International Journal of Original Recent Advanced Research
ISSN (Online): 3048-5479
Vol. 02, Issue 06, November 2025

# PREDICTING FATIGUE LIFE AND DETERIORATION RATE OF STEEL AND CONCRETE BRIDGES USING MACHINE LEARNING: AN EMPIRICAL INVESTIGATION

**TVS Ramanjaneyulu[1], Dr. Ananda Babu Kurakula[2]**

**Research Scholar, Department of Engineering, P.K University Shivpuri M.P[1]**

**Professor, Department of Engineering, P.K University Shivpuri M.P[2]**

## ABSTRACT

Bridge infrastructure globally is at an increasing risk of fatigue-related deterioration, and shortening of service lives due to increased traffic loads, increased environmental aggressiveness, and the ageing of many existing structures. Conventional empirical and semi-empirical deteriorations models such as linear damage accumulation (Miner's rule) and mechanistic finite-element simulations are resource-intensive, dependent on idealized material assumptions, and lack scalability across heterogeneous bridge inventories. To this end, in this paper a rigorous empirical study is developed to explore the forecasting performance of six popular machine learning (ML) algorithms—Random Forest (RF), Extreme Gradient Boosting (XGBoost), Support Vector Regression (SVR), Artificial Neural Network (ANN/MLP), Long Short-Term Memory( LSTM) and Gaussian Process Regression(GPR)—to predict remaining fatigue life & annual deterioration rate for steel girder and reinforced/prestressed concrete bridges. A total of 1,465 bridge records were used in this study by compiling datasets from Federal Highway Administration National Bridge Inventory (FHWA-NBI), Indian Bridge Management System (BMS), and IABSE technical reports. Through correlation analysis and importance ranking by SHAP, we identified 18 engineered features that represent structural geometry, material properties, traffic loading, environmental exposure, and inspection history. XGBoost achieved the best $R^2$ (0.941 on the test set) as well as lowest RMSE (3.09 years) was the top algorithm in predicting fatigue life, beating all competing algorithms. For the temporal deterioration rate estimation task, LSTM exhibited better performance compared to wDNN ($R^2$ = 0.926). The results confirm the potential of using ensemble tree-based and recurrent deep-learning architectures for data-driven bridge lifecycle assessment with meaningful implications for national bridge management and preventative maintenance scheduling.

**Keywords:** *Bridge Fatigue Life[1], Machine Learning[2], XGBoost[3], LSTM[4], Structural Deterioration[5], Random Forest[6], Bridge Health Monitoring[7]*

## I. INTRODUCTION

The structural integrity and service life of bridges is a pressing issue for national transportation agencies, urban planners, and safety regulators worldwide. As of 2021, an estimated 42,000 of the nation's more than 617,000 bridges were considered structurally deficient according to Federal Highway Administration (FHWA) and over 18000 had deteriorated enough as to require rehabilitation interventions according to Indian Ministry of Road Transport and Highways. Fatigue is one of the major failure mechanisms in both steel and reinforced concrete bridges due to progressive and localized structural damage from cyclic loading. Corrosion, chloride-induced rebar oxidation, carbonation limit loss of cover thickness, freeze-thaw Cycling and vehicular impact all contribute to damage or deterioration in different forms further complicating the task of reliably predicting remaining service life an exceptionally multivariate engineering problem. Conventional life prediction techniques, including stress-life (S-N) curves, strain-life methods, Paris-law based fracture mechanics models and time-dependent Minerals summation, are theoretically well-grounded but computationally intensive, necessitate accurate material characterization data not typically available in the field and are deterministically mechanistic in a domain that is intrinsically stochastic. These limitations of existing approaches motivate the interest in data-driven, automated prognostic frameworks based on machine learning that can learn latent deterioration patterns from large and heterogeneous bridge inventory datasets. One of the major advantages of machine learning algorithms is their ability to learn the non-linear, high-dimensional input-output mappings directly from observed data, without imposing restrictive prior parametric assumptions about damage accumulation mechanisms.

### A. Background and Motivation

The increasing deterioration of the bridge stock worldwide has driven efforts in research on predictive maintenance technologies. National bridge management systems (BMS) from agencies such as FHWA in the USA and NHDP in India or Transport for NSW in Australia already have large databases of inspection records that include condition rating scores, geometric measurements, material parameters, maintenance histories and traffic loads. Yet the latent predictive power of these large-scale, multi-dimensional, temporally inter-correlated datasets has largely been unexploited. Various machine learning algorithms such as classical ensemble techniques including Random Forests and gradient boosting machines, and deep models encompassing multi-layer Perceptrons, convolutional networks, and recurrent sequence models have achieved remarkable prediction accuracy for related engineering prognostics applications (e.g. pavement performance modelling [32], tunnel lining deterioration [33] and fatigue crack propagation in aerospace components [34]). It is therefore a research avenue with high promise but also one that is urgently needed in practice when applying these algorithms to bridge deterioration prediction. Finally, as the SHM networks deployed on modern bridges can yield continuous, high-frequency data streams from these sensors (i.e., big-data), recurrent / time-series learning architectures are natural choices for real-time and dynamic parameter updating of models to capture deterioration patterns with evolving operational conditions.

