

Article

Gradient-Boosted Survival Models for Corrosion Risk-Based Inspection of Gas Transmission Pipelines

Anna V. Shmonina * and Alexey S. Dikov 

Institute of Nuclear Physics ARKAE, Almaty 050032, Kazakhstan; dikov@inp.kz

* Correspondence: a.shmonina@inp.kz

Abstract

Prediction of the time to failure of pipeline materials supports inspection scheduling within risk-based inspection (RBI) frameworks. We compare three survival models, namely the Cox proportional hazards model (Cox PH), the random survival forest (RSF), and XGBoost with the survival:cox objective, on 823 corrosion incidents from the PHMSA database (1986–2025) retained after an engineering-validity screen. The comparison turns on a point that is easy to miss at the level of implementation. For the survival:cox objective, the predict() method of XGBoost returns the hazard ratio $\exp(\hat{\eta})$, whereas the Breslow estimator of the baseline cumulative hazard expects the linear predictor $\hat{\eta}$ and applies the exponential internally. Chaining the two evaluates $\exp(\exp(\hat{\eta}))$. Both maps are strictly increasing, so the concordance index is identical under either scale, to the last representable digit in our computation, and the defect does not appear in discrimination-based validation, while the estimated survival function is destroyed. A simulation with a known survival function isolates the magnitude: substituting the hazard ratio for the linear predictor raises the mean absolute deviation of $\hat{S}(t | x)$ from 0.073 to 0.398 and the integrated Brier score (IBS) from 0.154 to 0.395, at an unchanged concordance index of 0.723. On the pipeline corpus, the erroneous scale inflates the IBS to 0.356 ± 0.043 over 20 random splits. On the correct scale, XGBoost-Cox attains $C = 0.871 \pm 0.017$ and $IBS = 0.0614 \pm 0.0056$ and exceeds both reference models on both axes simultaneously (Cox PH 0.853 ± 0.018 and 0.0717 ± 0.0067 ; RSF 0.807 ± 0.020 and 0.0751 ± 0.0030 ; two-sided Wilcoxon $p = 1.9 \times 10^{-6}$ throughout). It also attains the lowest expected calibration error at every horizon from 10 to 40 years. No post hoc calibration is required, and adding one degrades the result. Out-of-time validation preserves the discrimination, whereas the absolute probabilities require periodic re-estimation of the baseline hazard. The feature-importance hierarchies agree with the ISO 8044 classification of corrosion factors.

Keywords: survival analysis; time-to-failure modelling; gradient boosting; Breslow estimator; predictor scale; concordance index; integrated Brier score; material degradation; pipeline corrosion; PHMSA



Academic Editor: Nicanor Cimpoesu

Received: 2 July 2026

Revised: 2 August 2026

Accepted: 5 August 2026

Published: 7 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Direct losses from metal corrosion are estimated at 3–4% of the gross domestic product of developed industrial countries. When equipment downtime, accidents, and environmental damage are included, the total effect can exceed 6% of GDP [1]. For the United States, a comparable estimate gives 276 billion USD in annual direct costs at the end of the 1990s and 550 billion USD when indirect effects are included [2]. For China, the corresponding share is approximately 3.3% of GDP [3]. A substantial part of these costs can be avoided by

timely and accurate failure prediction. Accurate models reduce the number of redundant inspections and lower the risk of a missed defect, which makes the estimation of failure probability a question of direct economic value.

The scale of these losses follows from the structure of the underlying process. Corrosion of metal in contact with soil or water rarely reduces to a single mechanism. Uniform surface dissolution, localized pitting, stress corrosion cracking, microbially influenced damage, and corrosion under insulation can develop simultaneously within the same pipe section [4,5]. These mechanisms can reinforce one another, and the resulting damage rate depends on dozens of interrelated factors, including the composition of the environment, the temperature, the condition of the protective coating, the performance of cathodic protection, and microbial activity. This multifactorial character distinguishes corrosion from simpler service-life problems and makes it a suitable object for validating methods designed for heterogeneous, interacting variables. A methodology validated on corrosion data can later be transferred to other complex degradation processes with a similar data structure, such as fatigue under cyclic loading, radiation embrittlement, and high-temperature creep.

The growing interest in machine learning for corrosion prediction is motivated by the limitations of empirical models. Ensembles of decision trees, including Random Forest, gradient boosting, and XGBoost, systematically outperform classical regression relationships in predicting the corrosion rate of low-alloy steels in marine and atmospheric environments [6,7]. For pipeline transport, similar ensemble algorithms have been applied to the estimation of internal corrosion from pressure, temperature, and CO₂ concentration [8], while reviews of stress corrosion cracking report limited validation of such models on field data [9]. Within the RBI methodology, machine learning is used to automate screening risk assessment and reaches equipment classification accuracies above 90% [10], including problems specific to hydrogen technologies [11]. These studies formulate the task as regression of the corrosion rate or as threshold risk classification rather than as censored time-to-event analysis, and they do not provide a probabilistic estimate of PoF(t) over an arbitrary horizon.

The task has a statistical structure that is common to all time-dependent degradation processes. A population of objects with individual characteristics moves toward a failure event that, for some objects, has not yet occurred at the time of observation, so the data are right-censored. Survival analysis is the branch of statistics developed for exactly this data structure. Its machine-learning extensions, such as random survival forests and gradient boosting with a Cox loss, capture nonlinear feature interactions without prescribing a specific physical degradation law in advance. The integrity management of gas transmission pipelines, formalized in the RBI standards API RP 580 [12] and API RP 581 [13], is a convenient testbed for these methods because an extensive open registry of operational incidents exists and the material and environmental characteristics are clearly defined. Corrosion accounts for approximately one-quarter of the reported incidents on gas transmission pipelines in the registry studied here, which motivates the practical interest in data-driven prediction of corrosion failure probability.

Modern survival analysis originates from the semi-parametric Cox proportional hazards model, in which the coefficients are estimated by maximizing the partial likelihood and the baseline hazard is recovered separately in a non-parametric way [14]. Random survival forests relax the proportional-hazards assumption and capture nonlinear feature interactions through an ensemble of trees [15]. Gradient-boosting models, in particular XGBoost with the survival:cox objective, combine the Cox partial likelihood with an additive expansion over decision trees and are becoming a standard choice in applied work [16]. These methods have been applied to pipeline integrity and to broader material reliability problems [17–20]. Direct comparisons of different architectures on a single dataset, with

a unified validation scheme and explicit calibration checks, remain rare [21,22]. In addition, the interaction between gradient-boosted survival models and post hoc recovery of the baseline hazard by the Breslow method is poorly studied. This question matters for obtaining correct absolute probabilities rather than only a correct risk ranking.

This work addresses a specific practical gap. Gradient-boosted survival models are widely reported to rank objects well while producing absolute failure probabilities that are unsuitable for engineering decisions, and this pattern is usually attributed to the architecture of the booster. The pattern is reproduced exactly on the present corpus, and its origin is an interface-level mismatch of predictor-scale conventions between the boosting library and the baseline-hazard estimator. Once the two conventions are aligned, the contradiction disappears. XGBoost-Cox becomes superior to Cox PH in discrimination and in calibration at the same time, and no post hoc correction is required. The result is established on 823 PHMSA corrosion incidents (1986–2025) and is supported by a simulation in which the true survival function is known by construction. The feature importances recovered independently by the three models are consistent with the normative classification of corrosion factors in ISO 8044 [23], which serves as an interpretability check. Throughout the paper, the transferability of the approach to other degradation problems is treated as a hypothesis to be tested rather than as an established property (Sections 3.7 and 4.4).

The distinction matters for how such results are read. A body of work reports gradient-boosted survival models that discriminate well and calibrate poorly, and proposes post hoc corrections in response, ranging from isotonic and spline-based recalibration of predicted probabilities at fixed horizons [24,25] to temporal recalibration of the baseline of an existing Cox model [26]. Such corrections are appropriate when the deficiency is genuine. When the deficiency originates instead in a scale-convention mismatch, a monotone post hoc transformation can appear to repair the model while acting only as an approximate inverse of a spurious exponential. A logarithmic compression illustrates the mechanism, since $\log(1 + \exp(\hat{\eta}))$ approaches $\hat{\eta}$ for $\hat{\eta}$ well above zero; the approximation fails for $\hat{\eta} < 0$, and the resulting procedure is unstable across splits. Supplementary Material S2 documents this behaviour on the present corpus. Explicit reporting of the predictor scale is therefore a precondition for interpreting published calibration results for boosted survival models.

The aim of this study is to determine whether a gradient-boosted survival model can provide, at the same time, the ranking accuracy required for inspection prioritization and the calibrated absolute probabilities required for quantitative risk computation within RBI frameworks. The contributions are threefold. First, we identify and quantify a scale-convention mismatch in the common XGBoost-Cox pipeline with post hoc Breslow recovery, and show by simulation against a known survival function that it inflates the integrated Brier score by a factor of 2.6 while leaving the concordance index unchanged. Second, we report a corrected benchmark of three survival models on the pipeline corpus, in which XGBoost-Cox exceeds Cox PH in discrimination and in calibration simultaneously, without any post hoc correction. Third, we present a validation protocol that combines repeated random splits, a simulation with known ground truth, permutation controls, out-of-time and leave-region-out designs, and an explicit treatment of the observation window, including left truncation and the role of calendar covariates.

2. Materials and Methods

2.1. Data Source and Corrosion Subset

PHMSA incident reports for gas transmission and gathering pipelines for the period 1986–2025 were obtained from the open registry maintained by the U.S. Department of Transportation [27]. Three reporting periods with different column schemas (1986–2001, 2002–2009, and 2010–2025) were harmonized into a single table with 4293 records. Filtering

by the cause field CORROSION produced 964 records. Duplicate entries and entries with critical missing values in modelling-relevant fields were removed. The remaining records were screened against engineering-validity bounds. The pipeline age had to be consistent with the installation date, the wall thickness had to lie within API 5L ranges, the diameter had to be at least 2 in, and the incident and maximum operating pressures had to be strictly positive and below 4000 psi. Values violating these bounds were treated as invalid and were either removed or, for isolated implausible covariate entries, set to missing and handled by the imputation described in Section 2.2. The working corpus contains $N = 823$ incidents and covers 38 U.S. states and the federal Outer Continental Shelf waters of the Gulf of Mexico. The stationarity of the corrosion incident share $f_{\text{corr}}(t) = N_{\text{corr}}(t)/N_{\text{all}}(t)$ was verified with the non-parametric Mann–Kendall trend test [28,29] ($Z = -1.11$; $p = 0.27$; Sen’s slope of -0.001 per year).

