# Segmented XGBoost Model for Accurate Chloride-Induced Corrosion Rate Prediction in Reinforced Concrete

# Abstract: This study presents a robust machine learning framework for predicting corrosion rates in reinforced concrete structures exposed to chloride environments. Leveraging a segmented Extreme Gradient Boosting (XGBoost) architecture and Virtual Sample Generation (VSG) to handle data imbalance, the model achieves an R² score of 0.902 with a Root Mean Square Error (RMSE) of 0.0014 µA/cm². Segmentation based on corrosion magnitude allows specialized models to focus on distinct degradation mechanisms. Nearly 49% of predictions fall within a 10% error margin, with strong predictive accuracy across both low and high corrosion regimes. The results demonstrate that incorporating advanced data preprocessing, interaction features, and targeted modeling significantly enhances the reliability of corrosion forecasting tools, with practical implications for the durability design of reinforced concrete infrastructure.

# 1. Introduction

# Reinforced concrete structures exposed to aggressive environments, such as those influenced by marine conditions or deicing salts, are particularly vulnerable to chloride-induced corrosion of embedded steel reinforcement. The ingress of chloride ions through concrete pores initiates corrosion once the chloride concentration at the rebar level exceeds a critical threshold, leading to expansive rust formation, cracking, delamination, and eventual loss of structural capacity. This degradation process not only compromises the structural integrity and safety of infrastructure but also imposes significant economic burdens related to repair, rehabilitation, and service disruptions.

# Accurate prediction of corrosion rates in such environments is crucial for enabling durability-based design, optimizing maintenance schedules, and extending the service life of reinforced concrete assets. Conventional models used for corrosion rate estimation often rely on empirical or semi-empirical formulations with simplifying assumptions. These models typically consider a limited set of input parameters and assume linear or weakly nonlinear interactions, which may fail to capture the full complexity of corrosion processes influenced by multiple interacting environmental, material, and design variables.

To overcome these limitations, this study introduces an advanced machine learning-based predictive framework that leverages segmented Extreme Gradient Boosting (XGBoost) regression. This approach enables the model to account for distinct behavioral patterns across different ranges of the target variable by training separate regressors within defined segments, thereby improving accuracy and interpretability. Furthermore, to address data scarcity and enhance the model’s generalization capacity, a virtual data augmentation technique is employed to synthetically expand the dataset while preserving the underlying statistical distributions and physical consistency. By integrating segmentation with XGBoost and virtual sample generation, the proposed method aims to capture complex nonlinear relationships, mitigate data imbalance, and better represent rare but critical corrosion behaviors, ultimately offering a robust and scalable tool for corrosion rate prediction in chloride-laden environments.

Chloride-induced corrosion of steel reinforcement remains one of the most critical durability challenges in reinforced concrete (RC) structures, particularly in marine environments and regions subject to deicing salts. This degradation mechanism significantly reduces the structural integrity and service life of concrete infrastructure. Accurate prediction of corrosion rates is essential for service life modeling, maintenance planning, and durability-based design. A wide range of models, from empirical equations to advanced machine learning approaches, have been developed to predict corrosion rates. This review examines key contributions in the field, highlighting traditional models, mechanistic and probabilistic approaches, and recent advances in data-driven modeling.

1. **Reinforced Concrete Chloride-Induced Corrosion Models**

Chloride-induced corrosion is an electrochemical process initiated when chloride ions penetrate the concrete cover and exceed a threshold concentration at the reinforcement surface. This breaks down the passive oxide layer on steel, initiating corrosion that leads to expansive rust formation. The rate of corrosion is governed by several parameters, including: 1) Chloride concentration, 2) Concrete cover depth, 3) Water-to-cement ratio, 4) Relative humidity and temperature, 5) Oxygen availability, and 6) Electrical resistivity of concrete. The complex interplay among these parameters makes accurate modeling of corrosion rates highly challenging.

Understanding concrete composition and durability mechanisms remains foundational. Neville (2011) and Mehta & Monteiro (2014) provide authoritative resources on concrete materials science. Thomas (2013) highlighted the role of supplementary cementitious materials (SCMs) in enhancing chloride resistance.

