates and assume stationary transitions. In parallel, reliability-based life-cycle formulations optimise maintenance against lifetime safety and whole-life cost [10–12], and probabilistic scheduling of bridge-network interventions has been studied extensively [11], [12]. Our decision layer inherits this scheduling view but drives it with calibrated, data-driven multi-horizon risks rather than with parametric reliability profiles.

2.2. Machine Learning for Bridge Condition

ML methods learn flexible, high-dimensional relationships between bridge attributes and condition. Lee et al. used artificial neural networks for backward prediction in a bridge management system [13]; Kim and Yoon identified environmental drivers of deterioration in cold regions [14]; and Bektas et al. applied classification trees to NBI ratings [15]. Gradient-boosted trees are especially effective on tabular inventory data: Lim and Chi applied XGBoost with SHAP to bridge damage estimation [16], and Assaad and El-adaway compared several learners for deck deterioration [17]. Zhang and Qiu [18] provide the most complete integration to date, unifying prediction, distribution-free uncertainty, calibration and interpretability with a consequence-aware priority index, and demonstrating that the resulting ranking averts substantially more consequence-weighted deterioration than condition- or age-based heuristics at equal budget. That work explicitly identifies multi-horizon formulations, survival analysis for remaining-useful-life and explicit budget-constrained optimisation as natural extensions; the present paper develops precisely those three directions and quantifies what each is worth.

2.3. Multi-Step Forecasting, Conformal Prediction and Calibration

The forecasting literature distinguishes recursive (iterated) from direct (per-horizon) strategies, and documents the bias–variance trade-off between them; large-scale comparisons find direct strategies more robust when the horizon is long, at the price of losing coherence across horizons [19]. Modern global forecasters instead condition a single model on the horizon [20], [21]—an idea we transfer to inventory-level deterioration and strengthen with a monotone constraint that encodes cumulative damage.

Conformal prediction produces prediction sets with finite-sample, distribution-free coverage under exchangeability [22], [23]; inductive (split) conformal and its regression analysis [22], [24], together with normalised and quantile-based variants [25], make it practical for modern learners. Extensions to sequential and multi-step settings [26], to distribution shift [27] and to covariate shift [28] are active areas. Separately, probability calibration corrects over- or under-confident scores; Platt scaling [29], isotonic regression and their evaluation are standard [30], [31]. For interpretability, SHAP attributes model output to inputs with

game-theoretic consistency [32], with an exact polynomial-time algorithm for trees [33]. Discrete-time survival and random survival forests provide the RUL vocabulary we adopt [34]. To our knowledge these tools have not previously been combined to certify a whole bridge deterioration trajectory or to schedule a multi-year programme.

3. Benchmark and Problem Formulation

3.1. NBI Schema and Prediction Targets

The NBI records, for every public highway bridge, structural attributes and 0–9 condition ratings for the deck (Item 58), superstructure (Item 59) and substructure (Item 60); ratings of 4 or below denote Poor, and FHWA classifies a bridge as Poor if any of the three component ratings is at most 4 [2]. Bridges are re-inspected on a nominal 24-month cycle.

From each bridge's record at the analysis year t0 we address two tasks at each horizon $h \in \{2, 4, 6, 8, 10\}$ years: (a) regression of the future deck condition rating, and (b) classification of the first-passage event "the bridge has entered Poor condition by t0 + h". The first-passage definition is the quantity maintenance planning actually needs (when will intervention become necessary), it is monotone in h by construction, and it connects the classification head directly to a survival function.

3.2. A Reproducible Longitudinal Benchmark with Counterfactuals

We construct a benchmark of 12,000 bridges whose feature schema mirrors the FHWA coding guide (Table I) and whose deterioration dynamics reproduce documented behaviour. Static attributes are drawn from realistic distributions: year built (a mixture reproducing the interstate-era and 1980s construction booms), material and design family, spans, structure length, deck width, skew, functional class, average daily traffic (ADT), truck percentage, deck protection and detour length. Each bridge is aged from construction with a component-specific accumulated-damage process,

$$C_k(a) = 9 - r_k \cdot a_{pk} \cdot \exp(\Sigma_j \beta_{kj} z_j) \qquad (1)$$

Where C_k is the latent condition of component k at effective age a, the parameters r_k and $p_k > 1$ set the base rate and the convex acceleration of damage, and the covariates z_j modulate the rate exponentially. Coefficient signs and magnitudes follow established engineering knowledge: decks deteriorate fastest; truck loading and freeze–thaw with de-icing salt or marine chloride exposure accelerate deterioration; prestressed concrete is durable while steel corrodes and timber decays; and deck protection slows deck deterioration [5], [9], [14]. Stochastic rehabilitation events, more likely as condition worsens, reset the effective age of affected components and produce the saw-tooth trajectories seen in real inspection histories. Latent conditions are perturbed by inspection noise and rounded to integers on the 0–9 scale, and observations are sampled biennially. A persistent, unobserved workmanship term is drawn per bridge so that history does not fully reveal the deterioration rate.

Two features of the generator are essential for this study. First, the future is simulated under a do-nothing policy, which is the quantity a bridge-management deterioration model is required to forecast, with a random walk in the deterioration rate and occasional unrecorded minor maintenance so that long horizons are genuinely uncertain. Second, the latent randomness is drawn once and reused across scenarios, so that for every bridge we obtain the do-nothing trajectory and the trajectory that would follow rehabilitation in each candidate period. These are valid potential outcomes and permit exact, non-fabricated policy evaluation—something no observational inventory can supply.

