Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

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

Bipul Bhattarai

bipul@contact.bipulbhattarai.com.np

University Of South Dakota https://orcid.org/0009-0001-4810-4007

Naina Maharajan

naina.maharjan@usd.edu

University Of South Dakota https://orcid.org/0009-0002-5302-3209

Sulav Dahal

sulav.dhala@mavs.uta.edu

University Of Texas at Arlington https://orcid.org/0009-0008-5736-5293

Research Article

Keywords: microbiologically influenced corrosion, quorum sensing, Random Forest, SHAP, biofilm, sulfate-reducing bacteria, machine learning, pipeline integrity

Posted Date: July 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10192769/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

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

Bipul Bhattarai
Dept. of Biomedical Engineering
University of South Dakota
Vermillion, SD, USA
bipulbhattarai70@gmail.com

Sulav Dahal
Dept. of Information Systems & Operations Management
University of Texas at Arlington
Arlington, TX, USA
sulav.dahal@mavs.uta.edu

Naina Maharjan
Dept. of Biomedical Engineering
University of South Dakota
Vermillion, SD, USA
naina.maharjan@coyotes.usd.edu

***Abstract*—Microbiologically 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, H_2S) 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 cross-validated 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.**

***Index Terms*—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]. Sulfate-reducing bacteria (SRB, e.g., *Desulfovibrio* spp.) and acid-producing bacteria (APB, e.g., *Pseudomonas aeruginosa*) are the primary culprits [3], [5].

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 [4]. 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 [4]. Similarly, *P. aeruginosa* QS governs the switch to sessile biofilm growth [5].

Despite this evidence, no published ML model for MIC incorporates QS signals as features. Existing approaches use only physicochemical parameters [10], [11] or pipeline operational data [6], without capturing microbial community dynamics. 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 [8] and colon cancer histological grading via deep learning [9] — yet this paradigm has not been applied to MIC with QS features.

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 [6], [7]; (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.

II. Related Work

ML for Corrosion. Chae et al. trained Random Forest on 365 compiled experiments for CO_2 corrosion (R^2=0.739) [12]. Zounemat-Kermani et al. applied meta-learning to sewer MIC using hydraulic parameters [10]. Yarveisy et al. proposed a Bayesian network for MIC likelihood integrating operational and microbiological data [11]. None of these incorporated QS signals.

XGBoost and LLM for MIC. Do et al. [6] demonstrated XGBoost with LLM-embedded tabular features for MIC prediction in 2D coating screening, achieving 98.5% accuracy. Bhattarai et al. [7] 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. [4] showed that QS modulation in *D. vulgaris* controls 162 genes including sulfate

reduction pathways (dsrABC, apsAB) and biofilm formation genes. Jia et al. [5] 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 [6]. (3) Full-text extraction from more than 15 key papers [3]–[5], [14]–[17] provided verified corrosion rates (range: 0.5–2,400 μm/year). 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 ($S_{syn} = f_{QS} \cdot S_{bio}$).

Biofilm community: SRB, IOB, APB, methanogen presence; biofilm/EPS presence; community complexity ($C = \sum organism_i$); biofilm aggression score ($S_{bio} = 0.35 \cdot srb + 0.25 \cdot bio + 0.20 \cdot eps + 0.10 \cdot apb + 0.10 \cdot iob$).

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 [18].

IV. Results

A. Model Comparison

TABLE I
Classification Performance (5-Fold CV + Test Set)

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 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 tree-based models outperforming deep learning on small tabular datasets [19].

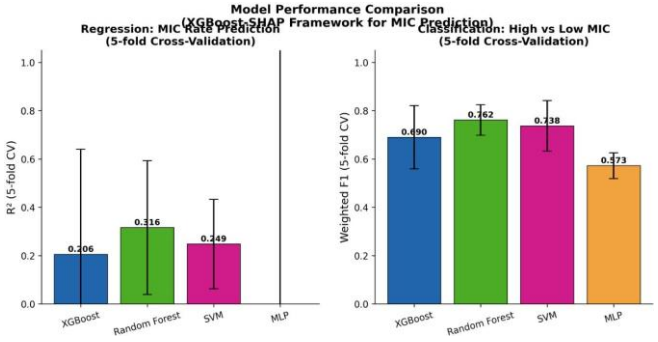


Fig. 1. Model comparison: regression R^2 (left) and classification F1 (right) across 5-fold CV. Random Forest achieves the highest F1=0.762 for MIC risk classification.