Bentz (2007) studied the influence of curing conditions on hydration and porosity—two factors directly impacting chloride diffusivity. These works inform feature selection for data-driven models and ground machine learning efforts in physical phenomena.

Standardization efforts guide corrosion assessment in practice. ACI (2019) provides recommendations for corrosion protection in reinforced concrete. The fib Model Code (2006) offers probabilistic durability design methodologies, while ISO 9224 (2012) categorizes atmospheric corrosivity for environmental classification.

These codes offer boundary conditions and validation benchmarks for ML models, ensuring their relevance to structural engineering practice.

* 1. **Empirical and Semi-Empirical Models**

**Early Corrosion Prediction Models**

Initial corrosion models were primarily based on experimental observations and field data. Andrade and Alonso (1996) proposed empirical equations linking corrosion current density to environmental factors. Liu and Weyers (1998) introduced a mechanistic model using Fick’s second law to estimate corrosion initiation time and propagation rate. Vu and Stewart (2000) developed probabilistic life-cycle cost models incorporating corrosion as a function of exposure conditions.

Chloride-induced corrosion, particularly in marine environments and areas with deicing salts, remains a critical durability concern. Andrade and Alonso (2001) highlighted the value of real-time corrosion measurements, while Alonso et al. (2000) introduced the concept of chloride threshold values for steel depassivation. Angst (2011) reviewed methods for defining critical chloride content, calling for improved mechanistic understanding. Glass and Buenfeld (1997) proposed probabilistic thresholds accounting for material and environmental variability. Liu and Weyers (1998) further modeled time-to-cracking from chloride ingress, and Broomfield (2007) synthesized corrosion mechanisms and mitigation strategies. These efforts underline the complex, multifactorial nature of chloride-induced corrosion, shaped by factors like cement chemistry, exposure class, and reinforcement layout.

Empirical models, while useful for early assessments, often fall short in capturing nonlinear interactions, lack generalizability, and are highly sensitive to experimental conditions.

**2.2 Mechanistic and Diffusion-Based Models**

Mechanistic models aim to simulate chloride transport and corrosion processes. Tuutti (1982) introduced a two-phase service life model: initiation (chloride ingress) and propagation (active corrosion). Life-365 (2001) expanded this by incorporating time-dependent diffusion and surface chloride concentration. Although Fickian diffusion models are widely applied, they simplify complex transport mechanisms like binding, capillary suction, and convection, and may not capture localized or post-initiation corrosion.

Wang et al. (2015) proposed a mechanistic-empirical model combining diffusion kinetics with deterioration trends for bridge deck design. Medeiros and Helene (2009) emphasized the effects of surface treatments on chloride ingress and suction, which are vital for realistic modeling.

**2.3 Probabilistic and Reliability-Based Approaches**

To address uncertainties in exposure and material properties, probabilistic models such as Monte Carlo simulations (Stewart and Mullard, 2007) and Bayesian updating (Yang et al., 2014) have been adopted to predict service life and failure probability. Vu and Stewart (2000) integrated stochastic chloride ingress into structural reliability assessments. Ye et al. (2007) further quantified initiation time uncertainties. These methods highlight the limitations of deterministic models and the benefits of combining physics-based and statistical approaches.

**2.4 Data-Driven and Machine Learning Models**

Advances in computing and data availability have enabled machine learning (ML) models to capture complex, nonlinear relationships in corrosion prediction.

**2.4.1 Artificial Neural Networks (ANNs)**

ANNs have shown strong performance in predicting corrosion-related parameters. Chithra and Bhuvaneshwari (2016) used inputs such as cover depth, w/c ratio, and chloride content to estimate corrosion current density. Taffese et al. (2017) integrated ANN with finite element models for RC bridge service life prediction. Liu et al. (2020) applied ANN to chloride ingress modeling, while Zhang et al. (2018) and Gharachorlou & Ramezanianpour (2021) enhanced performance using genetic algorithms and hybrid methods. Despite their predictive strength, ANN models often suffer from low interpretability and require large datasets.

