XGBoost-Based Predictive Modeling Using Sentinel-2 Satellite Imagery and Empirical Data of Jakarta River Water Quality

Rijaldi R.M.[1], Prayoga G.[1*], Firmansyah F.S.[1], Faskayana Y.S.[2], and Liyantono[1]

[1]Center for Environmental Research, IPB University, Indonesia

[2]The Graduate School of Agriculture, Tokyo University of Agriculture and Technology, Japan

*gatotprayoga16@gmail.com

Abstract: *Machine learning methods have shown strong potential in estimating water quality parameters. Satellite remote sensing provides an efficient approach for monitoring water quality over extended periods and spatial variations, thereby reducing reliance on resource-intensive field monitoring. Therefore, this study aims to develop predictive modeling using XGBoost regression and Sentinel-2 satellite imagery to estimate river water quality, specifically Total Suspended Solids (TSS) and Total Dissolved Solids (TDS). The dataset consisted of 152 samples obtained from 60 river sampling sites across 12 river sections, collected during different periods between 2021 and 2024. Each in situ measurement was paired with harmonized reflectance values from Sentinel-2 bands to construct the training data for the XGBoost models. Using Google Earth Engine (GEE), reflectance values from Sentinel-2 bands were extracted for each sampling point and acquisition date. The derived spectral features were then used to train XGBoost regression models for each parameter. The predictive performance of the models was assessed using R^2, RMSE, and MAE. The XGBoost regression model achieved $R^2 = 0.761$, RMSE = 18.08 mg/L, and MAE = 15.21 mg/L for TSS, while for TDS it obtained $R^2 = 0.575$, RMSE = 85.71 mg/L, and MAE = 69.42 mg/L. These results indicate that the integration of Sentinel-2 imagery with XGBoost regression provides a robust predictive capability for TSS, whereas the estimation of TDS remains less accurate but still offers informative insights for water quality assessment. This study demonstrates the possibility of estimating river water quality using remote sensing. This approach could enhance the effectiveness and efficiency of water quality management. However, several significant limitations, such as the limited availability of Sentinel-2 imagery that coincided precisely with the in-situ sampling dates, may have introduced some temporal discrepancies in the dataset.*

Keywords: water quality, remote sensing, machine learning, XGBoost, Jakarta

Introduction

Water quality is a fundamental determinant of ecosystem health, human well-being, and socio-economic development. Clean and sufficient water resources are vital for drinking, agriculture, fisheries, industry, and recreation. Conversely, degraded water quality threatens biodiversity, poses public health risks, and reduces economic productivity (UNESCO, 2022). The river water quality in the Province of DKI Jakarta is currently facing serious challenges due to high levels of pollution originating from domestic activities, industrial discharges, and urban surface runoff. Major rivers such as Ciliwung, Pesanggrahan, Angke, and Sunter have been reported to be in a heavily polluted condition according to the DKI Jakarta Environmental Agency (2024).

This water quality issue is closely related to the high levels of Total Suspended Solids (TSS) resulting from sedimentation and anthropogenic activities that increase water turbidity. Elevated TSS concentrations reduce light penetration, lower primary productivity, and accelerate river siltation. Meanwhile, high levels of Total Dissolved Solids (TDS), largely originating from domestic and industrial effluents, can increase water salinity and electrical conductivity, ultimately degrade chemical quality and reduce the suitability of rivers as raw water sources (Billota & Brazier 2008). The combination of elevated TSS and TDS thus serves as a critical indicator of river water quality degradation in Jakarta.

Accurate monitoring of water quality is crucial to support such planning efforts and aligns with the Sustainable Development Goals (SDGs), particularly Goal 6 on clean water and sanitation (United Nations, 2023). However, conventional water quality monitoring methods still rely on in situ sampling and laboratory analysis, which provide accurate data but are costly, labor-intensive, and limited in both spatial and temporal coverage (Gholizadeh et al. 2016). These constraints are particularly challenging in metropolitan contexts such as Jakarta, where rivers face multiple and dynamic anthropogenic pressures.

Remote sensing offers a complementary approach, enabling repeated, large-scale observations that capture spatial variations and temporal dynamics often missed by traditional programs. Sentinel-2, with its high spatial resolution (10–60 m), five-day revisit cycle, and multispectral coverage, has been widely applied to estimate optically active parameters such as turbidity and suspended solids (Drusch et al., 2012; Vanhellemont & Ruddick, 2018).

Recent advances in machine learning further improve water quality estimation. Ensemble-based methods such as Extreme Gradient Boosting (XGBoost) can capture complex nonlinear relationships between spectral reflectance and water quality indicators while minimizing overfitting (Chen & Guestrin, 2016). Integrating Sentinel-2 imagery with XGBoost enables predictive modeling with reduced dependence on resource-intensive field monitoring.

