IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Instrumentation Mesure Métrologie
Vol. 25, No. 5, October, 2026, pp. 249-260
Journal homepage: http://iieta.org/journals/i2m

Development of Machine Learning-Based Near-Infrared Models for Predicting COD, BOD, and TSS in River and Irrigation Water Contaminated by POME

Mustaqimah[1,2,3*], Diswandi Nurba[1,2], Aris Munandar[1,3]

[1] Department of Agricultural Engineering, Faculty of Agriculture, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia
[2] Agricultural Mechanization Research Centre (PUSMEPTAN), Universitas Syiah Kuala, Banda Aceh 23111, Indonesia
[3] Institute of Agricultural and Environmental Development Research (LKPPL), Universitas Syiah Kuala, Banda Aceh 23111, Indonesia

Corresponding Author Email: mustaqimah@usk.ac.id

Copyright: ©2026 The authors. This article is published by IIETA and is licensed under the CC BY 4.0 license (http://creativecommons.org/licenses/by/4.0/).

https://doi.org/10.18280/i2m.250501

Received: 12 June 2026
Revised: 26 August 2026
Accepted: 8 September 2026
Available online: 31 October 2026

Keywords:
near-infrared spectroscopy, machine learning, palm oil mill effluent, chemical oxygen demand, biochemical oxygen demand, total suspended solids

ABSTRACT

Palm oil mill effluent (POME) significantly degrades water quality by increasing chemical oxygen demand (COD), biochemical oxygen demand (BOD), and total suspended solids (TSS). This study investigates near-infrared spectroscopy (NIR) combined with machine learning as a rapid measurement approach for these parameters. A dataset of 45 samples, including POME, river, and irrigation water, was analyzed. Spectral data (780–2500 nm) were preprocessed using Multiplicative Scatter Correction (MSC), Standard Normal Variate (SNV), and Savitzky–Golay (SG) filtering. Predictive models, including eXtreme Gradient Boosting (XGB), Ridge, Lasso, and Support Vector Regression (SVR), were developed using a calibration dataset and evaluated using a hold-out validation dataset. Among the evaluated models, SVR combined with SG preprocessing achieved the best performance for COD, with a coefficient of determination (R^2) of 0.94, root mean square error of prediction (RMSEP) of 90.26, and residual predictive deviation (RPD) of 3.41, and also showed strong predictive capability for TSS. In comparison, XGB with MSC yielded the highest accuracy for BOD, with R^2 of 0.97, RMSEP of 30.70, and RPD of 3.29. These results suggest that the integration of spectral preprocessing with machine learning offers a reliable and rapid approach for water quality assessment.

1. INTRODUCTION

Palm oil mill effluent (POME) is one of the most significant environmental pollutants generated by the palm oil industry, particularly in major producing countries such as Indonesia and Malaysia [1]. Due to its high organic load, oil and grease content, and suspended solids, untreated or poorly treated POME can severely degrade water quality when discharged into surrounding ecosystems. Key indicators of this pollution include chemical oxygen demand (COD), biochemical oxygen demand (BOD), and total suspended solids (TSS), which are widely used to evaluate the level of organic contamination in water bodies [2]. Elevated concentrations of these parameters can lead to oxygen depletion, eutrophication, and adverse impacts on aquatic organisms and ecosystem stability [3].

Conventional analytical methods for determining COD, BOD, and TSS are well established but present several limitations in terms of efficiency and practicality. COD measurement typically involves the use of strong oxidizing agents under controlled laboratory conditions, while BOD analysis requires a standard incubation period of five days [4]. Similarly, TSS determination involves filtration and gravimetric analysis, which are labor-intensive and time-consuming. These methods are not suitable for real-time monitoring or rapid assessment, especially in dynamic environmental systems where pollutant levels may fluctuate significantly. In response to these limitations, spectroscopic techniques have emerged as promising alternatives for rapid and non-destructive water quality analysis. Among these, near-infrared spectroscopy (NIR) has gained increasing attention due to its ability to provide fast measurements with minimal or no sample preparation [5]. NIR spectroscopy operates in the wavelength range associated with overtones and combination bands of molecular vibrations, allowing indirect characterization of chemical and physical properties of complex samples [6, 7].

Recent studies have demonstrated the potential of NIR and related spectroscopic methods for predicting water quality parameters, including COD, BOD, and TSS. For example, Qi et al. [7] and Ye et al. [8] reported that NIR spectroscopy combined with chemometric techniques could achieve satisfactory prediction accuracy for COD in wastewater systems. Similarly, UV-Vis and NIR-based approaches have been successfully applied for rapid monitoring of organic pollutants in surface water and industrial effluents. Despite these promising developments, predicting aggregate parameters such as COD, BOD, and TSS using NIR spectroscopy remains challenging. Unlike specific chemical compounds, these parameters represent cumulative properties of multiple constituents, including dissolved organic matter

and suspended particles. As a result, their relationship with spectral features is often indirect and highly nonlinear. This complexity necessitates the use of advanced data analysis techniques capable of capturing hidden patterns within spectral datasets.

Machine learning techniques have been increasingly adopted to enhance the predictive performance of spectroscopic models [7, 9]. Algorithms such as Support Vector Regression (SVR), artificial neural networks (ANN), and deep learning models have demonstrated strong capabilities in modeling complex and nonlinear relationships between spectral inputs and target variables. These approaches enable more accurate and robust predictions compared to traditional regression methods [10]. Furthermore, recent advancements in artificial intelligence have facilitated the development of real-time and automated water quality monitoring systems. AI-enabled sensors and integrated monitoring platforms have shown great potential for continuous assessment of environmental conditions, allowing for early detection of pollution events and improved decision-making in environmental management [11].

Nevertheless, most existing studies have focused on relatively homogeneous samples or controlled laboratory environments, which limits the generalizability of the developed models. In real-world conditions, water bodies are often influenced by multiple sources of contamination, resulting in highly heterogeneous sample compositions. For instance, the interaction between raw POME, river water, and agricultural runoff can lead to complex variations in organic matter and suspended solids, significantly affecting COD, BOD, and TSS levels [12, 13]. In addition, the variability of environmental samples poses significant challenges for model robustness and transferability. Models developed using limited or homogeneous datasets often fail when applied to new conditions, highlighting the importance of incorporating diverse sample types during model development. Recent studies have emphasized that the inclusion of heterogeneous datasets can significantly improve model performance and generalizability in environmental monitoring applications [14].

Therefore, there is a critical need to develop robust and reliable predictive models that can effectively handle heterogeneous environmental samples. In the context of POME contamination, such models are essential for accurately assessing water quality in complex and dynamic ecosystems [15]. Despite the growing body of research on spectroscopy and machine learning, studies that specifically address mixed water samples consisting of raw POME, river water, and irrigation water remain limited [16, 17]. This study aims to fill this research gap by developing machine learning-based models using NIR spectroscopy to predict COD, BOD, and TSS in mixed water samples contaminated by POME. Unlike previous studies that focus on homogeneous or controlled samples, this research incorporates diverse sample sources to better represent real environmental conditions [18, 19]. The integration of NIR spectroscopy with advanced machine learning algorithms is expected to improve prediction accuracy and model robustness [20, 21]. The outcomes of this study are anticipated to contribute to the development of rapid, non-destructive, and cost-effective tools for environmental monitoring, particularly in palm oil-producing regions. In addition, the proposed approach may support more efficient pollution control strategies and sustainable management of water resources affected by industrial effluents.

2. MATERIAL AND METHODS

2.1 Sample collection

A total of 45 water samples were collected, consisting of three categories: raw POME, river, and irrigation water, with 15 samples in each category. The POME samples were obtained from a palm oil mill located in Nagan Raya Regency, Aceh, Indonesia, specifically from the effluent treatment pond. River and irrigation water samples were collected from areas surrounding the palm oil mill, where contamination from POME discharge was suspected. The inclusion of multiple sample types was intended to represent heterogeneous environmental conditions and improve the robustness of the developed models. The total number of samples (n = 45) was considered adequate for machine learning modeling, particularly in spectroscopic-based environmental analysis, where relatively small but representative datasets are commonly used to develop reliable predictive models [6]. In the context of wastewater and POME monitoring, previous studies have shown that NIR spectroscopy combined with chemometric and machine learning approaches can successfully predict key parameters such as BOD and COD with good accuracy, even with limited sample sizes, provided that sufficient variability is captured [22]. POME itself is known to exhibit high variability in organic load and composition, making it suitable for data-driven modeling approaches that leverage spectral fingerprints. The samples were treated as a continuous concentration gradient rather than discrete classes, covering the full analytical range from low to high concentration. Samples from irrigation water are predicted at low concentration. Meanwhile, the extremely high concentration is from raw POME.