**2.4.2 Support Vector Machines (SVMs)**

SVMs, as shown by Pan et al. (2019), can outperform linear regression models in predicting corrosion using design and environmental features, particularly in small, high-dimensional datasets. However, their performance depends on kernel tuning and feature scaling.

**2.5 Decision Trees and Ensemble Methods**

Ensemble methods such as Random Forests and Gradient Boosting Machines (GBMs) provide improved generalization and robustness. Muthukannan et al. (2020) found Random Forests outperformed ANN and SVM in corrosion loss prediction. These methods can rank feature importance but still face interpretability challenges.

**2.6 Extreme Gradient Boosting (XGBoost) and Variants**

XGBoost (Chen & Guestrin, 2016) offers speed, scalability, and the ability to model feature interactions and missing data. Zhao et al. (2021) and Su et al. (2020) applied XGBoost to corrosion prediction, achieving high accuracy and improved interpretability using SHAP values (Lundberg & Lee, 2017). Altunkaynak and Yildiz (2020) used ensemble models to estimate permeability, and Yang et al. (2021) and Xu et al. (2022) improved predictions further through data segmentation, which allows models to specialize across subsets with different behavior patterns.

**2.7 Hybrid and Augmented Modeling Approaches**

Feature engineering is critical for enhancing model performance. Guyon and Elisseeff (2003) and Brownlee (2020) emphasized the role of domain knowledge in feature selection. To address data scarcity and imbalance, augmentation techniques such as SMOTE (Chawla et al., 2002), instance selection (Liu & Motoda, 2001), and VSG (noise-injection or Gaussian sampling) are used to expand datasets, especially for rare corrosion scenarios.

Hybrid models integrate physics-based insights with ML. Raissi et al. (2019) proposed Physics-Informed Machine Learning (PIML) frameworks that impose physical constraints during training, improving robustness and interpretability. Segmenting corrosion data into low, medium, and high ranges has been shown to improve predictive accuracy by accounting for heteroscedasticity.

**2.8 Comparative Studies and Benchmarking**

Several studies benchmarked ML models for corrosion prediction. Taffese et al. (2017) found ensemble methods consistently outperformed ANN and SVM. Mohammed et al. (2022) compared five ML models and highlighted the importance of data balance and feature selection. A segmented XGBoost model augmented with virtual samples achieved an R² of 0.902, with 49% of predictions within a 10% error margin—outperforming previous models.

To overcome current limitations, this study proposes a segmented XGBoost framework enhanced by virtual data augmentation. By training regressors on specific intervals of corrosion severity and synthetically expanding the dataset, the model captures distinct behaviors, improves generalization, and mitigates class imbalance. This approach is particularly effective for rare but severe corrosion events.

Model credibility also hinges on proper evaluation. Willmott and Matsuura (2005) recommend MAE over RMSE for robustness to outliers, while Kuhn and Johnson (2013) discussed validation techniques like cross-validation and grid search. Tools such as LIME (Ribeiro et al., 2016) and Bayesian dropout (Gal & Ghahramani, 2016) further aid interpretability and uncertainty quantification.

**Conclusion**

The evolution of corrosion prediction has shifted from empirical and mechanistic models to hybrid AI-driven frameworks. The segmented XGBoost approach, supported by virtual augmentation and rigorous feature engineering, offers a scalable, interpretable, and accurate solution for modeling chloride-induced corrosion. Yet, challenges remain regarding long-term field validation, code integration, and data availability. Future research should focus on transparent AI, uncertainty quantification, and fusion with physics-based models to advance durability assessment in civil infrastructure.

To overcome these limitations, this study introduces an advanced machine learning-based predictive framework that leverages segmented Extreme Gradient Boosting (XGBoost) regression. This approach enables the model to account for distinct behavioral patterns across different ranges of the target variable by training separate regressors within defined segments, thereby improving accuracy and interpretability. Furthermore, to address data scarcity and enhance the model’s generalization capacity, a virtual data augmentation technique is employed to synthetically expand the dataset while preserving the underlying statistical distributions and physical consistency. By integrating segmentation with XGBoost and virtual sample generation, the proposed method aims to capture complex nonlinear relationships, mitigate data imbalance, and better represent rare but critical corrosion behaviors, ultimately offering a robust and scalable tool for corrosion rate prediction in chloride-laden environments.