The time-to-event variable T is defined as the pipeline age at the moment of the recorded incident. The event indicator δ is given by the binary SIGNIFICANT field, which separates significant from non-significant incidents (597 versus 226 records, imbalance ratio $IR = 2.64$). This formulation reduces the problem to a standard right-censored survival analysis setting.

Three properties of this formulation should be stated explicitly. First, the corpus consists of incident records only, because the PHMSA registry does not enumerate the population of in-service pipeline segments that have never experienced a reportable incident. The estimand is therefore the conditional distribution of pipeline age at a significant corrosion incident, given that a reportable corrosion incident occurred. This quantity is suited to comparative risk ranking among incident-prone assets rather than to absolute population-level failure rates. Second, treating non-significant incidents as right-censored observations assumes that, conditional on the covariates, a non-significant incident at age t carries no information about the age at which a significant incident would have occurred. This non-informative censoring assumption cannot be tested from the registry alone. Its implications, together with the competing-risks alternative, are discussed in Section 4.4. Third, pipes installed before the observation window opened in 1986 enter the registry only if they survived to that date, which constitutes left truncation (Figure 1). A delayed-entry analysis that quantifies this effect is reported in Section 3.6. The time origin T_0 is fixed as the year of installation, so the time scale T is pipeline age. Because T_0 varies across installation cohorts, the out-of-time and cohort-extrapolation analyses in Section 3.7 test whether the models capture material ageing rather than epoch-specific manufacturing effects. A sensitivity analysis in Section 3.8 confirms that removing the records that violate the validity bounds shifts the headline metrics by less than one standard deviation.

The composition of the corpus drifts across the reporting decades. The share of incidents classified as significant rises monotonically from 0.38 in the 1980s to 0.87 in the 2020s, and the median pipeline age at incident rises from 24 to 52 years. This drift reflects both the ageing of the network and the tightening of the reporting criteria across the three regimes. A decade-by-decade composition table is provided in Supplementary Material S15. The practical consequence of this drift for model transfer is quantified by the out-of-time validation in Section 3.6.

2.2. Feature Engineering

Each record is described by fifteen covariates, namely SYSTEM_TYPE, STATE, CLASS, SHORE, PRESSURE_INCIDENT, PRESSURE_MAX_OP, MATERIAL, INSTALLATION_YEAR, DIAMETER, THICKNESS, MANUFACTURED_YEAR, PIPE_COATED, PROT, PROT_YEAR, and LOCATION. The categorical feature STATE is one-hot encoded with 38 levels. Missing values in the PROT and PROT_YEAR fields are treated as an

informative NO_DATA category because their non-random occurrence carries predictive information about installations that precede the introduction of mandatory cathodic-protection registration.

Observation process in the PHMSA registry: delayed entry (left truncation)

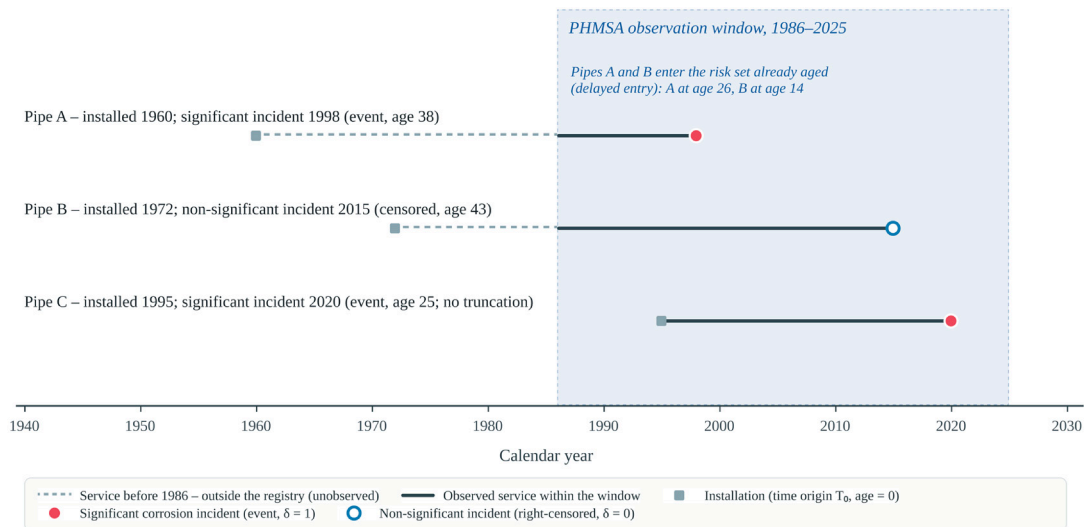


Figure 1. Observation process in the PHMSA registry. Pipes installed before 1986 enter the observation window at a positive age (delayed entry, left truncation); significant incidents are events ($\delta = 1$) and non-significant incidents are treated as right-censored observations ($\delta = 0$).

Missingness is concentrated in two fields. PROT_YEAR, the year of cathodic-protection installation, is absent in 37.1% of the records, and MANUFACTURED_YEAR is absent in 31.7%. The remaining numeric covariates are nearly complete (THICKNESS: 1.1%, PRESSURE_INCIDENT: 0.5%, DIAMETER: 0.1%, and PRESSURE_MAX_OP: 0.1%). Numeric covariates are imputed with the median computed within each training fold, which prevents information transfer across the train/test boundary. Categorical covariates receive an explicit NO_DATA level, which allows the models to treat missingness itself as potentially informative. Section 3.4 shows that PROT_NO_DATA acts as an age proxy. No records are deleted because of missingness beyond the physically implausible entries described in Section 2.1. The full missingness table is given in Supplementary Material S14, and a sensitivity analysis restricted to records with a directly recorded installation year is reported in Section 3.8.

2.3. Survival Models

The Cox proportional hazards model parameterizes the conditional hazard as $h(t | x) = h_0(t) \cdot \exp(\beta^T x)$ and estimates the coefficient vector β jointly with the baseline hazard $h_0(t)$ by maximizing the partial likelihood [14], combined with the Breslow estimator for the baseline cumulative hazard [30]:

$$L(\beta) = \prod_{i:\delta_i=1} \left[\frac{\exp(\beta^T x_i)}{\sum_{j \in R(T_i)} \exp(\beta^T x_j)} \right] \quad (1)$$

Random survival forests aggregate $B = 500$ survival trees grown on bootstrap subsamples with random feature selection at each split node [15]. The implementation uses the scikit-survival library, version 0.22.2 [31]. XGBoost-Cox, version 0.22.2, optimizes the Cox partial likelihood with respect to a linear predictor $\hat{\eta}(x)$ parameterized by an additive expansion over regression trees [16]. The baseline cumulative hazard

is recovered post hoc by applying the Breslow estimator [30] to that linear predictor, which requires the scale convention of Section 2.5 to be observed. All three models output a risk score $\hat{\eta}(x)$ and a survival function $\hat{S}(t | x)$ for each object. The hyperparameters were tuned by five-fold cross-validation on the training set. For XGBoost-Cox they are `max_depth` = 4, `learning_rate` = 0.03, `n_estimators` = 500, `subsample` = 0.8, `colsample_bynode` = 0.8, `reg_lambda` = 1.0, and `min_child_weight` = 5.

The choice of a right-censored survival formulation over a continuous accelerated-degradation model (ADM) follows from the data structure rather than from preference. An ADM, or a time-dependent degradation regression, requires repeated measurements of a degradation state variable, such as wall-loss trajectories from successive in-line inspections, from which a continuous damage path is reconstructed. The PHMSA registry records a single event per asset with no inspection history, so the degradation path is unobserved and the ADM likelihood cannot be identified from these data. What is observed is the age at a terminal event under right censoring, which is exactly the structure addressed by survival analysis. When wall-thinning trajectories become available, an ADM is the appropriate refinement, and the present formulation corresponds to its first-passage description.

For the Cox PH baseline, a ridge penalty of $\alpha = 0.1$ is applied on standardized covariates with Breslow tie handling; constant and degenerate one-hot columns are removed before fitting. The random survival forest aggregates $B = 500$ trees as specified above. All train/test splits are stratified by the event indicator and use the fixed seeds 0–19. The headline results are insensitive to the tuning configuration. An alternative XGBoost setting with 300 trees, a learning rate of 0.05, and per-tree column subsampling of 0.8 reproduces the metrics within noise. All computations were carried out in Python 3.10.20. The Cox proportional hazards model, the random survival forest, the Breslow estimator of the baseline cumulative hazard and the evaluation metrics are taken from scikit-survival 0.22.2 [31]; the gradient-boosted survival model from XGBoost 3.2.0 [16]; and the delayed-entry Cox analysis of Section 3.6 from lifelines 0.30.0. Preprocessing, resampling and the isotonic calibration of Supplementary Material S17 use scikit-learn 1.3.2, and the numerical work uses NumPy 1.26.4 and pandas 2.2.2.

2.4. Validation Metrics

Discrimination is measured by Harrell's concordance index [32]. Calibration is measured by the integrated Brier score (IBS),

$$IBS = \frac{1}{\tau} \int_0^{\tau} \mathbb{E} \left[(I\{T > t\} - \hat{S}(t|x))^2 \right] dt \quad (2)$$

which is the time-averaged squared deviation of the predicted survival function from the censoring-corrected empirical survival over the test horizon [33,34]. The evaluation grid is fixed once for all models and all configurations. Test objects whose observed time exceeds the largest training time are excluded, since the inverse-probability-of-censoring weights are not identified beyond the support of the training data, and the grid is then 100 equally spaced points between the smallest retained test time, floored at 5 years, and the largest retained test time, capped at 70 years. The number of objects excluded by this rule is reported per split in Supplementary Material S1. Both metrics are reported with bootstrap 95% confidence intervals computed from 500 resamples. Ranking stability is verified over 20 independent random 80/20 training/test splits stratified on the SIGNIFICANT field. The significance of pairwise IBS differences across splits is assessed with the Wilcoxon signed-rank test on the paired differences [35].

