Feature importance analysis for enhanced interpretability of spectrophotometric Machine Learning (ML) models in water quality monitoring

Análisis de importancia de los parámetros en el aprendizaje automático como mejora en la interpretabilidad del modelado espectrofotométrico para monitoreo de calidad del agua

Laura Hernández-Alpízar[1], José Andrés Gómez-Mejía[2]

Hernández-Alpízar, L; Gómez-Mejía, J. A. Feature importance analysis for enhanced interpretability of spectrophotometric Machine Learning (ML) models in water quality monitoring. *Tecnología en Marcha*. Vol. 39 Nº especial sobre Inteligencia Artificial. Febrero, 2026. Pág. 276-284.

https://doi.org/10.18845/tm.v39i5.8521

1 Associate professor. Research Center on Environmental Protection (CIPA), Costa Rica Institute of Technology. Costa Rica.
lahernandez@itcr.ac.cr
https://orcid.org/0000-0002-9193-8429

2 Environmental engineer. Chemistry School, Costa Rica Institute of Technology. Costa Rica.
jagomezm@ieee.org
https://orcid.org/0009-0005-1769-7283

Keywords

UV-Vis spectroscopy; nitrate; water; spectral interference; Random Forest; XGBoost.

Abstract

Ultraviolet-visible (UV-Vis) spectrophotometry for real-time NO_3^- quantification in water is commonly affected by spectral interferences from Dissolved Organic Matter (DOM). This study evaluates the use of machine learning (ML) models for this task, using feature importance analysis as a method to enhance chemical interpretability and detect spectral interferences. Four algorithms were compared using a dataset of 29 surface water samples: PCA-Random Forest (PCA-RF), PCA-XGBoost, full-spectrum RF (All-RF), and full-spectrum XGBoost (All-XGB). Leave-one-out cross-validation (LOOCV) showed no significant performance differences among the models ($p = 0.182$), with mean RMSE values between 0.6 and 0.8 mg / L. Nonetheless, feature importance analysis revealed that PCA-based models depend on variance rather than chemical relevance, which limits their reliability. The full-spectrum XGBoost model demonstrated superior spectral interpretability, successfully identifying both the NO_3^- absorption peak (≈ 220 nm) and the DOM interference correction peak (≈ 260 nm). This suggests that XGBoost could be advantageous for continuous water monitoring systems due to its ability to identify spectral interferences.

Palabras clave

Espectroscopía UV-Vis; nitrato; agua; interferencia espectral; Random Forest; XGBoost.

Resumen

La espectrofotometría ultravioleta-visible (UV-Vis) para la cuantificación de NO3$^-$ en tiempo real en el agua es comúnmente afectada por interferencias espectrales por parte de la materia orgánica disuelta (DOM). Este estudio evalúa la interpretabilidad de los modelos de aprendizaje automático (ML) para esta tarea, enfocándose en el análisis de importancia de características como método para mejorar la interpretabilidad química y detectar interferencias espectrales. Se compararon cuatro algoritmos utilizando un conjunto de datos de 29 muestras de agua superficial: PCA-Random Forest (PCA-RF), PCA-XGBoost, RF de espectro completo (All-RF) y XGBoost de espectro completo (All-XGB). La validación cruzada (LOOCV) no mostró diferencias significativas en el rendimiento entre los modelos ($p = 0.182$), con valores medios de RMSE entre 0.6 y 0.8 mg / L. El análisis de importancia de características reveló que los modelos basados en PCA dependen de la varianza en lugar de la relevancia química, lo que limita su fiabilidad. El modelo XGBoost con el espectro completo mostró una interpretabilidad espectral superior, identificando con éxito tanto el pico de absorción de NO_3^- (≈ 220 nm) como el pico de corrección de interferencia de DOM (≈ 260 nm). Esto sugiere que el uso de XGBoost podría ser adecuado para sistemas de monitoreo continuo del agua debido a su capacidad para identificar las interferencias espectrales.

