American Journal of
Biomedical Science & Research

www.biomedgrid.com
ISSN: 2642-1747

Review Article

Machine Learning in Nitrate Pollution: Source Identification and Health Risk Assessment in Aquatic Systems

Zhuoran Zhang, Jiacheng Li*

School of Environment, Beijing Normal University, China

***Corresponding author:** Jiacheng Li, School of Environment, Beijing Normal University, Xinjiekouwai St, Haidian District, Beijing, P.R. China.

To Cite This article: *Zhuoran Zhang, Jiacheng Li*, Machine Learning in Nitrate Pollution: Source Identification and Health Risk Assessment in Aquatic Systems. Am J Biomed Sci & Res. 2025 29(3) AJBSR.MS.ID.003808,* **DOI:** 10.34297/AJBSR.2025.29.003808

Received: December 01, 2025; **Published:** December 09, 2025

Abstract

With the rapid development of urbanization, industry, and agriculture, nitrate pollution in water has become a global environmental issue threatening water resource security and human health. Nitrate is chemically stable and highly mobile, primarily originating from excessive application of agricultural fertilizers and discharge of industrial and domestic wastewater. Long-term ingestion of water with excessive nitrate levels significantly increases the risk of diseases such as blue baby syndrome and hypertension. Accurately identifying nitrate sources, predicting their spatial distribution, and assessing health risks are crucial for pollution prevention and control and human health. Traditional hydrochemical and isotopic techniques have limitations in regional-scale applications, while machine learning methods, with their powerful data processing and pattern recognition capabilities, provide new approaches for such research. This review systematically discusses the application and progress of typical machine learning models in accurately predicting the spatial distribution of nitrate, highlighting their significant advantages in handling nonlinear relationships and spatial variability. Furthermore, it explores health risk assessment methods for nitrogen pollution, quantitatively revealing the health risks posed by nitrate through drinking water pathways to adults and children by calculating health risk indices, and finds that children generally face higher health threats. Research indicates that combining advanced machine learning prediction models with health risk assessment can provide a scientific basis for precise source identification, spatial simulation, risk management, and remediation strategy formulation for groundwater nitrate pollution.

Keywords: Nitrate pollution, Machine learning, Spatial prediction, Health risk assessment, Pollution source apportionment

Introduction

With the acceleration of urbanization leading to population growth and the rapid development of industry and agriculture, the impact of human activities on the water environment has become increasingly significant, and many regions are facing challenges from nitrate pollution [1]. Long-term consumption of water with excessive nitrate levels harms human health, significantly increasing the risk of diseases such as blue baby syndrome, hypertension, diabetes, and methemoglobinemia [2]. Nitrate, as a typical pollutant in the water environment, is characterized by chemical stability and high mobility, primarily originating from the excessive use of nitrogen fertilizers in agriculture, and improper discharge of industrial wastewater and domestic sewage [3]. Studies indicate that the absorption and utilization efficiency of nitrogen fertilizers by crops is only about 30% [4], with a large amount of unabsorbed nitrogen retained in the shallow soil layer, entering groundwater through leaching, thereby causing groundwater

nitrate pollution [5]. Research points out that factors such as the extensive application of agricultural fertilizers, land use patterns, and wastewater discharge have a key influence on changes in groundwater nitrate concentration [6]. Traditional hydrochemical analysis and isotopic techniques are common methods for studying groundwater nitrate pollution, useful for determining the distribution and sources of nitrate. However, these two methods have certain limitations in large-scale regional applications. To more accurately identify the sources of NO_3^- in groundwater, the SIAR model and linear mixing models have been gradually widely applied to estimate the contribution proportion of nitrate sources, combined with GIS software to analyze pollution sources and their spatial distribution characteristics [7].