**3. Data Preparation and Feature Engineering**

**3.1 Exploratory Data Analysis**

The dataset included key variables influencing chloride-induced corrosion in reinforced concrete: cover thickness, reinforcement diameter, water–cement (w/c) ratio, chloride ion concentration, temperature, relative humidity (RH), and exposure time. Histogram and boxplot analyses revealed limited variability in cover depth (50–65 mm) and rebar diameter (16–20 mm), with w/c ratios narrowly distributed around 0.40. Environmental conditions were tightly controlled, centered around 293 K and 80% RH.

Chloride ion content exhibited a right-skewed distribution (mode ≈ 6 kg/m³; max = 12 kg/m³), reflecting aggressive exposure conditions. Most samples represented short-term exposures (<2 years), and corrosion rates ranged from 0.05 to 0.7 µA/cm² with a moderate right skew.

**3.2 Feature Behavior and Modeling Implications**

Boxplots confirmed that most features were either discretely distributed or skewed. Chloride content and exposure time showed high variability and skewness, suggesting the need for transformation (e.g., log) prior to modeling. The observed nonlinear behavior supported the use of tree-based models, which are well-suited for capturing complex feature interactions.

**3.3 Correlation and Scatter Plot Insights**

Pearson correlation analysis identified RH (r = +0.70) and temperature (r = +0.63) as the strongest linear predictors of corrosion rate, followed by reinforcement diameter (r = +0.57). Cover thickness and w/c ratio were moderately negatively correlated, consistent with their protective roles. Chloride content showed only a weak linear correlation (r = +0.15), indicating potential nonlinear or threshold effects. A high correlation between RH and rebar diameter (r = +0.83) raised concerns about multicollinearity.

Scatter plots reinforced these observations. Corrosion rates increased with RH, temperature, and chloride content, while cover thickness and w/c ratio had weaker, inverse relationships. Exposure time exhibited a nonlinear (inverse-U) pattern, with peak corrosion occurring around one year.

**3.4 Engineered Interaction Features**

To embed domain-specific relationships, two interaction terms were introduced:

* Humidity\_Temp (RH × Temperature): Reflects synergistic effects of environmental severity on corrosion kinetics.
* Cement\_Cover (w/c × Cover Thickness): Captures combined effects of material permeability and protective barrier thickness.

These engineered features enhanced model interpretability and predictive accuracy by aligning with physical mechanisms.

**3.5 Virtual Sample Generation (VSG)**

To address class imbalance, particularly the underrepresentation of low corrosion scenarios, 100 synthetic samples were generated via bootstrapping and Gaussian noise injection (μ = 0, σ = 0.01). Engineered features were recalculated for consistency. This approach improved model generalization across the corrosion spectrum, especially under protective or newly constructed conditions.

**3.6 Segmented Modeling Strategy**

**3.6-1 Corrosion Regime Segmentation**

To capture distinct corrosion mechanisms, the dataset was segmented based on corrosion rate:

* Low-corrosion regime: < 0.15 µA/cm² (initiation phase)
* High-corrosion regime: ≥ 0.15 µA/cm² (propagation phase)

Each segment was modeled independently using a dedicated XGBoost regressor.

**3.6-2 Model Configuration**

Both models used the following hyperparameters:

* n\_estimators: 200
* learning\_rate: 0.05
* max\_depth: 4
* random\_state: 42

This configuration balanced model complexity and generalization. SHAP analysis was used to monitor feature importance and improve interpretability.

**3.6-3 Justification and Performance Gains**

Segmented modeling enabled the algorithm to specialize in each corrosion regime:

* In the low-corrosion regime, small variations in protective features (e.g., cover thickness, w/c ratio) were influential, helping detect incipient deterioration under mildly aggressive conditions.
* In the high-corrosion regime, dominant contributors such as chloride ion content and the engineered Humidity\_Temp feature were most predictive, capturing rapid degradation in severe environments.