2.5. Predictor Scale in the XGBoost-Cox to Breslow Pipeline

The Cox partial likelihood is defined with respect to a linear predictor $\eta(x)$ that enters the hazard as $h(t | x) = h_0(t) \cdot \exp(\eta(x))$. Libraries differ in what they return for that quantity under the same method name. In scikit-survival, the `predict()` method of `CoxPHSurvivalAnalysis` returns $\hat{\eta}(x)$. In XGBoost, for objective = “survival:cox”, `predict()` returns the hazard ratio $\exp(\hat{\eta}(x))$; the linear predictor is obtained only when the flag `output_margin = True` is set. The documented rationale is that the hazard ratio is the quantity of direct interest for that objective. The consequence is that two libraries routinely combined in the same pipeline adopt opposite conventions.

The Breslow estimator that recovers the baseline cumulative hazard $\hat{H}_0(t)$ expects the linear predictor and applies the exponential internally. Supplying the default output of the XGBoost `predict()` method to it therefore evaluates the survival function on the wrong scale. Writing $r(x)$ for the quantity actually passed to the estimator, the intended and the realized survival functions are

$$\hat{S}(t | x) = \exp(-\hat{H}_0(t) \cdot \exp(\hat{\eta}(x))) \quad (3)$$

when the linear predictor is supplied, against

$$\hat{S}_{\text{err}}(t | x) = \exp(-\hat{H}_0(t) \cdot \exp(\exp(\hat{\eta}(x)))) \quad (4)$$

when the hazard ratio is supplied in its place, both $\exp$ and the composition $\exp \circ \exp$ are strictly increasing, so the ordering of objects induced by $r(x)$ is identical under the two conventions and the concordance index takes exactly the same value. Discrimination-based validation is blind to the substitution. The partial likelihood, by contrast, is not invariant under such a map. It evaluates $\exp(\eta(x_i))$ against the sum of $\exp(\eta(x_j))$ over the risk set and therefore depends on the differences $\eta(x_i) - \eta(x_j)$, being invariant only under the addition of a constant, which cancels between numerator and denominator and is absorbed by the baseline hazard. The scale of $\hat{\eta}$ is thus identified by the fit up to a shift, and it is the quantity the Breslow estimator expects; substituting $\exp(\hat{\eta})$ for it is an error rather than a reparameterization. The estimated survival function, in contrast, is destroyed. On the studied corpus, the quantities supplied to the estimator span $[0.004, 8.6 \times 10^3]$, and the exponential of the upper end exceeds the double-precision range, which pins $\hat{S}(t | x)$ at numerical zero for a large fraction of the test objects. The saturation of gradient-boosted survival models reported for this pipeline is an artifact of the interface rather than a property of the booster.

The magnitude of the induced bias is isolated by simulation, because on real data the true survival function is unknown and genuine model error cannot be separated from scale error. The generating process is a Cox model with a constant baseline hazard. For $n = 3000$ objects, eight covariates are drawn independently from the standard normal distribution and the linear predictor is $\eta = x^T \beta$ with $\beta = (0.8, -0.6, 0.4, -0.3, 0.2, 0, 0, 0)$, centred at zero, so that three covariates carry no signal. Event times are drawn from the exponential distribution with rate $\lambda \cdot \exp(\eta)$ and $\lambda = 0.02$, which gives the true survival function $S(t | x) = \exp(-\lambda \cdot t \cdot \exp(\eta))$ in closed form. Censoring times are drawn independently from the uniform distribution on the interval from 0 to 120 years, and the observed time is the minimum of the two, so the censoring is random and independent of the covariates by construction. The data are split 70/30; an XGBoost-Cox model with the hyperparameters of Table 1 is fitted on the training part, and the baseline hazard is recovered by the Breslow estimator in three configurations: from the generating linear predictor, from the estimated margin $\hat{\eta}$, and from the estimated hazard ratio $\exp(\hat{\eta})$ with the same mean shift and clip to ± 20 used on the pipeline corpus. Metrics are evaluated on a grid of 60 points from $t = 1$ to the 90th percentile of the test times. Because $S(t | x)$ is

known, the mean absolute deviation $|\hat{S}(t | x) - S(t | x)|$ over that grid is available as a direct measure of bias alongside the integrated Brier score.

Table 1. Hyperparameter configuration of the three survival models (single frozen configuration used for all reported results).

Model	Configuration
XGBoost-Cox	objective survival:cox; 500 trees; learning rate 0.03; max depth 4; subsample 0.8; colsample_bynode 0.8; min_child_weight 5; reg_lambda 1.0; reg_alpha 0; selected by five-fold CV on the training set
Cox PH	ridge penalty $\alpha = 0.1$ on standardized covariates; Breslow tie handling
RSF	$B = 500$ trees; min_samples_split = 10; min_samples_leaf = 5; max_features = $\sqrt{n}$ (sqrt); bootstrap subsampling; fixed random state
Splits/seeds	$20 \times (80/20)$, stratified by event indicator; seeds 0–19

On the pipeline corpus, five configurations were compared over 20 independent random splits sharing a common stratified partition, a common preprocessing pipeline, and a common IBS time grid. Two of them are the hazard ratio supplied to the Breslow estimator and the linear predictor supplied directly; the remaining three involve post hoc rescaling of the score and are reported in Supplementary Material S2. All fitting, including recovery of the baseline hazard, is confined to the training partition. The absence of information leakage across the train/test boundary was verified by a permutation check in which the training-set event indicators are randomly reassigned, with the pair of event indicator and observation time permuted jointly, so that the marginal outcome distribution is preserved and only its association with the covariates is destroyed. The concordance index then falls to 0.484 ± 0.053 , and the integrated Brier score rises to 0.167 ± 0.014 , against 0.871 ± 0.017 and 0.0614 ± 0.0056 on the intact labels (Supplementary Material S1).

3. Results

3.1. Exploratory Characterization of the Corrosion Corpus

Table 2 summarizes the descriptive statistics of the corrosion corpus. The median pipeline age at incident is 40 years, the median diameter is 16 inches, and the median operating pressure is 784 psi. Three regions, Texas (197 incidents), Louisiana (182), and the federal Gulf of Mexico waters (170), together account for 65.4% of the 823 incidents.

Table 2. Descriptive statistics of the numerical covariates of the working corrosion corpus after removal of duplicate entries, invalid event and time records, and physically implausible values ($N = 823$).

Field	<i>n</i> _Valid	Mean	Median	Q1	Q3	Min	Max
PRESSURE_INCIDENT (psi)	819	745.45	784	521.00	925.0	1.5	3176
PRESSURE_MAX_OP (psi)	822	1056.54	1024	809.25	1300.0	50	3377
DIAMETER (in)	822	16.34	16.0	10.75	20.0	2.0	48.0
THICKNESS (in)	814	0.34	0.31	0.25	0.4	0.125	1.2
PIPE_AGE (years)	823	39.93	40	27.0	51.0	1	99
INSTALLATION_YEAR	823	1966.65	1968	1958.0	1977.0	1908	2018

3.2. Verification of the Predictor Scale and Simulated Bias

The convention stated in Section 2.5 was verified directly on the fitted booster. For every test object, the default output of `predict()` was compared with the output obtained under `output_margin = True`. The two are related by an exact exponential: the maximum absolute discrepancy is $\max |\log \text{predict}(X) - \hat{\eta}| = 5.4 \times 10^{-8}$ over the test set, which is the resolution of the double-precision logarithm (Figure 2a). The linear predictor spans approximately $[-5.3, 9.1]$ on this corpus and is centred near 1.5 (Figure 2b), so the default output spans $[0.005, 8.9 \times 10^3]$. Figure 2c shows why a logarithmic compression of the default output appears to repair the pipeline. The sign-preserving map $\text{sign}(r) \cdot \log(1 + |r|)$ reproduces $\hat{\eta}$ to within 0.001 for $\hat{\eta} > 5$ and to within 0.04 for $\hat{\eta}$ between 2 and 5, and diverges from it on the 18.2% of objects with $\hat{\eta} < 0$, where the mean absolute discrepancy is 2.03. A check of this kind costs one additional call to `predict()` and belongs before any reconstruction of the baseline hazard. Table 3 quantifies the consequence of omitting it, in a simulation where the true survival function is known by construction. Supplying the estimated margin to the Breslow estimator gives a mean absolute deviation of the survival function of 0.073, which is the genuine estimation error of the booster; supplying the hazard ratio in its place raises it to 0.398, a factor of 5.5, and the integrated Brier score from 0.154 to 0.395. The concordance index is 0.723 under both, because the substitution is a strictly increasing transformation. Figure 3 shows the same result over time.

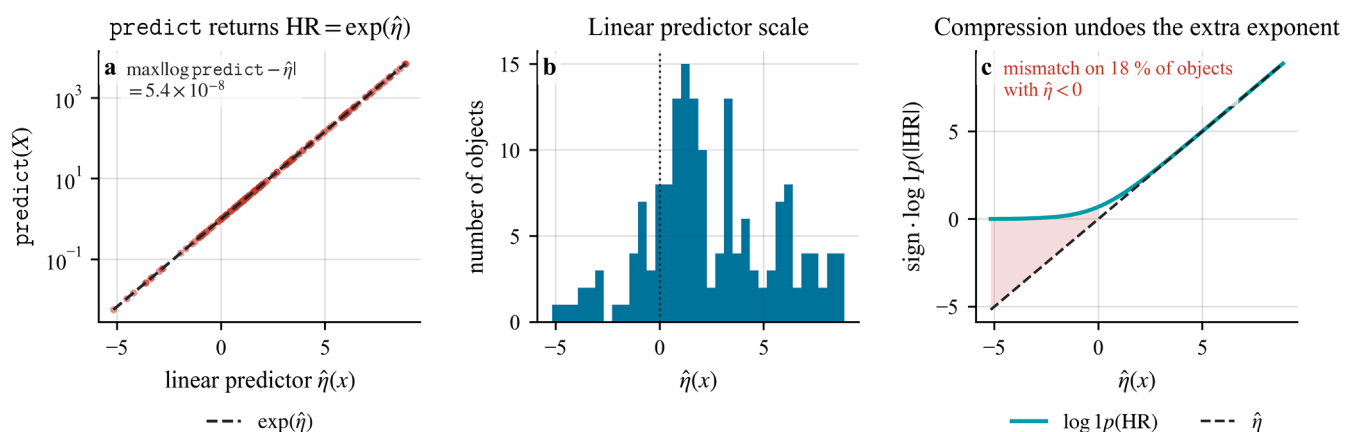


Figure 2. Predictor scale of XGBoost-Cox on the pipeline corpus: (a) default output of `predict()` against the linear predictor obtained with `output_margin = True`, logarithmic ordinate; (b) distribution of the linear predictor $\hat{\eta}(x)$; (c) sign-preserving logarithmic compression of the default output against $\hat{\eta}$, showing agreement for $\hat{\eta} > 0$ and divergence for $\hat{\eta} < 0$.