Table I. NBI-Schema Feature Groups in the Benchmark

Group	Representative features (NBI item)
Structural	material/design (43), spans (45), length (49), deck width (52), skew
Traffic	ADT (29), truck % / ADTT (109), functional class (26)
Environment	climate/exposure region, freeze–thaw, chloride exposure
Maintenance	deck protection (108), reconstruction flag and age (106)
Condition	deck/super/sub ratings at t0, t0−2, t0−4 (58/59/60), trend
Consequence	ADT exposure, criticality, detour length, deck area, cost

The benchmark reproduces salient features of the real inventory (Fig. 2): mean bridge age at the analysis year is 44.8 years against a roughly 44-year national average [1]; deck ratings concentrate at 6–7 (Satisfactory–Good) as in the NBI, with 41.5% of bridges in that band and a median of 7; 14.3% of bridges are Poor at t0, consistent with the elevated deficiency of aging cohorts (the national figure is about 7.5% across all ages [1]); and mean condition declines with age fastest under freeze–thaw-plus-salt exposure, the ordering expected from chloride-induced corrosion. Poor prevalence grows from 15.4% at h = 2 to 37.0% at h = 10. Its purpose is a realistic, shareable testbed, not a claim about any specific bridge population.

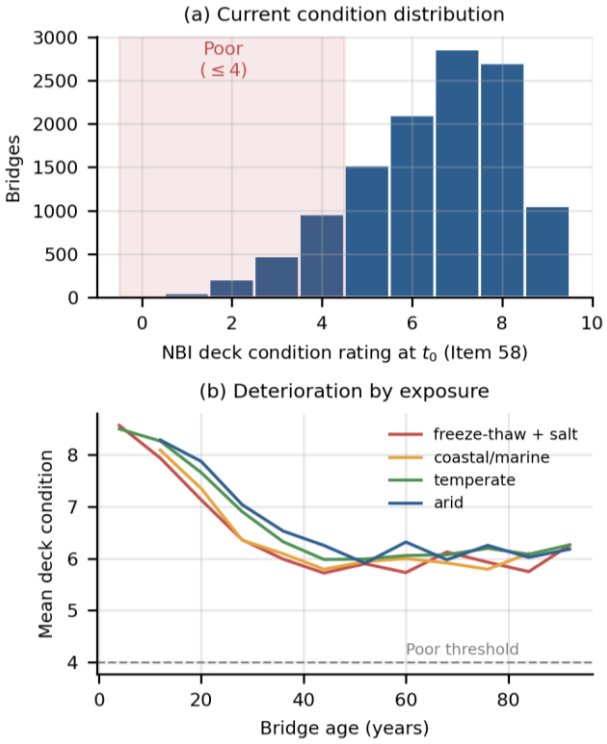

Figure 2. Benchmark realism: (a) the deck-rating distribution concentrates at 6–7 with a Poor tail; (b) mean condition declines with age, fastest under freeze–thaw-plus-salt exposure.

3.3. Protocol

Predictors comprise all attributes available at t0—structural, environmental, traffic, current and two lagged component ratings, short-term condition trend, and maintenance history—so no future information is used. After removing bridges with incomplete inspection histories, 11,963 bridges remain, split 60/20/20 into training, calibration and test sets, stratified on the h = 10 Poor label. The calibration split is reserved for conformal prediction, probability calibration and the difficulty model, and is never used for model fitting. Consequence and cost enter only the decision layer and are excluded from the predictors.

4. Methodology

4.1. Horizon-Conditioned Learner with a Coherence Constraint

Let x denote the feature vector observed at the analysis year and h the forecast horizon. Rather than fitting one model per horizon, we stack the training data over horizons, append h as an explicit feature, and fit a single gradient-boosted ensemble f(x, h). Cumulative damage under a do-nothing policy implies that the expected condition is non-increasing in h and the first-passage risk is non-decreasing in h. We impose these directly as monotone constraints on the horizon feature (sign −1 for the regression head, +1 for the classification head), which tree ensembles support exactly. Two consequences follow.

Coherence. For any x, the predicted path over h is monotone by construction, so the induced discrete-time hazard is non-negative and the implied survival function is valid. Writing R(h) for the calibrated first-passage risk and S(h) = 1 − R(h) for the survival function, the hazard on the biennial grid is $\lambda(h) = 1 - S(h)/S(h-2)$, and the restricted mean time to Poor over a 12-year window is $2[1 + \Sigma\, S(h)]$.

Horizon generalisation. Because h is an input rather than an index over models, the HCM can be queried at horizons absent from the training grid, whereas a per-horizon bank must be retrained.

We compare the HCM against (i) Direct-PH, an independent model per horizon (ridge/logistic, random forest, histogram gradient boosting, XGBoost), and (ii) Recursive, a single two-year-step model for the three components rolled forward with its own predictions, the classical iterated strategy [19]. Numeric features are standardised and categoricals one-hot encoded; classifiers use the logistic objective and all seeds are fixed.

4.2. Trajectory-Level Conformal Prediction

Split conformal prediction on the calibration set of size n yields, for each horizon h, a radius q(h) equal to the empirical quantile of the absolute residual at level $\lceil (n+1)(1-\alpha) \rceil / n$, and the band $f(x, h) \pm q(h)$ covers the truth with probability at least $1-\alpha$ marginally [22], [24]. The event a planner cares about is the intersection over horizons, which these bands do not control. We evaluate four constructions at the 90% level:

(a) naive per-horizon bands; (b) a Bonferroni band using level α/H per horizon, valid but conservative; (c) a max-norm simultaneous band whose nonconformity score is the largest scale-standardised absolute residual over the five horizons, the scale at each horizon being its mean calibration residual, so that horizons are placed on a common footing before the maximum is taken; and (d) a locally adaptive variant that replaces this constant scale by a difficulty model fitted on one half of the calibration split to predict the absolute residual from the features and the horizon, with the conformal quantile taken on the other half. Constructions (c) and (d) inherit exact finite-sample joint coverage under exchangeability because a single scalar score is conformalised.

4.3. Coherence-Preserving Calibration

Tree ensembles can produce distorted class probabilities [30]. We fit isotonic regression per horizon on the calibration split. Because each horizon is calibrated independently, the monotonicity of the risk path is not preserved; we therefore apply a cumulative-maximum projection that replaces the calibrated risk at horizon h by the running maximum of the calibrated risks up to h. This is the isotonic projection of the calibrated path onto the monotone cone under the sup-norm, so it moves probabilities by the minimum amount required to restore coherence.