Despite progress, applications in tropical and highly urbanized rivers remain limited. Jakarta's rivers are heavily polluted by domestic and commercial wastewater, industrial discharges, agricultural runoff, solid waste, and leakage from septic systems (Apip et al., 2015). These are poorly represented in satellite-based water quality studies. To address this, the present study aims to develop predictive modeling using XGBoost regression and

Sentinel-2 satellite imagery to estimate river water quality, specifically Total Suspended Solids (TSS) and Total Dissolved Solids (TDS).

Literature Review

a. Total Suspended Solids (TSS) and Total Dissolved Solids (TDS)

Optically active parameters in water, such as total suspended solids (TSS) interact with light to alter the energy spectrum of reflected solar radiation (Ngamile et al., 2025). Total Suspended Solids (TSS) are an important parameter in assessing river water quality because they reflect the amount of suspended solids, which include mud, sediment, and microscopic organisms. High TSS concentrations can increase air turbidity, reduce light penetration, and negatively impact primary productivity and aquatic life (Billota & Brazier 2008).

In addition to TSS, monitoring and evaluating parameters such as TDS is essential to gain a comprehensive understanding of how natural environmental changes and human activities impact water bodies (Adjovu et al. 2023). Although TDS is not generally classified as a primary water pollutant, it serves as an important indicator of water quality. TDS consists of a mixture of salts, metals, metalloids, and dissolved organic matter. Organic fractions of TDS are typically released through the growth and decomposition of biological materials, such as plant roots and microbial activity in aquatic systems. TDS concentration shows a close relationship with salinity, as elevated salinity typically corresponds to higher electrical conductivity driven by dissolved chemical constituents. Natural processes such as rock weathering and runoff, together with anthropogenic pressures from household and industrial discharges, significantly affect TDS levels in freshwater systems (Effendi, 2003). Elevated TDS levels are often associated with increased electrical conductivity and may also correspond to reduced dissolved oxygen concentrations (Butler & Ford 2018).

Remote sensing techniques make it possible to have spatial and temporal view of surface water quality parameters and more effectively and efficiently monitor the waterbodies and quantify water quality issues. Most studies have primarily focused on optically active variables, such as total suspended solids (TSS). However, research on total dissolved solids (TDS) remains limited because TDS itself is not optically active. Nevertheless, in many rivers, dissolved solids concentrations (TDS) may exhibit empirical correlations with other factors, particularly suspended sediments (TSS) (Butler & Ford 2018).

b. Water Quality Monitoring and Its Challenges

Water quality monitoring plays a vital role in managing freshwater resources, protecting aquatic ecosystems, and ensuring compliance with environmental regulations. Conventional monitoring relies on in situ sampling and laboratory analyses, which provide reliable and standardized measurements. However, these methods are costly, time-consuming, and often limited in spatial and temporal resolution (Gholizadeh et al. 2016).

Such limitations become more pronounced in rapidly growing metropolitan areas like Jakarta, where rivers are exposed to diverse and fluctuating pollution sources. Frequent field campaigns across multiple sites are difficult to maintain, leading to spatial gaps and temporal discontinuities in data records. Furthermore, environmental heterogeneity in urban river systems reduces the representativeness of point-based measurements. These challenges underscore the need for complementary approaches that can capture water quality variability more efficiently and cost-effectively.

c. Remote Sensing for Water Quality Assessment

Remote sensing has emerged as a powerful tool for large-scale and repeated water quality assessment. Satellite platforms provide synoptic observations that enable spatially continuous monitoring, complementing traditional sampling efforts. The Sentinel-2 mission, operated by the European Space Agency, offers 13 spectral bands in the visible, near-infrared (NIR), and shortwave infrared ranges, with high spatial resolution (10–60 m) and a five-day revisit time (Drusch et al., 2012). These characteristics make Sentinel-2 particularly suitable for inland and coastal water applications.

Numerous studies have demonstrated the potential of Sentinel-2 for monitoring optically active parameters such as turbidity, chlorophyll-a, suspended sediments, and water color (Dekker et al., 2002; Binding et al., 2010; Zheng et al., 2015; Vanhellemont & Ruddick, 2018). For instance, Vanhellemont and Ruddick (2018) applied Sentinel-2 data to estuarine waters and showed its capability in capturing turbidity variations, while Virdis et al. (2022) validated its applicability for riverine water quality assessment in large systems.

Despite these advantages, remote sensing of inland waters remains challenging in tropical regions. High turbidity, mixed pollution sources, and frequent cloud cover complicate atmospheric correction and reduce image availability (Gholizadeh et al., 2016). Moreover, some water quality indicators, such as nutrients and dissolved solids, are not optically active, limiting their detectability via multispectral sensors. These constraints necessitate methodological innovations, including the integration of advanced algorithms and multi-sensor datasets.

d. Machine Learning Approaches

Machine learning has gained increasing attention in environmental modeling due to its ability to capture complex, nonlinear, and high-dimensional relationships. Algorithms such as support vector regression, random forests, and gradient boosting have been successfully applied to predict water quality parameters from remotely sensed data (Bui et al., 2020).