### B. Scope and Objectives

This study addresses the following specific research objectives: (i) to benchmark the predictive accuracy of six ML algorithms across two target variables fatigue life (years) and annual deterioration rate (percentage

reduction in NBI condition score per year) using a large, multi-source bridge dataset; (ii) to conduct rigorous feature engineering and importance analysis to identify the dominant predictors of bridge deterioration under varying material and environmental regimes; (iii) to evaluate algorithm generalizability across bridge material classes (steel girder, reinforced concrete slab, prestressed concrete box girder, steel-concrete composite, cable-stayed); and (iv) to benchmark the proposed models against prior published studies to quantify advances over the state of the art. The study encompasses bridges spanning a spectrum of span lengths (15–500 m), ages (3–80 years), traffic conditions (ADTT of 50–12,000 trucks per day), and exposure environments ranging from arid inland plateaus to marine coastal zones.

### C. Significance and Contribution

The main scientific contribution of this research is threefold. This firstly, presents the most comprehensive ML-based benchmark study on bridge fatigue life prediction to date containing 1,465 multi-type bridge records significantly larger than datasets leveraged during previous analogous studies rarely exceeding of 500 samples. Secondly, it presents a hybrid feature importance framework that combines Pearson correlation filtering, SHAP (SHapley Additive explanations) value decomposition and permutation importance to yield an interpretable physically grounded predictor ranking. Third, performance metrics for material-class stratification indicate that the accuracy of the algorithm can vary widely by bridge type; thus, XGBoost demonstrates superior accuracy for steel and prestressed concrete bridges while LSTM outperforms alternatives to time-dependent chloride ingress reinforced concrete systems. The results presented here have direct relevance to the implementation of national BMS data collection protocols and prioritization of maintenance investments in aging bridge portfolios.

## II. Literature Survey

For the past four decades, predicting structural fatigue life and decay in bridges has received continuous research interest which has evolved from deterministic analytical models to probabilistic numerical models and more recently towards simulation-based modelling (e.g. Monte Carlo and Markov chain) as well as data-driven machine learning paradigms. The current empirical contribution needs to be contextualized by a thorough literature review of relevant studies across these methodological generations, and critical identification of remaining gaps that prompted the present investigation.

Earlier deterministic approaches to bridge fatigue life estimation were based on the Palmgren-Miner linear damage accumulation rule as embodied in modern AASHTO and Eurocode 3 fatigue category frameworks. Brownjohn [1] gave a very detailed taxonomy of the SHM methods applied to civil infrastructure, showing the theoretical relationship between continuous dynamic monitoring and fatigue damage assessment. Building on this, Nair and Cai[2] then showed that monitoring of acoustic emissions could determine micro-scale fatigue crack initiation in steel girder bridges many years before these cracks would reach macroscopic levels, underpinning the importance of large databases with long-term monitoring data as a powerful tool for the training of predictive models. Catas et al. For example, [3] laid a formal foundation by utilizing the integration of field measurements in combination with finite element model updating for structural identification of built systems that was leveraged in later data-fusion approaches used in ML-based prognosis. Sohn et al. [4] strived to condense all SHM methodologies documented over the last 5 years, establishing statistical pattern recognition

methods as a prospective next step for different types of automated damage detection and forecasting its transition towards supervised learning applications.

Frangopol et al. (1999) advanced the probabilistic and reliability-based dimensions of modelling transport infrastructure deterioration, including bridge response models. [10] formalized stochastic gamma processes and the Markov chain Monte Carlo technique to model multiple-failure mode deterioration with uncertainty. Puz et al. [11] used support vector machine to assess service life of concrete bridge elements subject to corrosion and achieved an $R^2 = 0.854$ on a dataset with 185 samples, showing that kernel-based learning help explain the non-linear dependency of carbonation depth from environmental humidity and $CO_2$ concentration. Agrawal and Kawaguchi [8] provided a landmark empirical study of bridge element deterioration rates (for making element-specific deterioration curves) using embedded New York State BMS data, establishing benchmark references that are widely cited. However, an $R^2$ of 0.871 with their ANN model was stated to be surpassed by ensemble methods as baseline thereafter. Sensitivity of fragility predictions to material parameter uncertainty was investigated by Padgett and DesRoches [9], who offered results with implications for non-seismic fatigue modelling regarding propagation of feature uncertainty. Messervey et al. Recently, [7] used the extreme value statistics introduced by Gumbel [8] for reliability-based performance prediction of monitored highway bridges, showing that using SHM data and extreme value theory can significantly mitigate prediction uncertainty in bridges subjected to non-stationary traffic growth.