All samples were collected in clean high-density polyethylene (HDPE) bottles, transported to the laboratory, and stored under controlled conditions prior to analysis. Before measurement, samples were homogenized to ensure consistency. Specifically, samples were stored at 4 ℃ in airtight HDPE bottles and allowed to equilibrate at room temperature before analysis. Prior to each spectral measurement, samples were thoroughly mixed using a vortex mixer at maximum speed for 60 s to ensure homogeneity and minimize TSS settling. Spectral acquisition was initiated within 30 s after mixing to minimize particle sedimentation effects. Each sample was scanned in triplicate and the three spectra were averaged to form the final spectrum used for modeling, reducing the influence of random measurement noise.

2.2 Reference analysis of water quality parameters

The concentrations of COD, BOD, and TSS were determined using standard laboratory methods. COD was measured using the dichromate reflux method. BOD was determined using the five-day incubation method (BOD_5). TSS was measured using the gravimetric method. All analyses were conducted at the Laboratory of Basic Chemistry, Department of Chemistry, Faculty of Mathematics and Natural Sciences, Universitas Syiah Kuala, Banda Aceh, Indonesia. The laboratory analysis results were used as reference values for model prediction. All procedures refer to the standard methods outlined in Standard Methods for the Examination of Water and Wastewater, ensuring the accuracy and reliability of the measurements. The observed concentration ranges of

the analyzed parameters were as follows: COD: 1–1000 mg/L; BOD: 0.38–350 mg/L; and TSS: 8.7–709 mg/L. These wide ranges reflect the variability of contamination levels across different sample types and provide a robust dataset for model development.

2.3 Near-infrared spectral acquisition

NIR spectra were collected using a Fourier Transform Near-Infrared (FT-NIR) spectrometer, namely the Thermo Nicolet Antaris II FT-NIR Spectrometer [23], equipped with Integrated Thermo Operation® software. The instrument operates within a wavelength range of 780–2500 nm. The FT-NIR instrument was configured at 32 scans, a spectral resolution of 2.0 cm^{-1}, and an optical gain of 4×. Spectra were acquired in transflectance mode using a liquid sample module. Background calibration was performed at the beginning of each measurement session using a certified Spectralon diffuse reflectance standard, and instrument validation was confirmed prior to sample analysis. Prior to spectral acquisition, samples were equilibrated to room temperature to minimize temperature-induced spectral variations, as temperature changes can affect molecular vibrations and shift spectral features [24]. Each sample was scanned multiple times, and the resulting spectra were averaged to enhance signal stability and reduce noise. This averaging process helps minimize unwanted variations and improves the signal-to-noise ratio [25]. The acquired spectra contain information related to molecular vibrations and physicochemical characteristics of the samples, which can be correlated with water quality parameters through chemometric analysis [7, 26].

2.4 Spectral preprocessing

To enhance the quality of the spectral data and improve model performance, several preprocessing techniques were applied. Multiplicative Scatter Correction (MSC) was used to correct scattering and baseline effects, while Standard Normal Variate (SNV) was employed to normalize spectral variations by removing additive and multiplicative noise [22]. Savitzky–Golay (SG) filtering was applied to smooth spectral signals and reduce high-frequency noise through local polynomial fitting, using the following parameters: derivative order (m) = 1 (first derivative), polynomial order (p) = 3, and window length (w) = 11 data points. These parameters were selected to effectively remove baseline drift while enhancing spectral contrast. For MSC, the reference spectrum was computed as the mean spectrum of the training set. For SNV, normalization was applied using the mean and standard deviation of each individual spectrum. Both MSC and SNV were applied over the full wavelength range (780–2500 nm) without exclusion of water absorption bands. All preprocessing parameters were fitted exclusively on the calibration set and subsequently applied to the validation set without refitting. These preprocessing techniques are widely used in NIR spectroscopy to improve signal quality and model robustness [23]. Furthermore, the methods were applied both individually and in combination to determine the optimal preprocessing strategy for each predictive model, as different preprocessing approaches may yield varying levels of prediction accuracy [27]. The preprocessing methods and the parameters applied to each are summarized in Table 1.

Table 1. Summary of spectral preprocessing parameters

Preprocessing Method	Parameter	Value/Description
Without Correction (Raw)	–	No preprocessing applied; original spectra used directly
Multiplicative Scatter Correction (MSC)	Reference spectrum	Mean spectrum of training set
	Wavelength range	780–2500 nm (full range, no exclusion)
Standard Normal Variate (SNV)	Normalization basis	Mean and standard deviation of each individual spectrum
	Wavelength range	780–2500 nm (full range, no exclusion)
Savitzky–Golay (SG)	Derivative order (m)	1 (first derivative)
	Polynomial order (p)	3
	Window length (w)	11 data points

Note: All preprocessing parameters were fitted exclusively on training data (n = 31) and subsequently applied to the test set (n = 14) without refitting, to prevent data leakage.

2.5 Machine learning model development

Machine learning models were developed to predict COD, BOD, and TSS from NIR spectral data. These models learn the relationship between spectral inputs and target variables through model training and validation processes, enabling the prediction of sample properties [21]. The algorithms used in this study included eXtreme Gradient Boosting (XGB), Ridge Regression, Lasso Regression, and SVR, which are commonly applied in spectroscopic data analysis. In particular, machine learning techniques are well-suited for handling nonlinear relationships between spectral data and water quality parameters, thereby improving predictive performance. The dataset was used for both training and evaluation to assess model generalization ability. Hyperparameter optimization was performed using 5-fold cross-validation on the training set through the caret package in R. For SVR, a Radial Basis Function (RBF) kernel was used with the following search ranges: cost $C \in \{0.1, 1, 10, 100\}$, epsilon $\varepsilon \in \{0.01, 0.1, 0.5\}$, and gamma $\gamma \in \{0.001, 0.01, 0.1\}$. For XGB, the tuned parameters were: max_depth $\in \{3, 5, 7\}$, learning rate (η) $\in \{0.01, 0.1, 0.3\}$, nrounds $\in \{50, 100, 200\}$, and subsample $\in \{0.7, 0.8, 1.0\}$. For Ridge and Lasso, the regularization parameter lambda was tuned over the range {0.001, 0.01, 0.1, 1, 10, 100}. The full spectral range (780–2500 nm) was used as input features without dimensionality reduction. Model performance was evaluated using root mean square error of calibration (RMSEC) for the calibration set and root mean square error of prediction (RMSEP) for the validation set [28]. All models were implemented in the R programming language using relevant packages for machine learning and data analysis.

2.6 Model validation and evaluation

The dataset was randomly divided into two subsets using a 70:30 ratio, with a fixed random seed (set.seed(123) in R). The calibration set (70%, n = 31) was used for model development and training, including hyperparameter optimization through 5-fold cross-validation using the caret package in R. The validation set (30%, n = 14) was reserved for evaluating the

predictive performance of the developed models using samples that were not included in model training or hyperparameter optimization. Throughout this manuscript, the term "calibration set" refers to the dataset used for model training and development, whereas "validation set" refers to the hold-out dataset used to evaluate model predictive performance. No separate test set or hold-out validation was used in this study. No stratification by sample type was applied during the random data split, which represents a limitation of this study. Future research should include a truly independent dataset collected from different sampling dates or locations to further assess model generalizability [29]. Model performance was evaluated using the following statistical metrics: coefficient of determination (R^2), root mean square error (RMSE, mg/L), and residual predictive deviation (RPD). Higher R^2 values and lower RMSE values indicate better predictive performance. No transformation was applied to the response variables; all COD, BOD, and TSS values were used in their original mg/L scale. All RMSEC and RMSEP values are reported in mg/L. Preprocessing parameters were fitted on training data only and applied to the test set without refitting. The extremely low RMSEC values observed for XGB models (approaching zero in some cases) reflect severe overfitting to training data—a known behavior of ensemble tree models on small datasets—and should be interpreted in conjunction with the corresponding RMSEP values [30].