Machine Learning in Nitrate Source and Spatial Distribution Analysis

Existing studies, through field investigations of land use patterns and fertilization conditions, combined with groundwater sample analysis and meteorological and geospatial data, have systematically explored the response relationship between different land use types and groundwater nitrogen concentration. Using geographical detectors and machine learning methods, the main factors influencing groundwater nitrate-nitrogen can be identified, and its spatial distribution predicted. Machine learning, as a rapidly developing branch of artificial intelligence in recent years, simulates human thinking through computer software and hardware technology, enabling functions such as information collection, feature extraction, and classification modeling, providing new perspectives for studying scientific issues like nitrogen metabolism, cycling, and utilization. Due to its relatively low computational cost and high generalization ability, machine learning has gradually been applied to simulation studies of groundwater nitrate-nitrogen concentration [8]. Machine learning algorithms such as Random Forest, Logistic Regression, and Neural Networks have been progressively applied to predict the spatial distribution of groundwater nitrate. For example, *Sarkar et al.* [9] demonstrated that the Random Forest algorithm can effectively capture the complex relationships between environmental factors and nitrate concentration when processing large-scale data, possessing high prediction accuracy and stability. Existing machine learning models primarily comprehensively consider environmental factors such as climate, soil characteristics, human activities, and hydrogeological parameters [10].

Among the numerous machine learning methods, Random Forest usually performs better. For instance, *Band et al.* [11] compared the performance of four machine learning models—Cubist regression, Support Vector Machine, Random Forest, and Bayesian Artificial Neural Network—in modeling groundwater nitrate-nitrogen concentration, with results showing that Random Forest modeling performed the best. Another study [12] used three machine learning methods—Deep Neural Network (DNN), Extreme Gradient Boosting (EGB), and Multiple Linear Regression (MLR)—to predict groundwater nitrate pollution in the Mazandaran Plain in northern Iran, ultimately determining that the EGB method had the highest performance in nitrate concentration prediction, with distance to industrial areas, population density, groundwater depth, and evaporation rate being the key factors influencing nitrate concentration. Neural networks based on hydrological characteristics and oxygen content inversing the carbonate system and nutrient concentration models can be used to estimate key variables related to biogeochemistry in the global ocean, including the concentrations of three nutrients: NO_3^-, PO_4^{3-}, and Si(OH)4 [13]. Researchers constructed seven neural network models, as shown in Figure 2, with selected input variables including hydrological and biogeochemical parameters, spatial parameters, and temporal parameters. Based on a multilayer perceptron artificial neural network, the optimal structure was determined after multiple tests. [13] (Figure 1).

Nitrogen cycle processes such as nitrogen fixation, nitrification, nitrate reduction, denitrification, and anammox have potential significant importance for regulating nitrogen transport in rivers. In the process of nitrogen transformation, microbial communities provide key biogeochemical functions; however, current understanding of their diversity and internal community structure relationships remains insufficient. Developing new mathematical models to mine microorganisms with denitrification and phosphorus removal functions helps fully utilize microbial resources and discover new nitrogen and phosphorus metabolic pathways [14]. Traditional statistical methods can no longer meet the needs of large-sample microbiome studies. For example, Principal Component Analysis, Redundancy Analysis, and Canonical Correspondence Analysis screen key microorganisms by analyzing the correlation between dominant microorganisms and environmental factors and comparing changes in microbial communities before and after treatment, leading to discrepancies in results from different studies [15]. The predictive potential of omics and machine learning opens new avenues for environmental pollution management and status assessment, helping to reveal the impact of external factors on microbial community diversity and dominant microbial metabolic pathways in activated sludge systems, thereby promoting environmental bioremediation processes.

Leveraging Machine Learning for Health Risk Control Strategies

Applying machine learning methods to investigate and identify NO_3^- pollution high-risk areas has important theoretical and practical value and has gained widespread attention; however, research on nitrogen transformation processes and pollution risk assessment remains relatively limited. Currently, many scholars have established various risk assessment models based on the impacts of different pollutants on human health, among which the health risk assessment model proposed by the U.S. National Academy of Sciences is the most commonly used [10] (Figure 2).