B. SHAP Feature Importance

The SHAP beeswarm (Fig. 2) and importance bar chart (Fig. 3) reveal the top five features by mean $|\phi|$: 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.

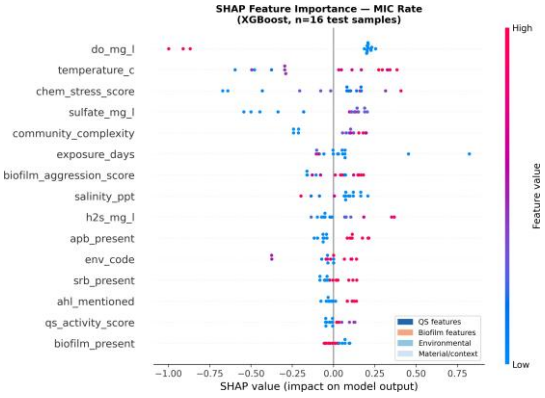


Fig. 2. SHAP beeswarm. Features ordered by mean $|\phi|$. DO, temperature, and chemical stress dominate. QS-composite features appear in top 10.

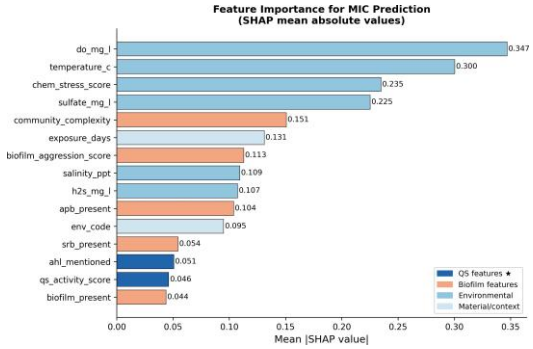


Fig. 3. Mean $|\phi|$ importance (top 15 features). Blue: QS. Orange: biofilm. Light blue: environmental.

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. [4] 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.

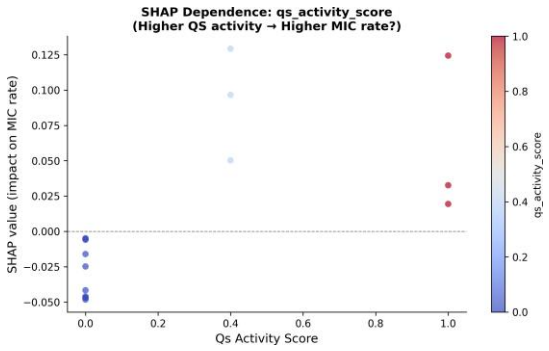


Fig. 4. SHAP dependence: QS activity score vs. MIC risk. Higher QS activity increases predicted MIC risk [4].

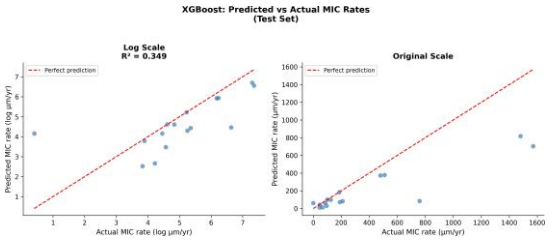


Fig. 5. Predicted vs. actual MIC rates (log scale). Moderate R^2=0.316 motivates binary risk classification.

V. Discussion

Why QS Features Matter. QS signals directly control the transition from low-corrosion planktonic growth to high-corrosion sessile biofilm formation [4]. 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 [4]. 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 [16]. Temperature (2nd) governs both microbial metabolism and electrochemical kinetics. Chemical stress score (3rd) reflects H_2S, sulfate, and pH — the primary SRB metabolic products [1].

Comparison to Prior Work. Our RF (F1=0.762, AUC=0.846) is comparable to Chae et al.'s RF (R^2=0.739 on 365 records) [12] despite a much smaller dataset, validating the curated literature approach. The MLP's poor performance (CV F1=0.573) is expected for small tabular datasets [19]. The multi-modal sensing paradigm used in stress classification [8] and cancer detection [9] motivates future integration of continuous electrochemical sensor streams with QS community features for real-time MIC monitoring.

Limitations. The 78-record dataset restricts generalizability. 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 [6], develop a multi-class severity classifier, and integrate the QS digital twin framework [7] 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 QS-derived 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 QS-derived biofilm community features.