3. RESULTS AND DISCUSSION

3.1 Near-infrared spectra features

The resulting spectra illustrate the differences in chemical characteristics of each sample type, which are related to variations in the levels of organic matter and dissolved solids in them. Figure 1 displays the spectral features of NIR scans of POME, river water, and irrigation water samples. In general, the POME spectrum shows a higher absorption intensity than river water and irrigation water, reflecting a greater concentration of organic compounds and pollutant load. Conversely, river water and irrigation water show a relatively lower and smoother absorption pattern, indicating a lighter level of pollution. These variations in absorption patterns are an important basis for the process of extracting spectral features and developing prediction models for COD, BOD, and TSS levels using machine learning algorithms.

The spectral image shows several main peaks related to the molecular characteristics of POME, river and irrigation water samples. It should be noted that in aqueous systems, NIR spectra are dominated by water overtone and combination bands (particularly around 1450 nm and 1950 nm), which makes unambiguous assignment of spectral features to specific organic compounds difficult without complementary analytical evidence. The following tentative assignments are therefore offered with caution. The peaks with ranges of 1091–1126, 1270–1320, and 1640–1685 nm may suggest the presence of stretching vibrations and overtones of methyl groups (CH_3) originating from organic compounds such as lipids, fatty acids, and other hydrocarbon compounds commonly found in POME wastewater [31]. The absorption intensity in these regions is higher in POME than in river water and irrigation water, indicating a higher organic matter content.

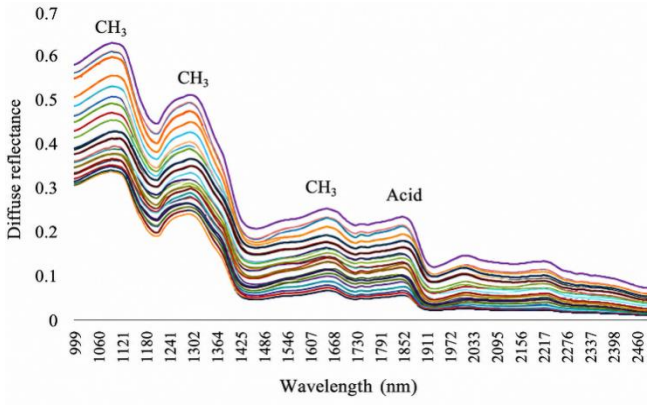


Figure 1. Spectral features of palm oil mill effluent (POME), river water and irrigation water

Meanwhile, the peak with a range of 1836–1870 nm may be tentatively related to the combination of O–H and C=O vibrations of carboxylic acids (acid groups), which may suggest the presence of polar compounds, such as organic acids and free fatty acids. These assignments should be interpreted cautiously, as spectral changes in aqueous NIR spectra are frequently influenced by variations in water structure and light scattering rather than direct absorption by specific compounds [32, 33]. This band can also be influenced by the contribution of water molecules, considering that the area is close to the H_2O absorption band. Overall, the presence of these peaks illustrates the differences in chemical composition between sample types and is an important spectral feature that contributes to the development of prediction models for COD, BOD, and TSS levels based on NIR and machine learning.

3.2 Calibration model of chemical oxygen demand contents

Table 2 presents the calibration and validation performance of the developed machine learning models for predicting COD concentrations from NIR spectral data under different spectral preprocessing techniques. The models evaluated include XGB, Ridge regression, Lasso regression, and SVR, each combined with raw spectra as well as MSC, SNV, and SG preprocessing. Model performance is assessed using R^2, RMSEC, RMSEP, and RPD, providing a comprehensive comparison of model accuracy and robustness across different preprocessing strategies.

The performance of machine learning models for predicting COD concentrations using NIR spectral data is presented in Table 2. Overall, the results demonstrate that the combination of appropriate preprocessing techniques and machine learning algorithms significantly influences model accuracy. Among all evaluated models, SVR combined with SG preprocessing exhibited the best performance, achieving a validation R^2 of 0.94 and the lowest RMSEP of 90.26. Additionally, this model produced the highest RPD value of 3.41, indicating excellent predictive capability (see Figure 2). In contrast, SVR models using raw spectra showed poor generalization performance (R^2 = 0.54), highlighting the importance of spectral preprocessing. This finding confirms that noise reduction and signal enhancement techniques such as SG filtering play a crucial role in improving model robustness. These results are consistent with previous studies reporting that SG preprocessing enhances spectral signal quality and improves prediction accuracy in spectroscopic models [34].

Table 2. Calibration and validation statistics for the prediction of chemical oxygen demand (COD)

Model	Pre-Processing	Calibration		Validation		RPD
		RMSEC (mg/L)	R^2	RMSEP (mg/L)	R^2	
XGB	Raw	0.02	1.00	203.50	0.97	1.28
	MSC	4.85	1.00	134.99	0.88	2.50
	SNV	10.02	1.00	105.30	0.90	2.47
	SG	0.00	1.00	239.99	0.89	1.28
Ridge	Raw	170.76	0.73	144.73	0.76	1.80
	MSC	206.28	0.57	203.83	0.58	1.66
	SNV	207.54	0.60	126.54	0.73	2.05
	SG	112.29	0.88	103.19	0.93	2.99
Lasso	Raw	165.65	0.74	144.52	0.76	1.80
	MSC	152.22	0.77	160.11	0.75	2.11
	SNV	125.03	0.86	148.70	0.88	1.75
	SG	140.37	0.84	94.82	0.91	3.25
SVR	Raw	99.17	0.91	202.74	0.54	1.28
	MSC	106.65	0.91	222.18	0.97	1.52
	SNV	64.00	0.96	206.54	0.96	1.26
	SG	**29.63**	**1.00**	**90.26**	**0.94**	**3.41**

Note: eXtreme Gradient Boosting (XGB); Support Vector Regression (SVR); root mean square error of calibration (RMSEC); coefficient of determination (R^2); root mean square error of prediction (RMSEP); residual predictive deviation (RPD); Multiplicative Scatter Correction (MSC); Standard Normal Variate (SNV); Savitzky–Golay (SG). Bold values indicate the best-performing model based on validation performance.

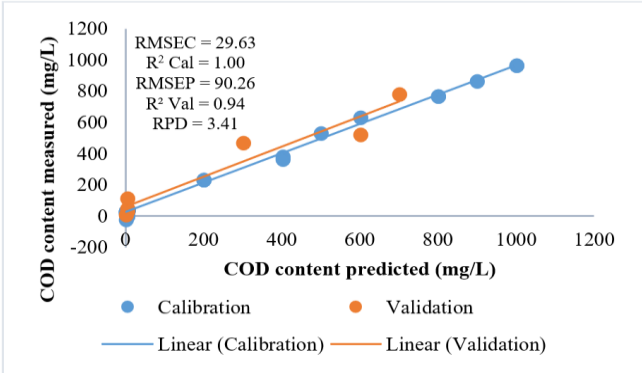


Figure 2. Calibration and validation scatter plot of measured versus predicted COD content (mg/L) by SVR-SG model
Note: Chemical oxygen demand (COD); Support Vector Regression (SVR); Savitzky–Golay (SG).

The distribution of COD values (1–1000 mg/L) in this dataset reflects the substantial difference in organic load between raw POME and environmental water samples. This distributional characteristic has important implications for model performance, as the relatively few high-COD POME samples may exert disproportionate influence on the calibration, potentially leading to higher prediction errors at the extremes of the concentration range. The dominance of low-COD samples in the dataset may also bias the model toward better performance in the lower concentration range, which should be considered when interpreting the overall RMSEP values.

