Explainable Discrete-Time Survival Learning for Bridge Deck Deterioration and Risk-Based Preservation Prioritization

Shuxin Zhang[1, *], Lei Qiu[2]

[1] University of California, Berkeley, California, The United States

[2] Ningbo University of Technology, Ningbo, Zhejiang, China

[*]shuxin_zhang@berkeley.edu

Abstract

Highway agencies must decide which bridge decks to treat before their condition degrades, yet most data-driven deterioration models either impose restrictive parametric hazard shapes or recast the problem as snapshot classification of the current condition rating, which discards the censoring structure of periodic inspections. This paper presents an explainable survival machine-learning framework built directly on the sojourn-time structure of National Bridge Inventory (NBI) deck condition ratings. From a 23-year (1992–2014) nationwide panel of 150,144 concrete highway bridge decks we extract 48,901 right-censored sojourn spells in the "Good" state (condition rating 7), of which 16,182 end in an observed downgrade. We propose DT-GBH, a discrete-time gradient-boosted hazard model that expands every spell into person-period records and learns the annual conditional downgrade probability, so that interval-censored annual reporting, a fully flexible baseline hazard and non-proportional covariate effects are all handled natively. Against Kaplan–Meier, Cox proportional hazards, componentwise boosted Cox, random survival forest and XGBoost Cox/AFT baselines, DT-GBH gives the best discrimination (C-index 0.766, 95% CI [0.758, 0.776]; Uno's C 0.769; five-year AUC 0.784) and the lowest integrated Brier score (0.143). Because the model is a tree ensemble on the hazard logit, TreeSHAP provides exact additive attributions, and we introduce a time-resolved attribution view that exposes how the dominant drivers change with the years already spent in the state. Two methodological findings are reported: a uniform five-year retrospective window of deterioration-history features adds 0.033 C-index over inventory attributes alone, whereas calendar "panel-position" features inflate apparent accuracy by 0.013 and must be excluded; and transfer to unseen states degrades every model to C ≈ 0.66, where the machine-learning advantage over Cox disappears. Finally, predicted five-year downgrade probabilities are combined with traffic exposure and a normalized treatment-cost model into a Preservation Prioritization Index, which is scored retrospectively against transitions actually observed in held-out data: averaged over budget levels from 5% to 50% it recovers 82.0% of the benefit attainable by a clairvoyant ranking, against 51.6% for predicted risk alone and only 20.6% for the worst-first heuristic, which is barely better than random selection.

Keywords

Bridge management, deterioration modelling, survival analysis, discrete-time hazard, gradient boosting, SHAP, explainable machine learning, National Bridge Inventory, preservation prioritization.

1. Introduction

Bridge decks are the primary component of a highway bridge that deteriorates fastest, and they absorb a disproportionate share of preservation budgets [1]. Under the U.S. National Bridge Inventory (NBI) programme every public bridge is inspected on a regular cycle and its deck, superstructure and substructure are each assigned an integer condition rating (CR) from 9 (excellent) to 0 (failed) [2], [3]. Because preservation treatments such as sealing, thin overlays and joint repair are far cheaper and far more effective while a deck is still in good condition, the decision-relevant quantity for an asset manager is not the current rating but the remaining time in the current rating-that is, a time-to-event quantity.

Two largely separate literatures address this problem. The first models condition evolution as a stochastic process: homogeneous and non-homogeneous Markov chains [4], [5], [6], semi-Markov and Weibull sojourn-time models [7], [8], [9], [10], and duration or proportional-hazard models estimated from inspection records [11], [12], [13], [14], [15]. These models are transparent and explicitly probabilistic, but they impose strong functional assumptions-exponential or Weibull sojourn distributions, proportional hazards, additive linear predictors-and are usually estimated with a handful of covariates and on a single state's data.

The second literature applies supervised machine learning to NBI records: classification and regression trees[17],[18],[19], random forests and gradient boosting [20], [21], [50], neural networks and hybrid pipelines [22], [24]. These studies report high accuracy, but they almost always predict the rating observed at a fixed snapshot. Doing so discards two features that are intrinsic to inspection data: the outcome is a duration, and that duration is right-censored because the observation window is finite and because rehabilitation removes assets from the risk set. Classification metrics computed on snapshots therefore cannot answer "in how many years", and models trained this way are silently biased by the censoring mechanism.

Survival machine learning is the natural bridge between the two literatures, yet its use in bridge management is still thin. Lu and Guler [25] compared a random survival forest with an AFT-Weibull model on Pennsylvania deck data; Li et al. [26] and Huang et al. [27] fitted Weibull and Cox models to Chinese bridge-management records; Fleischhacker et al. [16] built a Bayesian time-in-condition-rating (TICR) model on nationwide NBI data. None of this work (i) adopts a discrete-time formulation matched to the annual NBI reporting cycle, (ii) supplies exact local and global explanations of the learned hazard, or (iii) closes the loop from a predicted hazard to a budget-constrained preservation decision that is scored against outcomes actually observed in the data.

This paper addresses those three gaps. Our contributions are:

1) A leakage-controlled cohort construction procedure that converts the NBI annual condition panel into right-censored sojourn spells with an identical retrospective window and identical minimum potential follow-up for every spell, and that isolates the calendar "panel-position" variables which would otherwise inflate measured accuracy.

2) DT-GBH, a discrete-time gradient-boosted hazard model for condition-state sojourn times, which yields a fully flexible baseline hazard, non-proportional covariate effects, calibrated survival curves and a natural treatment of the one-year interval censoring of annual reporting.