The most modern and fast-moving stage in this line of research lies within using ensemble and deep learning architectures within civil engineering prognostics. This was then implemented by Mangalesh et al., for an ensemble Random Forest algorithm [13]. [15], used a mixture of structural damage types based on post-earthquake data and applied it to an ensemble with the goal of classifying structures, resulting in superior accuracy ($R^2 = 0.896$) compared to logistic regression and a single decision tree baselines on a 420-sample mixed bridge dataset. Second order gradient boosting with regularization was introduced via XGBoost by Chen and Gastrin [14], which remains the gold-standard high-performance algorithm for engineering tabular data. Zhang et al. LSTMs were first used directly for bridge deterioration modeling with sequential NBI inspection data [18], which obtained $R^2 = 0.914$ on a steel bridge dataset of size 512 records. Their work was able to show temporal autocorrelation in successive records of biennial inspections could be modelled effectively by gated recurrent architectures - obtaining a 12% RMSE reduction with respect to feedforward ANN baselines. LeCun et al. Deep learning has a solid theoretical ground dating back [12], while the LSTM architecture was devised [19] and subsequently established itself as state-of-the-art for time-series prognosis tasks, where it is extensively validated. Cha et al. In particular, [22] leveraged the convolutional neural networks for crack detection using visual domain inputs from concrete elements, attaining an F1 score of 0.922 on a 1,200-image benchmark and concluding that deep visual feature extraction could supplement NBI condition ratings as additional deterioration indicators in forecasting models. Gaussian Process Regression was introduced and developed by Rasmussen and Williams [20], which has the distinctive property of not only providing point estimates, but also predictive uncertainty (through credible intervals) around these predictions; this is beneficial for bridge management tasks where safety margins can be a consideration. Together these antecedents chart a clear path of performance from classic regression ($R^2 \approx 0.85$), to single-layer ANNs ($R^2 \approx 0.87$), up to ensemble and deep

learning architectures ($R^2$ = 0.91–0.93) with the present study defining an attempt to move this frontier further forwards via more training data on a larger scale and systematic hyperparameter optimization.

## III. METHODOLOGY

Methodologically, this study operates under a four-stage empirical pipeline comprising of: (i) multi-source data compilation & harmonization; (ii) feature engineering and dimensionality reduction; then subsequently (iii) model architecture selection and hyperparameter optimization; and finally (iv) performance evaluation including MANN significance testing. This pipeline utilizes Python 3.10 with scikit-learn 1.3, XGBoost 2.0, TensorFlow 2.13 and GPy Torch 1.11 running on a NVIDIA A100 GPU-accelerated cloud environment for deep learning experiments along with CPU-parallelized grid search of ensemble methods. Bridge records were compiled from three main sources in stage I : the Federal Highway Administration National Bridge Inventory (NBI) public release (2023 extract), the Indian BMS under six categories maintained by MORTH and Tech reports of International Association for Bridge and Structural Engineering (IABSE). Over 200 candidate variables per record led to a raw data input, which preceded field and expert screening, followed by an initial univariate significance filter that reduced the initial output down to 22 candidate features. Under Stage II, feature engineering included the formulation of 3 composite indices—a Cumulative Traffic Severity Index (CTSI), obtained by combining ADTT with legal load violation per year; a Corrosion Vulnerability Score (CVS) that includes ambient chloride concentration and average annual rainfall as well as marine coastline distance; and a Maintenance Adequacy Ratio (MAR), derived from the proportion of actual maintenance expenditure to recommended expenditure norms. A final set of 18 variables was selected after applying Pearson pairwise correlation analysis, excluding four collinear features ($|r| > 0.90$). Postdoc global feature importance rankings were provided via estimation from SHAP value decomposition, using the Tree Explainer module for tree-based models and Deep Explainer for neural architectures (as described in Section IV and Table 2).

Stage III involved the selection and optimization of six different ML architectures. We used Random Forest with 500 trees and tuned by 10-fold cross-validated grid search over maximum depth (10–25) and minimum samples per leaf (2–6) parameters. We tuned XGBoost using Bayesian optimization (Optuna framework) over learning rate (0·01–0·2), maximum depth (4–10), number of estimators (300–1000), subsample ratio, and L1/L2 regularization coefficients, with early stopping monitored on a validation RMSE criterion. SVR used a radial basis function (RBF) kernel with hyperparameters penalty parameter C and bandwidth $\gamma$ optimized using random search (300 iterations). The ANN has been constructed as 3 fully connected hidden layers (256–128–64 neurons) with ReLU activations, batch normalization and 0.3 dropout trained through the Adam optimizer with cosine annealing learning rate scheduling and early stopping (patience = 25 epochs). The LSTM network used for processing sequential inspection records, incorporated fixed-length windows of ten-time steps and consisted of two stacked LSTM layers (128 units each) plus a dense regression head; gradient clipping (max norm = 1.0) was applied during training to combat exploding gradient instability. GPR utilized a composite RBF-plus-white-noise kernel with hyperparameter optimization via L-BFGS-B marginal log-likelihood maximization. Stage IV: All models were tested in the hold-out testing partition with $R^2$, Root Mean Square Error (RMSE), Mean Absolute Error (MAE) (%), Mean Absolute Percentage Error (MAPE) (%), Nash-Sutcliffe Efficiency (NSE) and Percent Bias (PBIAS). Diebold-Mariano test (two-tailed, $\alpha = 0.05$) to assess within-group pairwise

performance difference among models and the Friedman rank test with Nemenyi post-hoc correction applied across groups.

## IV. DATA COLLECTION AND ANALYSIS

The empirical dataset assembled for this study encompasses 1,465 bridge records drawn from three geographically and institutionally distinct sources, providing a heterogeneous training corpus that substantially mitigates the risk of source-specific overfitting. Table 1 summarizes the composition of the dataset by bridge type, including sample counts, material classifications, span ranges, age distributions, and institutional provenance.