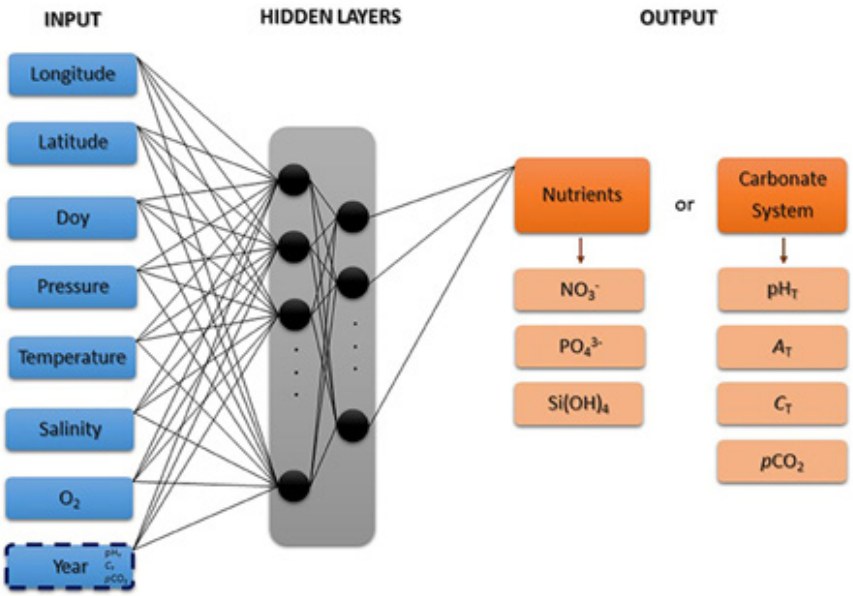

Figure 1: Multilayer perceptron neural network model for inversing nutrient concentrations [NO_3−, PO_4^{3}−, and $Si(OH)_4$] [13].

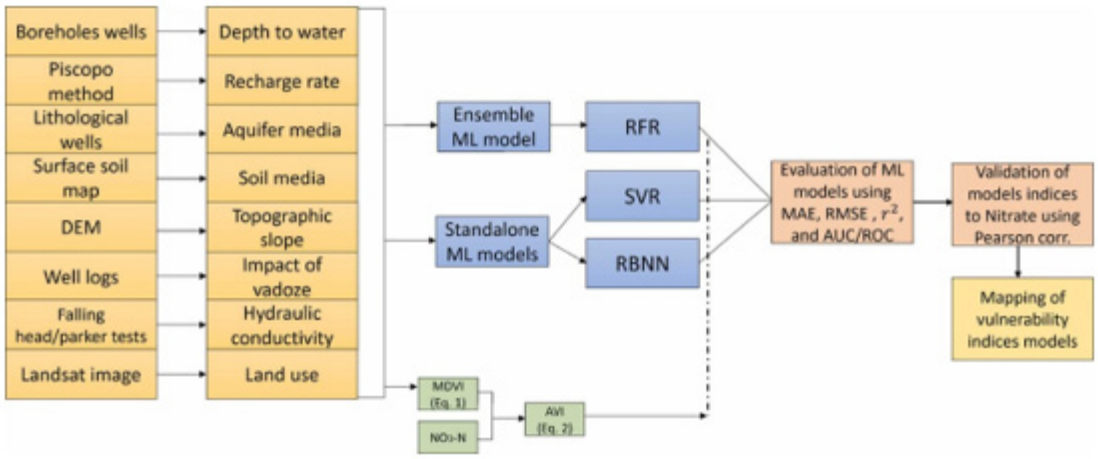

Figure 2: The general methodology for evaluating groundwater vulnerability [10].

On the basis of predicting pollution distribution, using health risk assessment methods to quantitatively assess the potential impact of NO_3^- pollution in groundwater on the health of different populations, calculating its non-carcinogenic health risk index (HI), and identifying high-risk groups can provide scientific basis and decision support for groundwater resource management, pollution prevention and control, and population health protection. *Lee et al*. [16] showed that the Geographically Weighted Random Forest model (GWRF) performed excellently in predicting the spatial distribution of groundwater NO_3^- on Jeju Island. Compared to the traditional classical linear mixing model, the Geographically Weighted Random Forest model fully considers spatial variability, yielding more accurate prediction results without spatial bias. SHAP analysis showed that key factors affecting groundwater nitrate-nitrogen concentration include average elevation, heavy clay proportion, agricultural land proportion, and urban land proportion. These results indicate that combining geographically weighted structures with machine learning models has broad prospects in groundwater modeling using geospatial data. Therefore, these findings can provide references for formulating targeted groundwater nitrate-nitrogen pollution control strategies [16] (Figure 3).