3) An explanation layer that applies TreeSHAP on the hazard logit and a time-resolved attribution view showing how driver importance evolves with the time already spent in the state.

4) A Preservation Prioritization Index (PPI) that combines the predicted horizon-specific downgrade probability, traffic exposure and a normalized treatment-cost model, together with a retrospective budget-constrained evaluation against observed transitions.

5) A validation protocol with three holdout designs-random, unseen states and later entry years-plus a replication on two neighbouring condition states, from which we report both the gains and the limits of the approach.

2. Related Work

2.1. State- and Duration-based Deterioration Models

Markov-chain formulations dominate bridge-management practice because transition matrices plug directly into life-cycle optimisation [4], [6], [49]. Their memoryless assumption is, however, contradicted by inspection data: hazard rates rise with the time already spent in a state [8], [15]. Semi-Markov and Weibull sojourn models relax this [7], [9], [13], [14], and duration models with covariates [11], [12] allow heterogeneity, but all of them fix the hazard family in advance. Recent work extends duration modelling to nationwide data with Bayesian hierarchical priors [16] and to generalized gamma lifetimes [48].

2.2. Machine Learning on NBI Records

Tree ensembles are now routinely applied to NBI condition ratings [19], [20], [21], [50], often with environmental covariates [22] and increasingly with SHAP used post hoc to interpret the fitted classifier [21]. Two systematic issues recur. First, the target is the rating itself, so the learning problem is ordinal classification on a snapshot rather than prediction of a transition time. Second, variable selection is frequently driven by convenience rather than by an explicit causal or physical argument [23], and maintenance actions that censor the deterioration process are rarely modelled [24].

2.3. Survival Machine Learning and Explanation

Outside civil engineering, survival machine learning is mature: random survival forests [32], gradient-boosted Cox models [34], and neural extensions such as DeepSurv and Cox-Time [36], [37]. Discrete-time formulations, which model the conditional hazard in each interval with a binary regression, date back to Allison [30] and are treated comprehensively by Tutz and Schmid [31]; tree-based discrete-time models have been developed for competing risks [33]. This family is a particularly good match for NBI data, where ratings are reported once per year and the exact transition instant is unobserved. On the explanation side, SHAP [38], [39],[40] provides additive attributions with uniqueness guarantees and an exact polynomial-time algorithm for tree ensembles, although the caution of Rudin [41] about post hoc explanation of opaque models remains pertinent; we address it by keeping the explained quantity-the annual hazard logit-directly interpretable. Time-dependent explanations of survival models have been studied through SurvSHAP(t) [51] and SurvLIME [52], which attribute a model's predicted survival curve at each time point for a fixed asset. Our time-resolved view is complementary and cheaper: because the discrete-time formulation already carries the period index as an input, the attribution at each elapsed period is an ordinary TreeSHAP value on the hazard logit and requires no surrogate model or curve decomposition.

3. Data and Cohort Construction

3.1. Source Panel

We use the nationwide NBI concrete-deck panel released with [16], which contains, for 150,144 distinct concrete highway bridge decks in 47 states, the deck condition rating reported in each year from 1992 to 2014 together with time-invariant inventory attributes: year built, year reconstructed, deck structure type, structural material, type of design and construction, type of wearing surface, membrane type, deck protection, deck area, number of lanes on the structure, designated inspection interval, average daily truck traffic (ADTT), distance from seawater,

functional class, maintenance responsibility, and a nine-level climatic region. All results below are computed from these observed records; no condition data are simulated.

3.2. Sojourn Spells, Events and Censoring

Let $c_i(y)$ denote the deck rating of bridge i in year y. A spell in state s starts in the first year t_0 for which $c_i(t_0) = s$ and $c_i(t_0 - 1) > s$, so that entry into the state is observed rather than inferred; this removes the left truncation that affects assets already sitting in the state at the start of the panel. Sojourn time is measured in years since t_0. The event of interest is the first downgrade,

$$T_i = \min\{ k \geq 1 : c_i(t_0 + k) < s \}, \quad (1)$$

and a spell is right-censored at the last year with a valid rating when (i) the panel ends, (ii) the rating becomes unavailable, or (iii) the rating increases, which indicates a rehabilitation or re-decking action that removes the deck from the natural-deterioration risk set. Case (iii) is a competing event; we treat it as independent censoring and return to this assumption in Section VI. At most one spell per bridge is retained, so observations are independent across bridges.

3.3. Leakage-controlled Window

Because the panel is a finite window, the calendar position of a spell determines how much follow-up it can possibly have. A model given the entry year can therefore predict the administrative censoring limit rather than the physics of deterioration, and the resulting discrimination is optimistic. We control this by construction: entry is restricted to $t_0 \in [1992 + W, 2014 - H]$ with $W = H = 5$, so that every spell has exactly W years of complete retrospective record and at least H years of potential follow-up, and every deterioration-history feature is computed inside that same W-year window. The entry year itself, and any feature that is a function of it, is placed in a separate "panel-position" block that is excluded from the main model and used only in the ablation of Section V-C.

3.4. Cohort and Covariates

The primary cohort is state $s = 7$ ("Good"), the most populous rating and the state in which preventive treatment is most cost-effective. It contains 48,901 spells from 47 states, 16,182 of which (33.1%) end in an observed downgrade; the median observed sojourn is 7 years and the median age at entry is 30 years. Cohorts for $s = 8$ and $s = 6$ are built identically and used for replication in Section V-D. Table 1 lists the covariate blocks; categorical variables are one-hot encoded with levels below 1% frequency pooled, giving 56 design columns.