4.4. Multi-Horizon Scheduling

The consequence of failure is a normalised score combining traffic exposure (log ADT), network criticality (functional class), detour penalty and asset value (log deck area), following the structure used in risk-based prioritisation [18]:

$$C_i = 0.40\, e_i + 0.30\, n_i + 0.15\, d_i + 0.15\, a_i \qquad (2)$$

The weights are illustrative and are policy inputs. Rehabilitation cost scales with deck area and a material-complexity factor. To be risk-averse we inflate the calibrated risk by the normalised conformal half-width,

$$\mathrm{RiUCB}(h) = \min\{1, R_i(h) + \lambda \cdot u_i(h) \cdot [1 - R_i(h)]\} \qquad (3)$$

With $\lambda \geq 0$. The benefit of treating bridge i in period t discounts the residual risk it removes,

$$B_i(t) = C_i \cdot \Sigma_{h \geq t}\, \delta^h\, [\mathrm{RiUCB}(h) - R_{\mathrm{post}}] \qquad (4)$$

With annual discount $\delta = 0.97$ and a small residual post-rehabilitation risk of 0.02 (in the benchmark, treated bridges are Poor in 0.06% of subsequent periods, so this is conservative). The programme then solves

$$\max \Sigma_i \Sigma_t\, B_i(t)\, x_{it} \quad \text{s.t.} \quad \Sigma_i\, c_i\, x_{it} \leq B_t\ \forall t, \quad \Sigma_t\, x_{it} \leq 1\ \forall I \qquad (5)$$

A multiple-choice knapsack that we solve by LP relaxation (HiGHS) followed by greedy repair. Baselines all consume the same per-period budgets: age-first, condition-first (worst current minimum rating), risk-only, RUL-first (shortest predicted restricted mean time to Poor), the single-horizon risk×consequence index of [18], a cost-effectiveness version of that index, a multi-horizon benefit/cost greedy, and an ablation with $\lambda = 0$. A perfect-foresight oracle that optimises (5) using the true counterfactual benefits provides an upper bound.

Evaluation uses ground truth, not model output: for an assignment we compute the discounted, consequence-weighted number of Poor bridge-years removed relative to the do-nothing trajectory, expressed as a percentage of the total consequence-weighted Poor bridge-years in the network.

5. Experiments and Results

Models are implemented with scikit-learn [35], XGBoost [36] and LightGBM-class histogram boosting [37], [38-40]; SHAP values use TreeSHAP [32], [33]. All randomness is seeded. Conformal prediction, calibration and the difficulty

model use only the calibration split.

5.1. Multi-Horizon Accuracy

Table II reports test-set performance. Accuracy degrades smoothly with the horizon, as it must: the first-passage base rate rises from 15.4% to 37.0% and ten-year-ahead ratings are intrinsically uncertain. The HCM is the strongest learner at every horizon on both tasks, reaching ROC-AUC 0.952 at h = 2 and 0.903 at h = 10 and the lowest Brier score throughout, while cutting MAE relative to the best per-horizon model at h ≥ 4 (1.379 versus 1.394 rating points at h = 10). The gains are small but systematic, and they come for free: one model replaces five. The recursive strategy degrades sharply with the horizon (AUC 0.900 to 0.721; MAE 1.175 to 2.333), the classical accumulation of one-step error [19], confirming that direct-style conditioning is the right base strategy for decade-scale inventory forecasting.

Removing the monotone constraint (HCM-nc) changes accuracy negligibly (AUC 0.901 versus 0.903 at h = 10), so coherence is obtained essentially at zero predictive cost.

Table II. Multi-Horizon Test-Set Performance (XGBoost Base Learner)

Task / strategy	h=2	h=4	h=6	h=8	h=10
MAE, future deck rating					
Ridge (linear)	0.895	1.060	1.205	1.362	1.486
Random forest	0.771	0.951	1.102	1.267	1.412
Direct-PH	0.760	0.930	1.100	1.253	1.394
Recursive	1.175	1.470	1.702	2.049	2.333
HCM (proposed)	0.776	0.923	1.080	1.239	1.379
ROC-AUC, first passage to Poor					
Logistic	0.939	0.926	0.908	0.896	0.881
Random forest	0.950	0.945	0.929	0.916	0.899
Direct-PH	0.947	0.941	0.931	0.916	0.901
Recursive	0.900	0.853	0.822	0.768	0.721
HCM (proposed)	0.952	0.945	0.933	0.919	0.903
Brier score, HCM	0.063	0.078	0.091	0.105	0.119
Poor base rate (%)	15.4	22.0	27.2	32.1	37.0

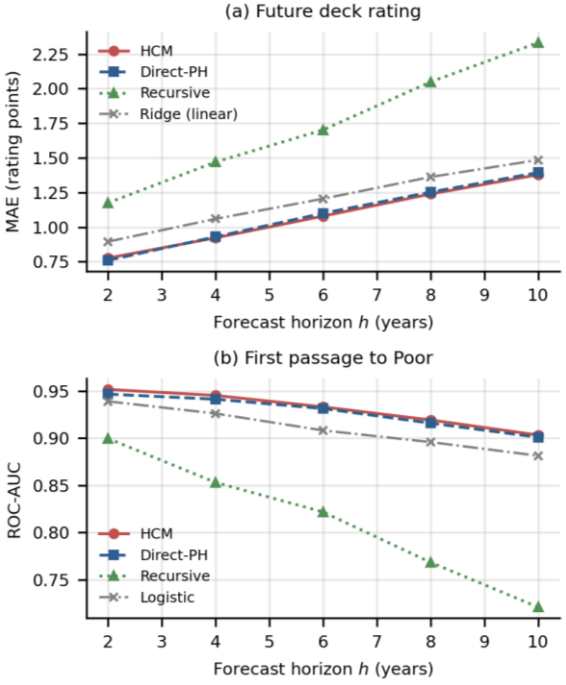

Figure 3. Accuracy versus forecast horizon: (a) mean absolute error of the future deck rating; (b) ROC-AUC of the first-passage classification. The horizon-conditioned model leads at every horizon; the recursive strategy degrades sharply.