A comparison of different machine learning algorithms reveals distinct performance patterns. The XGB model achieved perfect calibration performance ($R^2 = 1.00$) across all preprocessing methods; however, its validation performance varied significantly. For instance, XGB with MSC preprocessing achieved $R^2 = 0.88$ and RMSEP = 134.99, indicating moderate predictive performance. This discrepancy between calibration and validation results suggests potential overfitting, a common issue in powerful ensemble models such as XGB when applied to relatively small datasets [35]. Similarly, RF-type ensemble behavior (observed in XGB trends) has been reported to produce high training accuracy but may require careful tuning to generalize well. Linear models such as Ridge and Lasso regression showed relatively lower performance compared to nonlinear models. Although Ridge-SG achieved $R^2 = 0.93$, its performance was generally less stable across preprocessing methods. This is expected because linear models have limited capability in capturing complex nonlinear relationships between NIR spectra and water quality parameters. Previous studies have also reported that linear regression approaches often underperform compared to nonlinear machine learning models in spectroscopic applications [6].

Spectral preprocessing has a fundamental role in enhancing the performance of predictive models derived from NIR data. Among various preprocessing techniques, SG filtering is widely reported to provide superior performance, particularly when combined with regression-based models such as SVR and Ridge. This advantage is mainly due to its capability to smooth spectral signals by reducing high-frequency noise while preserving important spectral features associated with chemical constituents. Consequently, SG filtering enables better extraction of relevant information from NIR spectra, leading to improved prediction accuracy for water quality parameters such as COD and TSS [36]. Furthermore, SG filtering operates as a polynomial-based smoothing method that minimizes noise without significantly distorting the original signal structure, thereby maintaining meaningful absorption characteristics. This is particularly important in spectroscopic analysis, where subtle variations in spectral features are directly linked to chemical composition [37]. In contrast, raw spectral data are typically affected by various sources of interference, including noise, baseline drift, and scattering effects, which can obscure relevant chemical information and negatively impact model performance. Preprocessing techniques such as SG filtering help mitigate these issues by removing unwanted variations and enhancing signal quality, ultimately leading to more robust and reliable predictive models [38]. Although MSC and SNV preprocessing also contributed to performance improvement by correcting scatter-related variations, their effectiveness was relatively limited compared to SG filtering in this study [39]. This suggests that, for heterogeneous water samples such as POME and environmental water matrices, noise reduction plays a more dominant role than scatter correction in improving model accuracy. These findings highlight the

importance of selecting preprocessing strategies that are aligned with both the characteristics of the dataset and the nature of the target parameters, particularly when dealing with complex environmental samples [34].

An important consideration in interpreting the model performance is the potential influence of sample type differences on prediction accuracy. The dataset comprises three distinct sample types—raw POME, river water, and irrigation water—which differ substantially in their physicochemical characteristics and concentration ranges. Consequently, there is a risk that the models may partially learn to discriminate between sample classes rather than capturing the true quantitative relationship between spectral features and COD concentration. This phenomenon, known as class-based confounding, can artificially inflate validation performance metrics when the test set contains samples from all three classes. Nevertheless, the wide COD range covered by the dataset (1–1000 mg/L) and the heterogeneous nature of the samples suggest that the model has been exposed to sufficient spectral variability to learn genuine concentration-related patterns rather than class membership alone.

The use of hold-out validation provided a more realistic assessment of model performance by evaluating the models on the hold-out validation set, thereby reducing the risk of overfitting [40]. The satisfactory validation results indicate that the developed models are capable of maintaining predictive reliability despite the relatively limited dataset size (n = 45), which is a common constraint in environmental monitoring studies. A key strength of this study lies in the inclusion of heterogeneous water samples, such as raw POME, river water, and irrigation water, covering a wide range of contamination levels [3]. This diversity enhances the representativeness of the dataset while also introducing additional complexity due to variations in spectral characteristics across sample types. Despite this challenge, the strong performance of the SVR-SG model demonstrates that machine learning approaches can effectively capture complex and nonlinear relationships within heterogeneous datasets. This highlights the importance of incorporating diverse sample conditions during model development to improve generalizability and ensure practical applicability in real-world environmental monitoring scenarios.

The results of this study demonstrate that the integration of NIR spectroscopy with machine learning provides a rapid and non-destructive approach for COD prediction in POME-contaminated water. The strong predictive performance indicates that the models are capable of capturing spectral features associated with organic compounds contributing to oxygen demand, which are primarily reflected through overtone and combination bands of C–H and O–H bonds in the NIR region [6]. Compared to conventional laboratory methods, which are time-consuming and require chemical reagents, the proposed approach offers significant advantages in terms of analysis speed, cost-efficiency, and environmental sustainability. These benefits are particularly relevant for applications requiring frequent or continuous monitoring. Therefore, the developed method shows strong potential for real-time and in-situ monitoring of organic pollution in POME systems. This capability can support more efficient environmental management and early detection of contamination events in complex water systems [8].

3.3 Calibration model of biochemical oxygen demand contents

Table 3 summarizes the calibration and validation results of machine learning models used to predict BOD concentrations from NIR spectral data, incorporating different preprocessing approaches. The models assessed comprise XGB, Ridge regression, Lasso regression, and SVR, applied to both raw spectra and preprocessed data using MSC, SNV, and SG preprocessing methods. Model performance is assessed using the R^2, RMSEC, RMSEP, and RPD, allowing for a comprehensive evaluation of predictive accuracy and model robustness across different preprocessing strategies.

The performance of the developed machine learning models for predicting BOD concentrations is presented in Table 3. Overall, the results indicate that both the choice of algorithm and spectral preprocessing method significantly influence model accuracy and generalization capability. Among all evaluated models, XGB combined with MSC preprocessing exhibited the best predictive performance, achieving a validation R^2 of 0.97, RMSEP of 30.70, and RPD of 3.29. This result indicates excellent predictive ability, as RPD values greater than 3 are generally considered indicative of robust and reliable models for quantitative prediction (see Figure 3).

Table 3. Calibration and validation statistics for the prediction of biochemical oxygen demand (BOD)

Model	Pre-Processing	Calibration		Validation		RPD
		RMSEC (mg/L)	R^2	RMSEP (mg/L)	R^2	
XGB	Raw	0.08	1.00	151.59	0.02	0.81
	MSC	**3.77**	**1.00**	**30.70**	**0.97**	**3.29**
	SNV	5.91	1.00	82.02	0.77	1.86
	SG	0.80	1.00	50.74	0.94	2.80
Ridge	Raw	47.71	0.76	97.63	0.28	1.26
	MSC	67.12	0.58	60.10	0.60	1.68
	SNV	56.75	0.61	121.36	0.63	1.25
	SG	32.46	0.89	83.43	0.87	1.70
Lasso	Raw	45.79	0.78	92.44	0.35	1.33
	MSC	50.74	0.75	76.67	0.35	1.32
	SNV	40.05	0.80	86.62	0.75	1.76
	SG	47.46	0.80	90.15	0.91	1.58
SVR	Raw	35.61	0.87	122.59	0.14	1.00
	MSC	37.12	0.87	88.21	0.92	1.15
	SNV	17.72	0.97	110.92	0.48	1.37
	SG	37.27	0.91	97.14	0.88	1.46

Note: eXtreme Gradient Boosting (XGB); Support Vector Regression (SVR); root mean square error of calibration (RMSEC); coefficient of determination (R^2); root mean square error of prediction (RMSEP); residual predictive deviation (RPD); Multiplicative Scatter Correction (MSC); Standard Normal Variate (SNV); Savitzky–Golay (SG). Bold values indicate the best-performing model based on validation performance.

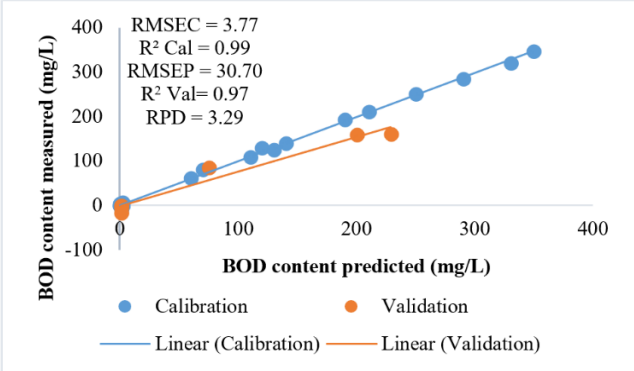


Figure 3. Calibration and validation scatter plot of measured versus predicted BOD content (mg/L) by XGB-MSC model
Note: Biochemical oxygen demand (BOD); eXtreme Gradient Boosting (XGB); Multiplicative Scatter Correction (MSC).