Table 1. covariate blocks used by the models

Block	Variables
Inventory (numeric)	age at entry, effective age since reconstruction, reconstructed flag, log deck area, log ADTT, lanes on structure, designated inspection interval, log distance to seawater
Inventory (categorical)	deck structure type, structural material, design and construction type, wearing surface, membrane, deck protection, functional class, maintenance responsibility, climatic region
Deterioration history (5-yr window)	rating before entry, rating 5 years before entry, 5-year rating drop, number of downgrades and of upgrades in the window, mean rating in the window, consecutive years above the state within the window
Panel position (excluded)	calendar entry year, number of prior observations, panel index

3.5. Descriptive Analysis

Fig. 1 shows Kaplan–Meier estimates [29] of the probability of remaining in CR 7, stratified by age at entry and by climatic region. Sojourn is markedly shorter for older decks and differs systematically across climate zones, and the curves are clearly non-exponential: the slope steepens over the first decade, which is inconsistent with a homogeneous Markov chain and motivates a flexible baseline hazard.

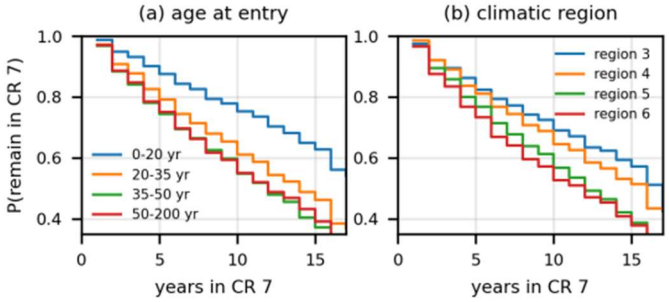

Fig. 1 Kaplan–Meier probability of remaining in deck condition rating 7, training portion of the primary cohort, stratified by (a) age at entry and (b) NBI climatic region.

4. Methodology

4.1. Discrete-time Hazard Formulation

NBI ratings are reported once per year, so a downgrade observed in year $t_0 + k$ only tells us that the transition occurred inside the interval $(t_0 + k - 1, t_0 + k]$. Rather than approximate this with a continuous-time likelihood, we adopt the discrete-time hazard [30], [31]

$$h(k \mid x) = \Pr(\, T = k \mid T \geq k, x \,), \quad k = 1, 2, \ldots \tag{2}$$

from which the survival and cumulative-incidence functions follow exactly:

$$S(t \mid x) = \prod_{k=1}^{t} [\, 1 - h(k \mid x) \,]. \tag{3}$$

Expanding each spell i into T_i person-period records with binary labels $y_{ik} = 1\{k = T_i, \delta_i = 1\}$, where δ_i indicates an observed downgrade, the discrete-time likelihood factorises as a Bernoulli likelihood,

$$\mathscr{L} = \prod_i \prod_k h(k|x_i)^{y_{ik}} [1 - h(k|x_i)]^{1-y_{ik}}, \quad k = 1, \ldots, T_i, \tag{4}$$

so any probabilistic binary learner with a log-loss objective estimates the hazard consistently. Right censoring is handled automatically: a censored spell simply contributes T_i records all labelled zero.

4.2. DT-GBH

We parameterise the hazard with a gradient-boosted tree ensemble on the logit scale,

$$h(k \mid x) = \sigma(\, F(x, k) \,), \;\; F(x, k) = \sum_m \eta\, f_m(x, k), \tag{5}$$

where σ is the logistic function, f_m are regression trees fitted by LightGBM [35] to the negative gradient of the log-loss, and the elapsed period index k enters as an ordinary input feature. Two properties follow directly. First, because k is a free splitting variable, the baseline hazard is fully non-parametric in discrete time, with no exponential or Weibull assumption. Second, because trees can split jointly on k and on covariates, the model represents non-proportional (time-varying) covariate effects without any explicit interaction terms. This is the key structural difference from gradient-boosted Cox models, which retain a proportional-hazards factorisation.

Predicted survival curves are obtained by evaluating the ensemble at each k and taking the product in (3). For ranking we use the negative restricted mean sojourn time,

$$r(x) = -\sum_t S(t \mid x), \quad t = 1, \ldots, t_{max}, \tag{6}$$

with $t_{max} = 15$ years, which summarises the whole predicted curve rather than a single horizon. Hyper-parameters are learning rate 0.05, 63 leaves, minimum 200 records per leaf, feature and row subsampling 0.8, L2 penalty 1.0, and the number of boosting rounds is chosen by early stopping on a validation split.

4.3. Explanation Layer

Because F(x, k) is a tree ensemble, TreeSHAP [38], [39] computes exact Shapley attributions in polynomial time, and the local additivity property

$$F(x, k) = \varphi_0(k) + \sum_j \varphi_j(x, k) \tag{7}$$

holds on the log-odds scale of the annual downgrade probability-an engineering-meaningful quantity, unlike the arbitrary risk score of a Cox-type model. Global importance is the mean absolute attribution over a sample of test person-periods. We additionally define a time-resolved attribution profile by fixing the period index and averaging over assets,

$$\Phi_j(k) = E_x \mid \varphi_j(x, k) \mid, \tag{8}$$

which reveals whether a driver acts mainly on early-life or late-life hazard-information that a single global importance ranking cannot convey and that a proportional-hazards model cannot represent at all.