Introduction

Ultraviolet-Visible (UV-Vis) spectroscopy, based on the Beer-Lambert law, directly relates analyte concentration (substance of interest) with light absorbance at a specific wavelength [1]. This characteristic makes it desirable for continuous monitoring, especially when interferences are absent. Interferences can occur when other compounds absorb light at the same wavelength as the target analyte [2]. This is the case of nitrates (NO_3^-) quantification by UV-Vis spectroscopy

which is constrained by spectral overlap of Dissolved Organic Matter (DOM) [3]. When DOM interference is present, a dual-wavelength correction method (Standard Methods 4500 NO3 B) can be used to evaluate UV absorbance at 220 nm and 275 nm [4]. The spectroscopy method is not suitable for NO_3^- quantification if two times the peak intensity at 275 is more than 10% of the peak intensity at 220 nm [4]. This correction could be difficult for natural waters with highly variable and intricate organic compositions [5]. Linear methods, such as Partial Least Squares Regression (PLSR), assume linear relations and cannot grasp the complexity of a complete spectrum, often failing to consider the intricate relationships between analyte absorbance and interferences [6]. Machine learning (ML) algorithms can address the limitations of linear methods and empirical corrections by capturing full-spectrum information, enabling analysis of greater complexities [7].

Feature importance analysis is a methodology used to quantify the predictive influence that each input variable (feature) on the training of a ML model. This could enhance model interpretability. Despite its potential for understanding complex relationships in the data, this technique is still underutilized in spectrophotometric modeling [8]. If a ML algorithm is trained with the full spectrum, every absorbance at each wavelength is incorporated as an independent variable, so that the absorbances of analyte and interferences are completely considered in the spectral predictions [9], [10], [11].

A feature importance analysis of spectroscopic NO_3^- prediction by ML selected models is here studied to evaluate their application for spectral data analysis, considering chemical interpretability as a desirable characteristic.

Methodology

Data acquisition and preprocessing

An open dataset containing 29 samples of surface water was used [12]. It contained the target variable NO_3^- concentration, Total Organic Carbon (TOC) which is used as an indicator of DOM and absorbance measurements from 200.38 nm to 759.17 nm, in intervals of approximately 0.03 nm. Nonetheless, only the UV spectrum (from 200 nm to 400 nm) was selected to ignore non-chemical interference present in the visible spectrum. Therefore, the dataset contains 29 rows and 518 columns. Due to the high dimensionality, a Principal Component Analysis was performed on the spectral data, looking to represent 99% of the total variance.

Spectral data analysis

The general spectrum of the dataset was analyzed to determine UV absorbance response related to NO_3^- and DOM. A Pearson correlation analysis of UV absorbance and the concentration was performed to visually identify the relative absorbance peaks at specific wavelengths for each chemical species.

Modeling

Four machine learning methods were applied: Random Forest and XGBoost with principal components (PCA-RF and PCA-XGB), and Random Forest and XGBoost with the full spectra input (All-RF and All-XGB). For model evaluation, leave-one-out cross-validation (LOOCV) was applied, where a single data point is used for testing and the remaining data is used for training. The Root Mean Squared Error (RMSE) was calculated to assess model performance in every iteration. For the PCA-based models, the optimal number of factors or components was determined by iteratively evaluating LOOCV performance across the range of components

that explain 99% of the variance. To ensure model simplicity, the optimal number of principal components was selected using the "1 standard error rule", which recommends choosing the simplest model whose RMSE is within one standard error of the best one [13].

Statistical Analysis

The non-parametric Friedman test was applied to determine if there was a significant difference between the performance of the models (α=0.05). The null hypothesis of the test states that there is no significant difference in the RMSE among the four models being compared. Lastly, the models were trained with the complete dataset to obtain the feature importance analysis.

Software

The data analysis was performed in a Google Colab notebook, using the Python language (version 3.12.11) and the packages: 'pandas' (2.2.2), 'numpy' (2.0.2), 'sklearn' (1.6.1), 'scipy' (1.16.2) and 'xgboost' (3.0.5).

Results

Spectral data analysis

NO_3^- concentrations range from 0.21 mg / L up to 7.05 mg / L, with a median value of 3.05 mg / L. This range of target variables indicates sufficient variability for model training. TOC concentrations, related to DOM, range from 3.27 mg / L to 8.32 mg / L, with a median 6.99 mg / L, which indicates the presence of organic matter in the samples (Figure 1).