Among these methods, Extreme Gradient Boosting (XGBoost) has shown particular promise. Developed as an optimized implementation of gradient boosting, XGBoost offers scalability, efficiency, and robust handling of overfitting (Chen & Guestrin, 2016). Its ensemble-based architecture allows the integration of multiple weak learners to achieve high predictive accuracy. Several studies have demonstrated its applicability in water quality prediction, such as Ananta et al. (2024), who successfully applied XGBoost to estimate of water feasibility based on household water quality parameters.

The combination of satellite observations and machine learning is particularly valuable for monitoring optically active parameters, where spectral signatures are highly complex. By leveraging non-linear regression capabilities, machine learning can outperform traditional empirical models and enhance prediction reliability across diverse water conditions.

e. Research Gaps and Context for Jakarta

Although remote sensing and machine learning approaches have been widely applied in temperate and subtropical regions, their implementation in tropical and highly urbanized rivers remains limited. Studies in Southeast Asia are still scarce, despite the region's vulnerability to water pollution driven by rapid urbanization and industrialization.

Jakarta's rivers represent a challenging case study. They are characterized by relatively are heavy polluted, strong seasonal variability, and multiple pollution sources, including domestic wastewater, industrial effluents, solid waste and runoff from agricultural land (Apip et al., 2015). These conditions complicate the detection of water quality signals from satellite imagery, especially under frequent cloud cover. Furthermore, the scarcity of consistent and spatially distributed in situ datasets hinders model calibration and validation.

Addressing these challenges requires integrated approaches that harmonize field data and satellite observations, while also leveraging advanced algorithms such as XGBoost. This study responds to these gaps by developing and validating predictive models tailored to Jakarta's urban rivers, thereby extending the application of satellite-based water quality monitoring to tropical megacities.

Methodology

a. Study Area

The study was conducted in Jakarta, Indonesia, focusing on 12 major river sections that traverse the metropolitan area. These rivers are subject to multiple anthropogenic pressures, including domestic wastewater discharges, industrial effluents, and urban runoff, reflecting the city's heterogeneous pollution sources. To provide a clearer geographical context, the study area and spatial distribution of sampling sites is illustrated in Figure 1 below.