5.2. Trajectory Coherence and Horizon Generalisation

Fig. 4(a) quantifies the failure that motivates the HCM. Under a bank of independent per-horizon classifiers, the predicted probability of having entered the Poor state decreases somewhere along the path for 32.2% (random forest), 44.8% (XGBoost) and 53.5% (histogram boosting) of test bridges; the recursive model violates coherence for 43.8%. On the regression side the corresponding rates reach 82.9%. The HCM violates coherence for 0.0% of bridges by construction, and even the unconstrained variant is already close (0.7%), which shows that the constraint is not fighting the data—it is removing residual estimation noise that would otherwise be handed to a decision maker as a physically impossible forecast. Fig. 4(b) shows representative paths.

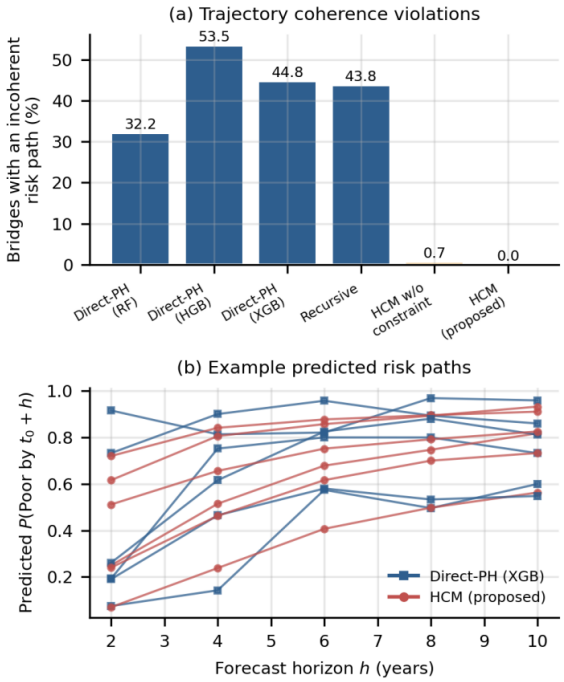

Figure 4. Trajectory coherence: (a) share of test bridges whose predicted first-passage risk path is non-monotone; (b) representative paths for bridges where the per-horizon bank is incoherent.

The practical consequence appears in the survival view. Converting each risk path to a discrete-time hazard, 44.9% of bridges receive at least one negative hazard under the per-horizon bank—an undefined survival function—against 0.0% under the HCM. Both achieve a concordance index of 0.887 for ranking the true first-passage time, and the HCM's restricted mean time to Poor is nearly unbiased (9.36 versus 9.31 years on average, MAE 1.54 years), so validity is gained without sacrificing discrimination (Fig. 6(b)).

To test horizon generalisation we retrained the HCM on $h \in \{2, 4, 8, 10\}$ and queried it at the unseen horizon $h = 6$. Zero-shot performance essentially matches a model that saw $h = 6$ in training (AUC 0.934 versus 0.933; MAE 1.084 versus 1.080) and matches linear interpolation between the $h = 4$ and $h = 8$ per-horizon models (AUC 0.932), which the HCM achieves without any post-hoc interpolation rule and with a coherence guarantee that interpolation does not provide.

5.3. Trajectory-Level Uncertainty

Table III and Fig. 5 report the central uncertainty result. Naive per-horizon split conformal delivers exactly what it promises marginally—coverage 0.900 to 0.911 across horizons—yet the whole 10-year trajectory is covered for only 73.6% of bridges, far below the 90% a planner would assume. A Bonferroni correction restores joint validity (93.8%) but overshoots and widens the band by 51% (6.97 versus 4.62 rating points). The max-norm simultaneous band attains 91.5% joint coverage at width 6.46, and the locally adaptive variant attains 89.9%—statistically indistinguishable from the nominal level—at width 6.28, i.e. 9.9% narrower than Bonferroni at higher effective efficiency. The ordering is reproduced under the per-horizon model bank (73.6% versus 72.5% naive joint coverage), so it is a property of the coverage target rather than of the learner.

Table III. Conformal Bands: Marginal vs. Trajectory Coverage (Target 90%)

Construction	Mean marginal cov.	Joint trajectory cov.	Mean width
Per-horizon (naive)	0.904	0.736	4.62
Bonferroni	0.982	0.938	6.97
Max-norm	0.973	0.915	6.46
Max-norm + adaptive	0.970	0.899	6.28

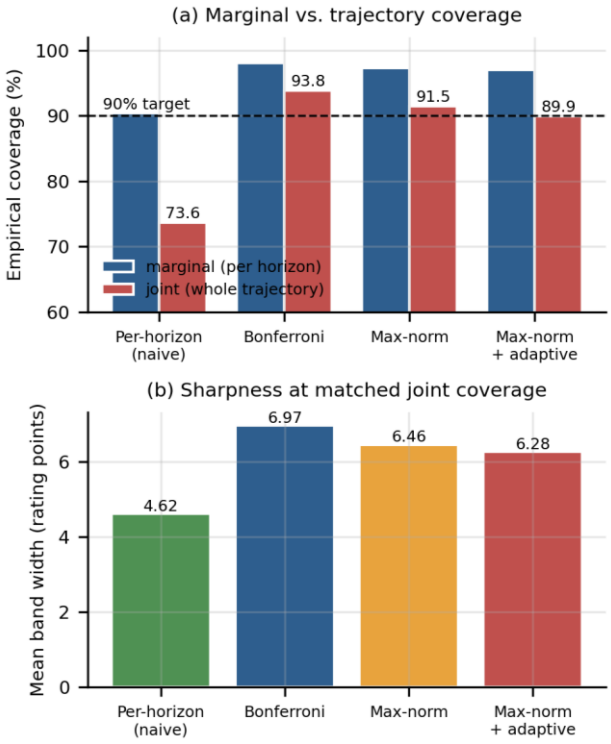

Figure 5. Trajectory-level uncertainty: (a) marginal coverage is on target for every construction, but joint coverage of the whole path is not; (b) band width at matched joint coverage.