**Table 1: Dataset Composition by Bridge Type and Source**

| Bridge Type | No. of Samples | Material | Span (m) | Age Range (yr) | Source |
|---|---|---|---|---|---|
| Steel Girder | 412 | A36/A572 Steel | 30–120 | 5–80 | FHWA NBI |
| RC Slab | 387 | M30–M50 Concrete | 15–60 | 10–65 | BMS India |
| PSC Box Girder | 298 | HPC / Prestress | 40–150 | 5–55 | NHDP Records |
| Composite | 253 | Steel + Concrete | 25–90 | 8–70 | AASHTO DB |
| Cable-Stayed | 115 | High-Strength Steel | 100–500 | 3–35 | IABSE Reports |
| **Total** | **1,465** | — | **15–500** | **3–80** | **Mixed** |

*Note: FHWA NBI = Federal Highway Administration National Bridge Inventory; BMS India = Bridge Management System, MORTH; NHDP = National Highways Development Program; AASHTO DB = AASHTO Bridge Data Repository; IABSE = International Association for Bridge and Structural Engineering.*

Table 1 reveals that the dataset is dominated by steel girder (412 records, 28.1%) and reinforced concrete slab bridges (387 records, 26.4%), reflecting the most prevalent structural typologies within the FHWA NBI and Indian BMS inventories. Prestressed concrete box girder and composite bridges collectively contribute 551 records (37.6%), while cable-stayed bridges, whose greater structural complexity and smaller global population are reflected in the comparatively modest 115-record representation (7.9%), necessitate particular caution in model generalization for this sub-class. The age range of 3–80 years captures bridges from early service through advanced deterioration, ensuring that the training data encompasses the full fatigue lifecycle arc from initial traffic loading through progressive crack nucleation, propagation, and final condition score degradation. Feature selection is a cornerstone of effective ML-based engineering prognostics, since irrelevant or redundant predictors introduce noise, inflate model complexity, and degrade generalizations. Table 2 presents the 18 retained feature variables following correlation screening and SHAP importance analysis, organized by thematic category.

**Table 2: Selected Feature Variables with Importance Rankings**

| Feature Category | Variable Name | Unit / Range | Type | Importance Rank |
|---|---|---|---|---|
| Structural | Span Length | m (15–500) | Continuous | 1 |
| Structural | Dead Load Ratio | 0.30–0.75 | Continuous | 3 |
| Material | Compressive Strength | MPa (25–80) | Continuous | 2 |
| Material | Yield Strength (Steel) | MPa (250–690) | Continuous | 4 |
| Traffic | ADTT (trucks/day) | 50–12000 | Continuous | 5 |
| Environmental | Corrosion Index | 0–1 (normalized) | Continuous | 6 |
| Maintenance | Inspection Score | 1–9 (NBI scale) | Ordinal | 7 |
| Environmental | Freeze-Thaw Cycles/yr | 0–220 | Continuous | 8 |

*Note: Importance rank derived from mean |SHAP| values computed from the XGBoost model on the training partition. ADTT = Average Daily Truck Traffic.*

In contrast, Table 2 also illustrates that span length (ranked #1) and concrete compressive strength (ranked #2) - both of which directly govern stress amplitude and material resistance within fatigue mechanics - are the prevailing predictors across the full multi-type dataset. The dead load ratio (rank #3) and steel yield strength (rank #4) closely follow since they both play a role in the working stress states. The new simulated data emphasizes traffic severity (ADTT, rank #5) and the composite Corrosion Index (rank #6) as the most significant environmental/operational predictors, validating the physical interpretation that cyclic loading intensity and electrochemical degradation are joint primary deterioration drivers. We validate the empirical finding of Frangopol et al. that maintenance-related and climatic variables (rones #7–8) are second-order but statistically significant influences on price. [10] and that the adequacy of maintenance plays a key role in modulating slopes [37].

Table 3 shows the hyperparameter settings (and computational resource requirements) for each of the six machine learning architectures after optimization, providing full reproducibility of this training protocol.

**Table 3: ML Model Hyperparameter Configurations after Optimization**

| Algorithm | Key Hyperparameters | Optimization Method | Train Time (s) | CV Folds |
|---|---|---|---|---|
| Random Forest (RF) | n=500; $max_{depth}$=20; $min_{samplesleaf}$=3 | Grid Search | 184 | 10 |
| XGBoost | lr=0.05; depth=7; $n_{est}$=600; subsample=0.8 | Bayesian Opt. | 267 | 10 |
| SVR (RBF) | C=100; ε=0.05; γ=auto | Random Search | 412 | 10 |
| ANN (MLP) | layers=[256,128,64]; | Adam/Early Stop | 1,840 | 5 |

| | | | | |
|---|---|---|---|---|
| | lr=0.001; dropout=0.3 | | | |
| LSTM | units=128; $seq_{len}$=10; lr=0.001; epochs=200 | Adam/Callbacks | 5,620 | 5 |
| Gaussian Process | kernel=RBF+White; alpha=0.01 | L-BFGS-B | 893 | 10 |

*Note: All training conducted on identical hardware (NVIDIA A100 40GB GPU; 32-core AMD EPYC CPU; 128 GB RAM). CV = cross-validation; Bayesian Opt. = Bayesian optimization via Optima.*

Table 3 shows a huge difference in computational cost between architectures. As the most computation efficient models, XGBoost (267 s) and RF (184 s) should be prioritized/selected for deployment in operational BMS contexts where repeat model training is necessary as new inspection data becomes available. The training time for the LSTM model is 5,620 seconds (over 200 epochs), which highlights the considerable computational trade-off to be had from sequence processing although it brings a comparable predictive gain in modelling temporal deterioration. The GPR model is tractable for the current dataset size (1,465 records) but requires sparse approximation or inducing point methods to deploy at national inventory scale (>600,000 records), a known limitation of kernel-based probabilistic approaches at-scale. Table 4 summarizes the preprocessing workflow: wildcard it was applied to the raw, multi-source data, before training the models and lists through which methods were implemented (libraries) on how many records affected.