4.4. Preservation Prioritization Index

Let $\hat{p}_i(H) = 1 - S(H \mid x_i)$ be the predicted probability that deck i leaves the state within H years. We define exposure e_i as the product of ADTT and deck area, normalised by its network median, and a normalized treatment cost with a fixed (mobilisation) share α and an area-proportional share,

$$c_i = \alpha + (1 - \alpha)\, A_i / \text{median}(A), \tag{9}$$

so that the ratio, not the absolute magnitude, of costs matters. The index is the expected benefit per unit cost,

$$PPI_i = \hat{p}_i(H) \cdot e_i / c_i. \tag{10}$$

Under a budget equal to a fraction β of the cost of treating the whole network, a ranking rule π selects assets greedily in decreasing score until the budget is exhausted. Its realised benefit is evaluated with the transitions actually observed in the holdout data,

$$B(\pi, \beta) = \sum_i 1\{ T_i \le H, \delta_i = 1 \} \cdot e_i, \quad i \in \pi(\beta), \tag{11}$$

and we report the targeting efficiency $TE(\pi, \beta) = B(\pi, \beta) / B(\pi^*, \beta)$, where π^* is the clairvoyant ranking that knows the observed outcomes. This measures the quality of the targeting decision only; it assumes a treatment applied within the horizon prevents the downgrade, which is stated as an assumption and not estimated from the data.

4.5. Baselines, Metrics and Validation

Baselines are: a covariate-free Kaplan–Meier reference; ridge-penalised Cox proportional hazards [28]; componentwise gradient-boosted Cox [45]; a random survival forest [32]; XGBoost with the Cox objective; and XGBoost with the accelerated-failure-time objective, to which the annual interval censoring is passed explicitly as $(T_i - 1, T_i]$ for events and $[T_i, \infty)$ for censored spells [34]. Survival curves for the XGBoost Cox model use a Breslow baseline estimated on the training split. Discrimination is measured by Harrell's C [42], by Uno's inverse-probability-of-censoring-weighted C [43] and by time-dependent AUC at 3, 5 and 8 years; probabilistic accuracy by the integrated Brier score over 2–10 years [44]. Confidence intervals for C are percentile bootstrap intervals with 200 resamples of the test set. To keep the IPCW censoring estimator strictly positive, follow-up is administratively truncated before computing weighted metrics. Implementation uses scikit-survival [45],[46], scikit-learn [47], LightGBM [35] and XGBoost [34].

Three holdout designs are used: (i) random, a 70/15/15 split over bridges; (ii) unseen states, in which a random 25% of states is withheld entirely, testing geographic transfer; and (iii) later entry years, training on spells entering before 2005 and testing on spells entering from 2005 onward, testing temporal extrapolation.

5. Results

5.1. Predictive Accuracy

Table 2. discrimination and calibration on the random holdout of the primary cohort (CR 7)

Model	C-index	95% CI	Uno's C	AUC(3)	AUC(5)	AUC(8)	IBS
Kaplan-Meier	0.500	--	0.500	0.500	0.500	0.500	0.184
Cox PH	0.687	[0.675, 0.698]	0.688	0.691	0.696	0.708	0.168
Boosted Cox (CW)	0.666	[0.655, 0.676]	0.669	0.663	0.670	0.678	0.172
Random Survival Forest	0.710	[0.699, 0.721]	0.714	0.703	0.712	0.746	0.161
XGBoost-Cox	0.759	[0.750, 0.768]	0.761	0.759	0.777	0.793	0.147
XGBoost-AFT	0.761	[0.752, 0.770]	0.762	0.767	0.781	0.794	--
DT-GBH (proposed)	0.766	[0.758, 0.776]	0.769	0.765	0.784	0.800	0.143

Table 2 reports the random-holdout results. All covariate models beat the Kaplan–Meier reference by a wide margin, confirming that NBI attributes carry substantial information about sojourn length. Cox proportional hazards reaches C = 0.687; the three tree-ensemble survival models add roughly seven points of concordance, and DT-GBH is the strongest on every metric, with C = 0.766 [0.758, 0.776], Uno's C = 0.769, five-year AUC = 0.784 and integrated Brier score 0.143 against 0.168 for Cox. The bootstrap intervals of DT-GBH and of the two XGBoost variants

overlap, so the margin over the best continuous-time boosting baseline is small; the substantive gap is between the linear proportional-hazards model and the ensemble family, and it indicates strong non-linearity and interaction among the inventory attributes. That DT-GBH also attains the lowest integrated Brier score is the more useful result for asset management, because the Brier score rewards well-calibrated absolute probabilities rather than mere ranking.