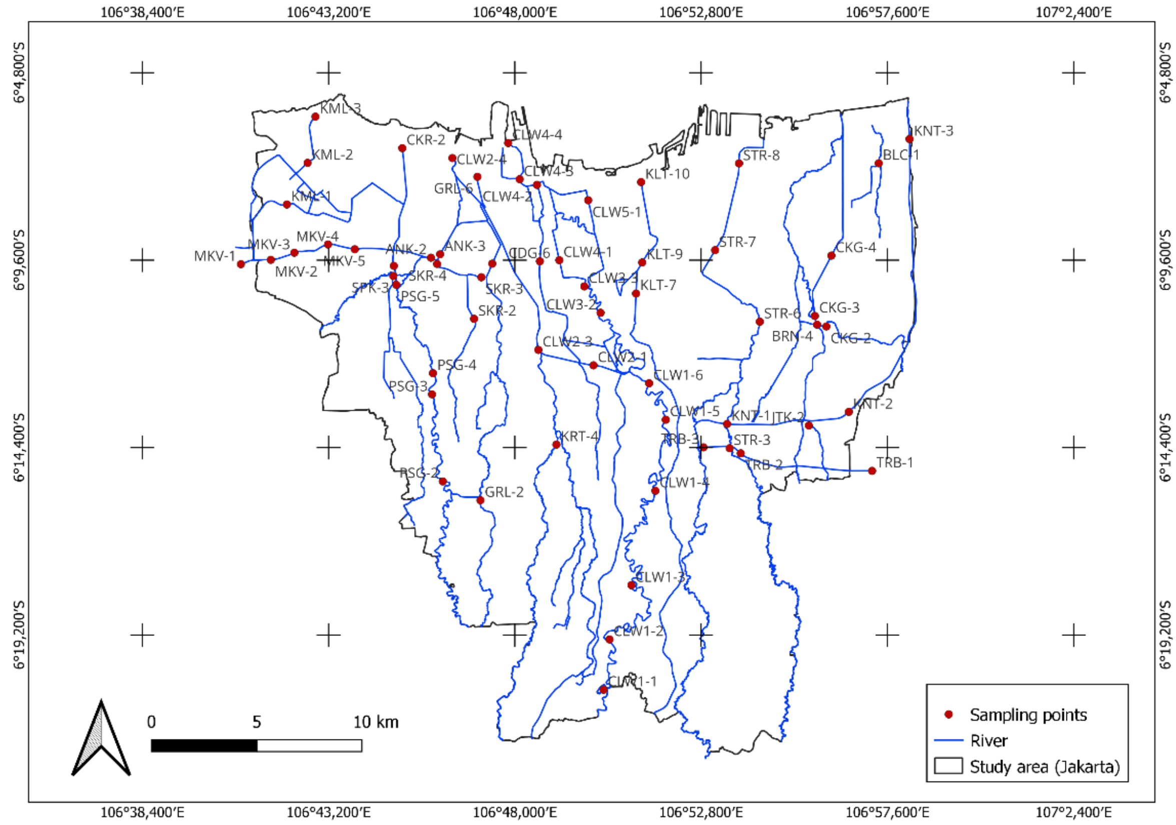


Figure 1: Map of the study area.

b. Data Collection

This study employed a combination of satellite-based remote sensing data and field-based empirical measurements to develop predictive models for river water quality, specifically Total Suspended Solids (TSS) and Total Dissolved Solids (TDS). Two primary datasets were utilized: (i) Sentinel-2 Level-2A multispectral imagery and (ii) in-situ measurements of Total Suspended Solids (TSS, mg/L) and Total Dissolved Solids (TDS, mg/L).

Sentinel-2 Level-2A surface reflectance imagery was accessed and processed via Google Earth Engine (GEE) platform. Sentinel-2 provides multispectral observations with 13 bands in the visible, near-infrared (NIR), and shortwave infrared (SWIR) regions, with spatial resolutions ranging from 10 to 60 meters and a revisit time of 5 days at the equator. The Level-2A product, which includes atmospheric correction using the Sen2Cor algorithm was chosen to ensure consistency and reliability of reflectance values. To minimize cloud contamination, a cloud mask based on the Scene Classification Layer (SCL) was applied.

The dataset consisted of 152 samples obtained from 60 river sampling sites across 12 river sections (Figure 1), collected during different periods between 2021 and 2024. At each sampling event, water samples were analyzed in the laboratory to determine concentrations of Total Suspended Solids (TSS, mg/L) and Total Dissolved Solids (TDS, mg/L) following standard protocols. These empirical measurements paired with Sentinel-2 reflectance values acquired on the same date as the field sampling, using the spatial coordinates of each site. This approach ensured precise temporal alignment between satellite-derived spectral information and ground-based measurements of TSS and TDS. The integration of these two datasets allowed the final data matrix to capture both the spectral variability observed by remote sensing and the empirical variability in water quality, thereby providing a robust foundation for training and evaluating the machine learning models.

c. Model Development

This study employed XGBoost (eXtreme Gradient Boosting) regression models to predict water quality parameters (TSS and TDS) from Sentinel-2 multispectral imagery. The modeling framework consisted of systematic data preprocessing, feature engineering, model training, and rigorous validation procedures to ensure robust predictive performance. The input features comprised 10 Sentinel-2 spectral bands (B2-Blue, B3-Green, B4-Red, B5-RedEdge1, B6-RedEdge2, B7-RedEdge3, B8-NIR, B8A-RedEdge4, B11-SWIR1, B12-SWIR2) and six derived spectral indices (NDVI, NDWI, NDTI, Blue/Red ratio, Red/Blue ratio, Red/Green ratio), totaling 16 potential predictor variables. Data quality was ensured through outlier detection using the Interquartile Range (IQR) method with a threshold multiplier of 2.0, which identified and removed approximately 5-7% of extreme values that could bias model training.

Feature selection was conducted using Pearson correlation analysis to identify the spectral variables most relevant to each target parameter. Variables with higher absolute correlation coefficients were retained for subsequent modeling. This dimensionality reduction approach mitigated multicollinearity issues and reduced computational complexity while maintaining predictive accuracy. Prior to model training, feature scaling was applied using RobustScaler transformation, which standardizes features by removing the median and scaling according to the IQR. This approach was preferred over standard normalization as it demonstrates greater robustness to outliers and ensures that all features contribute equally to the model without being dominated by variables with larger numerical ranges.

The dataset was partitioned into training (80%) and independent test (20%) sets using stratified random sampling with a fixed random seed (42) to ensure reproducibility. The XGBoost regressor was configured with hyperparameters optimized for small to medium-sized datasets: 50 estimators with a learning rate of 0.1, maximum tree depth of 3, minimum child weight of 5, subsample ratio of 0.8, column subsample ratio of 0.8, gamma value of 0.5, L1 regularization (alpha) of 1.0, and L2 regularization (lambda) of 2.0. These conservative hyperparameter settings were specifically designed to prevent overfitting, which is a common challenge when working with limited training samples.