Table 3. Bias induced by the predictor-scale substitution in a simulation with a known survival function (3000 objects, eight standard normal covariates of which three are noise, constant baseline hazard $\lambda = 0.02$, independent uniform censoring, 70/30 split). The first row supplies the generating linear predictor to the Breslow estimator, the second the estimated margin $\hat{\eta}$, and the third the default output $\exp(\hat{\eta})$ of the XGBoost `predict()` method under the same safeguard applied to the pipeline corpus. The mean absolute deviation of the survival function is taken over the evaluation grid and over all objects.

Scale Supplied to the Breslow Estimator	C-Index	IBS	Mean $ \hat{S} - S $
Generating linear predictor (truth)	0.739	0.145	0
Estimated linear predictor $\hat{\eta}$ (margin)	0.723	0.154	0.073
Estimated hazard ratio $\exp(\hat{\eta})$	0.723	0.395	0.398

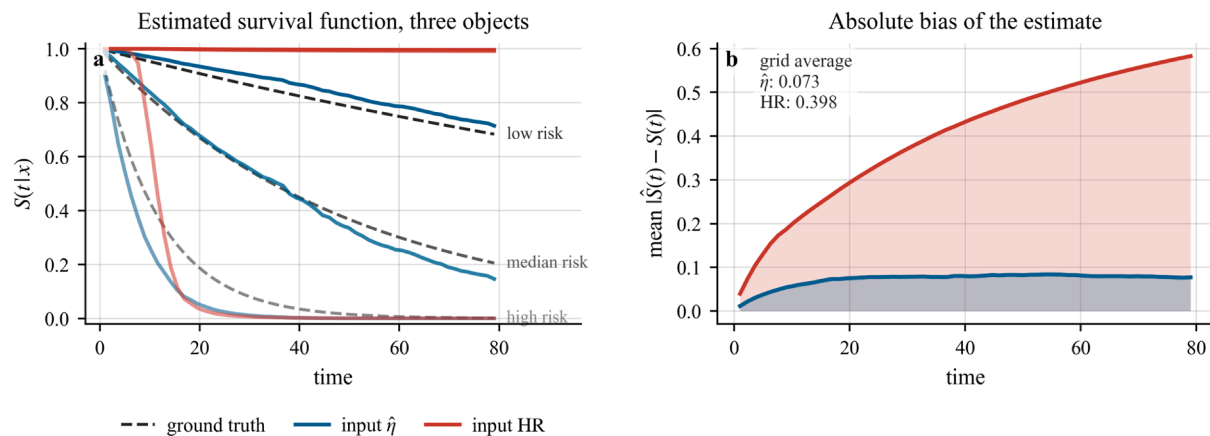


Figure 3. Consequence of the scale substitution in the simulation with a known survival function: (a) estimated survival functions for a low-risk, a median-risk, and a high-risk object against the generating truth; (b) mean absolute deviation of the estimated survival function as a function of time.

3.3. Discrimination and Calibration of the Survival Models

Table 4 and Figure 4 report the scale configurations on the pipeline corpus over 20 independent random splits sharing a common stratified partition, a common preprocessing pipeline, and a common IBS time grid. Direct use of the hazard ratio overflows the exponential of the Breslow estimator, so that configuration is evaluated with the safeguard adopted in the original pipeline, a mean shift followed by a clip to ± 20 . The clip prevents the overflow at the cost of the ranking: the hazard-ratio range of $[0.005, 8.9 \times 10^3]$ collapses onto the two boundaries for the majority of objects, the concordance index falls to 0.619 ± 0.026 , and the integrated Brier score is 0.356 ± 0.043 . Supplying the linear predictor instead gives 0.871 ± 0.017 and 0.0614 ± 0.0056 on the same splits, a reduction in the IBS by a factor of 5.8 with an improvement on every split (Wilcoxon $W = 0$, $p = 1.9 \times 10^{-6}$). The invariance of the concordance index to the scale substitution, argued algebraically in Section 2.5, holds exactly in the computation: over the 20 splits, the concordance index obtained from the unclipped hazard ratio equals the one obtained from the linear predictor to the last representable digit, with a maximum absolute difference of zero. The collapse to 0.619 is therefore attributable to the clip alone and not to the scale.

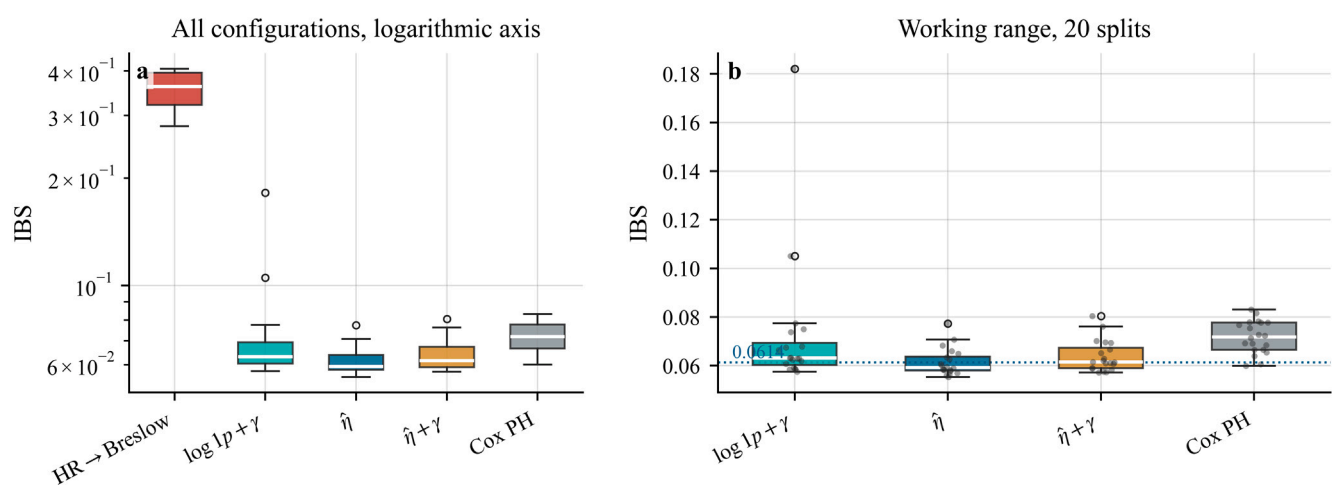


Figure 4. Integrated Brier score by configuration over 20 random splits: (a) all configurations on a logarithmic ordinate; (b) the working range, with the individual splits overlaid.

Table 4. Discrimination and calibration of the three survival models over 20 independent random splits ($N = 823$), all fitted on the same partitions with the same preprocessing pipeline and the same IBS time grid. The first four rows differ only in the quantity supplied to the Breslow estimator. The random survival forest reconstructs the survival function internally and does not use that estimator. Values are mean $\pm$ standard deviation. The p -values are two-sided Wilcoxon signed-rank tests of the paired IBS differences against the linear-predictor configuration.

Configuration	C-Index	IBS	p (IBS)
Hazard ratio $\exp(\hat{\eta})$, centred and clipped to ± 20	0.619 ± 0.026	0.356 ± 0.043	1.9×10^{-6}
Linear predictor $\hat{\eta}$	0.871 ± 0.017	0.0614 ± 0.0056	reference
Linear predictor $\hat{\eta}$ with Cox rescaling	0.871 ± 0.017	0.0638 ± 0.0064	9.5×10^{-6}
log1p compression with Cox rescaling	0.871 ± 0.017	0.0722 ± 0.0280	1.9×10^{-6}
Cox PH	0.853 ± 0.018	0.0717 ± 0.0067	1.9×10^{-6}
Random survival forest	0.807 ± 0.020	0.0751 ± 0.0030	1.9×10^{-6}

Against the two reference models evaluated under the identical protocol, XGBoost-Cox, on the correct scale, is superior on both axes at once. The concordance index is 0.871 ± 0.017 against 0.853 ± 0.018 for Cox PH and 0.807 ± 0.020 for the random survival forest, with median paired gains of $+0.017$ and $+0.068$ (Wilcoxon $W = 0$, $p = 1.9 \times 10^{-6}$ in both cases). The integrated Brier score is 0.0614 ± 0.0056 against 0.0717 ± 0.0067 and 0.0751 ± 0.0030 , with median paired gains of -0.0104 and -0.0150 ($W = 0$, $p = 1.9 \times 10^{-6}$ in both cases). The two reference models rank in opposite order on the two axes: Cox PH discriminates better than the forest (median $+0.053$, $p = 3.8 \times 10^{-6}$) while the forest calibrates marginally worse (median $+0.0043$, $p = 0.017$). The discrimination–calibration trade-off reported for gradient-boosted survival models does not arise once the predictor scale is correct. Post hoc rescaling of the score does not improve on the plain linear predictor: a single-parameter Cox rescaling raises the IBS to 0.0638 ± 0.0064 ($p = 9.5 \times 10^{-6}$), and the two-step compression-and-rescaling procedure raises it to 0.0722 ± 0.0280 , with two of the 20 splits exceeding 0.10. Both are documented in Supplementary Material S2.

Table 5 and Figure 5 resolve calibration by prediction horizon. Decile-binned calibration diagrams were computed at t of 10, 20, 30, and 40 years, with the observed survival in each bin estimated by the Kaplan–Meier method and the number of bins adapted to the size of the test partition. XGBoost-Cox attains the lowest expected calibration error at every horizon, with 0.012, 0.030, 0.053, and 0.058, against 0.014, 0.030, 0.058, and 0.067 for Cox PH and 0.018, 0.063, 0.095, and 0.125 for the random survival forest. The ordering by expected calibration error reproduces the ordering by integrated Brier score, so the aggregate result is not an artifact of averaging over the horizon. The forest degrades fastest with the horizon, which is consistent with the coarse resolution of its internally reconstructed survival function at long times. Calibration in the large shows that all three models are mildly conservative. At $t = 30$ years, the mean predicted survival is 0.726 for XGBoost-Cox, 0.732 for Cox PH, and 0.752 for the forest, against a Kaplan–Meier value of 0.782 on the same test partitions, so the downward, safe-side bias reaches 0.056 in absolute survival. The level values behind this comparison are tabulated in Supplementary Material S3.