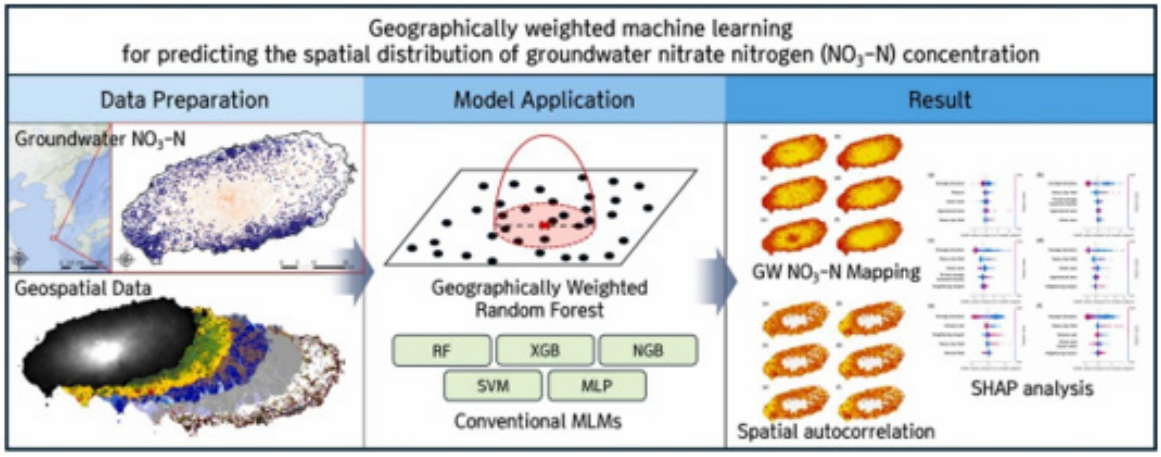

Figure 3: Technical framework of geographically weighted machine learning for predicting spatial distribution of groundwater nitrate-nitrogen concentration [16].