Training separate models reduced bias from regime averaging, enhanced feature interpretability, and allowed learning of nonlinear thresholds (e.g., chloride concentration limits). Compared to a unified model, the segmented approach improved:

* Accuracy across both low and high corrosion conditions,
* Sensitivity to protective features in early-stage corrosion, and
* Interpretability by yielding regime-specific feature rankings aligned with corrosion science.

**3.7 Conclusion of the Modeling Framework**

The final modeling framework integrated:

* Segmented XGBoost models to reflect regime-specific behaviors,
* Engineered interaction features to capture key physical dependencies,
* Virtual Sample Generation to correct class imbalance,
* Log-transformation of the skewed target variable, and
* SHAP-based interpretation to ensure transparency.

This approach delivered robust, interpretable, and physically grounded predictions of corrosion rate, offering a valuable tool for durability assessment and risk-informed design in chloride-exposed reinforced concrete structures.

# 4. Evaluation Metrics and Predictive Accuracy

# The segmented XGBoost framework developed in this study was rigorously evaluated using standard regression metrics to assess its ability to predict corrosion rates in reinforced concrete subjected to chloride exposure. The model's quantitative performance was evaluated based on three key indicators: the coefficient of determination (R²), root mean squared error (RMSE), and mean absolute error (MAE).

# Coefficient of Determination (R² = 0.900): This value indicates that the model explains 90% of the observed variance in corrosion rates, a notable achievement given the inherent complexity and variability of material degradation processes in real-world environmental conditions.

# Root Mean Squared Error (RMSE = 0.0014 µA/cm²): The low RMSE demonstrates that the model consistently produces predictions with minimal deviation from the actual measured values, even in cases of very low corrosion current densities.

# Mean Absolute Error (MAE = 0.0237 µA/cm²): The model’s average prediction error remains within a narrow band, underscoring its reliability for practical applications, particularly in structural health monitoring and durability forecasting where conservative thresholds are commonly used.

# These results reflect the model’s ability to capture complex, nonlinear relationships between corrosion rates and influential design/environmental parameters, including temperature, relative humidity, chloride ion concentration, and mix composition.

# 4.2 Predictive Robustness and Error Distribution

# Beyond central tendency metrics, the robustness of the model was evaluated through percentile-based error statistics and a detailed distribution of percent errors. This analysis aimed to verify the model’s performance across the full spectrum of corrosion behaviors and ensure its applicability in both high-risk and low-risk structural scenarios.

# Mean Percent Error: 14.56%

# Median Percent Error: 10.06%

# The relatively low mean and median percent errors demonstrate not only high average accuracy but also reduced sensitivity to outliers or anomalous input conditions.

# A threshold-based breakdown of absolute percent errors provides further insights into model reliability:

| Absolute Percent Error Range | Percentage of Samples |
| --- | --- |
| ≤ 5% | 28.57% |
| ≤ 10% | 48.98% |
| ≤ 30% | 85.71% |

# Nearly half of the predictions fall within 10% of the actual corrosion rate, and over 85% lie within a 30% margin—well within the acceptable range for most engineering applications. These results confirm that the model not only performs well on average but maintains consistent accuracy across varying levels of corrosion severity.

# The symmetrical distribution of prediction errors, coupled with low bias, suggests a high degree of generalization and stability. This is particularly important for infrastructure applications, where false positives (over-predicted degradation) and false negatives (under-predicted degradation) can have costly or unsafe implications.

# 4.3 Advantages of the Segmented Modeling Framework

# The use of a segmented modeling approach—splitting the dataset into low- and high-corrosion regimes—proved instrumental in improving performance. By tailoring separate XGBoost regressors to each corrosion regime, the framework captured the unique mechanisms and feature interactions present at different stages of material deterioration.

# The low-corrosion model effectively captured subtle influences of protective parameters such as increased cover thickness and lower w/c ratios. These inputs, though often overshadowed in aggregate models, are critical for assessing incipient deterioration and maintaining passivity.

# The high-corrosion model responded sensitively to environmental stressors, identifying dominant drivers such as chloride concentration and synergistic humidity–temperature effects (captured through the engineered Humidity\_Temp interaction feature).