5.4. Calibration

Raw HCM probabilities are already close to calibrated (mean ECE 0.020, Brier 0.063–0.119 across horizons), reflecting the logistic objective; isotonic regression improves the intermediate horizons (ECE 0.0204 → 0.0141 at $h = 4$, 0.0214 → 0.0165 at $h = 8$) and is neutral elsewhere. The important effect is structural: per-horizon isotonic calibration destroys trajectory coherence for 13.9% of bridges, reintroducing exactly the pathology the constrained learner removed. The cumulative-maximum projection restores coherence for 100% of bridges while leaving calibration untouched (mean ECE 0.0184 → 0.0182, mean Brier 0.092). Fig. 6(a) shows the resulting reliability curves tracking the diagonal at $h = 2$, 6 and 10.

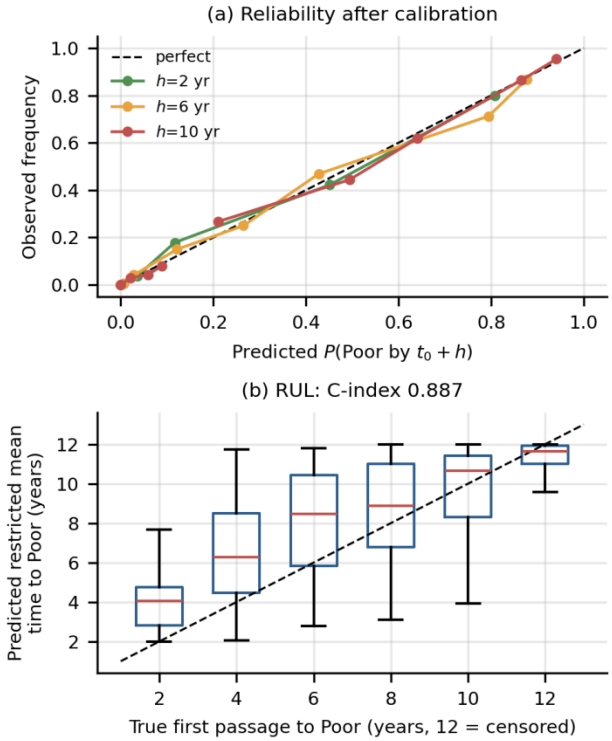

Figure 6. Calibration and remaining useful life: (a) reliability after coherence-preserving isotonic calibration; (b) predicted restricted mean time to Poor against the true first-passage time.

5.5. Horizon-Resolved Interpretability

Because the HCM exposes h as an input, SHAP attributions can be computed as a function of the forecast horizon—an analysis that a per-horizon bank cannot deliver consistently, since each horizon is a different model. Fig. 7(a) confirms the expected global ordering at h = 10: the current minimum component rating, years since reconstruction and the current and lagged deck ratings dominate, followed by age, deck protection, the other component ratings and traffic loading. Fig. 7(b) reveals the horizon effect. Relative to h = 2, the attribution share of present-state variables falls as the horizon lengthens (lagged deck rating ×0.78, minimum component rating ×0.89) while durability and exposure variables rise (age ×1.24, number of reconstructions ×1.37, freeze–thaw-plus-salt exposure ×1.48). This is the physically expected transition from a state-dominated short-term regime to a rate-dominated long-term regime, and it is direct evidence that the model has learned deterioration mechanisms rather than spurious correlations. Operationally it also tells an agency that short-horizon triage can rely on inspection records, whereas ten-year programming must weight environment and maintenance history.

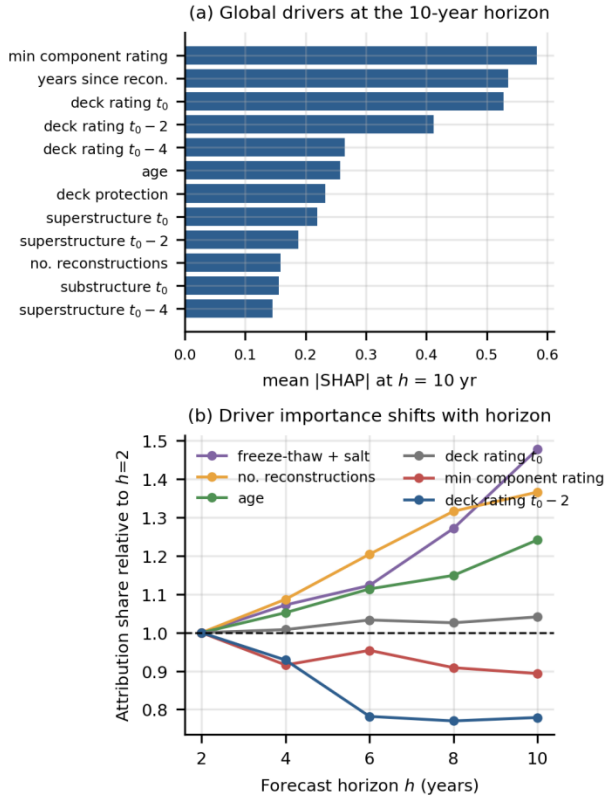

Figure 7. Horizon-resolved SHAP attribution: (a) global drivers at the ten-year horizon; (b) attribution share relative to h = 2, showing the shift from state-dominated to rate-dominated drivers.

5.6. Maintenance Programme Effectiveness

Table IV and Fig. 8 report the central decision result on ground-truth counterfactuals. At a 15% programme budget, MHSP averts 51.9% of the consequence-weighted Poor bridge-years, against 33.1% for the single-horizon risk×consequence index in the spirit of [18], 28.4% for RUL-first, 27.4% for risk-only, 22.8% for condition-first and 10.6% for age-first triage—the two heuristics closest to current practice. The advantage holds across the entire 5–30% budget range, and MHSP reaches 79.9% of the perfect-foresight upper bound (65.0%).