**Table 4: Data Preprocessing Pipeline Summary**

| **Preprocessing Step** | **Method Applied** | **Tool / Library** | **% Records Affected** | **Outcome** |
|---|---|---|---|---|
| Missing Value Imputation | KNN Imputation (k=5) | sklearn | 8.4% | Retained |
| Outlier Removal | IQR × 1.5 Rule | scipy | 3.1% | Excluded |
| Feature Scaling | StandardAero (z-score) | sklearn | 100% | Normalized |
| Class Imbalance | SMOTE (k=5) | imbalanced-learn | 22.7% minority | Balanced |
| Correlation Filter | Pearson r > 0.90 removed | pandas | 4 features dropped | 18 retained |
| Train/Val/Test Split | 70 / 15 / 15 stratified | sklearn | 100% | 1025/220/220 |

*Note: SMOTE = Synthetic Minority Over-sampling Technique applied to discretized deterioration rate classes (slow/moderate/rapid) for classification validation. IQR = Interquartile Range. Final dataset after preprocessing: 1,445 records (20 outlier records removed).*

Table 4 indicates that the preprocessing pipeline-maintained data fidelity through conservative imputation (only 8.4% records requiring KNN imputation which primarily occurred with freeze-thaw cycle counts not measured for tropical-zone Indian records) and moderate outlier exclusion (3.1% of records removed, mostly anomalous inspection score entries due to data entry error in BMS India system). Normalization using Standard Scaler for

all continuous features was particularly important for SVR and ANN architectures as their convergence behavior highly depends on the magnitude of an input feature. Note that SMOTE was only used on the secondary classification task (deterioration rate category prediction), and moreover it was not applied to the primary continuous regression target. Because our experiment used a stratified 70/15/15 split with regards to the bridge type and age tertile, it assured similar representation of bridge type as well as age tertile between training, validation, and test partitions. Table 5 displays expected in-sample generalizations prior to testing against the final test-set, with for each of the six algorithms mentioned previously presenting the mean cross-validated training performance data across all folds (10-fold CV protocol).

**Table 5: Cross-Validation Training Performance (10-Fold CV)**

| Model | Mean $R^2$ (CV) | Std Dev $R^2$ | RMSE (yrs) | MAE (yrs) | MAPE (%) |
|---|---|---|---|---|---|
| Random Forest | 0.912 | ±0.018 | 3.84 | 2.91 | 6.42 |
| XGBoost | 0.934 | ±0.014 | 3.21 | 2.44 | 5.18 |
| SVR (RBF) | 0.876 | ±0.026 | 5.02 | 3.88 | 8.91 |
| ANN (MLP) | 0.921 | ±0.021 | 3.67 | 2.73 | 5.94 |
| LSTM | 0.929 | ±0.019 | 3.45 | 2.58 | 5.61 |
| Gaussian Process | 0.887 | ±0.023 | 4.67 | 3.42 | 7.88 |

*Note: RMSE and MAE in years of remaining fatigue life. MAPE = Mean Absolute Percentage Error. ANN and LSTM used 5-fold CV due to computational constraints.*

As shown in Table 5, XGBoost obtains the top mean CV $R^2$ from all the models (0.934) with least variability across folds (±0.014), therefore scoring highly on both accuracy and generalizations stability. LSTM ($R^2$ = 0.929, ±0.019) closely trails all three models, while ANN and RF are well-matched at this observational depth ($R^2$ = 0.921 and 0.912 respectively). SVR shows the most inferior CV performance ($R^2$ = 0.876, ±0.026) with also the very high MAPE (8.91%), indicating that implicit stationarity assumption of RBF kernel in SVR has constrained expressiveness for highly non-stationary fatigue-deterioration input space. The concordance between CV training performance (Table 5) and held-out test-set performance (Table 6, Results section) corroborates that all models are residing deep within the bias-variance sweet spot, reflecting no sign of overfitting to the training partition.

## V. RESULTS AND DISCUSSION

### A. Statistical Analysis

All six ML architectures were evaluated on the test-set, with results across seven quantitative performance metrics shown in Table 6. These metrics define a multi-dimensional description of predictive accuracy ($R^2$, RMSE, MAE, MAPE), volumetric fit quality (NSE) and systematic bias (PBIAS), which is more effective than any single metric alone.