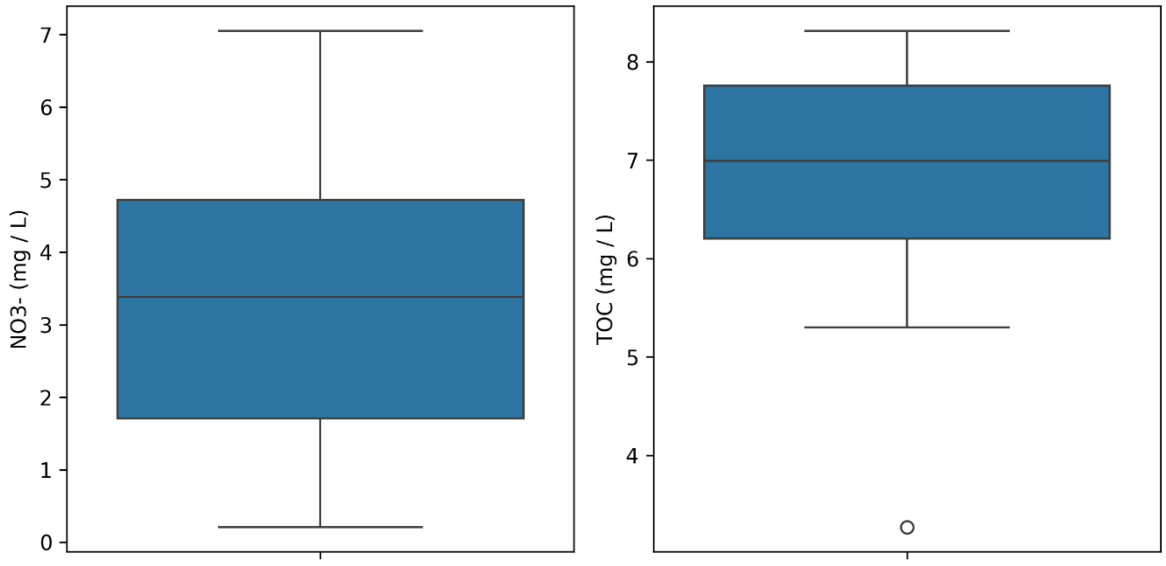

Figure 1. NO_3^- and TOC concentrations distribution of the surface water samples in the data set.

The absorbance spectra of the samples present a pronounced peak near 220 nm, and the relatively high the standard deviation around that area confirms the presence of NO_3^- with a high variability profile. Differences between the minimum (0.21 mg / L) and maximum (7.05 mg / L) NO_3^- concentration show the wide spectral range for model training (Figure 2).