In contrast, XGB models using raw spectra showed extremely poor validation performance (R^2 = 0.02), despite achieving perfect calibration accuracy (R^2 = 1.00). This discrepancy strongly suggests overfitting, where the model captures noise rather than meaningful patterns in the data. Such behavior is commonly observed in powerful ensemble learning algorithms when preprocessing is insufficient or when the dataset is relatively small [35]. The validation R^2 of XGB-Raw for BOD prediction (R^2 = 0.02) can be explained by the sensitivity of the XGB algorithm to unprocessed spectral data. Without preprocessing, raw NIR spectra contain substantial baseline drift, multiplicative scattering effects, and inter-sample path length variations that are unrelated to BOD concentration. XGB's decision tree architecture interprets these noise-driven spectral differences as arbitrary splits during training, leading the boosting mechanism to amplify irrelevant patterns. The resulting model memorizes noise-specific features of the training set and fails to generalize to the test set. MSC effectively normalizes these baseline and scattering effects, allowing XGB to focus on genuine BOD-correlated spectral patterns, which explains the dramatic improvement from R^2 = 0.02 (Raw) to R^2 = 0.97 (MSC).

The highly skewed distribution of BOD values (0.38–350 mg/L), with the majority of samples clustered at low concentrations and a small number of high-BOD POME samples at the upper tail, introduces additional modeling complexity. High-BOD samples contribute disproportionately to the overall RMSEP, and their limited representation in the dataset may reduce model reliability at extreme concentrations. This distributional imbalance reinforces the importance of collecting additional samples across the full BOD range in future studies to improve calibration robustness.

A comparison across models shows that nonlinear models outperform linear models in predicting BOD from NIR spectra [41, 42]. While Ridge and Lasso regression achieved moderate performance, their validation R^2 values remained relatively low and inconsistent across preprocessing methods. For example, Ridge regression achieved its best performance with SG preprocessing (R^2 = 0.87), while Lasso achieved R^2 = 0.91 with SG. However, both models exhibited unstable performance across different preprocessing methods, indicating limited robustness. On the other hand, nonlinear models such as XGB and SVR demonstrated better adaptability to complex spectral patterns. This finding is consistent with previous studies, which reported that nonlinear machine learning algorithms are more effective in modeling the complex relationships between spectral data and water quality parameters [6].

Spectral preprocessing played a critical role in determining model performance, with MSC providing the most substantial improvement, particularly for the XGB model. This behavior can be attributed to the ability of MSC to effectively correct multiplicative and additive scattering effects caused by variations in particle size, turbidity, and optical path length, which are prominent in heterogeneous water samples containing suspended solids [34]. By normalizing these variations, MSC enhances the consistency of spectral signals, allowing the XGB algorithm to more effectively capture meaningful patterns related to BOD. The superior performance of XGB combined with MSC also reflects the sensitivity of ensemble-based models to input data quality. XGB relies on iterative error minimization through boosting, which can amplify noise and irrelevant variations if the input spectra are not properly corrected. Therefore, the application of MSC helps stabilize the input features, leading to improved model generalization and predictive accuracy. Although SG filtering also improved model performance, particularly for Lasso and SVR models, its impact was slightly less pronounced for XGB in this case. This suggests that, for BOD prediction in heterogeneous water systems, correcting scattering effects is more critical than smoothing high-frequency noise. These findings highlight that the effectiveness of preprocessing techniques is strongly dependent on both the nature of the target parameter and the characteristics of the machine learning algorithm. Similar observations have been reported in previous studies emphasizing the importance of selecting preprocessing strategies tailored to both data properties and modeling approaches [35].

The application of hold-out validation ensures that the reported model performance is reliable and not dependent on a specific data split. The high validation performance of the XGB-MSC model (R^2 = 0.97) indicates strong generalization capability, even with a relatively small dataset. However, the large variation in performance across preprocessing methods, particularly for XGB, highlights the sensitivity of machine learning models to input data quality. This reinforces the importance of proper preprocessing and model validation in spectroscopic applications.

The relatively wide range of BOD values (0.38–350 mg/L) reflects substantial variability in organic contamination levels across different sample types, including raw POME, river water, and irrigation water. Such variability increases the complexity of the modeling task due to the nonlinear and heterogeneous nature of the underlying chemical and physical processes. However, it also enhances the practical relevance of the developed models by better representing real environmental conditions. The strong performance of the XGB-MSC model suggests that ensemble-based machine learning approaches are capable of effectively capturing complex patterns within heterogeneous datasets [3, 35]. The boosting mechanism in XGB enables the model to iteratively learn from residual errors generated in previous stages, thereby progressively improving predictive performance and adapting to complex data distributions. This sequential learning strategy allows XGB to effectively model nonlinear relationships and interactions among variables, which is particularly beneficial for predicting BOD, a parameter influenced by both biochemical and physicochemical processes [41]. Moreover, the ability of XGB to handle heterogeneous datasets makes it

well-suited for environmental applications, where data often originate from diverse sources and exhibit high variability. Studies have shown that incorporating a wide range of environmental conditions during model training enhances the robustness and generalization capability of machine learning models. For instance, ensemble models such as XGB demonstrate strong performance across varying water quality parameters, indicating their capacity to capture complex environmental patterns [43]. Overall, these findings support the notion that both the boosting mechanism of XGB and the use of diverse training datasets are key factors contributing to improved model generalization and predictive accuracy in environmental monitoring applications [44].

Similar to COD prediction, the possibility of class-based confounding should be considered when interpreting BOD prediction results. The three sample types in this study exhibit markedly different BOD ranges—POME samples typically show much higher BOD values compared to river and irrigation water—which means the model may have learned class-specific spectral signatures in addition to genuine BOD-related features. The strong performance of XGB-MSC (R^2 = 0.97) may therefore partly reflect the model's ability to distinguish POME from environmental water samples based on overall spectral intensity differences, rather than purely from BOD-correlated absorption features. Future studies should include a leave-one-type-out validation to rigorously evaluate the model's ability to predict BOD across previously the hold-out validation set sample types.

The results demonstrate that the integration of NIR spectroscopy with machine learning, particularly XGB combined with MSC preprocessing, provides a reliable and efficient approach for predicting BOD in POME-contaminated water. The high predictive performance achieved by the model indicates its capability to capture complex relationships between spectral features and biodegradable organic content [6, 35]. Compared to conventional BOD analysis, which requires a standard incubation period of five days, the proposed method offers a significantly faster alternative while maintaining a high level of accuracy. This substantial reduction in analysis time highlights the potential of the approach for near real-time monitoring applications. Such capability is particularly valuable in palm oil-producing regions, where continuous monitoring of effluent discharge is essential to prevent environmental degradation. Therefore, the proposed method can support more efficient decision-making and pollution management strategies. These findings are consistent with previous studies highlighting the advantages of spectroscopy-based techniques for rapid water quality assessment [7].

3.4 Calibration model of total suspended solids content

The predictive capability of the developed models was evaluated using NIR spectral data under different preprocessing strategies. Table 4 reports the calibration and validation performance of the machine learning models for estimating TSS concentrations. Overall, the results indicate that model performance is strongly influenced by the choice of preprocessing technique, with certain methods yielding improved accuracy and robustness in both calibration and validation stages. The evaluated models include XGB, Ridge regression, Lasso regression, and SVR, combined with raw spectra as well as MSC, SNV, and SG preprocessing methods. Model performance is assessed using R^2, RMSEC, RMSEP, and RPD, providing a comprehensive comparison of model accuracy and robustness across different preprocessing strategies.