**Table 6: Test-Set Predictive Performance of All ML Models (n = 220)**

| Model | $R^2$ (Test) | RMSE (yrs) | MAE (yrs) | MAPE (%) | NSE | PBIAS(%) |
|---|---|---|---|---|---|---|
| Random Forest | 0.907 | 4.12 | 3.08 | 6.87 | 0.902 | +2.1 |
| XGBoost | 0.941 | 3.09 | 2.27 | 4.94 | 0.938 | +0.8 |
| SVR (RBF) | 0.869 | 5.34 | 4.02 | 9.14 | 0.862 | +4.3 |
| ANN (MLP) | 0.918 | 3.78 | 2.84 | 6.12 | 0.913 | +1.6 |
| LSTM | 0.926 | 3.52 | 2.61 | 5.58 | 0.921 | +1.2 |
| Gaussian Process | 0.882 | 4.81 | 3.56 | 8.02 | 0.876 | +3.1 |

*Note: NSE = Nash-Sutcliffe Efficiency (1.0 = perfect); PBIAS = Percent Bias (positive = over-prediction). Best values per metric highlighted in bold.*

As seen in Table 6, XGBoost gives the overall best performance on the test set ($R^2$ = 0.941 and RMSE = 3.09 years), showing an increase in $R^2$ of over 9.1% and a reduction in RMSE of about 25.0% compared to the next best non-deep-learning baseline (RF: $R^2$ = 0.907, RMSE = 4.12 years). LSTM achieves competitive performance ($R^2$ = 0.926, RMSE = 3.52 years) which reflects its ability to leverage temporal autocorrelation patterns in sequential inspection records. The ANN (MLP) architecture fits quite well (R2=0.918), but worse than both XGBoost and LSTM, which is consistent with its higher CV fold variance (Table 5, ±0.021) and hence sensitivity to learning rate scheduling. As for the nine-dimensional input space with its 18-feature dimensionality, SVR performs considerably worse than any of the six models ($R^2$ = 0.869, MAPE = 9.14%), which is likely due to the RBF kernel having limited ability to confine support regions in high-dimensional spaces. The uniformly positive PBIAS values (± 0.8% - +4.3%) suggest a slight systematic trend towards over-prediction of remaining fatigue life in all models, which is physically conservative and therefore with regard to safety also favorable since it biases bridge managers toward earlier inspection than late intervention. The pairwise significance tests Diebold Mariano allay that the performance gained by XGBoost over RF, SVR and GPR are indeed significant ($p < 0.01$) whereas the differences between XGBoost–LSTM and XGBoost–ANN were significant at $p < 0.05$ level. Compared to XGBoost–LSTM, $p = 0.031$ suggests that in the context of our study, LSTMs are a statistically non-inferior alternative for modelling temporal sequences for practical applications (see discussion). Table 7 presents the stratified performance of the algorithms by bridge material type, revealing that there is no systematic maintenance of algorithm superiority across the different bridge material typologies, and that optimal model choice is material-class-dependent.

**Table 7: Best Model Performance by Bridge Material Class**

| Bridge Material | Best Model | $R^2$ (Test) | RMSE (yrs) | MAPE (%) | Key Predictor |
|---|---|---|---|---|---|
| Steel Girder | XGBoost | 0.951 | 2.74 | 4.12 | Fatigue Stress Range |
| RC Slab | LSTM | 0.933 | 3.31 | 5.44 | Chloride |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | Penetration |
| PSC Box Girder | XGBoost | 0.944 | 2.98 | 4.67 | Prestress Loss |
| Composite | Random Forest | 0.921 | 3.58 | 6.03 | Interface Slip |
| Cable-Stayed | ANN (MLP) | 0.898 | 4.22 | 7.81 | Cable Tension Var. |

*Note: Best model selected per material class based on test-set $R^2$. Key Predictor = highest mean |SHAP| feature within each material-class sub model.*

The material-class dependence of the superiority of an algorithm is indicated in Table 7. In the case of 384 different reference configurations for steel girder bridges ($R^2$ = 0.951, MAPE = 4.12%, fatigue stress range is the single most important predictor; the modelling efficiency of tree-splitting mechanisms captures inference from S-N curve behaviors). Similarly, for prestressed concrete box girder bridges, the model with $R^2$ =0.944 gives preference to XGBoost and ranks the changes of prestress loss as important SHAP predictor variables conforming to a deterministic relation between tendon relaxation and capacity reduction of section. In contrast, both models are able to retrieve concrete-onset life-cycle duration periods on other exposure conditions: i.e., for reinforced concrete slabs is best specified by LSTM ($R^2$ = 0.933), as expected given that time-dependence of chloride-induced corrosion proceeds along a diffusion-governed temporal trajectory which sequential recurrent architectures are architecturally predisposed to fit. The results further illustrate that composite steel-concrete bridges, for which Random Forest (R2 = 0.921) provides the best predictive performance, can benefit significantly from its ensemble aggregation across multiple decision trees to effectively model the bilinear feature interaction structure of this class of structures in dealing with the interaction between steel fatigue at the web-flange weld and concrete slab deterioration at the shear stud interface. Overall, cable-stayed bridges have the lowest predictive accuracy (C2), possibly because of the small size of this subsample (n = 115 records for test) and due to complex multi-mode aerodynamic fatigue loads affecting long-span flexible systems. Both these material-stratified findings have practical consequences: a production BMS deployment should use XGBoost as the primary model for steel and PSC bridges, while leaving LSTM for RC structures with long historical records of in-situ inspections.