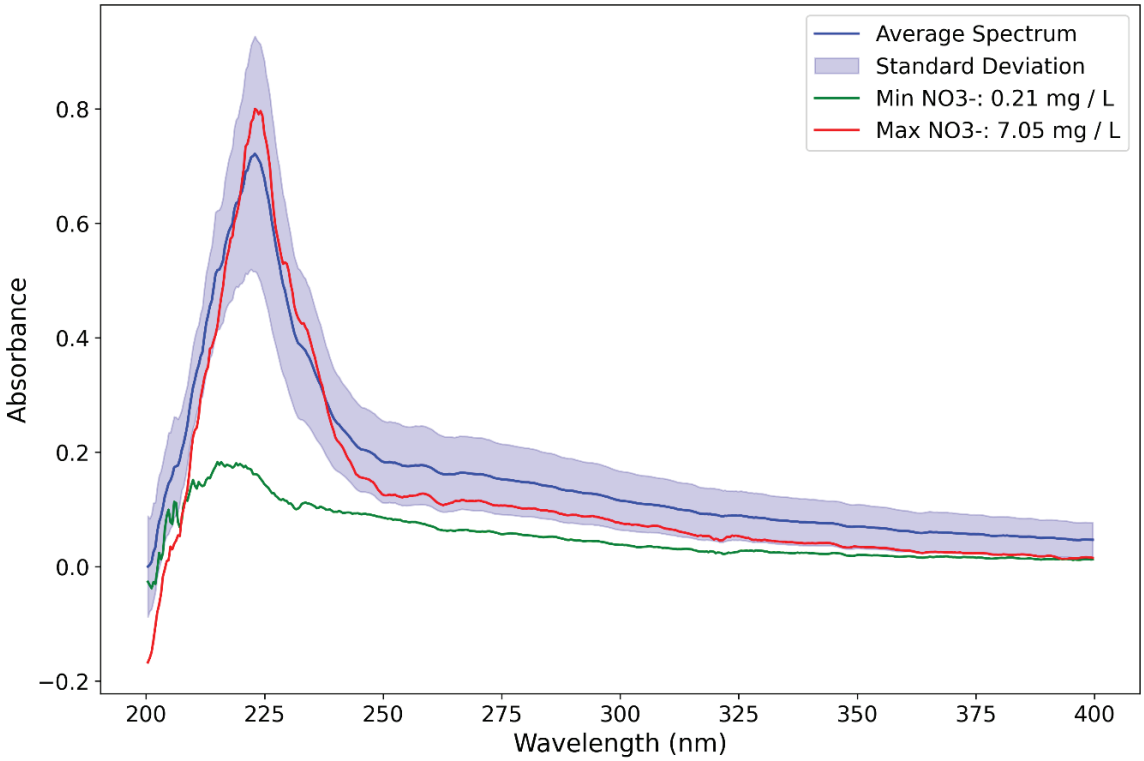

Figure 2. Full-Spectrum data range and variability for model training, defined by the average UV spectrum and NO_3^- concentration extremes.

The Pearson correlation analysis shows two different scenarios of UV spectra. The NO_3^- concentration shows a strong positive correlation (+ 0.9) centered at ≈ 220 nm, corresponding to its absorption peak. The TOC exhibits a correlation peak around 240 nm and a linear correlation from 300 to 400 nm (Figure 3). This figure shows a correlation overlap between the two variables, particularly around 210-240 nm, which is attributed to the known mutual interference signal from organic matter and the nitrate ion where both species absorb light in that spectral region [14].

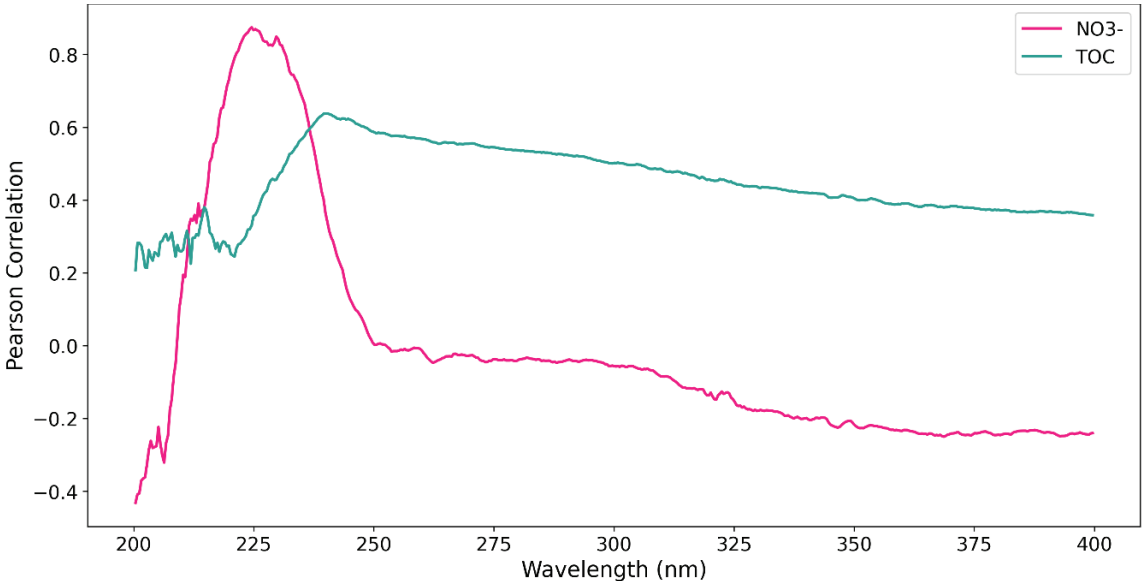

Figure 3. Pearson correlation between NO3- and TOC concentrations and UV absorbance.

Modeling and feature importance analysis

The mean LOOCV RMSE of the models is between 0.6 and 0.8 mg / L, with standard deviations around ± 0.5 and ± 0.7 mg / L. For the non-parametric Friedman test, a p-value of 0.182 was obtained. Therefore, the null hypothesis is rejected, which means that there is no significant difference between the models at 95% confidence interval.

