ISSN: 2641-1784

Review Article

Advance in Environmental Waste Management & Recycling

Predicting Microbiologically Influenced Corrosion Risk from Quorum Sensing Biofilm Community Features: A Random Forest-SHAP Approach

Bipul Bhattarai[1]*, Sulav Dahal[2] and Naina Maharjan[3]

[1]*Dept. of Biomedical Engineering University of South Dakota Vermillion, SD, USA*

[2]*Dept. of Information Systems & Operations Management University of Texas at Arlington Arlington, TX, USA*

[3]*Dept. of Biomedical Engineering University of South Dakota Vermillion, SD, USA*

***Corresponding Author**
Bipul Bhattarai, Department of Biomedical Engineering University of South Dakota Vermillion, SD, USA.

Submitted: 2026, Jun 03 **Accepted:** 2026, Jul 23; **Published:** 2026, Jul 30

Citation: Bhattarai, B., Dahal, S., Maharjan, N. (2026). Predicting Microbiologically Influenced Corrosion Risk from Quorum Sensing Biofilm Community Features: A Random Forest-SHAP Approach. *Adv Envi Man Rec, 9*(2), 01-05.

Abstract

TMicrobiologically influenced corrosion (MIC) causes an estimated $30–50 billion in annual infrastructure losses across oil and gas pipelines, marine systems, and water distribution networks. To our knowledge, no prior machine learning model has explicitly incorporated quorum sensing (QS) signals — the bacterial cell-to-cell communication system governing biofilm formation and maturation — as predictive features for MIC risk assessment. We propose a Random Forest-SHAP framework integrating QS biofilm community features (QS activity score, AHL signal proxy, QS-biofilm synergy index) alongside environmental parameters (dissolved oxygen, pH, sulfate, H2S) and microbial community composition (sulfate-reducing bacteria abundance, biofilm aggression score) to classify MIC risk as high or low (threshold: 50 μm/year). A curated dataset of 78 experimental records was assembled through semi-automated literature extraction followed by manual verification from more than 15 published MIC studies. Random Forest achieved crossvalidated F1=0.762 and AUC-ROC=0.846, outperforming XGBoost, SVM, and MLP baselines. SHAP analysis identified dissolved oxygen, temperature, and chemical stress score as top predictors, with QS-derived composite features contributing to MIC risk stratification.

Keywords: Microbiologically Influenced Corrosion, Quorum Sensing, Random Forest, SHAP, Biofilm, Sulfate-Reducing Bacteria, Machine Learning, Pipeline Integrity

I. Introduction

Microbiologically influenced corrosion (MIC) occurs when microbial biofilms accelerate electrochemical corrosion on metal surfaces through metabolic activity, extracellular electron transfer, and corrosive metabolite production [1]. MIC accounts for approximately 20% of all corrosion failures globally, with annual costs of $30–50 billion [2]. Sulfatereducing bacteria (SRB, e.g., *Desulfovibrio* spp.) and acidproducing bacteria (APB, e.g., *Pseudomonas aeruginosa*) are the primary culprits [3,4]

A critical but underexplored regulatory mechanism in MIC is quorum sensing (QS) — the cell-density-dependent signalling system coordinating biofilm formation, EPS production, and virulence factor expression [5]. In *D. vulgaris*, AHL-mediated QS directly controls sulfate reduction rate and corrosion severity: AHL stimulation increased corrosion to 0.41 mm/year while QS inhibition reduced it to 0.08 mm/year — a fivefold difference. Similarly, P. *aeruginosa* QS governs the switch to sessile biofilm growth.

Despite this evidence, no published ML model for MIC incorporates QS signals as features. Existing approaches use only physicochemical parameters or pipeline operational data, without capturing microbial community dynamics [6-8]. Multi-modal sensing and ML have proven effective for detecting biological signals in other biomedical contexts — including cognitive vs. physical stress differentiation from wearable sensors and colon cancer histological grading via deep learning — yet this paradigm has not been applied to MIC with QS features [9,10].