# This dual-model strategy eliminated the trade-offs typically encountered in unified models, which tend to average behavior across the corrosion spectrum, leading to overprediction in passive states and underprediction in aggressive environments.

# 4.4 Model Enhancements: Feature Engineering and Data Balancing

# The model’s performance gains were also attributable to two key enhancements:

# Feature Engineering: The introduction of interaction terms, such as Humidity\_Temp and Cement\_Cover, significantly improved the model's ability to represent real-world corrosion phenomena. These features captured compounded effects that individual variables could not explain independently, aligning data-driven learning with known physical mechanisms.

# Virtual Sample Generation (VSG): To address the underrepresentation of low-corrosion cases, synthetic samples were generated via bootstrapping and Gaussian noise injection. This ensured that protective scenarios were adequately represented during training, reducing systematic bias and improving prediction quality in data-sparse regions.

# 4.5 Practical Implications and Deployment Readiness

# The final model, combining segmentation, feature engineering, and data augmentation, delivered a robust and interpretable predictive framework for corrosion rate estimation. It demonstrated strong generalization across a range of mix designs, environmental exposures, and material properties. Compared to traditional regression or single-model approaches, the segmented XGBoost model offers:

# Higher accuracy across diverse corrosion regimes,

# Improved sensitivity to critical durability factors,

# Enhanced interpretability through regime-specific SHAP value analysis, and

# Greater readiness for deployment in practical engineering applications such as structural inspection, service life estimation, and durability design optimization.

# The segmented XGBoost approach offers a rigorous, domain-informed framework for predicting chloride-induced corrosion in reinforced concrete. Its strong predictive performance, supported by engineered features and data balancing techniques, affirms its potential for adoption in durability modeling and infrastructure asset management. This study underscores the importance of aligning machine learning methods with engineering insight to build models that are both accurate and interpretable.

# 5. Conclusion

# This study introduces a novel two-stage XGBoost-based framework for chloride-induced corrosion rate prediction in reinforced concrete. By combining segmentation, virtual data augmentation, and advanced feature interaction modeling, the proposed approach achieved high predictive accuracy and interpretability. The model provides a valuable tool for structural durability assessment and lifecycle service planning.

Accurate prediction of chloride-induced corrosion rates in reinforced concrete is vital for structural durability assessment and life-cycle management. While traditional models laid the groundwork, they fall short in capturing the complexity of real-world interactions. Machine learning, especially ensemble and hybrid approaches like segmented XGBoost with virtual data augmentation, offers promising improvements in predictive accuracy and robustness. Continued efforts toward data integration, model interpretability, and standardization will be essential for translating these advancements into practical, scalable solutions for infrastructure management.

# 8. Future Work

# Extend the model with additional long-term exposure data (>10 years).

# Incorporate electrochemical and microstructural variables for multi-scale modeling.

# Benchmark against physics-informed neural networks for hybrid modeling.

**References**

**Chloride-Induced Corrosion in Reinforced Concrete**

Andrade, C., & Alonso, C. (2001). On-site measurements of corrosion rate of reinforcements. *Construction and Building Materials, 15*(2–3), 141–145. https://doi.org/10.1016/S0950-0618(00)00063-7

Alonso, C., Andrade, C., Castellote, M., & Castro, P. (2000). Chloride threshold values to depassivate reinforcing bars embedded in a standardized OPC mortar. *Cement and Concrete Research, 30*(7), 1047–1055. https://doi.org/10.1016/S0008-8846(00)00265-9

Angst, U. M. (2011). Chloride induced reinforcement corrosion in concrete: Concept of critical chloride content–methods and mechanisms. *Cement and Concrete Research, 39*(12), 1122–1138. https://doi.org/10.1016/j.cemconres.2009.08.006

Broomfield, J. P. (2007). *Corrosion of steel in concrete: Understanding, investigation and repair* (2nd ed.). CRC Press.