The models trained with PCA output have their optimal performance with 2 components, giving the most internal importance to the second principal component (PC2) (Table 1).

Table 1. Internal importance analysis of the PCA-based models.

Principal component	PCA-based Model	
	RF (feature importance)	XGB (feature importance)
PC2	0.926	0.976
PC1	0.074	0.024

PCA decomposes the spectral information based on the explained variance of the dataset, with PC1 and PC2 each one individually explaining 86.6% and 6.39% of the total variance, respectively. Despite PC2 explaining the least variance, the two PCA-based models assigned the highest weight to it. The PC can be interpreted according to the wavelengths with the highest loadings: PC2 captures the NO_3^- peak absorption near ≈ 220 nm and PC1 captures the DOM signal [2] (Figure 4). Nevertheless, the PCA approach is based on non-chemical variance [15], which could limit its interpretability or be a performance constraint in cases with low natural chemical variability, particularly when the concentration of the analyte is in the range of background noise or other minor components [7].

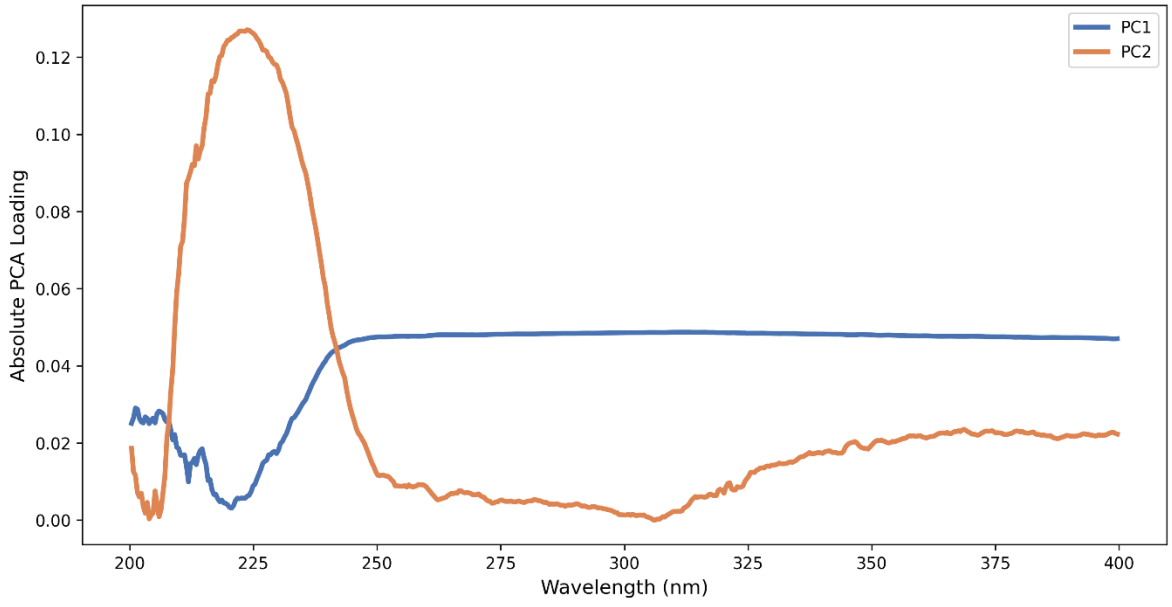

Figure 4. Wavelength relevance for the training of PCA-ML base models

Both ML algorithms trained for NO_3^- quantification with the full spectra showed a predictive importance weight in the NO_3^- signal around 220 nm. Nevertheless, XGB assigned a second importance weight near 260 nm, which normally works for DOM quantification (Figure 5). This result implies that XGB confirmed the DOM correction for spectroscopic NO_3^- determination that was deduced empirically [4]. It suggests that XGB, using the full spectra, has a better capacity to consider the DOM contents and the interference signal in the sample analysis than RF or the PCA-based models, for spectroscopy analysis.