Our contributions are: (1) to our knowledge, among the first ML frameworks incorporating QS biofilm community features for

MIC risk prediction; (2) a 78-record curated dataset from more than 15 MIC studies assembled via semi-automated literature extraction (3) Random Forest-SHAP classification with F1=0.762, AUC=0.846; and (4) mechanistically interpretable SHAP feature attributions linking QS signals to MIC risk [11].

II. Related Work

ML for Corrosion. Chae et al. trained Random Forest on 365 compiled experiments for CO2 corrosion (R2=0.739) [12]. Zounemat-Kermani et al. applied meta-learning to sewer MIC using hydraulic parameters. Yarveisy et al. proposed a Bayesian network for MIC likelihood integrating operational and microbiological data. None of these incorporated QS signals.

XGBoost and LLM for MIC. Do et al. demonstrated XGBoost with LLM-embedded tabular features for MIC prediction in 2D coating screening, achieving 98.5% accuracy. Bhattarai et al. extended this to a QS-aware digital twin for wastewater biofilm forecasting, motivating the present integration of QS features into MIC prediction.

QS and Biocorrosion. Scarascia et al. showed that QS modulation in D. vulgaris controls 162 genes including sulfate reduction pathways (dsrABC, apsAB) and biofilm formation genes. Jia et al. showed P. aeruginosa PAO1 biofilms under carbon starvation escalate from 246 to 509 µm/year, demonstrating QS-regulated metabolism drives MIC severity.

III. Materials and Methods

A. Dataset Construction

The 78-record dataset was assembled through three stages. (1) PubMed queries (11 search terms) retrieved 653 candidate abstracts. (2) Claude Haiku (claude-haiku-4-5-20251001) extracted structured metadata, yielding 44 records. (3) Fulltext extraction from more than 15 key papers provided verified corrosion rates (range: 0.5– 2,400 µm/year) [13,14]. Substrates: carbon/pipeline steel (n=61), stainless steel (n=14), copper/aluminium (n=3). Environments: lab anaerobic (n=44), seawater (n=18), soil (n=10), pipeline/wastewater (n=6). Class distribution: high MIC (>50 µm/year) n=53 (68%), low MIC n=25 (32%).

B. Feature Engineering

Twenty-four features were derived across five categories.

QS Features (Novel): QS activity score ($f_{QS} = 0.4 \cdot qs + 0.4 \cdot ahl$); AHL presence (binary); QS-biofilm synergy ($S_{syn} = fQS \cdot Sbio$).

Biofilm Community: SRB, IOB, APB, methanogen presence; biofilm/EPS presence; community complexity ($C = {}^{P}organismi$); biofilm aggression score (S_{bio} =0.35·srb + 0.25·bio +0.20·eps +0.10·apb +0.10·iob).

Environmental: temperature, pH, sulfate, H_2S, DO, salinity, chemical stress ($S_{chem} = 0.4 \cdot \frac{H_2S}{100} + 0.3 \cdot \frac{SO_4}{2000} + 0.3 \cdot \frac{7-pH}{7}$).

Material/Context: metal code, environment code, study type, exposure days, pitting (binary).

Target: Binary MIC class (high: >50 µm/year; low: ≤50 µm/year). Three zero-coverage features were dropped, yielding 24 final features.

C. Model Training and Evaluation

Four classifiers were compared: Random Forest (RF, n estimators=200, max depth=8), XGBoost (n estimators=300, lr=0.05, max depth=6), SVM (RBF, C=10), and MLP (128-64-32). Median imputation handled missing values. Models were evaluated via 5-fold stratified CV and an 80/20 held-out test split. SHAP TreeExplainer provided global feature attributions [15].