The performance of machine learning models for predicting TSS concentrations based on NIR spectral data is presented in Table 4. The results show that model performance varies considerably depending on both the selected algorithm and the applied spectral preprocessing method. Among all evaluated models, SVR combined with SG preprocessing achieved the best overall performance, with a validation R^2 of 0.85, RMSEP of 52.10, and RPD of 2.39. This indicates a good predictive capability, as RPD values between 2.0 and 3.0 are generally considered suitable for quantitative prediction in environmental applications (see Figure 4). In contrast, SVR models using MSC preprocessing showed very poor performance (R^2 = 0.03), highlighting the strong dependency of model performance on preprocessing strategy. This result emphasizes the importance of selecting appropriate preprocessing techniques to extract meaningful information from spectral data. The poor performance of SVR-MSC for TSS prediction (R^2 = 0.03) can be attributed to the fundamental mismatch between the MSC correction mechanism and the physical nature of TSS. TSS is primarily governed by light scattering from suspended particles, meaning that scattering-related spectral variations carry direct analytical information about suspended solid concentration. MSC was designed to remove multiplicative scattering effects by normalizing spectra against a reference. However, this normalization inadvertently removes the scattering information that correlates with TSS. This effectively destroys the signal of interest, leaving the SVR model with insufficient spectral information to build a meaningful calibration. In contrast, SG first-derivative transformation preserves scattering-related spectral differences while removing baseline offsets, which explains the superior performance of SVR-SG for TSS prediction.

The TSS distribution (8.7–709 mg/L) is similarly right-skewed, driven by the high suspended solid content of raw POME samples. The uneven distribution of samples across the concentration range may affect model generalizability, particularly at high TSS values where the relationship between light scattering and spectral response becomes increasingly nonlinear. Future studies should aim for a more balanced representation of samples across the full TSS range to improve model robustness.

A comparison across different algorithms reveals that nonlinear models again outperform linear models, although the performance gap is less pronounced than in COD and BOD prediction. The XGB model showed strong calibration performance ($R^2 \approx 1.00$), but its validation results were inconsistent. For example, XGB-SG achieved R^2 = 0.86, while XGB-MSC dropped significantly to R^2 = 0.21. This pattern suggests that XGB is highly sensitive to preprocessing and may suffer from overfitting when the input data are not optimally conditioned. Linear models such as Ridge and Lasso regression exhibited relatively lower and more stable performance, with validation R^2 values generally ranging between 0.39 and 0.81. However, their inability to capture nonlinear relationships limits their predictive accuracy compared to nonlinear approaches. These findings are consistent with previous studies indicating that nonlinear machine learning models are better suited for handling complex spectral data and heterogeneous environmental samples [6].

Table 4. Calibration and validation statistics for the prediction of total suspended solids (TSS)

Model	Pre-Processing	Calibration		Validation		RPD
		RMSEC (mg/L)	R^2	RMSEP (mg/L)	R^2	
XGB	Raw	10.20	1.00	91.57	0.82	2.29
	MSC	3.71	1.00	156.59	0.21	1.08
	SNV	25.42	0.99	185.92	0.38	0.28
	SG	10.76	1.00	182.68	0.86	0.68
Ridge	Raw	125.69	0.51	102.77	0.80	2.04
	MSC	143.73	0.42	124.82	0.39	1.35
	SNV	141.14	0.49	130.45	0.53	0.40
	SG	92.49	0.79	92.17	0.70	1.35
Lasso	Raw	129.90	0.49	115.19	0.81	1.82
	MSC	150.79	0.37	129.03	0.41	1.31
	SNV	100.58	0.75	198.35	0.50	0.27
	SG	130.54	0.66	77.62	0.71	1.60
SVR	Raw	132.07	0.49	114.02	0.83	1.84
	MSC	140.58	0.54	176.20	0.03	0.96
	SNV	152.95	0.59	42.61	0.60	1.23
	SG	**121.90**	**0.68**	**52.10**	**0.85**	**2.39**

Note: eXtreme Gradient Boosting (XGB); Support Vector Regression (SVR); root mean square error of calibration (RMSEC); coefficient of determination (R^2); root mean square error of prediction (RMSEP); residual predictive deviation (RPD); Multiplicative Scatter Correction (MSC); Standard Normal Variate (SNV); Savitzky–Golay (SG). Bold values indicate the best-performing model based on validation performance.

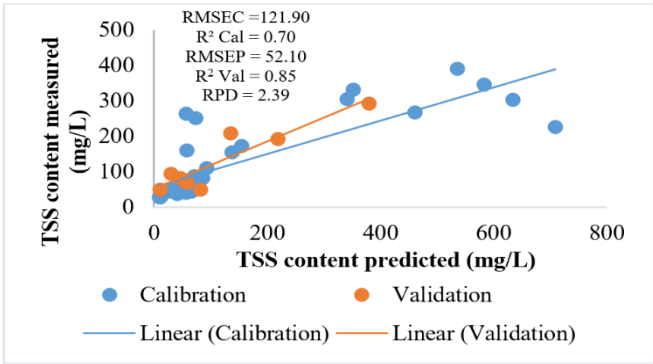


Figure 4. Calibration and validation scatter plot of measured versus predicted TSS content (mg/L) by the SVR-SG model
Note: Total suspended solids (TSS); Support Vector Regression (SVR); Savitzky–Golay (SG).

Spectral preprocessing plays a crucial role in improving model performance for TSS prediction, which is predominantly governed by physical rather than chemical characteristics of the samples. Among the evaluated preprocessing techniques, SG filtering consistently enhanced model performance, particularly for SVR and XGB models. This improvement can be attributed to the ability of SG filtering to effectively reduce high-frequency noise while preserving important spectral features related to light scattering caused by suspended particles. Such preservation is essential because these scattering-induced variations carry significant information about the physical properties of the sample [21]. In contrast, MSC and SNV preprocessing showed inconsistent effects across models. In some cases, such as SVR-MSC, preprocessing resulted in a significant decline in performance, suggesting that excessive correction of scattering effects may inadvertently remove relevant information linked to suspended solids. This highlights the importance of selecting preprocessing techniques that are aligned with the underlying nature of the target parameter. For TSS prediction, where scattering effects carry meaningful information, noise smoothing appears to be more beneficial than aggressive scatter correction. These findings are consistent with previous studies emphasizing the critical role of appropriate preprocessing strategies in spectroscopic modeling [34].

The use of hold-out validation provides a more reliable assessment of model robustness than calibration statistics alone, because performance is evaluated on samples excluded from model fitting. The satisfactory validation performance achieved by the SVR-SG model indicates its ability to generalize beyond the calibration dataset, despite the relatively limited sample size. This suggests that the model is capable of capturing the underlying relationships between spectral features and target variables rather than merely fitting noise [44, 45]. However, the variability in model performance observed across different preprocessing techniques highlights the sensitivity of spectroscopic machine learning models to data quality and feature conditioning. Spectral preprocessing plays a critical role in shaping the input feature space, and inappropriate preprocessing can either suppress relevant information or retain unwanted noise and artifacts, ultimately affecting model accuracy and stability. This sensitivity reflects the strong dependence of model performance on how well the preprocessing techniques preserve relevant spectral information while minimizing noise and distortions. Therefore, careful selection and optimization of preprocessing strategies are essential to ensure robust and reliable predictions in spectroscopic applications, particularly when dealing with heterogeneous environmental datasets [6, 34].

The potential for class-based confounding is particularly relevant for TSS prediction, given that TSS is predominantly governed by physical scattering characteristics rather than specific chemical absorption features. POME samples typically contain significantly higher suspended solid loads compared to river and irrigation water, creating distinct spectral signatures associated with turbidity and particle scattering. While the SVR-SG model achieved satisfactory validation performance, it cannot be excluded that part of this performance reflects discrimination between high-turbidity POME and lower-turbidity environmental water samples. A per-class RMSEP analysis and leave-one-type-out validation would provide more definitive evidence of the model's ability to generalize across different water matrices, and is recommended as a priority for future work.