The decomposition matters more than the headline. Making the single-horizon index cost-aware already lifts it from 33.1% to 50.2%: much of the benefit of an optimisation view comes simply from spending the budget on cost-effective interventions rather than on the highest-priority ones, since a fixed budget buys 770 treatments under a cost-aware rule versus 337 under a pure ranking. Adding multi-horizon timing on top contributes a further consistent gain (51.9% versus 50.2% at 15%, and 71.0% versus 68.1% at 30%), and the LP schedule improves on a greedy benefit/cost heuristic at every budget (51.9% versus 51.3% at 15%). The schedule is genuinely spread over the programme—219, 160, 149, 141 and 144 bridges in the five biennial periods at a 15% budget—confirming that the optimiser is exploiting timing, deferring bridges whose risk accrues late in order to release early budget for bridges that would otherwise fail sooner. Notably, MHSP and the cost-aware single-horizon rule detect a similar share of eventually-Poor bridges (57.2% each) yet avert markedly different consequence, which reiterates the point that detecting Poor bridges is not the same as protecting a network.

Table IV. Programme Effectiveness: Consequence-Weighted Poor Bridge-Years Averted (%)

Policy	5%	10%	15%	20%	30%
Age-first	3.4	6.7	10.6	14.6	23.4
Condition-first	6.1	14.1	22.8	30.9	46.7
Risk-only (h = 10)	8.7	17.8	27.4	37.1	53.9
RUL-first	9.4	18.7	28.4	37.7	54.7
BMPI, single horizon	11.8	23.1	33.1	41.4	55.5
BMPI/cost, single horizon	27.2	40.6	50.2	57.2	68.1
MH-greedy (benefit/cost)	28.0	41.7	51.3	59.1	70.0
MHSP, λ = 0 (ablation)	28.0	42.2	52.0	59.8	71.1
MHSP (proposed)	28.2	42.3	51.9	59.8	71.0
Oracle (perfect foresight)	35.7	52.9	65.0	74.3	87.4

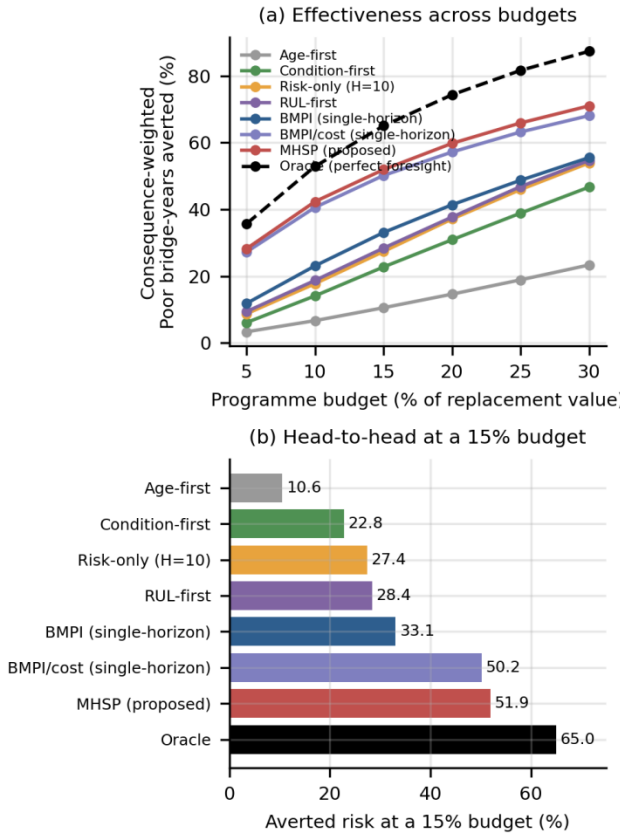

Figure 8. Maintenance programme effectiveness on ground-truth counterfactuals: (a) consequence-weighted Poor bridge-years averted versus programme budget; (b) head-to-head comparison at a 15% budget.

5.7. Risk Aversion and Distribution Shift

Sweeping the risk-aversion weight λ shows that mild inflation is neutral-to-beneficial and strong inflation is harmful: on the main test set, 51.96% (λ = 0), 51.93% (λ = 0.25), 51.04% (λ = 0.5) and 48.63% (λ = 1). This refines the small consistent gain reported for uncertainty inflation in single-horizon prioritisation [18]: once the risk path is conformally validated and isotonically calibrated, expected-benefit maximisation is already close to optimal, so the value of uncertainty quantification in a calibrated pipeline lies chiefly in defensible interval reporting rather than in re-ranking.

To probe the exchangeability assumption we trained and calibrated only on temperate and arid bridges and deployed on freeze–thaw-plus-salt and coastal bridges, a regional-transfer scenario (Fig. 9). Discrimination is essentially preserved (AUC 0.889 versus 0.891 in domain), while calibration and conformal validity degrade gracefully: mean ECE 0.0279 → 0.0302, Brier 0.1253 → 0.1353, joint trajectory coverage 0.917 → 0.904. Under shift the optimal λ moves away from zero (49.40% at λ = 0 versus 49.63% at λ = 0.25), and the in-domain optimum is likewise λ ≈ 0.25 (54.86% versus 54.35%). We therefore recommend λ = 0.25 as a default: it costs nothing in the exchangeable regime and hedges mildly when the model is transported.

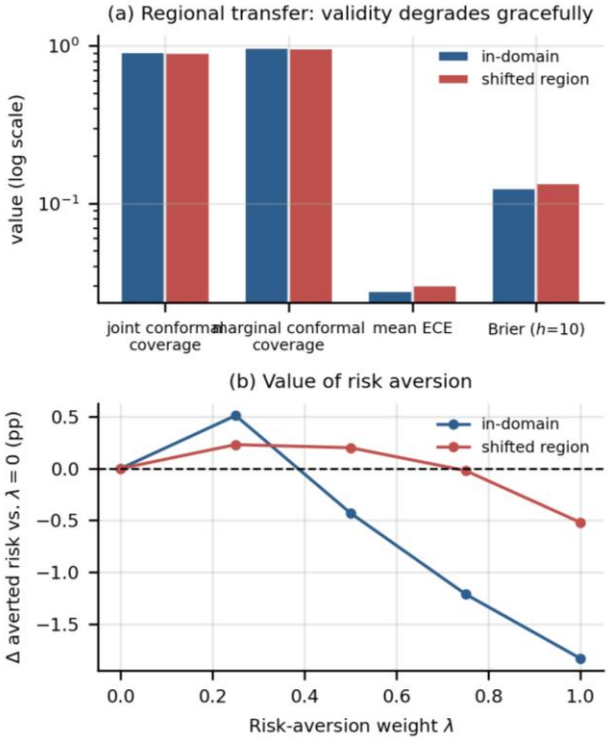

Figure 9. Regional transfer stress test: (a) calibration and conformal validity degrade gracefully when the model is trained on temperate/arid bridges and deployed on freeze–salt/coastal bridges; (b) the value of the risk-aversion weight λ.