Model performance was assessed using multiple complementary metrics. The coefficient of determination (R^2) quantified the proportion of variance explained by the model, while Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) measured prediction accuracy in the original units (mg/L). Equation 1, Equation 2, and Equation 3 are used to calculate performance metrics. To evaluate model generalizability and guard against overfitting, five-fold cross-validation was implemented on the entire cleaned dataset, providing robust estimates of model performance across different data subsets.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \qquad (1)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} \qquad (2)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i| \qquad (3)$$

Information:
n = number of observations/rows
Yi = actual value
Ŷi = predicted value

Overfitting was assessed by comparing training and test set R² values, with differences below 0.15 considered excellent, 0.15-0.25 considered good, and above 0.25 indicating potential overfitting concerns. Cross-validation stability was evaluated by calculating the absolute difference between test R² and mean cross-validation R², with values below 0.15 indicating stable model performance. Models achieving test R² values above 0.5 with stable cross-validation performance were deemed acceptable for publication, while R² values above 0.7 indicated strong predictive capability.

Feature importance was quantified using the gain metric from XGBoost, which measures the relative contribution of each feature to model predictions based on the improvement in accuracy it brings to the branches it is on. This analysis identified the most influential spectral variables for predicting each water quality parameter, providing insights into the physical relationships between remote sensing observations and in-situ water quality measurements.

All analyses were conducted in Python using scikit-learn for preprocessing and model evaluation, XGBoost for gradient boosting implementation, and standard scientific libraries including NumPy, pandas, and matplotlib for data manipulation and visualization. Model artifacts, including trained estimators, feature scalers, and selected feature lists, were serialized using Python's pickle protocol for deployment and future prediction applications. All the processes described including data collection and model development are summarized in Figure 2.