The TSS concentration range (8.7–709 mg/L) reflects substantial variability in suspended solids across different

sample types, including raw POME, river water, and irrigation water. This variability introduces significant modeling challenges due to differences in particle size distribution, turbidity, and optical properties, all of which influence light scattering behavior in the NIR region. Despite these complexities, the relatively strong performance of the SVR-SG model suggests that machine learning approaches are capable of capturing the indirect and nonlinear relationships between scattering-related spectral features and TSS levels. The ability of SVR to model complex feature spaces, combined with the noise-reducing capability of SG filtering, likely contributed to this performance. The inclusion of heterogeneous samples further enhances the robustness of the developed models by exposing them to a broader range of environmental conditions during training. This improves their generalizability and practical applicability in real-world monitoring scenarios. These findings support previous studies emphasizing that incorporating diverse datasets is essential for developing reliable spectroscopic models in environmental applications.

The results confirm that the integration of NIR spectroscopy with machine learning provides a feasible and efficient approach for the rapid prediction of TSS in POME-contaminated water systems. The ability of the model to capture scattering-related spectral features associated with suspended particles highlights its effectiveness in estimating TSS, which is primarily governed by physical properties rather than chemical composition. Compared to conventional gravimetric methods, which are labor-intensive and require extensive laboratory procedures, the proposed approach offers a significantly faster and more efficient alternative while maintaining satisfactory predictive performance. This advantage is particularly important for applications requiring near real-time monitoring, where timely information is critical for environmental management and pollution control. Therefore, the developed method has strong potential to support continuous monitoring systems in complex environmental settings. These findings are consistent with previous studies highlighting the benefits of spectroscopic techniques for rapid water quality assessment [7].

4. CONCLUSIONS

This study successfully developed machine learning-based models combined with NIR spectroscopy for the prediction of COD, BOD, and TSS in mixed water samples contaminated by POME. The results demonstrate that model performance is strongly influenced by both the choice of algorithm and spectral preprocessing techniques. Among the evaluated models, SVR combined with SG preprocessing showed the best performance for COD and TSS prediction, while XGB combined with MSC yielded the most accurate results for BOD prediction. The obtained validation results, with high R^2 and acceptable error values, indicate that the developed models are capable of reliably predicting water quality parameters across a wide range of concentrations. The use of heterogeneous samples, including raw POME, river water, and irrigation water, enhances the robustness and real-world applicability of the models. This approach allows the models to better represent environmental variability, which is often overlooked in studies using homogeneous datasets. Overall, the integration of NIR spectroscopy with machine learning provides a rapid, non-destructive, and cost-effective alternative to conventional laboratory methods for water quality assessment. This method has strong potential for real-time monitoring and environmental management in palm oil-producing regions. For long-term practical applications, the NIR-based models require periodic recalibration, retraining, and spectral drift monitoring to maintain robust prediction accuracy and reliability under changing environmental and operational conditions. A limitation of this study is the potential for heteroscedastic prediction errors across wide COD, BOD, and TSS concentration ranges, highlighting the need for future residual analysis, bias correction, and more rigorous validation to improve model generalizability across different water matrices. Future research should focus on increasing sample size, exploring advanced deep learning approaches, and validating the models under field conditions.

ACKNOWLEDGMENT

The authors are grateful to the Institute for Research and Community Service (LPPM) Universitas Syiah Kuala.

REFERENCES

[1] Sari, F.Y.A., Suryajaya, I.M.A., Christwardana, M., Hadiyanto, H. (2022). Cultivation of microalgae Spirulina platensis in palm oil mill effluent (POME) media with variations of POME concentration and nutrient composition. Journal of Bioresources and Environmental Sciences, 1(2): 57-62. https://doi.org/10.14710/jbes.2022.15052

[2] Khan, I.U., Rahman, M.A., Othman, M.H.D., et al. (2025). Innovative solutions for palm oil mill effluent treatment: A membrane technology perspective. ACS ES&T Water, 5(7): 3538-3562. https://doi.org/10.1021/acsestwater.4c00432

[3] Wu, T.Y., Mohammad, A.W., Jahim, J.M., Anuar, N. (2010). Pollution control technologies for the treatment of palm oil mill effluent (POME) through end-of-pipe processes. Journal of Environmental Management, 91(7): 1467-1490. https://doi.org/10.1016/j.jenvman.2010.02.008

[4] Aguilar-Torrejón, J.A., Balderas-Hernández, P., Roa-Morales, G., Barrera-Díaz, C.E., Rodríguez-Torres, I., Torres-Blancas, T. (2023). Relationship, importance, and development of analytical techniques: COD, BOD, and, TOC in water—An overview through time. SN Applied Sciences, 5: 118. https://doi.org/10.1007/s42452-023-05318-7

[5] Mustaqimah, Devianti, Munawar, A.A., Sufardi, S. (2024). Capability of short Vis-NIR band tandem with machine learning to rapidly predict NPK content in tropical farmland: A case study of Aceh Province agricultural soil dry land, Indonesia. Case Studies in Chemical and Environmental Engineering, 9: 100711. https://doi.org/10.1016/j.cscee.2024.100711

[6] Pasquini, C. (2018). Near infrared spectroscopy: A mature analytical technique with new perspectives—A review. Analytica Chimica Acta, 1026: 8-36. https://doi.org/10.1016/j.aca.2018.04.004

[7] Qi, X., Lian, Y., Xie, L., Wang, Y., Lu, Z. (2023). Water quality detection based on UV-Vis and NIR spectroscopy: A review. Applied Spectroscopy Reviews, 59(8): 1036-