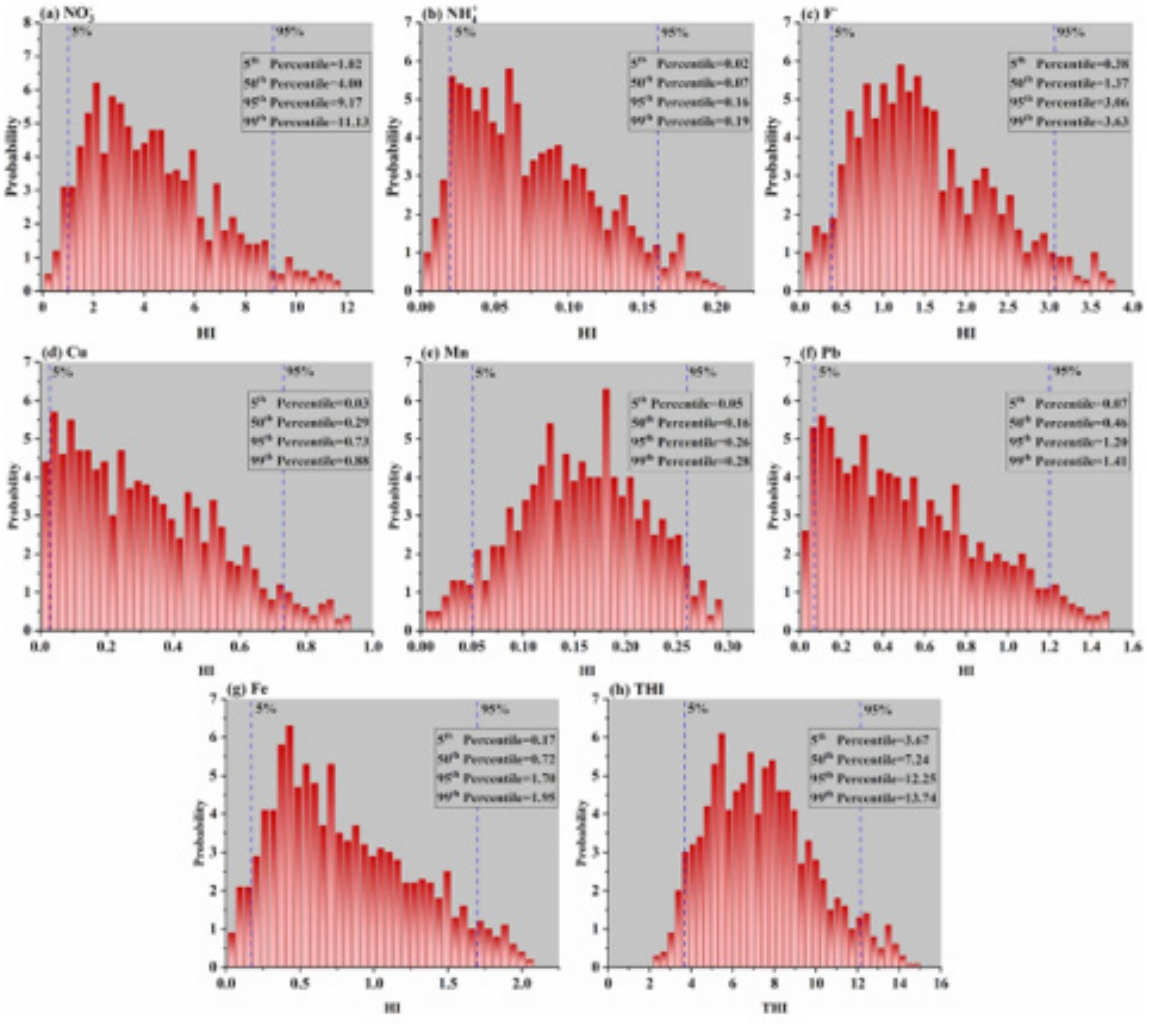

Figure 4: The probability distribution curve of HI and THI in children [17].

Related studies had insufficient extraction of feature information and subjective weights in principal component analysis. *Ruan et al.* [17] considered the uncertainty in the health risk assessment process and constructed a trapezoidal fuzzy number Monte Carlo stochastic simulation model, overcoming the inapplicability of triangular fuzzy numbers in handling highly discrete water quality data. According to the data in the figure, the Total Health Risk Index (THI) for children and adults was 2.15–15.03 and 0.86–7.261, respectively, far exceeding the safety threshold of 1. NO_3^- and F^- are the main risk control indicators for groundwater in the study area. The probability distribution of the total health risk index is overall left-skewed, indicating that most risk values fall in the medium-to-low risk zone. Combined with exposure parameters, this study selected the 95th percentile risk output value as the high-risk assessment value to provide a basis for decision-making. The 5th percentile values of the total health risk index for children and adults were 3.67 and 1.90, respectively, both exceeding the safety threshold of 1. This indicates that although most risk values are in the low-risk zone, there is still a probability exceeding 95% that high potential risks to human health exist. The total health risk index for children ingesting contaminated groundwater was twice that of adults, meaning that children face significantly higher health risks than adults, and the potential risk of children contracting diseases in the future requires high attention. The health risk index ranges for children and adults exposed to NO_3^- through drinking water were 0.07–11.76 and 0.03–5.76, respectively, with the vast majority of values exceeding 1, far above the acceptable level for humans. Among the child population, the 5th percentile value of the health risk index for nitrate exposure through drinking water was 1.02, and the 95th percentile value was as high as 9.55, both exceeding the safety threshold of 1, indicating significant risk; nitrate also poses potential non-carcinogenic risks to adults [17] (Figure 4).

Related research collected surface water and groundwater samples from a large rare earth mining area in southern China, analyzed their hydrochemical characteristics and stable isotopic composition, and used the health risk model recommended by the U.S. Environmental Protection Agency to assess the potential health risks of nitrate pollution to human health. Nitrate can harm the human body through two pathways: oral ingestion and dermal contact, with oral ingestion being the primary route. The chronic daily intake via oral ingestion (CDI_{Oral}) is calculated by formula (1):

$$CDI_{Oral} = (C_W \times I_R \times E_F \times E_D) / (B_W \times A_T) \quad (1)$$

Where, the Chronic Daily Intake (CDI) unit is μg/(kg·d); Nitrate Concentration in Water (C_W) unit is μg/L; Ingestion rate (I_R) unit is L/d (adult 2.2, child 1); Exposure frequency (E_F) unit is days/year (value 365); Exposure duration (E_D) unit is years (adult 70, child 10); Body weight (B_W) unit is kg (adult 70 kg, child 25 kg); Average Exposure Time (A_T) unit is days (adult 25550 days, child 3650 days).

This study used the Hazard Quotient (HQ) to assess non-carcinogenic risk, calculated by the following formula (2):

$$HQ = CDI / RfD \quad (2)$$

Where, the Reference Dose (RfD) for nitrate is 1600μg/(kg·d). A Hazard Quotient (HQ) >1 indicates potential non-carcinogenic risk [18] (Figure 5).