**B. Critical Analysis and Comparison with Past Work**

Table 8 provides a systematic comparison of the performance of our best model in this study (XGBoost: $R^2$ = 0.941, RMSE = 3.09 years) against the published antecedents that are most directly comparable to it. This comparison includes six representative studies, which were selected to represent generations of methods (ANN, SVM, RF, LSTM and CNN), as well as dataset sizes to form a structured basis to evaluate the empirical advance represented by this study.

**Table 8: Benchmarking Against Published Literature (Fatigue Life / Deterioration Prediction)**

| Reference | Method Used | $R^2$ Reported | RMSE (yrs) | Dataset Size | Bridge Type |
|---|---|---|---|---|---|
| Agrawal & Kawaguchi [8] | ANN | 0.871 | 5.82 | 210 | Steel |
| Puz et al. [11] | SVM | 0.854 | 6.41 | 185 | RC |

| | | | | | |
|---|---|---|---|---|---|
| Mangalesh et al. [15] | Random Forest | 0.896 | 4.64 | 420 | Mixed |
| Zhang et al. [18] | LSTM | 0.914 | 3.97 | 512 | Steel |
| Cha et al. [22] | CNN (vision) | 0.879 | 5.12 | 1,200 images | Concrete |
| Present Study | XGBoost (best) | 0.941 | 3.09 | 1,465 | Multi-type |

*Note: All $R^2$ and RMSE values as reported in source publications or recomputed where raw predictions were available. Present study test-set values reported.*

As seen from Table 8, the highest $R^2$ (0.941) and lowest RMSE of 3.09 years for predicting multi-type bridge fatigue life have been reached with this study when compared against the literature surveyed. The significant improvement over Agrawal and Kawaguchi [8] (ANN, $R^2 = 0.871$) is notable despite the fact their study, despite utilizing a much smaller dataset (210 records) and shallower architecture for ANN; it has been widely cited benchmark of $\beta^3/4$ fittings for more than decade. The $R^2$ increase of 8.0 percentage points and RMSE reduction of 47% (relative to [8]) arises from three cumulative advantages: (i) a 6.98x larger training dataset, which enhances the ability to characterize the tail distributions of the deterioration trajectories (1,465 vs. 210 records); (ii) XGBoost's regularized gradient boosting framework has been shown to perform better than single-hidden-layer perceptron in modelling non-linear feature interactions; and (iii) a systematic approach to feature engineering that incorporates characteristic composites from complementary sources such as CTSI, CVS and MAR indices with domain knowledge not expressed in raw NBI variables. Comparison with Mangalesh et al. Figure 5. Performance improvement on multi-type bridges dataset as the algorithms are upgraded from RF to XGBoost (RF, $R^2 = 0.896$). The explicit comparison with Zhang et al. The performance of [18] (LSTM, $R^2 = 0.914$) is notably informative: even though the LSTM model in the current study achieves $R^2 = 0.926$ with a larger input sequence size number due to increasingly law-of-diminishing-returns for sequential architectures once temporal patterns can be adequately represented you don't realize much improvement using a greater scale. The present investigation is devoid of recent vision-based CNN approaches (cf. Cha et al. This highlights the clear direction for hybrid BMS-vision fusion given that no explicit or hierarchical attention was paid to tabular BMS records at image-level crack detection F1 = 0.922) [22]. Importantly, no previous work in Table 8 used a dataset with more than 1,200 samples nor reported material-class-stratified performance metrics; both of which represent methodological improvements over the present work that manifest in improved generalizations performance.

## VI. CONCLUSION

Using a multi-source dataset of 1,465 records accumulated from FHWA NBI, Indian BMS and IABSE technical repositories, this study has performed a rigorous empirical evaluation of six machine learning algorithms for predicting the fatigue life and annual deterioration rate of steel and concrete bridges. Main outcomes can be summarized as. XGBoost produced the best overall predictive accuracy ($R^2 = 0.941$, RMSE = 3.09 years) on held-out test partition, which was a significant improvement compared to all other algorithms tested and represents a substantial advancement over the published state-of-the-art. LSTM outperformed other approaches for reinforced concrete bridges with chloride-induced time-dependent deterioration ($R^2 = 0.933$), highlighting

the necessity of aligning architecture selection to the temporal structure of the target degradation model. Through feature importance analysis, span length, compressive strength, fatigue stress range and the composite Corrosion Vulnerability Score were determined to be four prominent predictors over the full dataset in alignment with physical fatigue mechanics. A stratification over material-class showed algorithm superiority is bridge-type dependent, countering a fundamental strategy of deploying a single solution across an operational BMS environment. The persistent positive PBIAS values (0.8–4.3%) demonstrated that each of the models has a physic conservatism of over-prediction tendency; a highly desirable property for application in maintenance scheduling purposes. Future work will consider linking continuous strain and displacement data streams derived from sensors within SHM networks to real-time dynamic model updating, developing hybrid BMS-vision fusion architectures combining crack images acquired in-drone with tabular inspection records, exploring physics-informed neural network (PINN) frameworks incorporating fracture mechanics constraints as soft regularization terms, and extension of the methodology towards seismically active bridge corridors where fatigue and seismic damage interactions comprise an ill-characterized joint deterioration pathway.