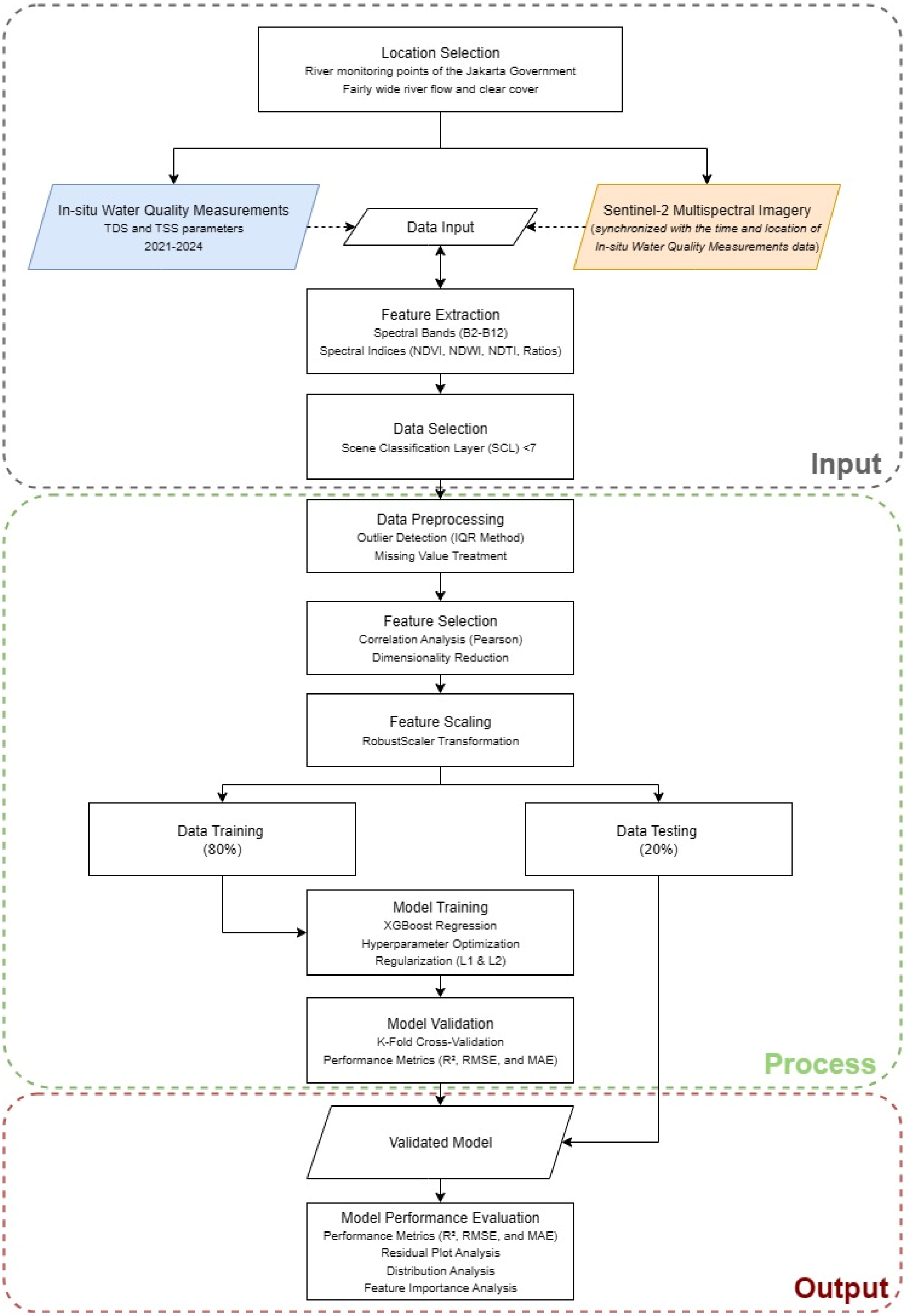

Location Selection
River monitoring points of the Jakarta Government
Fairly wide river flow and clear cover
In-situ Water Quality Measurements
TDS and TSS parameters
2021-2024
Data Input
Sentinel-2 Multispectral Imagery
(synchronized with the time and location of In-situ Water Quality Measurements data)
Feature Extraction
Spectral Bands (B2-B12)
Spectral Indices (NDVI, NDWI, NDTI, Ratios)
Data Selection
Scene Classification Layer (SCL) <7
Input
Data Preprocessing
Outlier Detection (IQR Method)
Missing Value Treatment
Feature Selection
Correlation Analysis (Pearson)
Dimensionality Reduction
Feature Scaling
RobustScaler Transformation
Data Training
(80%)
Data Testing
(20%)
Model Training
XGBoost Regression
Hyperparameter Optimization
Regularization (L1 & L2)
Model Validation
K-Fold Cross-Validation
Performance Metrics (R², RMSE, and MAE)
Process
Validated Model
Model Performance Evaluation
Performance Metrics (R², RMSE, and MAE)
Residual Plot Analysis
Distribution Analysis
Feature Importance Analysis
Output

Figure 2: Framework of the study.

Results and Discussion

a. Model Performance Evaluation for TSS Estimation

The initial TSS dataset comprised 151 water samples with corresponding Sentinel-2 spectral measurements, which were first subjected to outlier detection to ensure robustness in the modeling process. Outlier detection using the Interquartile Range (IQR) method with a threshold multiplier of 2.0 identified 11 extreme observations (7.3% of the dataset) with TSS concentrations ranging from 217.0 to 1505.0 mg/L. This preprocessing step resulted in 140 cleaned samples, improving the representativeness of the dataset by eliminating extreme observations that could bias the regression model. Following this, feature selection based on Pearson correlation coefficients identified 12 spectral variables exhibiting the strongest associations with TSS concentrations. The Red/Green ratio demonstrated the highest correlation (r = +0.524), followed by Red/Blue ratio (r = +0.497) and Normalized Difference Turbidity Index (NDTI, r = +0.496), indicating that visible wavelength ratios and turbidity-specific indices were the primary predictors. The refined dataset was then partitioned into training and testing subsets, with 112 samples (80%) allocated for training and 28 samples (20%) allocated for testing, with RobustScaler normalization applied to mitigate the influence of scale differences among spectral features. This stratification ensured that the model had sufficient variability for learning while maintaining statistical independence for assessing predictive generalization.

The training of the XGBoost regression model achieved strong predictive capacity for TSS concentrations, highlighting the effectiveness of ensemble learning in capturing nonlinear relationships between spectral features and suspended sediment content. On the training subset, the model achieved a root mean square error (RMSE) of 14.68 mg/L and an R^2 of 0.929, indicating that over 92% of the variance in measured TSS values was explained by the model. These values suggest that the spectral features selected were highly informative, allowing the algorithm to exploit both additive and interactive effects across wavelengths. Such high training accuracy reflects the well-documented ability of boosting algorithms to minimize residual errors by iteratively correcting weak learners and combining them into a strong predictive model. Importantly, the relatively low RMSE values indicate that the deviations between predicted and observed TSS concentrations were minor, supporting the potential of optical reflectance for suspended sediment quantification.

Independent testing results further validated the model's predictive performance under unseen conditions. The scatter plot of measured versus predicted TSS values revealed a moderate-to-strong linear relationship (Figure 3). On the 28 test samples, the model achieved an RMSE of 18.08 mg/L, an MAE of 15.21 mg/L, and an R^2 of 0.761, indicating that approximately 76% of the variability in TSS concentrations was successfully explained. The relatively modest increase in RMSE from training to testing (3.40 mg/L) and the small overfitting score of 0.169 demonstrate that the model maintained generalization capacity and did not excessively memorize training data. Cross-validation using a 5-fold strategy confirmed these findings, with a mean R^2 of 0.608 ± 0.159 and RMSE of 32.30 ± 10.39 mg/L, reflecting variability across partitions but consistent evidence of predictive strength.