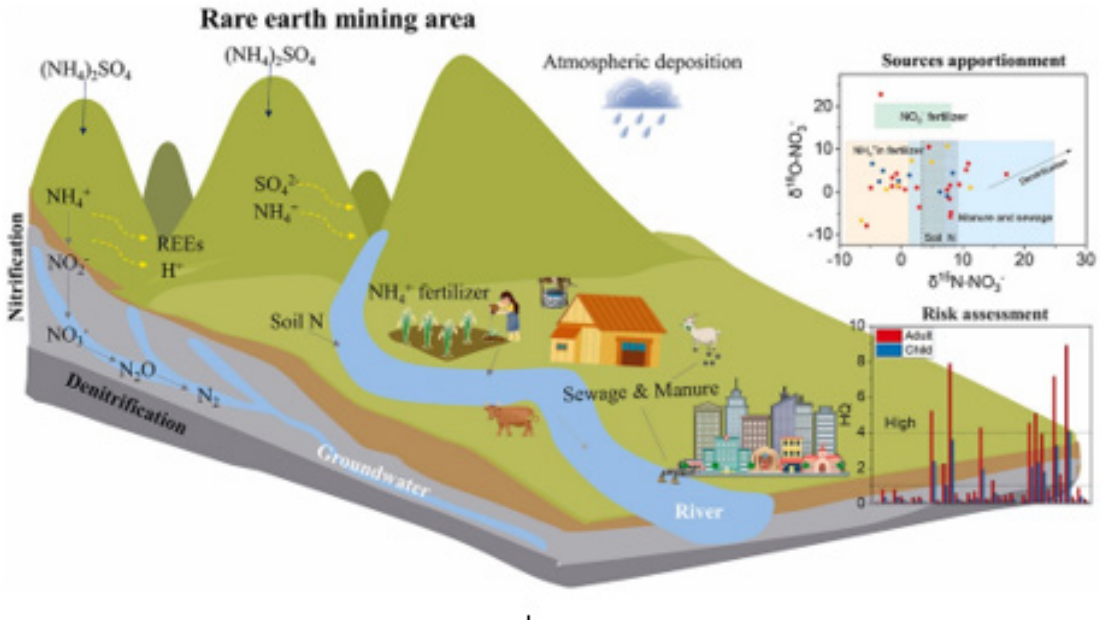

Figure 5: Schematic diagram of spatial distribution of nitrogen and its health risk [18].

Health risk assessment results showed that 31.4% of water samples posed medium to high non-carcinogenic risks, with high-risk areas mainly concentrated in the rare earth mining area. Furthermore, adults were more susceptible to non-carcinogenic health risks caused by nitrate than children. The nitrate nitrogen isotope ($\delta^{15}N–NO_3^-$) ranged from -6.43‰ to 17.09‰, and the nitrate oxygen isotope ($\delta^{18}O–NO_3^-$) ranged from -7.91‰ to 22.79‰, with this significant variability reflecting the comprehensive influence of multiple nitrogen sources and transformation processes. Additionally, rock weathering and dissolution affect groundwater chemical evolution, while precipitation and evaporation during the rainy and dry seasons did not cause significant changes in groundwater chemical composition. Adult health risk results showed that in the rainy and dry seasons, 27.69% and 52.31% of groundwater samples exceeded the acceptable limit for non-carcinogenic risk, respectively, while the corresponding proportions for children were 30.16% and 47.62%. The contribution percentages of nitrate, fluoride, and nitrate to the total risk were 61.29%, 28.71%, and 10.00% in the rainy season, and 68.84%, 20.85%, and 10.31% in the dry season, respectively. Contaminated water bodies can threaten human health through multiple exposure pathways. The human health risk assessment model links water environmental pollutant risks with human health, quantifies the degree of harm caused by water pollution to human health through various exposure pathways, and subsequently proposes targeted protection recommendations and measures. This study adopted the human health risk assessment method recommended by the U.S. Environmental Protection Agency (USEPA) to assess potential health effects caused by drinking groundwater. The calculation steps of this model are shown in (Figure 6) [19].