6. Discussion and Limitations

The framework gives agencies a transparent alternative to worst-first triage whose factors—predicted risk at each horizon, certified trajectory confidence, failure consequence and treatment cost—map onto concerns they already weigh informally. Three findings are directly actionable. Coherence should be imposed, not hoped for: it costs nothing in accuracy and is a precondition for any survival or RUL statement. Uncertainty should be reported at the trajectory level, because per-horizon bands overstate confidence in the object planners actually consume by roughly sixteen percentage points. And budget realism dominates ranking sophistication: cost-

awareness and timing together roughly double the consequence averted relative to a well-designed but cost-blind priority index.

Several limitations qualify these findings. The experiments use a synthetic, if carefully grounded, benchmark; its schema and mechanisms match the NBI and its aggregate statistics align with national figures [1], but absolute performance on a specific agency's data will differ and models should be re-fit and re-calibrated on operational inventories, which the pipeline supports directly. Counterfactual evaluation is only possible because the generator supplies potential outcomes; on real data the decision layer would have to be assessed by off-policy estimation or by prospective trial. The rehabilitation model is a single "as-new" action, whereas agencies choose among treatments of differing cost and effect; extending (5) to a treatment menu is a straightforward multiple-choice generalisation. Consequence weights are illustrative and should encode agency policy, possibly elicited by multi-criteria methods. Finally, conformal validity rests on exchangeability, which distribution shift over long horizons can violate; our transfer study shows graceful degradation, and adaptive conformal methods [27], [28] are a promising remedy.

7. Conclusion

We presented a multi-horizon framework that unifies coherent bridge deterioration forecasting, trajectory-level uncertainty quantification and budget-constrained maintenance scheduling. A horizon-conditioned gradient-boosted model with a monotone horizon constraint is the most accurate learner at every horizon (ROC-AUC 0.952–0.903 for two- to ten-year horizons) while eliminating trajectory incoherence entirely, where independent per-horizon models produce physically impossible paths for 32–54% of bridges and imply negative hazards for 44.9%. A locally adaptive max-norm conformal band delivers 89.9% simultaneous coverage of the whole ten-year trajectory—against 73.6% for nominally 90% per-horizon bands—while being 9.9% narrower than a Bonferroni construction, and a cumulative-maximum projection restores the coherence that per-horizon isotonic calibration destroys for 13.9% of bridges at no calibration cost. The resulting multi-horizon scheduling policy averts 51.9% of consequence-weighted Poor bridge-years at a 15% budget, far exceeding single-horizon risk×consequence prioritisation (33.1%), condition-first (22.8%) and age-first (10.6%) practice, and attaining 79.9% of a perfect-foresight bound. The results argue for treating deterioration prediction and maintenance programming as one coherent, uncertainty-aware, budget-constrained decision problem, and the pipeline transfers directly to operational National Bridge Inventory data.

References

[1] Zhang, S., & Qiu, L. (2024). Data-driven assessment of pavement performance and resilience under extreme weather using the Long-Term Pavement Performance Program. AI and Data Science Journal, 1(1), 57–66.

[2] Zhang, H. (2024). Graph-aware multi-task learning for early prediction of timing violations and routing congestion in digital IC physical design. World Journal of Information Technology, 2(1), 13–20.

[3] Wang, J., Tan, Y., Jiang, B., Wu, B., & Liu, W. (2025). Dynamic marketing uplift modeling: A symmetry-preserving framework integrating causal forests with deep reinforcement learning for personalized intervention strategies. Symmetry, 17(4), 610.

[4] Zhao, X., Liu, J., Wang, Y., & Wang, J. (2026). CryptoMamba-SSM: Linear complexity state space models for cryptocurrency volatility prediction. IEEE Open Journal of the Computer Society, 7, 226–243.

[5] Wang, J., Liu, J., Zheng, W., & Ge, Y. (2025). Temporal heterogeneous graph contrastive learning for fraud detection in credit card transactions. IEEE Access.

[6] Sun, T., Yang, J., Li, J., Chen, J., Liu, M., Fan, L., & Wang, X. (2024). Enhancing auto insurance risk evaluation with transformer and SHAP. IEEE Access, 12, 116546–116557.

[7] Zhang, H. (2025). Reinforcement learning approaches for layout optimization in electronic design automation with electromagnetic compatibility constraints. Frontiers in Robotics and Automation, 2(2), 77–93.

[8] Han, X., Yang, Y., Chen, J., Wang, M., & Zhou, M. (2025). Symmetry-aware credit risk modeling: A deep learning framework exploiting financial data balance and invariance. Symmetry, 17(3), 341.

[9] Chen, J., Cui, Y., Zhang, X., Yang, J., & Zhou, M. (2024). Temporal convolutional network for carbon tax projection: A data-driven approach. Applied Sciences, 14(20), 9213.

[10] Sun, T., Wang, M., & Chen, J. (2025). Leveraging machine learning for tax fraud detection and risk scoring in corporate filings. Asian Business Research Journal, 10(11), 1–13.

[11] Chen, J., Wang, M., & Sun, T. (2025). Intelligent tax systems and the role of natural language processing in regulatory interpretation. American Journal of Machine Learning, 6(4), 74–94.

[12] Chen, Z., Liu, J., & Chen, J. (2025). Machine learning methods for financial forecasting in enterprise planning: Transitioning from rule-based models to predictive analytics. Frontiers in Artificial Intelligence Research, 2(3), 541–564.