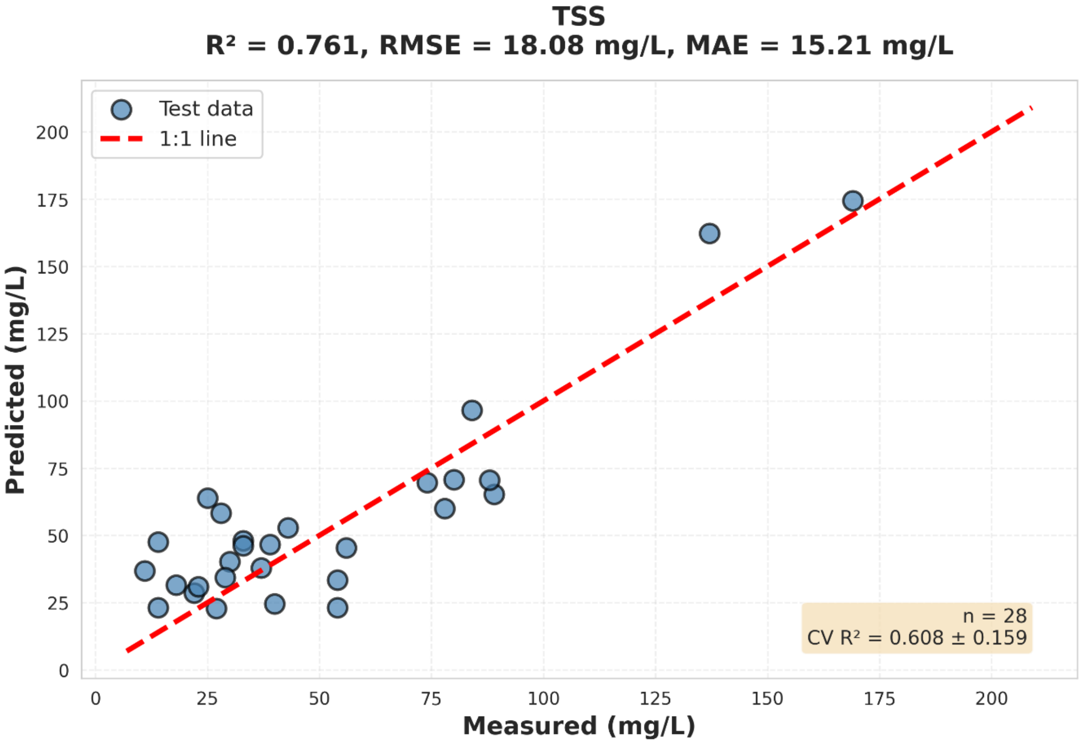
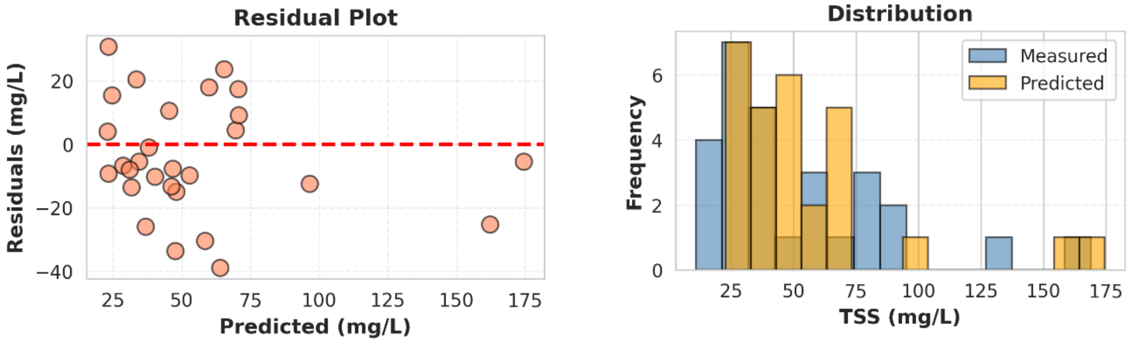


Figure 3: Scatter plot of measured versus predicted TSS values.

Comparison of measured and predicted TSS distributions revealed general alignment in central tendencies, with both distributions exhibiting right-skewed characteristics typical of water quality parameters. The predicted distribution demonstrated slightly reduced variance compared to measured values, indicating a tendency toward regression to the mean that a common behavior in predictive models.