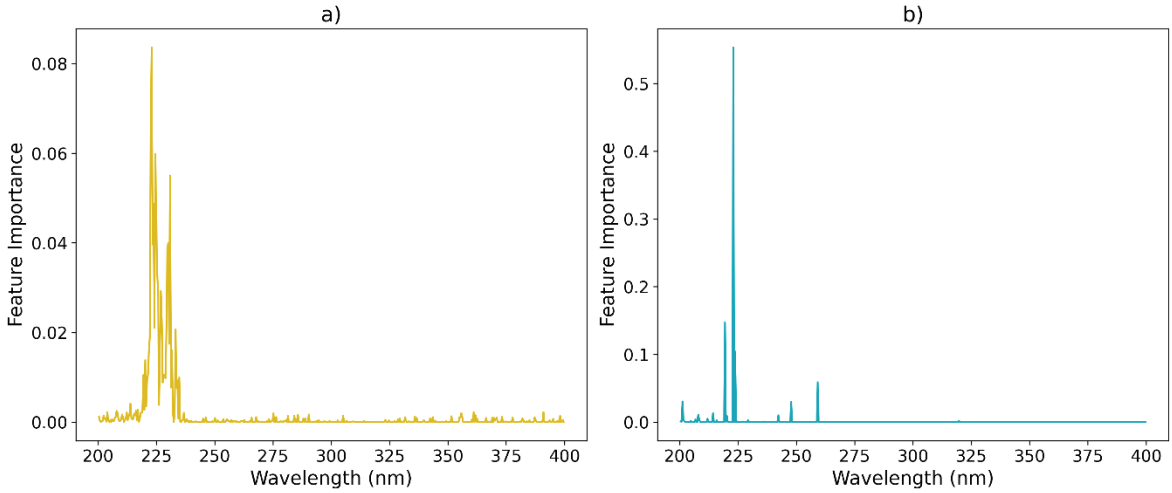

Figure 5. Relationship between feature importance and the full UV spectrum, for the models: a) Random Forest, b) XGBoost.

Table 2 shows a closer result of the feature importance analysis. RF mostly relies on the absorbance measurements around 220 nm, while XGB gives also significant importance to the absorbance at 259.17 nm.

Table 2. Wavelengths with the highest feature importance in the full-spectra ML models.

Model trained with the full spectrum			
RF		XGB	
Top 5 Wavelengths (nm)	Feature importance	Top 5 Wavelengths (nm)	Feature importance
222.97	0.084	222.97	0.553
222.57	0.066	219.42	0.147
224.55	0.060	223,76	0.104
230.87	0.055	259.17	0.059
223.76	0.049	201.18	0.030

The results of the XGBoost model rely on the boosting technique, which is an additive sequential learning process where each new tree is trained to correct the previous ones [16]. This refinement allows XGB to progressively assign greater weight to the most relevant wavelengths in this case. That is also the reason why its feature importance graph has few and clear peaks, which indicates a more refined feature importance distribution. On the contrary, RF builds each tree

independently on a random subset of data and features. Based on that, XGB could have more potential to distinguish interferences in spectrophotometric signals, which would be relevant in on-line monitoring systems [17].

Conclusions

Based on the LOOCV RMSE, no model was significantly better among the tested. Nonetheless, the XGB model could be the most advantageous model for chemical analysis and on-line determinations since it has superior chemical interpretability based on the feature importance. Also, the model performance could be improved with the inclusion of new and varied cases in on-line mode. PCA-based models may be less effective for this purpose because they primarily rely on explained variance, which can be limited in waters with low analyte concentrations. Additionally, PCA-based models do not effectively facilitate the identification of the specific DOM interference wavelength.

Future research should focus on integrating XGBoost models with monitoring systems enhanced by local interpretability methods (such as SHAP) to provide feature importance analysis for individual predictions.

Acknowledgements

The authors acknowledge the Vice-Rectorate of Research and Extension (VIE) of the Costa Rica Institute of Technology (TEC) for the funding of the research project: "*Desarrollo de una plataforma tecnológica escalable y modular para el registro de variables físicas y químicas asociadas a la calidad y abundancia del agua potable*", to which this article contributes.

References

[1] J. Villalobos-Villegas, A. Carrasquilla-Batista and L. Hernández-Alpízar, "Water quality monitoring station through nitrate measuring with IoT," in *2023 IEEE 5th International Conference on BioInspired Processing (BIP),* Alajuela, Costa Rica, 2023, doi: 10.1109/BIP60195.2023.10379419.