5.2. Generalisation across space and time

Table 3. generalisation under the two structured holdout designs

Protocol	Model	C-index	Uno's C	AUC(5)	IBS
Unseen states	Cox PH	0.658	0.661	0.668	0.157
Unseen states	XGBoost-Cox	0.657	0.666	0.645	0.161
Unseen states	DT-GBH (proposed)	0.656	0.666	0.641	0.163
Later entry years	Cox PH	0.659	0.658	0.680	0.119
Later entry years	XGBoost-Cox	0.684	0.684	0.706	0.114
Later entry years	DT-GBH (proposed)	0.692	0.692	0.716	0.113

Table 3 shows a clear asymmetry. Under temporal extrapolation the ordering of models is preserved and DT-GBH retains its advantage (C = 0.692 versus 0.659 for Cox), with the integrated Brier score even improving because the later cohort has shorter follow-up. Under geographic transfer, however, every model collapses to C ≈ 0.66 and the ensembles lose their edge entirely-DT-GBH reaches 0.656 against 0.658 for Cox, and its integrated Brier score is in fact slightly worse. The interpretation is that a large part of the non-linear signal that the ensembles exploit is state-specific: it reflects local design standards, materials practice, de-icing policy and inspector behaviour that do not transfer. This is a practically important negative result. It implies that a nationally trained deterioration model should not be deployed in a state absent from its training data without recalibration, and that reported accuracy from random splits over pooled national data systematically overstates transferability.

5.3. Feature-block Ablation and Panel-position Optimism

Table 4. feature-block ablation for DT-GBH (random holdout)

Feature configuration	p	C-index	Uno's C	IBS
A1 inventory only	49	0.733	0.734	0.153
A2 history only	7	0.646	0.652	0.169
A3 inventory+history	56	0.766	0.769	0.143
A4 +panel-position (leaky)	59	0.779	0.781	0.140

Table 4 separates the two feature families. Inventory attributes alone give C = 0.733; the seven deterioration-history variables alone give 0.646, which is well above chance from a very small feature set and confirms that the recent trajectory of a deck carries independent predictive content; combining them yields 0.766, an improvement of 0.033 over inventory data alone and of -0.0098 in integrated Brier score. This matters operationally because history features are free-they are already in every agency's inspection archive-whereas improving inventory data is expensive.

Adding the panel-position block raises C by a further 0.013, to 0.779. That gain is spurious: calendar entry year and the number of prior observations carry no physical information about deterioration, but they identify how much follow-up a spell can possibly have and therefore let the model anticipate the administrative censoring pattern. Before the window controls of Section III-C were imposed, an unrestricted version of the same cohort produced C = 0.831 by

the same mechanism. Studies that pool multi-year NBI panels and split at random should therefore report explicitly whether calendar-position variables enter the feature set.

5.4. Replication on Adjacent Condition States

Table 5. replication on adjacent entry states

Cohort	n	Events	Model	C-index	AUC(5)	IBS
CR8->CR<8	11109	6452	Kaplan-Meier	0.500	0.500	0.223
CR8->CR<8	11109	6452	Cox PH	0.683	0.715	0.190
CR8->CR<8	11109	6452	DT-GBH (proposed)	0.748	0.775	0.162
CR6->CR<6	53055	14394	Kaplan-Meier	0.500	0.500	0.186
CR6->CR<6	53055	14394	Cox PH	0.660	0.658	0.173
CR6->CR<6	53055	14394	DT-GBH (proposed)	0.739	0.747	0.147

Table 5 repeats the pipeline for entry states 8 and 6. The pattern is stable: DT-GBH improves C over Cox by six to eight points and reduces the integrated Brier score in both cohorts, and the absolute level of discrimination is highest for state 8, where the event rate is 58% and the deterioration signal is strongest. The framework is therefore not specific to the CR 7 sojourn.

5.5. Calibration

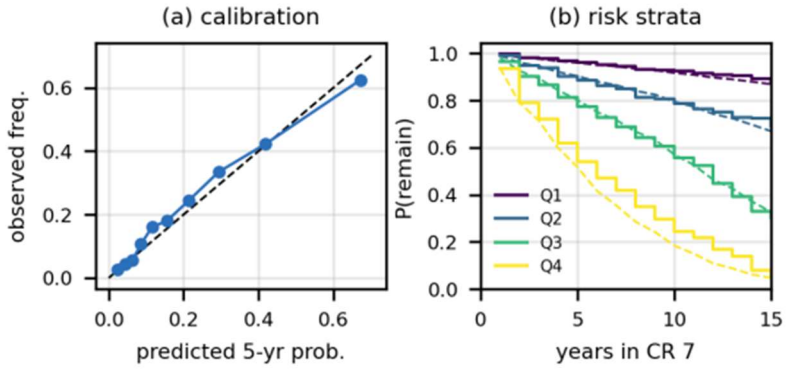

Fig. 2 (a) Predicted versus observed five-year downgrade probability by decile of predicted risk. (b) Observed Kaplan–Meier curves (solid) and mean predicted survival (dashed) by quartile of predicted risk, test split.

Fig. 2(a) plots observed against predicted five-year downgrade frequency by decile of predicted risk; the points track the diagonal across the full range, and the mean predicted probability (20.8%) is close to the observed frequency (22.0%) on the 6,554 test spells with sufficient follow-up, with a five-year Brier score of 0.1369 and a calibration slope of 0.900 (a slope of one denotes perfect calibration; the small shortfall indicates marginally over-dispersed predictions). Fig. 2(b) compares Kaplan–Meier curves with mean predicted survival within risk quartiles; the predicted curves separate in the right order and stay close to the empirical ones over the whole horizon, which is what makes the output usable as an input to a life-cycle cost model rather than merely as a ranking.

5.6. What the Model Has Learned

Fig. 3 ranks the twelve strongest attributions. The elapsed time in state and the age of the deck dominate, followed by age, log dist. to seawater, effective age. The ordering is engineering-plausible rather than merely statistical: the recent trajectory variables (the rating five years before entry and the number of downgrades inside the window) enter high in the ranking,

which is consistent with the ablation of Section V-C, and traffic loading (ADTT) and deck geometry contribute more than any single categorical design attribute.