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

References

[1] T. Gu, R. Jia, T. Unsal, and D. Xu, "Toward a better understanding of microbiologically influenced corrosion caused by sulfate reducing bacteria," *J. Mater. Sci. Technol.*, vol. 35, pp. 631–636, 2019.

[2] R. Jia, T. Unsal, D. Xu, Y. Lekbach, and T. Gu, "Microbiologically influenced corrosion and current mitigation strategies: A state of the art review," *Int. Biodeterior. Biodegrad.*, vol. 137, pp. 42–58, 2019.

[3] D. Xu and T. Gu, "Carbon source starvation triggered more aggressive corrosion against carbon steel by the *Desulfovibrio vulgaris* biofilm," *Int. Biodeterior. Biodegrad.*, vol. 91, pp. 74–81, 2014.

[4] G. Scarascia *et al.*, "Effect of quorum sensing on the ability of *Desulfovibrio vulgaris* to form biofilms and to biocorrode carbon steel in saline conditions," *Appl. Environ. Microbiol.*, vol. 86, no. 2, 2019.

[5] R. Jia *et al.*, "Microbiologically influenced corrosion of C1018 carbon steel by nitrate reducing *Pseudomonas aeruginosa* biofilm under organic carbon starvation," *Corros. Sci.*, vol. 114, pp. 180–189, 2017.

[6] T. Do *et al.*, "Utilizing XGBoost for the prediction of material corrosion rates from embedded tabular data using large language model," in *Proc. 2023 IEEE BIBM*, Istanbul, Turkey, 2023.

[7] B. D. S. Gurung, T. Do, S. Aryal, N. Maharjan, D. Bhandari, and B. Bhattarai, "Digital twin forecasting of quorum-sensing associated biofilm microorganisms in urban wastewater over a 30-week interval," in *Proc. 2025 IEEE BIBM*, Wuhan, China, 2025, pp. 7563–7565, doi: 10.1109/BIBM66473.2025.11356769.

[8] S. Dahal and B. Bhattarai, "Temporal dynamics analysis for cognitive vs physical stress differentiation using multi-modal wearable sensors," in *Proc. 2026 Int. Conf. Computing, Communication, Control and Cyber-Physical Systems*, 2026.

[9] S. Dahal and B. Bhattarai, "Two-stage deep learning pipeline for colon cancer detection and histological grading using transfer learning and ensemble methods," in *Proc. 2026 6th Int. Conf. Advanced Research in Computing (ICARC)*, 2026, pp. 1–6.

[10] M. Zounemat-Kermani and A. Aldallal, "Predicting microbiologically influenced concrete corrosion in self-cleansing sewers using meta-learning techniques," *Corrosion*, vol. 80, no. 4, pp. 338–348, 2024.

[11] R. Yarveisy, F. Khan, and R. Abbassi, "Data-driven predictive corrosion failure model for maintenance planning of process systems," *Comput. Chem. Eng.*, vol. 157, p. 107612, 2022.

[12] H. Chae *et al.*, "Machine learning prediction of corrosion rate for carbon steel in supercritical CO_2 environments," *Int. J. Greenhouse Gas Control*, 2022.

[13] T. L. Skovhus, R. B. Eckert, and C. Rodrigues, "Management and control of microbiologically influenced corrosion (MIC) in the oil and gas industry," *J. Biotechnol.*, vol. 256, pp. 31–45, 2017.

[14] L. Chen, B. Wei, and X. Xu, "Effect of sulfate-reducing bacteria (SRB) on the corrosion of buried pipe steel in acidic soil solution," *Coatings*, vol. 11, p. 625, 2021.

[15] L. Xu *et al.*, "Prevention of severe pitting corrosion of 13Cr pipeline steel by biocide and biofilm dispersant," *Front. Mater.*, vol. 11, p. 1407655, 2024.

[16] S. Permeh, K. Lau, and M. Duncan, "Effect of crevice morphology on SRB activity and steel corrosion under marine foulers," *Bioelectrochemistry*, vol. 142, p. 107922, 2021.

[17] T. Gu, K. Zhao, and S. Nesic, "A new mechanistic model for MIC based on a biocatalytic cathodic sulfate reduction theory," in *Proc. CORROSION 2009*, NACE, Houston, TX, 2009.

[18] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Adv. Neural Inf. Process. Syst.*, 2017, pp. 4765–4774.

[19] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why tree-based models still outperform deep learning on tabular data," in *Adv. Neural Inf. Process. Syst.*, 2022.