Table 5. Calibration by prediction horizon, averaged over 20 random splits. The expected calibration error is computed over bins of predicted survival, with the observed survival in each bin estimated by the Kaplan–Meier method. The reference column is the Kaplan–Meier estimate on the whole test partition.

Horizon, yr	Kaplan–Meier $\hat{S}(t)$	ECE, XGBoost-Cox	ECE, Cox PH	ECE, RSF
10	0.983	0.012	0.014	0.018
20	0.916	0.030	0.030	0.063
30	0.782	0.053	0.058	0.095
40	0.586	0.058	0.067	0.125

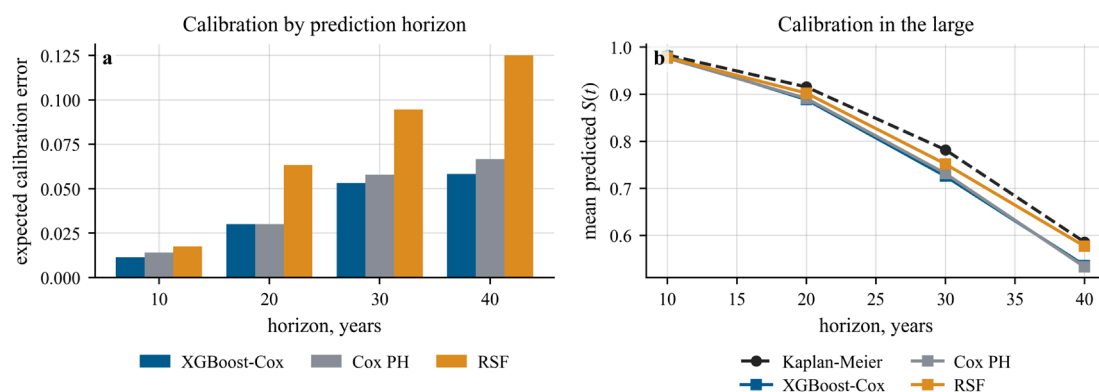


Figure 5. Calibration of the three survival models by prediction horizon, averaged over 20 random splits: (a) expected calibration error; (b) mean predicted survival against the Kaplan–Meier estimate on the same test partitions.

3.4. Feature Importance and ISO 8044 Mapping

Figure 6 presents the five most influential features of each survival model, recomputed on the screened 823-record corpus. RSF places INSTALLATION_YEAR first ($\Delta C = 0.049$), followed by MANUFACTURED_YEAR (0.014) and PROT_YEAR (0.014), while PRESSURE_INCIDENT (0.005) and the missing-protection-year indicator PROT_NO_DATA (0.001) contribute much less. Cox PH keeps INSTALLATION_YEAR as the dominant regressor ($\beta = 2.491$, HR = 12.07), followed by the geographic indicators STATE_IA (HR = 3.10), STATE_MO (HR = 2.65), and STATE_IL (HR = 2.61). PROT_NO_DATA ranks third by $|\beta|$ but carries a protective sign (HR = 0.36), so a missing protection-year record is associated with a lower hazard in this sample. This pattern is consistent with PROT_NO_DATA marking mainly transmission segments installed before protection-year record-keeping became mandatory, so the indicator correlates with pipe age rather than acting as an independent risk driver. XGBoost-Cox again places INSTALLATION_YEAR first (0.103), followed by PROT_NO_DATA (0.067) and SHORE_OFFSHORE (0.056), with STATE_WATER (0.048) and PROT_YES (0.046) completing the top five. RSF favours age-related covariates, whereas Cox PH and XGBoost-Cox favour protection-record and location covariates once age is accounted for. The union of the top-five features of the three models spans the four blocks of corrosion factors codified in ISO 8044 [23], namely material age (INSTALLATION_YEAR, MANUFACTURED_YEAR, PROT_YEAR), environment aggressiveness (STATE_IA, STATE_MO, STATE_IL, SHORE_OFFSHORE, STATE_WATER), protection state (PROT_NO_DATA, PROT_YES), and mechanical loading (PRESSURE_INCIDENT). The correspondence with ISO 8044 is presented as a qualitative plausibility check of the recovered importance hierarchy and not as a physical validation. The standard classifies corrosion factors into categories but does not prescribe a functional weighting or a pre-

dictive ranking, so the agreement only confirms that the influential covariates fall into recognized factor categories. A quantitative physical validation against soil mechanics or electrochemical benchmarks would require covariates that are absent from the registry and is left for future work.

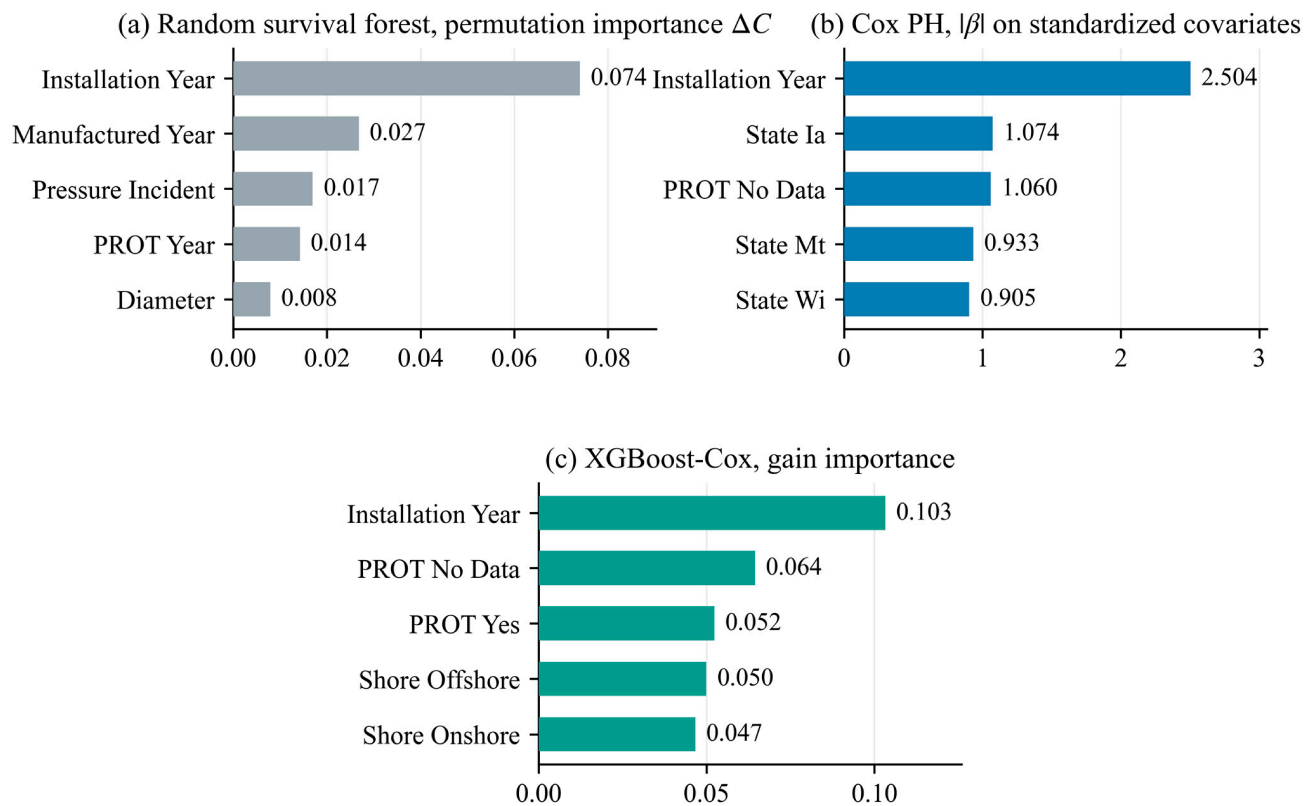


Figure 6. Top five influential features for (a) RSF; (b) Cox PH; (c) XGBoost-Cox.

3.5. The Observation Window and the Role of Calendar Covariates

The corpus consists of incident records, so the support of the outcome variable is determined jointly by the installation year and by the boundaries of the reporting window. A pipe installed in year I can appear in the registry only at an age between $\max(0, 1986 - I)$ and $2025 - I$. Table 6 quantifies the resulting geometry by installation cohort. For every cohort installed after 1950, the observed maximum age lies at or within four years of the arithmetic ceiling, so the window is saturated and the upper tail of the age distribution is fixed by the calendar rather than by the degradation process. The pre-1950 cohort is the single exception, with an observed maximum of 99 years against a ceiling of 125 years. Figure 7 shows the same structure at the level of individual records: the admissible region is a triangle bounded above by $2025 - I$ and below by $\max(0, 1986 - I)$, and the incidents fill it densely.

Because `INSTALLATION_YEAR` encodes both the physical age of the asset and its position relative to the observation window, its contribution has to be separated from the physics. Two specifications were compared over the same 20 splits. With the calendar covariates `INSTALLATION_YEAR` and `MANUFACTURED_YEAR` retained, XGBoost-Cox attains $C = 0.871 \pm 0.017$ and $IBS = 0.0614 \pm 0.0056$, and Cox PH attains 0.853 ± 0.018 and 0.0717 ± 0.0067 . With both removed, leaving twelve covariates that describe geometry, material, pressure, and protection, the discrimination falls to 0.769 ± 0.024 and 0.701 ± 0.032 , and the integrated Brier score rises to 0.0958 ± 0.0062 and 0.1184 ± 0.0077 . Roughly two-thirds of the concordance above chance therefore rests on calendar information. The ordering of the two models is preserved under both specifications, so the

comparison in Section 3.3 does not depend on the calendar covariates, but the absolute performance does, and the part of that performance attributable to window geometry rather than to degradation cannot be identified from incident records alone. A third specification addresses the status of `PRESSURE_INCIDENT`, which is the only one of the fifteen covariates recorded at the moment of the event; the remaining fourteen are asset attributes fixed before it. Removing it leaves both models essentially unchanged: XGBoost-Cox attains $C = 0.869 \pm 0.017$ and $IBS = 0.0611 \pm 0.0050$, and Cox PH attains 0.851 ± 0.018 and 0.0720 ± 0.0067 , so the concordance index shifts by 0.0015 and 0.0017, and the integrated Brier score by 0.0003 in either direction. No result reported in this paper depends on that covariate. The full covariate set is retained in the analyses above so that a single specification underlies every diagnostic, and the reduced set is available as the primary specification should the concurrent measurement be judged inadmissible in principle rather than on effect.