1060. https://doi.org/10.1080/05704928.2023.2294458
[8] Ye, B., Cao, X., Liu, H., et al. (2022). Water chemical oxygen demand prediction model based on the CNN and ultraviolet-visible spectroscopy. Frontiers in Environmental Science, 10: 1027693. https://doi.org/10.3389/fenvs.2022.1027693
[9] Devianti, Sufardi, Bulan, R., Nurlaili, Haris, O. (2023). Application of LW-NIR for rapid prediction of primary nutrients cropland by chemometrics: Comparison among preprocessing and machine learning algorithm approaches. Case Studies in Chemical and Environmental Engineering, 8: 100454. https://doi.org/10.1016/j.cscee.2023.100454
[10] Li, J., Liu, J., Tauqeer, I.M., Shao, Z. (2025). Water chemical oxygen demand prediction based on a one-dimensional multi-scale feature fusion convolutional neural network and ultraviolet-visible spectroscopy. RSC Advances, 15(24): 19431-19442. https://doi.org/10.1039/d5ra00933b
[11] Olawade, D.B., Wada, O.Z., Ige, A.O., Egbewole, B.I., Olojo, A., Oladapo, B.I. (2024). Artificial intelligence in environmental monitoring: Advancements, challenges, and future directions. Hygiene and Environmental Health Advances, 12: 100114. https://doi.org/10.1016/j.heha.2024.100114
[12] Novianta, M.A., Syafrudin, Warsito, B. (2024). Multiple linear regression modeling for analysis of factors affecting COD and BOD on river water quality in Yogyakarta, Indonesia. E3S Web of Conferences, 517: 16001. https://doi.org/10.1051/e3sconf/202451716001
[13] Lokman, A., Ismail, W.Z.W., Aziz, N.A.A. (2025). A review of water quality forecasting and classification using machine learning models and statistical analysis. Water, 17(15): 2243. https://doi.org/10.3390/w17152243
[14] Zhu, J.J., Yang, M.Q., Ren, Z.J. (2023). Machine learning in environmental research: Common pitfalls and best practices. Environmental Science & Technology, 57(46): 17671-17689. https://doi.org/10.1021/acs.est.3c00026
[15] Samawi, M.F., Lanuru, M., Mutmainnah. (2026). Assessment of the impact of palm oil mill effluent (POME) on water quality of the Oko-Oko River, Indonesia. Nature Environment and Pollution Technology, 25(2): 1837. https://doi.org/10.46488/nept.2026.v25i02.d1837
[16] Maguire, T.J., Dominato, K.R., Weidman, R.P., Mundle, S.O.C. (2021). Ultraviolet-visual spectroscopy estimation of nitrate concentrations in surface waters via machine learning. Limnology and Oceanography: Methods, 20(1): 26-33. https://doi.org/10.1002/lom3.10468
[17] Zhu, M.Y., Wang, J.W., Yang, X., et al. (2022). A review of the application of machine learning in water quality evaluation. Eco-Environment & Health, 1(2): 107-116. https://doi.org/10.1016/j.eehl.2022.06.001
[18] Wenger, S.J., Olden, J.D. (2012). Assessing transferability of ecological models: An underappreciated aspect of statistical validation. Methods in Ecology and Evolution, 3(2): 260-267. https://doi.org/10.1111/j.2041-210x.2011.00170.x
[19] Cao, H.L., Xie, X.J., Xiao, Z.Y., Liu, W.J. (2024). Transferability of machine learning models for geogenic contaminated groundwaters. Environmental Science & Technology, 58(20): 8783-8791. https://doi.org/10.1021/acs.est.4c01327
[20] Panchbhaia, K.G., Gautamb, P.B., Lanjewarc, M.G., Patle, L.B. (2025). Near-infrared spectroscopy coupled with machine learning for soil properties prediction. International Journal of Remote Sensing, 46(17): 6635-6667. https://doi.org/10.1080/01431161.2025.2541943
[21] Zhang, W., Kasun, L.C., Wang, Q.J., Zheng, Y., Lin, Z. (2022). A review of machine learning for near-infrared spectroscopy. Sensors, 22(24): 9764. https://doi.org/10.3390/s22249764
[22] Kim, S.Y., Ćurko, J., Gajdoš Kljusurić, J., et al. (2021). Use of near-infrared spectroscopy on predicting wastewater constituents to facilitate the operation of a membrane bioreactor. Chemosphere, 272: 129899. https://doi.org/10.1016/j.chemosphere.2021.129899
[23] Munawar, A.A., Devianti, Mustaqimah, et al. (2025). Rapid assessment of soil nutrient-related environmental risks and safety using near infrared spectroscopy and machine learning. Instrumentation Mesure Métrologie, 24(4): 285-296. https://doi.org/10.18280/i2m.240403
[24] Su, H., Li, L., Hu, X. (2024). Correction methods for visible near-infrared spectroscopy in water pH detection to mitigate temperature influence. Spectroscopy Letters, 57(10): 565-579. https://doi.org/10.1080/00387010.2024.2387162
[25] Tsenkova, R., Munćan, J., Pollner, B., Kovacs, Z. (2018). Essentials of aquaphotomics and its chemometrics approaches. Frontiers in Chemistry, 6: 363. https://doi.org/10.3389/fchem.2018.00363
[26] Agiomavriti, A., Bartzanas, T., Chorianopoulos, N., Gelasakis, A.I. (2025). Spectroscopy-based methods for water quality assessment: A comprehensive review and potential applications in livestock farming. Water, 17(16): 2488. https://doi.org/10.3390/w17162488
[27] Barra, I., Briak, H., Kebede, F. (2022). The application of statistical preprocessing on spectral data does not always guarantee the improvement of the predictive quality of multivariate models: Case of soil spectroscopy applied to Moroccan soils. Vibrational Spectroscopy, 121: 103409. https://doi.org/10.1016/j.vibspec.2022.103409
[28] Ibrahim, A., Alkarbi, N., Alsaadi, M., Almazrouei, A., Alshamsi, S., Hamouda, M.A. (2026). Satellite-based machine learning models for chlorophyll-a and TSS retrieval in Abu Dhabi's coastal waters. Frontiers in Marine Science, 13: 1787597. https://doi.org/10.3389/fmars.2026.1787597
[29] Ho, S.Y., Phua, K., Wong, L., Goh, W.W.B. (2020). Extensions of the external validation for checking learned model interpretability and generalizability. Patterns, 1(8): 100129. https://doi.org/10.1016/j.patter.2020.100129
[30] Hu, Y., Ghosh, C., Malakpour-Estalaki, S. (2023). A methodological framework for improving the performance of data-driven models: A case study for daily runoff prediction in the Maumee domain, USA. Geoscientific Model Development, 16: 1925-1936. https://doi.org/10.5194/gmd-16-1925-2023
[31] Desmiarti, R., Rosadi, M.Y.., Hazmi, A., Rahman, M.M., Naldi, N., Fajri, J.A. (2022). Biogas production from palm oil mill effluent using dielectric barrier discharge integrated with the aerated condition. Water, 14(22): 3774. https://doi.org/10.3390/w14223774
[32] Laubscher, A., Rose, L.J., Williams, P.J. (2023).

Determination of potential absorbance bands of fumonisin B1 in methanol with near infrared spectroscopy. Journal of Near Infrared Spectroscopy, 31(4): 211-223. https://doi.org/10.1177/09670335231183098

[33] Grabska, J., Ishigaki, M., Beć, K.B., Wójcik, M.J., Ozaki, Y. (2017). Correlations between structure and near-infrared spectra of saturated and unsaturated carboxylic acids. Insight from anharmonic density functional theory calculations. The Journal of Physical Chemistry A, 121(18): 3437-3451. https://doi.org/10.1021/acs.jpca.7b02053

[34] Rinnan, Å., Berg, F.V.D., Engelsen, S.B. (2009). Review of the most common pre-processing techniques for near-infrared spectra. Trends in Analytical Chemistry, 28(10): 1201-1222. https://doi.org/10.1016/j.trac.2009.07.007

[35] Chen, T., Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In KDD'16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA, pp. 785-794. https://doi.org/10.1145/2939672.2939785

[36] Vestergaard, R.J., Vasava, H.B., Aspinall, D., et al. (2021). Evaluation of optimized preprocessing and modeling algorithms for prediction of soil properties using VIS-NIR spectroscopy. Sensors, 21(20): 6745. https://doi.org/10.3390/s21206745

[37] Roger, J.M., Mallet, A., Marini, F. (2022). Preprocessing NIR spectra for aquaphotomics. Molecules, 27(20): 6795. https://doi.org/10.3390/molecules27206795

[38] Tong, P., Du, Y., Zheng, K., Wu, T., Wang, J. (2015). Improvement of NIR model by fractional order Savitzky-Golay derivation (FOSGD) coupled with wavelength selection. Chemometrics and Intelligent Laboratory Systems, 143: 40-48. https://doi.org/10.1016/j.chemolab.2015.02.017

[39] Cen, H.Y., He, Y. (2007). Theory and application of near infrared reflectance spectroscopy in determination of food quality. Trends in Food Science & Technology, 18(2): 72-83. https://doi.org/10.1016/j.tifs.2006.09.003

[40] Collins, G.S., Reitsma, J.B., Altman, D.G., Moons, K.G. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. Annals of Internal Medicine, 162(1): 55-63. https://doi.org/10.7326/m14-0697

[41] Ching, P., Zou, X., Wu, D., So, R., Chen, G. (2022). Development of a wide-range soft sensor for predicting wastewater BOD_5 using an eXtreme gradient boosting (XGBoost) machine. Environmental Research, 210: 112953. https://doi.org/10.1016/j.envres.2022.112953

[42] Chen, H.Z., Xu, L.L., Ai, W., Lin, B., Feng, Q.X., Cai, K. (2020). Kernel functions embedded in support vector machine learning models for rapid water pollution assessment via near-infrared spectroscopy. Science of The Total Environment, 714: 136765. https://doi.org/10.1016/j.scitotenv.2020.136765

[43] Savira, A., Andrianingsih, A. (2026). Machine learning analysis of Jakarta Bay water quality: Comparing models. Sinkron, 10(1): 110-120. https://doi.org/10.33395/sinkron.v10i1.15540

[44] Alnuwaiser, M.A., Javed, M.F., Khan, M.I., Ahmed, M.W., Galal, A.M. (2022). Support vector regression and ANN approach for predicting the ground water quality. Journal of the Indian Chemical Society, 99(7): 100538. https://doi.org/10.1016/j.jics.2022.100538

[45] Lotfi, K., Bonakdari, H., Ebtehaj, I., et al. (2019). Predicting wastewater treatment plant quality parameters using a novel hybrid linear-nonlinear methodology. Journal of Environmental Management, 240: 463-474. https://doi.org/10.1016/j.jenvman.2019.03.137