[2] Y. Guo *et al*, "Advances on Water Quality Detection by UV-Vis Spectroscopy," *Appl. Sci.,* vol. 10, *(19),* pp. 6874, 2020, doi: 10.3390/app10196874.

[3] M. F. Silva *et al*, "Usability of simplified UV–Vis spectrophotometric methods for the determination of nitrate in the presence of organic matter and chloride as interfering factors," *Wat. Pract. Tech.,* vol. 19, *(3),* pp. 1061–1070, 2024, doi: 10.2166/wpt.2024.043.

[4] T. R. Holm, "NO3- nitrogen (nitrate)," in *Standard Methods for the Examination of Water and Wastewater*, R. B. Bair, A. D. Eaton and E. W. Rice, Eds. Washington DC: American Public Health Association, 2017, pp. 1–2.

[5] Q. Huang *et al*, "Exploring the Impact of Dissolved Organic Matter on Nitrate Detection: Developing a Lab Experiment Using Standard Ultraviolet Spectrophotometry," *J. Chem. Educ.,* vol. 101, *(5),* pp. 2030–2038, 2024, doi: 10.1021/acs.jchemed.3c00958.

[6] T. J. Maguire *et al*, "Ultraviolet-visual spectroscopy estimation of nitrate concentrations in surface waters via machine learning," *Limnol Oceanogr Methods,* vol. 20, *(1),* pp. 26–33, 2022, doi: 10.1002/lom3.10468.

[7] Y. Lyu *et al*, "Development of statistical regression and artificial neural network models for estimating nitrogen, phosphorus, COD, and suspended solid concentrations in eutrophic rivers using UV–Vis spectroscopy," *Environ. Monit. Assess.,* vol. 195, *(9),* pp. 1114, 2023, doi: 10.1007/s10661-023-11738-0.

[8] J. Park *et al*, "Interpretation of ensemble learning to predict water quality using explainable artificial intelligence," *Sci. Total Environ.,* vol. 832, pp. 155070, 2022, doi: 10.1016/j.scitotenv.2022.155070.

[9] M. Cardia *et al*, "Machine Learning for the Estimation of COD from UV-Vis Spectrometer in Leather Industries Wastewater," *IJEPR,* vol. 11, pp. 10–19, 2023, doi: 10.11159/ijepr.2023.002.

[10] C. Chen *et al*, "Characteristic Wavelength Selection and Surrogate Monitoring for UV–Vis Absorption Spectroscopy-Based Water Quality Sensing," *Water,* vol. 17, *(3),* pp. 343, 2025, doi: 10.3390/w17030343.

[11] C. Fei *et al*, "Machine learning techniques for real-time UV-vis spectral analysis to monitor dissolved nutrients in surface water," in *AI and Optical Data Sciences II,* 2021, doi: 10.1117/12.2577050.

[12] J. Jiang and S. Tang. , 2022, "Spectral Water Quality Data," Mendeley Data, doi: 10.17632/d4vzbcxxcy.1.

[13] Y. Chen and Y. Yang, "The One Standard Error Rule for Model Selection: Does It Work?" *Stats,* vol. 4, *(4),* pp. 868–892, 2021, doi: 10.3390/stats4040051.

[14] S. M. Teague, "UV absorbing organic constituents," in *Standard Methods for the Examination of Water and Wastewater*, R. B. Bair, A. D. Eaton and E. W. Rice, Eds. Washington DC: American Public Health Association, 2017, pp. 1–2.

[15] F. L. Gewers *et al*, "Principal Component Analysis: A Natural Approach to Data Exploration," *ACM Comput. Surv.,* vol. 54, *(4),* 2021, doi: 10.1145/3447755.

[16] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining,* San Francisco, California, 2016, doi: 10.1145/2939672.2939785.

[17] S. Hossain *et al*, "Development of an Optical Method to Monitor Nitrification in Drinking Water," *Sensors,* vol. 21, *(22),* 2021, doi: 10.3390/s21227525.

Declaración sobre uso de Inteligencia Artificial (IA)

Para la revisión gramatical y ortográfica de este artículo, empleamos la herramienta de IA *ChatGPT*. Esta nos permitió identificar errores y mejorar la fluidez del texto. No obstante, realizamos una revisión final para garantizar que el artículo cumpliera con los estándares de calidad de la revista.