Table 6. Geometry of the observation window by installation cohort. The possible maximum age is $2025 - I$, evaluated at the lower edge of each cohort. The gap is the difference between the possible and the observed maximum.

Installation Cohort	<i>n</i>	Observed Max Age, yr	Possible Max Age, yr	Mean Age, yr	Gap, yr
1900–1950	112	99	125	64.9	26
1950–1960	136	75	75	49.9	0
1960–1970	202	63	65	41.4	2
1970–1980	228	55	55	32.9	0
1980–1990	85	41	45	23.3	4
1990–2000	37	31	35	18.6	4
2000–2026	23	23	25	11.6	2

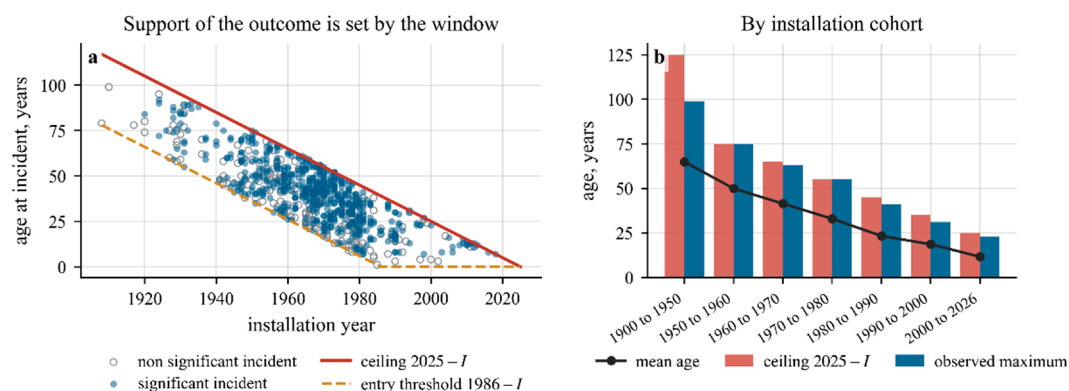


Figure 7. Support of the outcome variable in the PHMSA corpus: (a) age at incident against installation year, with the admissible region bounded by the reporting window; (b) observed and arithmetically possible maximum age by installation cohort.

3.6. Left Truncation

Pipes installed before 1986 enter the observation window at a positive age, which is a left-truncation structure (Figure 1). The naive risk set, which admits every object from age zero, over-represents early ages as a result. A delayed-entry Cox model, in which each object enters the risk set at the age it had reached in 1986, was fitted on the numeric covariates for comparison. Figure 8 reports the standardized coefficients under both risk sets. Four of the five coefficients shift by less than 0.016 in absolute value, which is within

the estimation noise. The exception is `INSTALLATION_YEAR`, whose coefficient falls from 1.280 to 1.044, a reduction of 18.4%. The direction is the expected one, since part of the apparent effect of the installation year is the entry structure of the window and correcting for delayed entry removes it. The physical covariates, namely diameter, wall thickness, and the two pressure fields, are insensitive to the correction. The XGBoost objective does not admit delayed entry, so this diagnostic is available only for the Cox baseline; its result should be read as an indication of the truncation sensitivity of the calendar covariates in the boosted model rather than as a correction applied to it. The coefficient table and the fitting details are given in Supplementary Material S11.

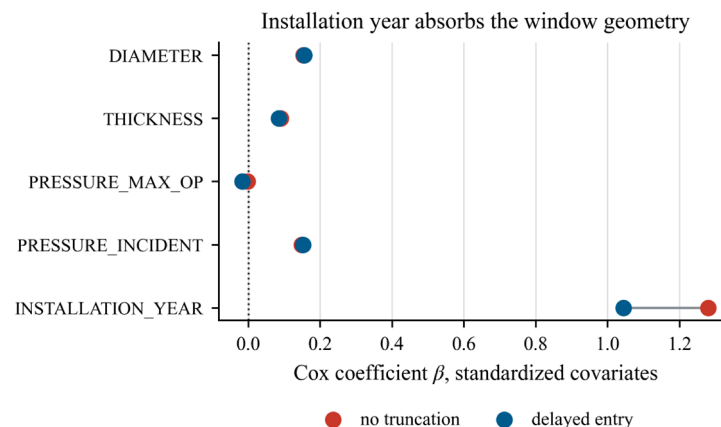


Figure 8. Standardized Cox coefficients under the naive risk set and under delayed entry at the age reached in 1986. Only `INSTALLATION_YEAR` shifts materially.

3.7. Out-of-Time Validation and Cohort Extrapolation

Random stratified splits provide statistically valid but temporally uninformed validation. Two complementary out-of-time designs were therefore evaluated on the corrected scale. The first mimics deployment: the model is trained on incidents reported up to a cutoff year and tested on later incidents. With the cutoff at the corpus median (2007; $N_{\text{train}} = 431$, $N_{\text{test}} = 392$), the concordance index is 0.869, and with the cutoff at 2010 ($N_{\text{train}} = 517$, $N_{\text{test}} = 306$), it is 0.852. Discrimination is therefore preserved at the level of the random-split baseline of 0.871 ± 0.017 . The absolute probability level does not transfer: the integrated Brier score on the later period rises to 0.271 and 0.215 respectively, a fourfold degradation relative to the random-split value. The cause is a shift in the reporting regime rather than in the ranking, since the share of incidents classified as significant rises from 0.38 in the late 1980s to above 0.85 after 2010. Re-estimating the baseline cumulative hazard on the later period while retaining the booster trained on the earlier one returns the integrated Brier score to 0.065 in both designs. That figure is an in-sample estimate of the baseline on the test era and therefore bounds from below what re-estimation can achieve in deployment, but its distance from 0.271 identifies the baseline level, and not the learned ranking, as the component that fails to transfer. The second design is a cohort-extrapolation stress test that splits on the installation year rather than the incident year. Training on pipes installed up to 1968 and predicting for younger cohorts gives a concordance index of 0.613, and training on pipes installed up to 1985 gives 0.542, which is close to chance. This reflects genuine non-stationarity of manufacturing technology together with the dominant role of `INSTALLATION_YEAR` established in Section 3.5. The two designs bound the applicability of the method: interpolation across calendar time within observed cohorts is reliable for ranking, whereas extrapolation to unobserved manufacturing generations is not. The four designs are tabulated in Supplementary Material S4.

3.8. Robustness Checks and Additional Diagnostics

Four diagnostic experiments corroborate the reported results. First, variance inflation factors across the eight numeric covariates confirm the absence of severe multicollinearity. The largest value is 3.48 for PRESSURE_MAX_OP, followed by 2.96 for INSTALLATION_YEAR and 2.46 for MANUFACTURED_YEAR, all below the conventional threshold of 5.0, and the pairs that carry them are the physically expected ones, namely the two pressure fields and the three year fields (Supplementary Material S5). Second, partial-dependence plots on the linear-predictor scale for the top five XGBoost-Cox features show a monotone but strongly nonlinear response to INSTALLATION_YEAR spanning roughly eleven units of log hazard, a functional form that the linear structure of Cox PH cannot represent, while the four one-hot covariates that follow it contribute shifts below one unit (Supplementary Material S6). Third, a Weibull accelerated-failure-time model, fitted as an independent parametric baseline, attains a concordance index of 0.849 ± 0.020 , below XGBoost-Cox and comparable to Cox PH, and provides a third interpretable reference point (Supplementary Material S7). Fourth, INSTALLATION_YEAR is recorded directly for all 823 retained records, so no imputed installation date enters the analysis and the dominant covariate carries no imputation artifact (Supplementary Material S9).

Two further checks address geographic structure. A regional stratification of the test-set discrimination across the three largest PHMSA territories, Texas, Louisiana, and the federal offshore waters, yields concordance indices of 0.895, 0.855, and 0.887 (Supplementary Material S10). A leave-region-out analysis, which trains with an entire dominant region excluded and tests on that region, yields 0.890 for Texas, 0.851 for Louisiana, and 0.824 for the offshore waters, with integrated Brier scores of 0.059, 0.068, and 0.110 (Supplementary Material S12). Discrimination therefore survives the removal of a dominant region, whereas calibration degrades on the offshore cohort, whose age distribution differs most from the onshore corpus. Asset-level uncertainty was quantified by refitting the pipeline on 100 bootstrap resamples of the training set. At $t = 30$ years, the 95% band on the predicted survival is 0.005 wide for a low-risk asset, 0.035 for a median-risk asset, and 0.013 for a high-risk asset, so on the probability scale the bands contract toward both extremes of the risk spectrum and the useful resolution is concentrated in the middle of the range (Supplementary Material S13).

Finally, two checks address the composition of the corpus and the formulation of the outcome. A pre-screening sensitivity analysis quantifies the effect of the engineering-validity screen: on the unscreened corpus ($N = 839$, which retains sixteen records with implausible covariate values) the concordance index is 0.871 ± 0.014 and the integrated Brier score is 0.0609 ± 0.0045 , both within one standard deviation of the screened-corpus values of 0.871 ± 0.017 and 0.0614 ± 0.0056 , so the reported conclusions are not driven by the screening decision (Supplementary Material S16). A reformulation of the problem as binary classification of incident severity, in which the pipe age moves from the outcome to the covariates and the incident-only sampling ceases to be a limitation, is reported in Supplementary Material S17. It attains an area under the ROC curve of 0.767 ± 0.038 and a Brier score of 0.161 against a no-skill reference of 0.198, and its net benefit exceeds the inspect-all strategy across the whole tested range of decision thresholds. A decision-curve analysis of the survival model itself, at the twenty-year horizon and with the event rate in the selected group estimated by the Kaplan–Meier method, keeps a positive net benefit across threshold probabilities from 0.05 to 0.50, whereas the inspect-all strategy turns negative above 0.08 (Supplementary Material S8).