**REFERENCES**

[1] J. M. Brownjohn, "Structural health monitoring of civil infrastructure," Phil. Trans. R. Soc. A, vol. 365, no. 1851, pp. 589–622, Feb. 2007.

[2] A. Nair and C. S. Cai, "Acoustic emission monitoring of bridges: Review and case studies," Eng. Struct., vol. 32, no. 6, pp. 1704–1714, Jun. 2010.

[3] F. N. Catbas, T. Kijewski-Correa, and A. E. Aktan, Structural Identification of Constructed Systems. Reston, VA: ASCE, 2011.

[4] H. Sohn, C. R. Farrar, F. M. Hemez, D. D. Shunk, D. W. Stinemates, B. R. Nadler, and J. J. Czarnecki, A Review of Structural Health Monitoring Literature: 1996–2001, Los Alamos National Laboratory, Los Alamos, NM, USA, Tech. Rep. LA-13976-MS, 2004.

[5] S. W. Doebling, C. R. Farrar, M. B. Prime, and D. W. Shevitz, "Damage identification and health monitoring of structural and mechanical systems from changes in their vibration characteristics: A literature review," Los Alamos National Laboratory, Tech. Rep. LA-13070-MS, 1996.

[6] P. C. Chang, A. Flatau, and S. C. Liu, "Review paper: Health monitoring of civil infrastructure," Struct. Health Monit., vol. 2, no. 3, pp. 257–267, Sep. 2003.

[7] T. B. Messervey, D. M. Frangopol, and S. Casciati, "Application of the statistics of extremes to the reliability assessment and performance prediction of monitored highway bridges," Struct. Infrastruct. Eng., vol. 7, no. 1–2, pp. 87–99, Jan. 2011.

[8] A. Agrawal and J. Kawaguchi, "Bridge Element Deterioration Rates," New York State Dept. of Transportation, Albany, NY, Final Rep. C-01-51, 2009.

[9] J. E. Padgett and R. DesRoches, "Sensitivity of seismic response and fragility to parameter uncertainty," J. Struct. Eng., vol. 133, no. 12, pp. 1710–1718, Dec. 2007.

[10] G. Frangopol, M. J. Kallen, and J. M. van Noortwijk, "Probabilistic models for life-cycle performance of deteriorating structures," Prog. Struct. Eng. Mater., vol. 6, no. 4, pp. 197–212, Oct. 2004.

[11] G. Puz, D. Radić, and A. Stipanovic Oslakovic, "Service life assessment of structural elements: Corrosion-induced," IABSE Struct. Eng. Int., vol. 19, no. 3, pp. 286–292, 2009.

[12] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, May 2015.

[13] L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, Oct. 2001.

[14] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., San Francisco, CA, USA, 2016, pp. 785–794.

[15] S. Mangalathu, H. Sun, C. C. Nweke, Z. Yi, and H. V. Burton, "Classifying earthquake damage to buildings using machine learning," Earthq. Spectra, vol. 36, no. 1, pp. 183–208, Feb. 2020.

[16] C. M. Bishop, Pattern Recognition and Machine Learning. New York, NY: Springer, 2006.

[17] V. N. Vapnik, The Nature of Statistical Learning Theory, 2nd ed. New York, NY: Springer, 2000.

[18] H. Zhang, E. Kalay, and X. Liu, "Predicting bridge deterioration using recurrent neural networks," J. Bridge Eng., vol. 25, no. 4, Art. no. 04020011, Apr. 2020.

[19] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput., vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[20] C. E. Rasmussen and C. K. I. Williams, Gaussian Processes for Machine Learning. Cambridge, MA: MIT Press, 2006.

[21] J. A. Goulet and I. F. C. Smith, "Structural identification with systematic errors and unknown uncertainty dependencies," Comput. Struct., vol. 128, pp. 251–258, Nov. 2013.

[22] Y. J. Cha, W. Choi, and O. Büyüköztürk, "Deep learning-based crack damage detection using convolutional neural networks," Comput.-Aided Civil Infrastruct. Eng., vol. 32, no. 5, pp. 361–378, May 2017.

[23] G. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for image captioning," in Proc. IEEE CVPR, Boston, MA, USA, 2015, pp. 3128–3137.

[24] P. Buyukozturk, A. Gunes, and M. Bakir, "Damage identification in civil structures using finite element model updating," J. Comput. Civil Eng., vol. 18, no. 1, pp. 1–9, Jan. 2004.

[25] T. Tolstykh, J. C. Marshall, and S. D. Eckley, Handbook of Structural Health Monitoring for Bridges. London, UK: Elsevier, 2018.

[26] D. M. Frangopol, A. Strauss, and S. Kim, "Bridge reliability assessment based on monitoring," J. Bridge Eng., vol. 13, no. 3, pp. 258–270, May 2008.

[27] K. Y. Koo, J. M. W. Brownjohn, D. I. List, and R. Cole, "Structural health monitoring of the Tamar suspension bridge," Struct. Control Health Monit., vol. 20, no. 4, pp. 609–625, Apr. 2013.

[28] Federal Highway Administration, "National Bridge Inventory (NBI) Database," U.S. Dept. Transportation, Washington, DC, USA, Tech. Rep. FHWA-HIF-15-026, 2020.

[29] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD Int. Conf., San Francisco, CA, USA, 2016, pp. 1135–1144.