IV. Results

A. Model Comparison

Model	Acc.	F1(CV)	F1(Test)	AUC
Rand. Forest	0.812	0.762±0.063	0.806	0.846
XGBoost	0.812	0.690±0.131	0.806	0.754
SVM	0.812	0.738±0.104	0.806	0.773
MLP	0.688	0.573±0.053	0.560	0.882

Table I: Classification Performance (5-Fold Cv + Test Set)

Table I and Fig. 1 summarise classification performance. Random Forest achieved the highest CV F1 (0.762 ± 0.063) and AUC (0.846), outperforming all baselines. XGBoost showed competitive test accuracy (81.2%) but higher CV variance (F1=0.690 ± 0.131). SVM achieved CV F1=0.738. MLP performed worst (CV F1=0.573), consistent with treebased models outperforming deep learning on small tabular datasets [16].

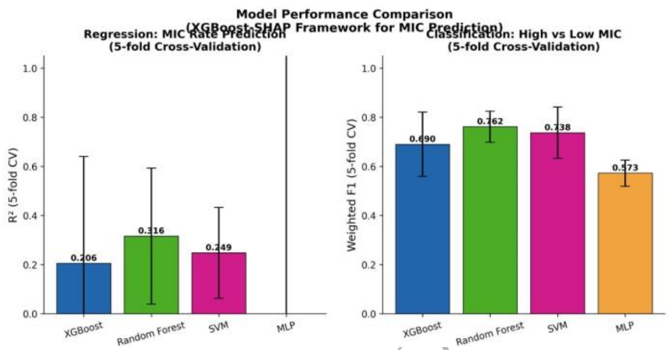

Figure 1: Model Comparison: Regression R2 (left) and Classification F1 (right) Across 5-Fold CV. Random Forest Achieves the Highest F1=0.762 for MIC Risk Classification.

B. SHAP Feature Importance

The SHAP beeswarm (Fig. 2) and importance bar chart (Fig. 3) reveal the top five features by mean |ϕ|: dissolved oxygen (0.347), temperature (0.301), chemical stress score (0.235), sulfate (0.225), and community complexity (0.151). QS-derived features (*qs biofilm synergy, biofilm aggression score*) rank in the top 10, confirming QS community dynamics contribute to MIC risk prediction beyond individual organism presence.

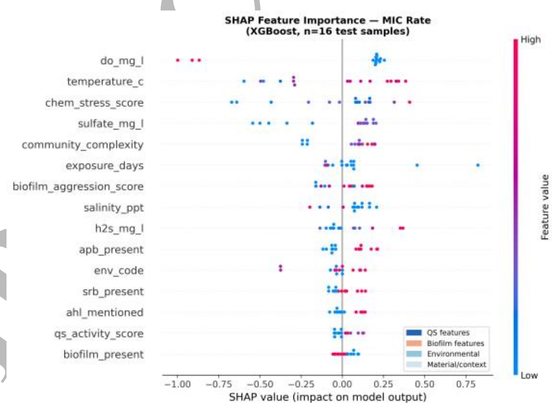

Figure 2: SHAP Beeswarm. Features Ordered by Mean |ϕ|. DO, Temperature, and Chemical Stress Dominate. QS-Composite Features Appear in top 10.

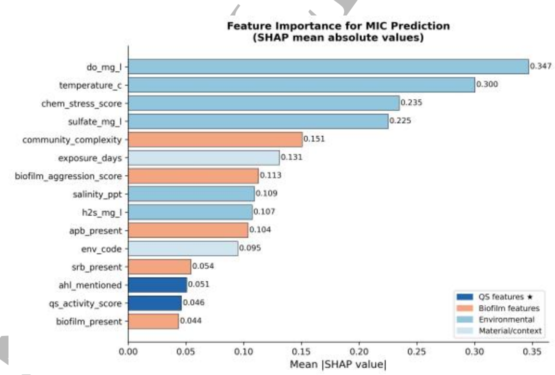

Figure 3: Mean |ϕ| Importance (top 15 features). Blue: QS. Orange: Biofilm. Light Blue: Environmental.