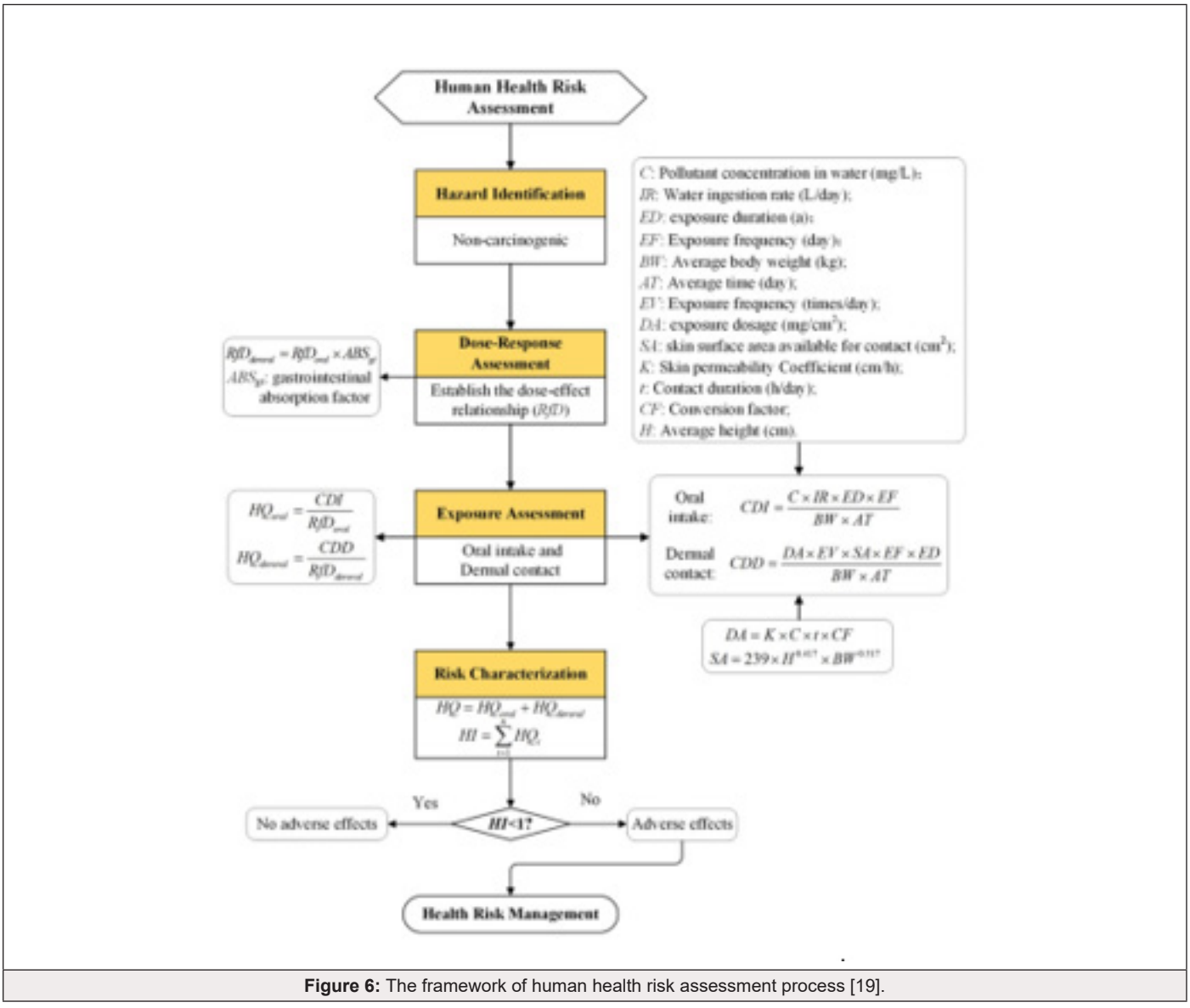

Figure 6: The framework of human health risk assessment process [19].

Conclusions

Through monitoring and machine learning methods, the sources and occurrence characteristics of nitrate pollution in the water environment can be accurately analyzed, and nitrate concentration prediction models can be constructed, providing scientific suggestions for nitrogen pollution prevention and control. The main means to control nitrate pollution include reducing anthropogenic nitrogen emissions, such as reducing fertilizer usage. Meanwhile, utilizing biogeochemical processes like denitrification can effectively mitigate nitrate pollution levels. Furthermore, health risk assessment helps accurately understand the risk of nitrate pollution in water, providing a scientific basis for water pollution treatment and health protection.

Acknowledgement

This work is supported by the Fundamental Research funds for the Central Universities and special fund of State Key Joint Laboratory of Environment Simulation and Pollution Control.

Conflict of Interest

None.