Glass, G. K., & Buenfeld, N. R. (1997). The presentation of the chloride threshold level for corrosion of steel in concrete. *Corrosion Science, 39*(5), 1001–1013. https://doi.org/10.1016/S0010-938X(98)00005-1

Liu, Y., & Weyers, R. E. (1998). Modeling the time-to-corrosion cracking in chloride contaminated reinforced concrete structures. *ACI Materials Journal, 95*(6), 675–681.

**Modeling Corrosion Rates in Concrete**

Vu, K. A. T., & Stewart, M. G. (2000). Structural reliability of concrete bridges including improved chloride-induced corrosion models. *Structural Safety, 22*(4), 313–333. https://doi.org/10.1016/S0167-4730(01)00003-1

Wang, X., Shi, X., & Xie, N. (2015). A mechanistic-empirical model to predict chloride-induced corrosion service life of concrete bridge decks. *Construction and Building Materials, 100*, 9–18. https://doi.org/10.1016/j.conbuildmat.2015.09.020

Medeiros, M. H. F., & Helene, P. (2009). Surface treatment of reinforced concrete in marine environment: Influence on chloride diffusion coefficient and capillary water absorption. *Construction and Building Materials, 23*(3), 1476–1484. https://doi.org/10.1016/j.conbuildmat.2008.07.023

Ye, H., Jin, N., Jin, X., & Zhang, C. (2007). A probabilistic model for predicting the initiation time of chloride-induced reinforcement corrosion in concrete structures. *Corrosion Science, 49*(11), 3926–3938. https://doi.org/10.1016/j.corsci.2007.04.004

**Machine Learning in Corrosion Prediction**

Chou, J.-S., & Pham, A. D. (2015). Smart artificial firefly colony optimization-based support vector regression for enhanced prediction of spalling behavior of concrete cover in fire-exposed RC beams. *Automation in Construction, 49*, 40–52. https://doi.org/10.1016/j.autcon.2014.10.007

Liu, R., Liu, Z., Xu, Y., & Lin, X. (2020). Prediction of chloride ingress in concrete using artificial neural network model. *Construction and Building Materials, 244*, 118356. https://doi.org/10.1016/j.conbuildmat.2020.118356

Zhang, L., Zhang, P., Zhao, J., & Du, C. (2018). Prediction of corrosion rate of steel in concrete using artificial neural network optimized by genetic algorithm. *Construction and Building Materials, 179*, 151–158. https://doi.org/10.1016/j.conbuildmat.2018.05.267

Gharachorlou, A., & Ramezanianpour, A. A. (2021). Machine learning-based predictive modeling for corrosion current density in reinforced concrete. *Construction and Building Materials, 278*, 122348. https://doi.org/10.1016/j.conbuildmat.2021.122348

**XGBoost and Segmented Modeling Approaches**

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). https://doi.org/10.1145/2939672.2939785

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

Yang, W., Chen, L., & Cai, H. (2021). A novel ensemble model with data segmentation for predicting concrete compressive strength. *Applied Sciences, 11*(3), 1248. https://doi.org/10.3390/app11031248

Su, Y., Cheng, L., & Wang, X. (2020). Performance prediction of structural concrete using XGBoost model. *Materials, 13*(7), 1538. https://doi.org/10.3390/ma13071538

Altunkaynak, A., & Yildiz, M. (2020). Estimation of concrete permeability using ensemble learning models. *Journal of Materials in Civil Engineering, 32*(12), 04020337. https://doi.org/10.1061/(ASCE)MT.1943-5533.0003443

Xu, L., Fang, Y., & Ma, X. (2022). Segmented ensemble learning for the prediction of concrete properties based on mixtures. *Materials Today Communications, 33*, 104527. https://doi.org/10.1016/j.mtcomm.2022.104527

**Feature Engineering and Data Augmentation**

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*, 1157–1182.

Brownlee, J. (2020). *Feature engineering for machine learning: Principles and techniques for data scientists*. Machine Learning Mastery.

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data, 6*(1), 1–48. https://doi.org/10.1186/s40537-019-0197-0

Liu, H., & Motoda, H. (2001). *Instance selection and construction for data mining*. Springer.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357. https://doi.org/10.1613/jair.953