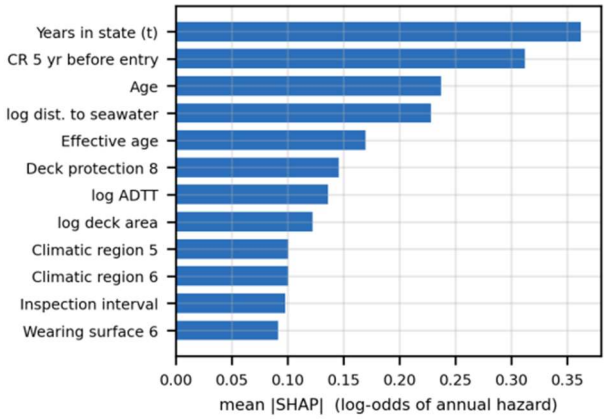

Fig. 3 Global TreeSHAP attribution for DT-GBH, mean absolute contribution to the log-odds of the annual downgrade hazard over test person-periods.

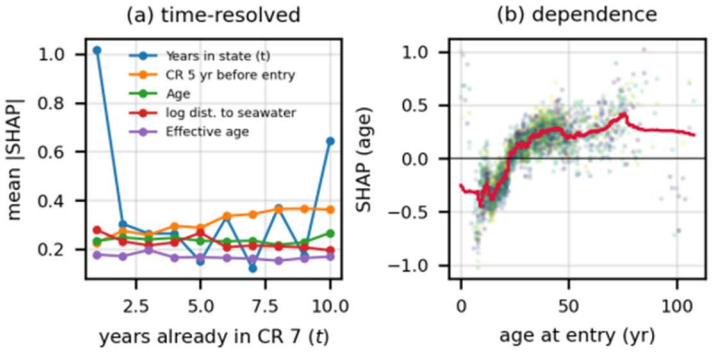

Fig. 4 (a) Time-resolved attribution $\Phi_j(k)$ for the five strongest features, as a function of the years already spent in CR 7. (b) SHAP dependence for age at entry; colour encodes the period index k.

Fig. 4(a) shows the time-resolved profiles of (8). They are not flat, which is direct evidence against proportional hazards: the contribution of some drivers grows with the years already spent in the state while others decay, so a single hazard ratio cannot describe them. Fig. 4(b) shows the dependence of the attribution on age at entry: the effect is close to monotone but distinctly non-linear, with the steepest increase in hazard over the first three decades of service and a flattening thereafter, a shape that a linear predictor in age would misrepresent at both ends. Attributions of this form can be reported per bridge, which is what an engineer needs in order to accept or override a recommendation.

5.7. Preservation Prioritisation

Table 6. targeting efficiency by budget share (α = 0.3)

Prioritisation rule	5%	10%	20%	30%	50%
PPI (proposed)	0.652	0.624	0.894	0.943	0.989
Risk only	0.339	0.374	0.539	0.622	0.706
Exposure only	0.535	0.714	0.844	0.912	0.978
Worst-first (oldest)	0.008	0.040	0.115	0.275	0.591
Random	0.071	0.077	0.147	0.201	0.345

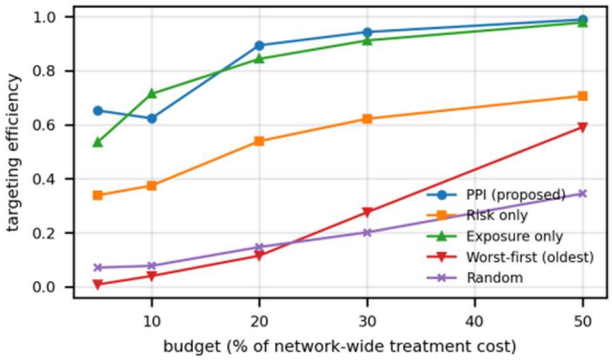

Fig. 5 Targeting efficiency against budget share for five prioritisation rules, evaluated on observed transitions in the test split with fixed-cost share $\alpha = 0.3$.

Table 6 and Fig. 5 evaluate the ranking rules of Section IV-D on the test split, scored against transitions that actually occurred. Averaged over the five budget levels, the PPI attains the highest targeting efficiency (0.820), followed by ranking on exposure per unit cost (0.797), predicted risk alone (0.516), the worst-first heuristic that treats the oldest decks (0.206) and random selection (0.168); the PPI is the best rule at 4 of the 5 budget levels. Three observations follow. First, neither ingredient of the index is sufficient on its own: predicted risk ignores that a downgrade on a large, heavily trafficked deck destroys more value, while exposure ignores whether a downgrade is imminent, and the product dominates both on average. Second, the advantage is not uniform. At the 10% budget the exposure-only rule is ahead of the PPI (0.714 versus 0.624), and the two curves cross again below 10%; when the budget only reaches a small number of very large structures, consequence weighting alone is a strong rule and the extra variance introduced by the predicted probability can hurt. We report this rather than tuning the index to hide it. Third, and most consequential for practice, the age-based worst-first heuristic performs very poorly once benefit is weighted by exposure and cost-at the tightest budgets it is worse than random selection-because age correlates with risk but carries no information about consequence, so it systematically directs money to small, lightly trafficked rural decks. Repeating the experiment with fixed-cost shares $\alpha \in \{0, 0.3, 0.6\}$ changes the absolute efficiencies but not the ordering of the rules.

