Title: Hybrid data-driven framework for interpretable prediction of nitrate and sulfate risks in coastal aquifers

Author names: Md. Abdullah-Al Mamun , Mst. Nazneen Aktar , Md Nashir Uddin , Mohaiminul Haider Chowdhury , Md. Saiful Islam , M. Safiur Rahman , Md. Ripaj Uddin , Md. Jakir Hossain , Anwar Zahid , Venkatramanan Senapathi , Abu Reza Md. Towfiqul Islam

Abstract: Coastal aquifers in southern Bangladesh are increasingly threatened by groundwater overexploitation, saltwater intrusion, and diffuse contamination, posing serious risks to safe water supply. While earlier studies have primarily examined salinity and overall water quality degradation, the spatial dynamics and controlling factors of nitrate (NO_3^-) and sulfate (SO_4^{2-}) contamination remain poorly constrained. This study introduces a novel explainable artificial intelligence (XAI) framework that integrates Conditional Inference Trees (CIT), Multiple Correspondence Analysis (MCA), and advanced machine learning (ML) algorithms to predict NO_3^- and SO_4^{2-} concentrations and elucidate their governing hydrogeochemical mechanisms. A total of 590 groundwater samples and 11 hydrochemical and environmental predictors were analyzed using CatBoost, Gradient Boosting Regression (GBR), and Artificial Neural Network (ANN) models. The tree-based CatBoost and GBR models were used for structured learning, while ANN captured complex non-linear dependencies, enabling a robust comparative assessment. Results show that saltwater intrusion and domestic wastewater discharge are the dominant drivers of SO_4^{2-} enrichment, whereas agricultural fertilizer inputs and salinity gradients control NO_3^- variability. SHAP-based interpretability analysis identified salinity, Ca^{2+}, and Na^+ as the key predictors for SO_4^{2-}, and salinity and K^+ for NO_3^-. MCA delineated contamination hotspots linked to anthropogenic activities and hydrogeochemical interactions. Learning curve analysis confirmed stable model performance without overfitting, while baseline model comparisons demonstrated that the proposed CatBoost and ANN frameworks achieved up to 4.5× higher predictive accuracy, underscoring their robustness and generalization capability. This study presents one of the first XAI-driven evaluations of NO_3^- and SO_4^{2-} contamination in coastal Bangladesh. By coupling interpretability with predictive modeling, it advances the understanding of multi-source groundwater contamination processes and provides a transferable decision-support framework for sustainable groundwater management in vulnerable coastal environments.

Key words: Coastal groundwater; Nitrate hotspot; Sulfate source; Multiple correspondence analysis; Conditional inference tree

10.1016/j.scitotenv.2025.181190

https://www.sciencedirect.com/science/article/abs/pii/S004896972502830X