References

1. Tsai S W, Alexandra Z, Li Y R, James F Browning, A Robert Hillman, et al. (2025) Controlling solvation in conducting redox polymers for selective electrochemical separation of nitrate from wastewater. Nat Commun 16(1): 10207.
2. Re V, Kammoun S, Sacchi E, R Trabelsi, K Zouari, et al. (2021) A critical assessment of widely used techniques for nitrate source apportionment in arid and semi-arid regions. Sci Total Environ 775: 145688.
3. Zhang M, Zhi Y, Shi J, Laosheng Wu (2018) Apportionment and uncertainty analysis of nitrate sources based on the dual isotope approach and a Bayesian isotope mixing model at the watershed scale. Sci Total Environ 639: 1175-1187.
4. Wang J Q, Deng H L (2022) Study on the source and control of groundwater nitrate nitrogen pollution in intensive agricultural areas around Zhangye city. Modern Agri Res 28: 129.
5. Lukas K, Lutz B, Martin B (2020) Nation-wide estimation of groundwater redox conditions and nitrate concentrations through machine learning. Environ Res Lett 15: 064004.
6. Zhang H, Fan R, Li C (2024) Spatial distribution and sources analysis of nitrate in the Beiluo River Watershed based on nitrogen and oxygen stable isotope. Environ Pollut Con 46: 43.
7. Xu B, Zhang Y (2018) Contamination characteristics and human health risk assessment of nitrate in shallow groundwater at Jinghui irrigation district in Shaanxi province China. J Arid Land Resour Environ 32: 70.
8. Rodriguez Galiano V, Mendes PM, Garcia Soldado JM, Mario Chica Olmo, Luis Ribeiro, et al. (2014) Predictive modeling of groundwater nitrate pollution using Random Forest and multisource variables related to intrinsic and specific vulnerability: a case study in an agricultural setting (Southern Spain). Sci Total Environ 476-477: 189-206.
9. Sarkar S, Mukherjee A, Gupta SD (2022) Predicting regional-scale elevated groundwater nitrate contamination risk using machine learning on natural and human induced factors. ACS ES&T Eng 2: 689.
10. Elzain HE, Chung SY, Senapathi V, Selvam Sekar, Seung Yeop Lee, et al. (2022) Comparative study of machine learning models for evaluating groundwater vulnerability to nitrate contamination. Ecotox Environ Saf 229: 113061.
11. Band SS, Janizadeh S, Pal SC, Indrajit Chowdhuri, Zhaleh Siabi, et al. (2020) Comparative analysis of artificial intelligence models for accurate estimation of groundwater nitrate concentration. Sensors 20(20): 5763.
12. Gholami V, Booij M (2022) Use of machine learning and geographical information system to predict nitrate concentration in an unconfined aquifer in Iran. J Clean Prod 360: 131847.
13. Sauzède R, Bittig HC, Claustre H, Jean Pierre Gattuso, Louis Legendre, et al. (2017) Estimates of water-column nutrient concentrations and carbonate system parameters in the global ocean: A novel approach based on neural networks. Front Mar Sci 4: 128.
14. Henze MH, Abbreviated R (1987) A general model for single-sludge wastewater treatment systems. Water Res 21(5): 505-515.
15. Xie E, Zhao X, Li K, Panwei Zhang, Xiuhua Zhou, et al. (2021) Microbial community structure in the river sediments from upstream of Guanting reservoir: potential impacts of reclaimed water recharge. Sci Total Environ 766: 142609.
16. Lee YH, Kim C, Jeong H, Dongho Kim, Byeongwon Lee, et al. (2025) Geographically weighted machine learning for predicting the spatial distribution of groundwater nitrate nitrogen (NO3-N) concentration. J Hydrol Reg Stud 62: 102867.
17. Ruan DM, Bian JM, Wang Y, Juanjuan Wu, Zhiqi Gu, et al. (2024) Identification of groundwater pollution sources and health risk assessment in the Songnen Plain based on PCA-APCS-MLR and trapezoidal fuzzy number Monte Carlo stochastic simulation model. J Hydrol 632: 130897.
18. Zhang QY, Shu W, Li FD, Ming Li, Jun Zhou, et al. (2022) Nitrate source apportionment and risk assessment: a study in the largest ion-adsorption rare earth mine in China. Environ Pollut 302: 119052.
19. Wang YH, Li PY (2022) Appraisal of shallow groundwater quality with human health risk assessment in different seasons in rural areas of the Guanzhong Plain (China). Environ Res 207: 112210.