4. Discussion

4.1. Origin of the Discrimination–Calibration Divergence

High discrimination does not guarantee correct calibration, because the two are independent axes of model quality [36]; see also [21,22]. Reports of gradient-boosted survival models that rank well and calibrate poorly are common enough that the pattern is often treated as a property of the architecture, and the previous version of this manuscript explained it by an invariance of the Cox partial likelihood under monotone transformations of the predictor. That explanation is mathematically incorrect, and we withdraw it. The partial likelihood evaluates $\exp(\eta(x_i))$ against the sum of $\exp(\eta(x_j))$ over the risk set, so it depends on the differences in the linear predictors and not on their ranks alone; the only transformation it does not see is the addition of a constant, which cancels between numerator and denominator. The quantity that is invariant under an arbitrary strictly increasing map is the concordance index, not the likelihood. The distinction carries the argument of this paper. Precisely because the partial likelihood identifies the scale of $\hat{\eta}$ up to a shift, the fitted linear predictor is a well-determined quantity and is the one the Breslow estimator expects. The linear predictor recovered from the fitted booster spans $[-5.3, 9.1]$, an ordinary range for a Cox-type model that poses no numerical difficulty for the estimator. The failure appears only after the exponential of that predictor, which spans four orders of magnitude, is passed to an estimator that exponentiates a second time.

The simulation of Section 3.2 separates the two candidate mechanisms, because it holds the fitted booster fixed and varies only the quantity supplied to the Breslow estimator. Supplying the estimated margin gives a mean absolute deviation of the survival function of 0.073 against the generating truth, which is the genuine estimation error of the booster. Supplying the hazard ratio in its place raises it to 0.398, a factor of 5.5, while the concordance index remains 0.723 in both cases. The scale substitution accounts for the bulk of the observed calibration failure, and the residual attributable to the booster itself is small. The divergence in the pipeline corpus therefore resolves in the opposite direction to the usual reading: with the scale corrected, XGBoost-Cox attains both the higher concordance index and the lower integrated Brier score, and the trade-off does not exist. A post hoc correction applied on top of the erroneous scale recovers most of the loss, because $\log(1 + \exp(\hat{\eta}))$ approaches $\hat{\eta}$ from above once $\hat{\eta}$ exceeds about two. Still, such a correction is an approximate inverse of an unintended operation rather than an estimate of a calibration map. It fails on the 18.2% of objects with $\hat{\eta} < 0$, where the mean absolute discrepancy is 2.03 on the log-hazard scale, and it carries five times the variance of the direct computation across splits (Supplementary Material S2). The implication is procedural. A pipeline that couples a boosting library to an external baseline-hazard estimator should assert the predictor scale explicitly, at the cost of one additional call to `predict()`, and should state which convention was used.

4.2. Practical Implications

The numerical results translate directly into guidance for model selection in time-to-failure estimation. On this corpus, XGBoost-Cox on the correct predictor scale dominates Cox PH on both axes, so it is the preferred choice for ranking assets in an inspection queue and for computing absolute failure probabilities within an RBI framework alike. Cox PH retains two advantages that neither metric captures, namely: directly interpretable coefficients on the log-hazard scale and support for delayed entry, which make it the appropriate instrument for the truncation diagnostic in Section 3.6. The random survival forest, which does not rely on the proportional-hazards assumption and does not use a post hoc baseline reconstruction, is the weakest of the three on this corpus on both axes, which indicates that the proportional-hazards structure is not the binding constraint

here; it remains useful as a reference against which that assumption can be checked on other corpora.

Two operational qualifications follow from the temporal results. The risk ranking produced by the model is stable across calendar eras and can be carried forward, whereas the absolute probability level cannot, because the reporting base rate has changed substantially over the four decades covered by the registry. A deployment should therefore re-estimate the baseline hazard on contemporaneous data at regular intervals while retaining the trained booster. The second qualification concerns cohorts. Predictions for manufacturing generations absent from the training data are not supported, as the extrapolation test of Section 3.7 shows directly, and a deployment should flag assets whose installation year falls outside the training range rather than score them.

4.3. Outlook

The methodology was demonstrated on pipeline corrosion as a convenient test case with publicly available data, clearly defined covariates, and a normative classification of degradation factors. Both the scale diagnostic and the validation protocol are architecture-agnostic, because they rely only on a real-valued risk score and a post hoc baseline-hazard estimator, and neither depends on the specific physical mechanism of degradation. In future work, we plan to apply the same approach to other complex degrading systems, in particular to materials under combined thermal and radiation loading, where survival modelling of property degradation faces the same statistical structure of right-censored time-to-failure observations and the same coupling between a flexible learner and a non-parametric baseline estimator.

4.4. Limitations

Five limitations should be acknowledged. First, PHMSA records incidents rather than inspections or the population at risk. The denominator of the population exposed to corrosion is unobserved, the estimand is conditional on the occurrence of a reportable corrosion incident, and the non-informativeness of censoring by non-significant incidents cannot be tested from the registry alone (Section 2.1). The severity classification in Supplementary Material S17 is the formulation of the same data that avoids this conditioning, at the cost of abandoning the time-to-event structure. Second, the observation window fixes the support of the outcome variable, and the calendar covariates absorb part of that geometry. Section 3.5 shows that removing them costs about two-thirds of the concordance above chance, and Section 3.6 shows that the installation-year coefficient falls by 18.4% under a delayed-entry risk set. The share of the reported discrimination that reflects degradation rather than window geometry is therefore bounded but not identified. Third, the corrosion subset contains no direct electrochemical parameters. Environmental indicators act as proxies, and coupling with field electrochemical measurements would strengthen the physical interpretation. Fourth, the covariates are static. An extension to time-varying inputs, such as seasonal corrosion-rate cycles and operating-pressure history, requires the time-varying extension of the Cox framework. A concrete pipeline would combine PHMSA records with gridded climate reanalysis at annual resolution. Fifth, the absolute probabilities are era-specific, as the out-of-time validation of Section 3.7 shows, and the transfer of the methodology to other degradation systems should be read as a testable hypothesis rather than as a demonstrated property.

A further caveat concerns competing risks. Non-corrosion failure mechanisms and decommissioning events do not appear in the corrosion-filtered corpus, and non-significant corrosion incidents are treated as non-informative right censoring. Under this convention, the fitted quantity is a cause-specific hazard of significant corrosion incidents. A

subdistribution (Fine–Grey) formulation [37] would yield different absolute probabilities if the competing events were dependent. A cause-specific comparison within the corpus is precluded by construction, because the CORROSION filter removes the competing-event records. Linking the corrosion subset back to the full PHMSA registry to enable a Fine–Grey analysis is a direction for future work.

5. Conclusions

This work compares three survival models, Cox proportional hazards regression, the random survival forest, and XGBoost with the survival:cox objective, on a corrosion subset of 823 incidents from the PHMSA pipeline incident database (USA) spanning 1986–2025 and retained after an engineering-validity screen of the covariates. The comparison is governed by a property of the interface between the two libraries involved. For the survival:cox objective, XGBoost returns the hazard ratio $\exp(\hat{\eta})$ from predict(), whereas the Breslow estimator expects the linear predictor $\hat{\eta}$ and applies the exponential itself; chaining the two evaluates $\exp(\exp(\hat{\eta}))$. Because both maps are strictly increasing, the concordance index is identical under either convention, and the defect is invisible to discrimination-based validation, while the estimated survival function is destroyed. A simulation against a known survival function places the magnitude at a factor of 5.5 in the mean absolute deviation of $\hat{S}(t \mid x)$, from 0.073 to 0.398, and a factor of 2.6 in the integrated Brier score, at an unchanged concordance index. On the pipeline corpus, the erroneous scale inflates the IBS to 0.356 ± 0.043 over 20 random splits, and prevention of the associated numerical overflow by clipping collapses the concordance index to 0.619 ± 0.026 . On the correct scale XGBoost-Cox attains $C = 0.871 \pm 0.017$ and $IBS = 0.0614 \pm 0.0056$ and exceeds Cox PH (0.853 ± 0.018 and 0.0717 ± 0.0067) and the random survival forest (0.807 ± 0.020 and 0.0751 ± 0.0030) on both axes simultaneously, with two-sided Wilcoxon $p = 1.9 \times 10^{-6}$ throughout, no post hoc calibration, and a measurable degradation when one is added. It also attains the lowest expected calibration error at every horizon from 10 to 40 years. The feature-importance hierarchies recovered independently by the three models span the corrosion-factor categories of ISO 8044. Out-of-time validation shows that the discrimination persists across calendar eras ($C = 0.85$ – 0.87), whereas the absolute probabilities require periodic re-estimation of the baseline hazard on contemporaneous data, and extrapolation across manufacturing cohorts is not supported. Within these boundaries, XGBoost-Cox on the correct predictor scale is suitable both for inspection prioritization and for quantitative risk computation within RBI frameworks, provided that the predictor scale is asserted explicitly rather than assumed.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/app16167884/s1>. Table S1: Leakage controls on 20 stratified splits; Table S2: Scale and rescaling configurations on the same 20 splits; Table S3: Mean predicted survival by horizon against the Kaplan–Meier reference; Table S4: Out-of-time validation and cohort extrapolation on the corrected predictor scale; Table S5: Standardized Cox coefficients under the naive risk set and under delayed entry; Table S6: Reading of Table S5; Table S7: Leave-region-out validation; Table S8: Bootstrap 95% bands on the predicted survival at $t = 30$ years; Table S9: Missingness profile of the covariates; Table S10: Corpus composition by decade and reporting regime; Table S11: Effect of the engineering-validity screen; Table S12: Records excluded by the engineering-validity screen; Table S13: Severity classification, four configurations over 20 random splits; Table S14: Severity classification, covariate composition; Table S15: Severity classification, forward transfer in time; Figure S1: Integrated Brier score by configuration over 20 random splits; Figure S2: Calibration diagrams at $t = 10, 20, 30, 40$ years; Figure S3: Kaplan–Meier curves of the temporal cohorts; Figure S4: VIF and Spearman correlation heatmap of numeric covariates; Figure S5: Partial dependence of the linear predictor on the five most influential features of XGBoost-Cox; Figure S6: Net benefit

of the inspection decision for the survival model at the twenty-year horizon; Figure S7: Bootstrap 95% bands on the predicted survival for representative assets; Figure S8: Exploratory diagnostics of the corrosion corpus; Figure S9: Severity classification, ROC curves and reliability diagram; Figure S10: Forward transfer in time, base-rate drift and calibration error; Figure S11: Net benefit of the inspection decision for the severity model; Figure S12: Severity classification, area under the ROC curve by covariate composition.