[13] Yue, X., Yang, J., & Liu, W. (2026). FraudDebate-Agent: A multi-agent LLM framework with an evidence-based debate mechanism for financial statement fraud detection. Mathematics, 14(15), 2695.

[14] Yue, X., & Zhu, R. (2024). An explainable hybrid machine learning framework for cash-flow-conditioned multi-horizon liquidity risk early warning: An SME-oriented benchmark study. World Journal of Management Science, 2(1), 63–72.

[15] Zi, B. (2024). Cloud-native distributed systems for real-time payment intelligence. AI and Data Science Journal, 1(1), 51–56.

[16] Zhang, S., & Qiu, L. (2023). An interpretable, uncertainty-aware framework for bridge deterioration prediction and risk-based maintenance prioritization. Academic Journal of Architecture and Civil Engineering, 1(2), 21–28.

[17] Jin, J., Xing, S., Ji, E., & Liu, W. (2025). Xgate: Explainable reinforcement learning for transparent and trustworthy API traffic management in IoT sensor networks. Sensors, 25(7), 2183.

[18] Xing, S., Wang, Y., & Liu, W. (2025). Multi-dimensional anomaly detection and fault localization in microservice architectures: A dual-channel deep learning approach with causal inference for intelligent sensing. Sensors, 25(11), 3396.

[19] Xing, S., & Wang, Y. (2025). Proactive data placement in heterogeneous storage systems via predictive multi-objective reinforcement learning. IEEE Access, 13, 117986–117998.

[20] Ji, E., Wang, Y., Xing, S., & Jin, J. (2025). Hierarchical reinforcement learning for energy-efficient API traffic optimization in large-scale advertising systems. IEEE Access.

[21] Xing, S., & Wang, Y. (2025). Cross-modal attention networks for multi-modal anomaly detection in system software. IEEE Open Journal of the Computer Society.

[22] Liu, Y., Ren, S., Wang, X., & Zhou, M. (2024). Temporal logical attention network for log-based anomaly detection in distributed systems. Sensors, 24(24), 7949.

[23] Ren, S., Jin, J., Niu, G., & Liu, Y. (2025). ARCS: Adaptive reinforcement learning framework for automated cybersecurity incident response strategy optimization. Applied Sciences, 15(2), 951.

[24] Li, P., Ren, S., Zhang, Q., Wang, X., & Liu, Y. (2024). Think4SCND: Reinforcement learning with thinking model for dynamic supply chain network design. IEEE Access, 12, 195974–195985.

[25] Hu, X., Zhao, X., Wang, J., & Yang, Y. (2025). Information-theoretic multi-scale geometric pre-training for enhanced molecular property prediction. PLOS ONE, 20(10), e0332640.

[26] Zhao, X., Sun, T., Ren, S., Yang, J., & Liu, Y. (2025). RAG-Based AI Agents for Enterprise Software Development: Implementation Patterns and Production Deployment. Frontiers in Artificial Intelligence Research, 2(3), 501–520.

[27] Sun, T., Wang, M., & Chen, J. (2025). Leveraging machine learning for tax fraud detection and risk scoring in corporate filings. Asian Business Research Journal, 10(11), 1–13.

[28] Bian, M., Li, P., Lin, Y., Wang, Y., & Teng, D. (2026). Assessing green public supply chains resilience under carbon neutrality goals: A multi-agent reinforcement learning framework. IEEE Access.

[29] Shang, W., Teng, D., Chen, T., Zhang, F., & Ding, J. (2026). Neuro-Elastic: A unified framework for hardware-aware adaptive quantization and dynamic sparsity in real-time edge intent prediction. IEEE Access.

[30] Teng, D. (2025). TEAS: Token-and energy-aware autoscaling for cost-efficient LLM serving. AI and Data Science Journal, 6(3), 1–10.

[31] Wang, B., Zhao, W., & Wang, Z. (2024). DAS-GNN: A scalable disagreement-aware graph neural framework for anomaly detection in large-scale networks. AI and Data Science Journal, 1(1), 67–75.

[32] Wang, Z., & Shen, Z. (2024). Flex-Talent: An explainable and fair machine learning framework for high-stakes leadership-pipeline evaluation. World Journal of Management Science, 2(1), 73–82.

[33] Lin, Y., Wang, Y., & Bian, M. (2024). XOR-EW: An explainable and cost-aware artificial intelligence framework for operational risk early warning in financial institutions. AI and Data Science Journal, 1(1), 94–102.

[34] Zhang, C., & Wang, Y. (2024). A data-driven multi-objective production scheduling framework for battery pack manufacturing under material arrival uncertainty. AI and Data Science Journal, 1(1), 86–93.

[35] Rhee, M., Zou, J., & Qin, Y. (2024). CLARITY: A closed-loop framework for duplicate-aware failure prediction and budgeted automated remediation in cloud infrastructure. AI and Data Science Journal, 1(1), 76–85.

[36] Wang, Y., & Zhang, C. (2023). AERIS: Adaptive executor-scoped resource management for privacy-preserving data processing in Apache Spark. Eurasia Journal of Science and Technology, 5(2), 25–33.

[37] Qin, Y., Zou, J., & Rhee, M. (2022). An AI-driven intelligent transaction execution framework for internet finance platforms: Cost-sensitive real-time integrity interdiction. Eurasia Journal of Science and Technology, 4(2), 36–44.

[38] Zou, J., Qin, Y., & Rhee, M. (2024). Beyond model scale: A task-aware reliability benchmark for large language models in clinical decision support. Innovation and Technology Studies, 1(1), 39–47.

[39] Bian, M., Wang, Y., & Lin, Y. (2024). Explainable supplier risk prediction for critical telecommunications infrastructure procurement: An empirical XGBoost–SHAP framework. Social Science and Management, 1(1), 64–77.

[40] Jiao, Y., & Liang, Y. (2024). An integrated process-mining and explainable machine-learning framework for bottleneck and internal-control risk detection in the multi-entity financial close. Social Science and Management, 1(3), 96–107.