6. Discussion and Limitations

Four limitations bound the conclusions. First, rating increases are treated as independent censoring, whereas they are in fact a competing event driven by the same covariates as deterioration; a cause-specific or discrete-time competing-risks extension [33] would remove this assumption and is a natural next step. Second, maintenance actions that do not change the rating are invisible in NBI data, so the estimated hazard is the hazard of an asset under whatever unrecorded maintenance regime prevailed [24], and it should be read as a conditional-on-practice hazard rather than a material-degradation law. Third, the condition rating is a coarse, partly subjective ordinal index; the observed geographic non-transferability in Table 3 is consistent with inspector-effect heterogeneity as well as with genuine environmental differences, and the two cannot be separated with these data alone. Fourth, the decision experiment measures targeting quality under a stated perfect-effectiveness assumption; a full life-cycle evaluation would require treatment-effectiveness and cost data that NBI does not contain.

The framework nonetheless has properties that matter for deployment. It consumes only data that every U.S. agency already reports; it produces calibrated absolute probabilities over a multi-year horizon rather than a ranking; its explanations are exact and are expressed on the

annual-hazard scale that engineers reason about; and it makes explicit two failure modes-panel-position leakage and geographic non-transferability-that are easy to overlook when accuracy is measured on random splits of pooled national data.

7. Conclusion

We presented an explainable survival machine-learning framework for bridge-deck condition deterioration and preservation prioritisation, built on 48,901 observed sojourn spells extracted from a 23-year nationwide NBI panel. The proposed discrete-time gradient-boosted hazard model matches the annual reporting cycle of the data, imposes no parametric hazard shape and no proportionality, and delivers the best discrimination (C-index 0.766) and the best probabilistic accuracy (integrated Brier score 0.143) among seven survival models. TreeSHAP attributions on the hazard logit, including a time-resolved view, make the learned deterioration structure inspectable; a five-year retrospective window of inspection-history features adds 0.033 C-index at no data-collection cost; and calendar panel-position features are shown to inflate apparent accuracy and are excluded. Coupling predicted horizon probabilities with traffic exposure and a normalized cost model into a Preservation Prioritization Index gives the highest mean targeting efficiency across budget levels (0.820) when scored against transitions observed in held-out data, although it is not uniformly best at every budget; the age-based worst-first heuristic in common use scores only 0.206, barely above random selection. Future work will address competing risks from rehabilitation, state-level recalibration for geographic transfer, and integration of the calibrated hazards into a life-cycle optimisation.

References

[1] American Society of Civil Engineers. (2025). 2025 report card for America's infrastructure. Reston, VA: ASCE.

[2] Federal Highway Administration. (1995). Recording and coding guide for the structure inventory and appraisal of the nation's bridges (Rep. FHWA-PD-96-001). Washington, DC: FHWA.

[3] Federal Highway Administration. (2022). Specifications for the national bridge inventory. Washington, DC: FHWA.

[4] Morcous, G. (2006). Performance prediction of bridge deck systems using Markov chains. Journal of Performance of Constructed Facilities, 20(2), 146–155.

[5] Madanat, S., Mishalani, R., & Ibrahim, W. H. W. (1995). Estimation of infrastructure transition probabilities from condition rating data. Journal of Infrastructure Systems, 1(2), 120–125.

[6] Madanat, S., Karlaftis, M. G., & McCarthy, P. S. (1997). Probabilistic infrastructure deterioration models with panel data. Journal of Infrastructure Systems, 3(1), 4–9.

[7] Ng, S.-K., & Moses, F. (1998). Bridge deterioration modeling using semi-Markov theory. In Structural safety and reliability: Proc. ICOSSAR'97, Kyoto, Japan. Rotterdam: Balkema.

[8] Sobanjo, J. O. (2011). State transition probabilities in bridge deterioration based on Weibull sojourn times. Structure and Infrastructure Engineering, 7(10), 747–764.

[9] Thomas, O., & Sobanjo, J. (2016). Semi-Markov models for the deterioration of bridge elements. Journal of Infrastructure Systems, 22(3), 04016010.

[10] Thompson, P. D., & Johnson, M. B. (2005). Markovian bridge deterioration: Developing models from historical data. Structure and Infrastructure Engineering, 1(1), 85–91.

[11] Mauch, M., & Madanat, S. (2001). Semiparametric hazard rate models of reinforced concrete bridge deck deterioration. Journal of Infrastructure Systems, 7(2), 49–57.

[12] Mishalani, R., & Madanat, S. (2002). Computation of infrastructure transition probabilities using stochastic duration models. Journal of Infrastructure Systems, 8(4), 139–148.

[13] Agrawal, A. K., Kawaguchi, A., & Chen, Z. (2010). Deterioration rates of typical bridge elements in New York. Journal of Bridge Engineering, 15(4), 419–429.

[14] Sobanjo, J., Mtenga, P., & Rambo-Roddenberry, M. (2010). Reliability-based modeling of bridge deterioration hazards. Journal of Bridge Engineering, 15(6), 671–683.

[15] Manafpour, A., Guler, I., Radlińska, A., Rajabipour, F., & Warn, G. (2018). Stochastic analysis and time-based modeling of concrete bridge deck deterioration. Journal of Bridge Engineering, 23(9), 04018066.

[16] Fleischhacker, A., Ghonima, O., & Schumacher, T. (2020). Bayesian survival analysis for US concrete highway bridge decks. Journal of Infrastructure Systems, 26(1), 04020001.