Author Contributions: Conceptualization, A.V.S.; methodology, A.V.S.; software, A.V.S.; formal analysis, A.V.S.; investigation, A.V.S.; data curation, A.V.S.; writing—original draft preparation, A.V.S.; writing—review and editing, A.S.D.; visualization, A.V.S.; supervision, A.S.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research has been funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Programme No BR24993119).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The PHMSA incident registry is openly available at <https://www.phmsa.dot.gov>. Processed datasets and the processing and experiment code can be provided by the corresponding author upon reasonable request.

Acknowledgments: All computations were performed with Python 3.10.20, scikit-survival 0.22.2, XGBoost 3.2.0, lifelines 0.30.0, NumPy 1.26.4, scikit-learn 1.3.2, and pandas 2.2.2, with fixed random seeds 0–19 for all reported results.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ADM	accelerated-degradation model
AFT	accelerated failure time
API	American Petroleum Institute
C-index	concordance index
Cox PH	Cox proportional hazards model
ECE	expected calibration error
GDP	gross domestic product
HR	hazard ratio
IBS	integrated Brier score
IPCW	inverse probability of censoring weighting
IR	imbalance ratio
ISO	International Organization for Standardization
KM	Kaplan–Meier estimator
PHMSA	Pipeline and Hazardous Materials Safety Administration
PoF	probability of failure
RBI	risk-based inspection
RSF	random survival forest
SD	standard deviation
VIF	variance inflation factor

References

1. Iannuzzi, M.; Frankel, G.S. The carbon footprint of steel corrosion. *npj Mater. Degrad.* **2022**, *6*, 101. [[CrossRef](#)] [[PubMed](#)]
2. Koch, G.H.; Brongers, M.P.H.; Thompson, N.G.; Virmani, Y.P.; Payer, J.H. *Corrosion Cost and Preventive Strategies in the United States*; Report No. FHWA-RD-01-156; Federal Highway Administration: McLean, VA, USA, 2002.
3. Hou, B.; Li, X.; Ma, X.; Du, C.; Zhang, D.; Zheng, M.; Xu, W.; Lu, D.; Ma, F. The cost of corrosion in China. *npj Mater. Degrad.* **2017**, *1*, 4. [[CrossRef](#)]

4. Knisz, J.; Eckert, R.; Gieg, L.M.; Koerdt, A.; Lee, J.S.; Silva, E.R.; Skovhus, T.L.; An Stepec, B.A.; Wade, S.A. Microbiologically influenced corrosion—more than just microorganisms. *FEMS Microbiol. Rev.* **2023**, *47*, fuad041. [\[CrossRef\]](#) [\[PubMed\]](#)
5. Cao, Q.; Pojtanabuntoeng, T.; Esmaily, M.; Thomas, S.; Brameld, M.; Amer, A.; Birbilis, N. A review of corrosion under insulation: A critical issue in the oil and gas industry. *Metals* **2022**, *12*, 561. [\[CrossRef\]](#)
6. Diao, Y.; Yan, L.; Gao, K. Improvement of the machine learning-based corrosion rate prediction model through the optimization of input features. *Mater. Des.* **2021**, *198*, 109326. [\[CrossRef\]](#)
7. Coelho, L.B.; Zhang, D.; Van Ingelgem, Y.; Steckelmacher, D.; Nowé, A.; Terryn, H. Reviewing machine learning of corrosion prediction in a data-oriented perspective. *npj Mater. Degrad.* **2022**, *6*, 8. [\[CrossRef\]](#)
8. Ben Seghier, M.E.A.; Höche, D.; Zheludkevich, M. Prediction of the internal corrosion rate for oil and gas pipeline: Implementation of ensemble learning techniques. *J. Nat. Gas Sci. Eng.* **2022**, *99*, 104425. [\[CrossRef\]](#)
9. Hussain, M.; Zhang, T.; Chaudhry, M.; Jamil, I.; Kausar, S.; Hussain, I. Review of prediction of stress corrosion cracking in gas pipelines using machine learning. *Machines* **2024**, *12*, 42. [\[CrossRef\]](#)
10. Rachman, A.; Ratnayake, R.M.C. Machine learning approach for risk-based inspection screening assessment. *Reliab. Eng. Syst. Saf.* **2019**, *185*, 518–532. [\[CrossRef\]](#)
11. Campari, A.; Vianello, C.; Ustolin, F.; Alvaro, A.; Paltrinieri, N. Machine learning-aided risk-based inspection strategy for hydrogen technologies. *Process Saf. Environ. Prot.* **2024**, *191*, 1239–1253. [\[CrossRef\]](#)
12. American Petroleum Institute. *API Recommended Practice 580: Risk-Based Inspection*, 3rd ed.; American Petroleum Institute: Washington, DC, USA, 2016.
13. American Petroleum Institute. *API Recommended Practice 581: Risk-Based Inspection Methodology*, 3rd ed.; American Petroleum Institute: Washington, DC, USA, 2016.
14. Cox, D.R. Regression models and life-tables. *J. R. Stat. Soc. Ser. B* **1972**, *34*, 187–220. [\[CrossRef\]](#)
15. Ishwaran, H.; Kogalur, U.B.; Blackstone, E.H.; Lauer, M.S. Random survival forests. *Ann. Appl. Stat.* **2008**, *2*, 841–860. [\[CrossRef\]](#)
16. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. [\[CrossRef\]](#)
17. Khan, F.; Yarveisy, R.; Abbassi, R. Cross-country pipeline inspection data analysis and testing of probabilistic degradation models. *J. Pipeline Sci. Eng.* **2021**, *1*, 308–320. [\[CrossRef\]](#)
18. Soomro, A.A.; Mokhtar, A.A.; Kurnia, J.C.; Lashari, N.; Lu, H.; Sambo, C. Integrity assessment of corroded oil and gas pipelines using machine learning: A systematic review. *Eng. Fail. Anal.* **2022**, *131*, 105810. [\[CrossRef\]](#)
19. Ben Seghier, M.E.A.; Keshtegar, B.; Tee, K.F.; Zayed, T.; Abbassi, R.; Trung, N.T. Prediction of maximum pitting corrosion depth in oil and gas pipelines. *Eng. Fail. Anal.* **2020**, *112*, 104505. [\[CrossRef\]](#)
20. Liu, Y.; Bao, Y. Review on automated condition assessment of pipelines with machine learning. *Adv. Eng. Inform.* **2022**, *53*, 101687. [\[CrossRef\]](#)
21. Huang, Y.; Bazzazzadehgan, S.; Li, J.; Arabshomali, A.; Li, M.; Bhattacharya, K.; Bentley, J.P. Comparison of machine learning methods versus traditional Cox regression for survival prediction in cancer using real-world data: A systematic literature review and meta-analysis. *BMC Med. Res. Methodol.* **2025**, *25*, 243. [\[CrossRef\]](#) [\[PubMed\]](#)
22. Birolo, G.; Rossi, I.; Sartori, F.; Rollo, C.; Fariselli, P.; Sanavia, T. Beyond Cox models: Assessing the performance of machine-learning methods in non-proportional hazards and non-linear survival analysis. *Comput. Biol. Med.* **2025**, *198*, 111176. [\[CrossRef\]](#) [\[PubMed\]](#)
23. ISO 8044:2024; Corrosion of Metals and Alloys—Vocabulary. International Organization for Standardization (ISO): Geneva, Switzerland, 2024.
24. Platt, J.C. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*; Smola, A.J., Bartlett, P., Schölkopf, B., Schuurmans, D., Eds.; MIT Press: Cambridge, MA, USA, 1999; pp. 61–74.
25. Niculescu-Mizil, A.; Caruana, R. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*; ACM: New York, NY, USA, 2005; pp. 625–632. [\[CrossRef\]](#)
26. van Houwelingen, H.C. Validation, calibration, revision and combination of prognostic survival models. *Stat. Med.* **2000**, *19*, 3401–3415. [\[CrossRef\]](#)
27. Pipeline and Hazardous Materials Safety Administration (PHMSA). Pipeline Incident Data; U.S. Department of Transportation. Available online: <https://www.phmsa.dot.gov> (accessed on 10 May 2026).
28. Mann, H.B. Nonparametric tests against trend. *Econometrica* **1945**, *13*, 245–259. [\[CrossRef\]](#)
29. Kendall, M.G. *Rank Correlation Methods*; Griffin: London, UK, 1948.
30. Breslow, N.E. Discussion on Professor Cox’s paper. *J. R. Stat. Soc. Ser. B* **1972**, *34*, 202–220. [\[CrossRef\]](#)
31. Pölsterl, S. scikit-survival: A library for time-to-event analysis built on top of scikit-learn. *J. Mach. Learn. Res.* **2020**, *21*, 1–6.
32. Harrell, F.E.; Lee, K.L.; Mark, D.B. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med.* **1996**, *15*, 361–387. [\[CrossRef\]](#)

33. Graf, E.; Schmoor, C.; Sauerbrei, W.; Schumacher, M. Assessment and comparison of prognostic classification schemes for survival data. *Stat. Med.* **1999**, *18*, 2529–2545. [[CrossRef](#)]
34. Kvamme, H.; Borgan, Ø. The Brier score under administrative censoring: Problems and a solution. *J. Mach. Learn. Res.* **2023**, *24*, 1–26. [[CrossRef](#)]
35. Wilcoxon, F. Individual comparisons by ranking methods. *Biom. Bull.* **1945**, *1*, 80–83. [[CrossRef](#)]
36. Van Calster, B.; McLernon, D.J.; van Smeden, M.; Wynants, L.; Steyerberg, E.W. Calibration: The Achilles heel of predictive analytics. *BMC Med.* **2019**, *17*, 230. [[CrossRef](#)] [[PubMed](#)]
37. Fine, J.P.; Gray, R.J. A proportional hazards model for the subdistribution of a competing risk. *J. Am. Stat. Assoc.* **1999**, *94*, 496–509. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.