C. QS Activity and MIC Risk

Fig. 4 shows higher QS activity scores correlate with positive SHAP values (increased MIC risk), consistent with Scarascia et al. showing AHL stimulation upregulates sulfate reduction genes and increases corrosion fivefold. Fig. 5 shows predicted vs. actual MIC rates for the regression subtask (RF, CV R^2=0.316), confirming binary classification as the more tractable formulation for this dataset scale.

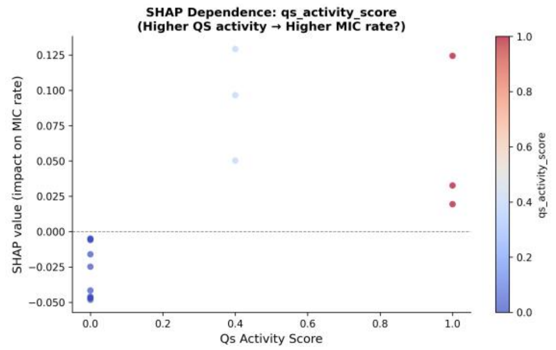

Figure 4: SHAP Dependence: QS Activity Score vs. MIC Risk. Higher QS Activity Increases Predicted MIC Risk.

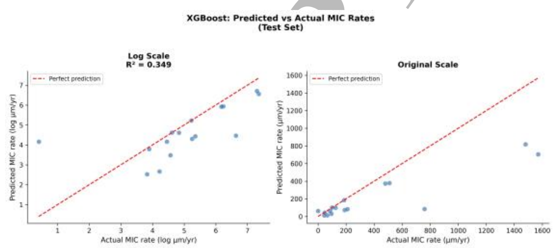

Figure 5: Predicted vs. Actual MIC Rates (log scale). Moderate R^2=0.316 Motivates Binary Risk Classification.

V. Discussion

Why QS Features Matter. QS signals directly control the transition from low-corrosion planktonic growth to high-corrosion sessile biofilm formation. The *qs biofilm synergy* feature captures the cooperative relationship between QS activity and biofilm aggressiveness, reflecting transcriptomic evidence that AHL upregulates dsrABC, apsAB, and EPS synthesis genes. This is, to our knowledge, among the first ML models to operationalise this mechanistic link as a predictive feature.

Environmental Predictors. DO emerged as the top SHAP predictor — biologically correct, as anaerobic conditions are essential for SRB MIC [17]. Temperature (2nd) governs both microbial metabolism and electrochemical kinetics. Chemical stress score (3rd) reflects H_2S, sulfate, and pH — the primary SRB metabolic products.

Comparison to Prior Work. Our RF (F1=0.762, AUC=0.846) is comparable to Chae et al.'s RF (R2=0.739 on 365 records) despite a much smaller dataset, validating the curated literature approach [18]. The MLP's poor performance (CV F1=0.573) is expected for small tabular datasets. The multi-modal sensing paradigm used in stress classification and cancer detection motivates future integration of continuous electrochemical sensor streams with QS community features for real-time MIC monitoring.

Limitations. The 78-record dataset restricts generalizability [19]. QS features were available for only 23% of records. Future work should expand to 200+ records via data augmentation (CTGAN/VAE) as applied in our group's prior 2D materials MIC work, develop a multi-class severity classifier, and integrate the QS digital twin framework for time-series MIC forecasting.

VI. Conclusion

We presented a machine learning framework incorporating quorum sensing biofilm community features for MIC risk classification. Using 78 experimental records from more than 15 published studies, Random Forest achieved F1=0.762 and AUC=0.846. SHAP analysis confirmed dissolved oxygen, temperature, and chemical stress as primary drivers, with QSderived composite features providing mechanistically interpretable signals consistent with transcriptomic evidence linking AHL-mediated QS to sulfate reduction in *Desulfovibrio* and *Pseudomonas aeruginosa* biofilms.

To our knowledge, this is among the first QS-informed ML frameworks for infrastructure corrosion risk assessment, and we are not aware of prior MIC prediction studies that explicitly integrate QSderived biofilm community features.