[17] Ghonima, O., Anderson, J. C., Schumacher, T., & Unnikrishnan, A. (2020). Performance of US concrete highway bridge decks characterized by random parameters binary logistic regression. ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering, 6(1). https://doi.org/10.1061/AJRUA6.0001031

[18] Bolukbasi, M., Mohammadi, J., & Arditi, D. (2004). Estimating the future condition of highway bridge components using national bridge inventory data. Practice Periodical on Structural Design and Construction, 9(1), 16–25.

[19] Bektas, B. A., Carriquiry, A., & Smadi, O. (2013). Using classification trees for predicting national bridge inventory condition ratings. Journal of Infrastructure Systems, 19(4), 425–433.

[20] Assaad, R., & El-adaway, I. H. (2020). Bridge infrastructure asset management system: Comparative computational machine learning approach for evaluating and predicting deck deterioration conditions. Journal of Infrastructure Systems, 26(3), 04020032.

[21] Li, Y., & Song, G. (2022). Ensemble-learning-based prediction of steel bridge deck defect condition. Applied Sciences, 12(11), 5442.

[22] Yang, C., Wang, X., & Nassif, H. (2024). Impact of environmental conditions on predicting condition rating of concrete bridge decks. Transportation Research Record. https://doi.org/10.1177/03611981241248647

[23] Chang, M., Maguire, M., & Sun, Y. (2017). Framework for mitigating human bias in the selection of explanatory variables for bridge deterioration modeling. Journal of Infrastructure Systems, 23(3), 04017002.

[24] Wang, F., Lee, C.-C., & Gharaibeh, N. G. (2022). Network-level bridge deterioration prediction models that consider the effect of maintenance and rehabilitation. Journal of Infrastructure Systems, 28, 05021009.

[25] Lu, M., & Guler, S. I. (2022). Comparison of random survival forest with accelerated failure time-Weibull model for bridge deck deterioration. Transportation Research Record, 2676(7), 296–311.

[26] Li, L., Lu, Y., & Peng, M. (2022). Deterioration model for reinforced concrete bridge girders based on survival analysis. Mathematics, 10(23), 4436.

[27] Huang, L., et al. (2022). A prognostic model for newly operated highway bridges based on censored data and survival analysis. Advances in Civil Engineering, 2022, 4667231.

[28] Cox, D. R. (1972). Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological), 34(2), 187–220.

[29] Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. Journal of the American Statistical Association, 53(282), 457–481.

[30] Allison, P. D. (1982). Discrete-time methods for the analysis of event histories. Sociological Methodology, 13, 61–98.

[31] Tutz, G., & Schmid, M. (2016). Modeling discrete time-to-event data. Cham, Switzerland: Springer.

[32] Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests. The Annals of Applied Statistics, 2(3), 841–860.

[33] Berger, M., Welchowski, T., Schmitz-Valckenberg, S., & Schmid, M. (2019). A classification tree approach for the modeling of competing risks in discrete time. Advances in Data Analysis and Classification, 13(4), 965–990.

[34] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794).

[35] Ke, G., et al. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30, 3146–3154.

[36] Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., & Kluger, Y. (2018). DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Medical Research Methodology, 18(1), 24.

[37] Kvamme, H., Borgan, Ø., & Scheel, I. (2019). Time-to-event prediction with neural networks and Cox regression. Journal of Machine Learning Research, 20(129), 1–30.

[38] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.

[39] Lundberg, S. M., et al. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1), 56–67.

[40] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144).

[41] Rudin, C. (2919). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215.

[42] Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., & Rosati, R. A. (1982). Evaluating the yield of medical tests. JAMA, 247(18), 2543–2546.

[43] Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B., & Wei, L. J. (2011). On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Statistics in Medicine, 30(10), 1105–1117.

[44] Graf, E., Schmoor, C., Sauerbrei, W., & Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. Statistics in Medicine, 18(17–18), 2529–2545.

[45] Pölsterl, S. (2020). scikit-survival: A library for time-to-event analysis built on top of scikit-learn. Journal of Machine Learning Research, 21(212), 1–6.

[46] Davidson-Pilon, C. (2019). lifelines: Survival analysis in Python. Journal of Open Source Software, 4(40), 1317.

[47] Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

[48] Lu, M., Hydock, J., Radlińska, A., & Guler, S. I. (2022). Reliability analysis of a bridge deck utilizing generalized gamma distribution. Journal of Bridge Engineering, 27(4), 04022006.

[49] Frangopol, D. M., Kong, J. S., & Gharaibeh, E. S. (2001). Reliability-based life-cycle management of highway bridges. Journal of Computing in Civil Engineering, 15(1), 27–34.

[50] Almarahlleh, N., Liu, H., Abudayyeh, O., & Almamlook, R. (2024). Predicting concrete bridge deck deterioration: A hyperparameter optimization approach. Journal of Performance of Constructed Facilities, 38(3). https://doi.org/10.1061/JPCFEV.CFENG-4714

[51] Krzyziński, M., Spytek, M., Baniecki, H., & Biecek, P. (2023). SurvSHAP(t): Time-dependent explanations of machine learning survival models. Knowledge-Based Systems, 262, 110234.

[52] Kovalev, M. S., Utkin, L. V., & Kasimov, E. M. (2020). SurvLIME: A method for explaining machine learning survival models. Knowledge-Based Systems, 203, 106164.