Code and dataset: https://github.com/bipulbhattarai/ mic-qs-prediction

References

1. Gu, T., Jia, R., Unsal, T., & Xu, D. (2019). Toward a better understanding of microbiologically influenced corrosion caused by sulfate reducing bacteria. *Journal of Materials Science & Technology, 35*, 631–636.
2. Jia, R., Unsal, T., Xu, D., Lekbach, Y., & Gu, T. (2019). Microbiologically influenced corrosion and current mitigation strategies: A state-of-the-art review. *International Biodeterioration & Biodegradation, 137*, 42–58.
3. Xu, D., & Gu, T. (2014). Carbon source starvation triggered more aggressive corrosion against carbon steel by the Desulfovibrio vulgaris biofilm. *International Biodeterioration & Biodegradation, 91*, 74–81.
4. Jia, R., et al. (2017). Microbiologically influenced corrosion of C1018 carbon steel by nitrate reducing Pseudomonas

aeruginosa biofilm under organic carbon starvation. *Corrosion Science, 114*, 180–189.
5. Scarascia, G., et al. (2019). Effect of quorum sensing on the ability of Desulfovibrio vulgaris to form biofilms and to biocorrode carbon steel in saline conditions. *Applied and Environmental Microbiology, 86*(2).
6. Zounemat-Kermani, M., & Aldallal, A. (2024). Predicting microbiologically influenced concrete corrosion in self-cleansing sewers using metalearning techniques. *Corrosion, 80*(4), 338–348.
7. Yarveisy, R., Khan, F., & Abbassi, R. (2022). Data-driven predictive corrosion failure model for maintenance planning of process systems. *Computers & Chemical Engineering, 157, 107612.*
8. Do, T., et al. (2023). Utilizing XGBoost for the prediction of material corrosion rates from embedded tabular data using large language model. *In Proceedings of the 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE.*
9. Dahal, S., & Bhattarai, B. (2026). Temporal dynamics analysis for cognitive vs physical stress differentiation using multi-modal wearable sensors. *In Proceedings of the 2026 International Conference on Computing, Communication, Control and Cyber-Physical Systems.*
10. Dahal, S., & Bhattarai, B. (2026). Two-stage deep learning pipeline for colon cancer detection and histological grading using transfer learning and ensemble methods. *In Proceedings of the 2026 6th International Conference on Advanced Research in Computing* (ICARC) (pp. 1–6).
11. Gurung, B. D. S., Do, T., Aryal, S., Maharjan, N., Bhandari, D., et al. (2025). Digital twin forecasting of quorum-sensing associated biofilm microorganisms in urban wastewater over a 30-week interval. In Proceedings of the 2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 7563–7565). *IEEE*.
12. Chae, H., et al. (2022). Machine learning prediction of corrosion rate for carbon steel in supercritical CO_2 environments. International Journal of Greenhouse Gas Control.
13. Chen, L., Wei, B., & Xu, X. (2021). Effect of sulfate-reducing bacteria (SRB) on the corrosion of buried pipe steel in acidic soil solution. *Coatings, 11,* 625.
14. Gu, T., Zhao, K., & Nesic, S. (2009). A new mechanistic model for MIC based on a biocatalytic cathodic sulfate reduction theory. In Proceedings of CORROSION 2009. *NACE International.*
15. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (pp. 4765–4774).
16. Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why tree-based models still outperform deep learning on tabular data. *In Advances in Neural Information Processing Systems*.
17. Permeh, S., Lau, K., & Duncan, M. (2021). Effect of crevice morphology on SRB activity and steel corrosion under marine foulers. *Bioelectrochemistry, 142*, 107922.
18. Skovhus, T. L., Eckert, R. B., & Rodrigues, C. (2017). Management and control of microbiologically influenced corrosion (MIC) in the oil and gas industry. *Journal of Biotechnology, 256*, 31–45.
19. Xu, L., et al. (2024). Prevention of severe pitting corrosion of 13Cr pipeline steel by biocide and biofilm dispersant. *Frontiers in Materials, 11,* 1407655.