Residual analysis provided critical insights into model assumptions and systematic biases. The residual plot displayed relatively random scatter around zero for predicted values below 60 mg/L, consistent with homoscedastic error variance and adequate model specification in this concentration range (Figure 4). However, a subtle funnel pattern emerged at higher predicted concentrations, with residual variance increasing proportionally with predicted TSS values. This heteroscedasticity suggests that prediction uncertainty scales with TSS magnitude, potentially reflecting limitations in the optical signal at elevated turbidity levels where multiple scattering effects and signal saturation may compromise spectral sensitivity. The absence of pronounced systematic trends or curvature in residuals indicates that the model's linear additive structure adequately captured the spectral-TSS relationships, though the increased scatter at higher concentrations warrants cautious interpretation of predictions exceeding 75 mg/L.



(a) (b)

Figure 4: (a) Residual plot of predicted TSS values and (b) distribution chart of measured and predicted TSS values

Feature importance rankings, derived from XGBoost's gain metric, revealed that spectral band ratios dominated the predictive framework, with NDTI exhibiting the highest importance (0.518), followed by the Red/Green ratio (0.219) and the Red/Blue ratio (0.067) (Figure 5). The dominance of these indices indicates their effectiveness in enhancing the spectral response of turbidity-related signals by minimizing atmospheric and illumination effects. In contrast, individual spectral bands such as blue (B2 = 0.036), green (B3 = 0.019), and red (B4 = 0.015) exhibited relatively low contributions, underscoring the limited discriminatory power of raw reflectance values for TSS estimation. NDWI (0.032) ranked fifth, capturing complementary information from green–NIR contrasts, while near-infrared (B8_NIR = 0.022) and shortwave infrared bands (B11_SWIR1 = 0.014; B12_SWIR2 = 0.023) showed minimal influence. This hierarchy suggests that ratio-based indices are more robust predictors of TSS than raw spectral bands, highlighting the importance of domain-informed feature engineering in water quality modeling.

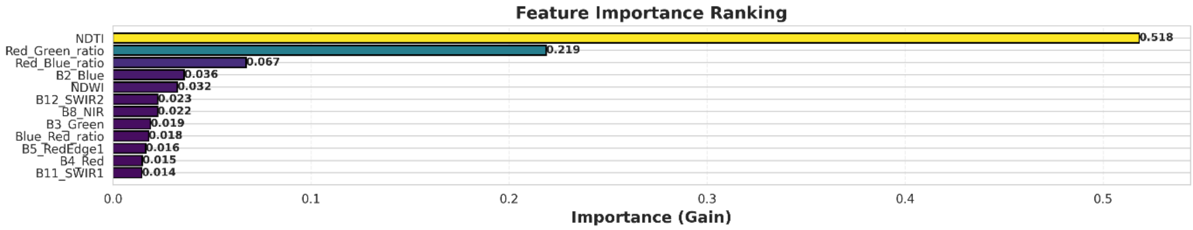


Figure 5: Feature importance ranking derived from XGBoost's gain metric on TSS prediction model.

b. Model Performance Evaluation for TDS Estimation

The TDS prediction model utilized the same initial dataset of 151 samples, with outlier removal identifying five extreme observations (3.3%) exhibiting TDS concentrations ranging from 964.0 to 6300.0 mg/L. This lower outlier proportion compared to TSS (7.3%) reflects the inherently lower variability in dissolved constituents relative to suspended particulates, though the presence of extreme values approaching 6300 mg/L suggests occasional contamination or saline intrusion events. Following outlier exclusion, 146 samples remained for model development. Feature selection yielded 14 spectral variables, two more than the TSS model, indicating that TDS prediction required a broader spectral signature to achieve comparable performance. Red/Blue ratio exhibited the strongest correlation (r = +0.492), followed by Blue/Red ratio (r = +0.368) and shortwave infrared bands B12_SWIR2 (r = +0.291) and B11_SWIR1 (r = +0.270). The lower correlation magnitudes compared to TSS features (maximum r = 0.524) reflect the fundamentally weaker optical activity of dissolved substances, which primarily influence water-leaving radiance through subtle absorption processes rather than direct backscattering. Data partitioning yielded 116 training samples and 30 testing samples, with standardization applied to ensure feature compatibility.

The TDS model exhibited moderate predictive capability, with training $R^2 = 0.812$ and RMSE = 62.97 mg/L, indicating reasonable capacity to learn spectral-TDS associations. Test performance, however, revealed substantial degradation, with R^2 declining to 0.575 and RMSE increasing to 85.71 mg/L. The MAE of 69.42 mg/L and MAPE of 65.09% highlight considerable prediction uncertainty, with typical errors exceeding 60% of observed values for low-TDS samples. Five-fold cross-validation results were particularly revealing, yielding mean $R^2 = 0.356 \pm 0.090$ and mean RMSE = 112.26 ± 16.51 mg/L. The substantial discrepancy between test R^2 (0.575) and cross-validation R^2 (0.356) indicates that the specific train-test partition fortuitously yielded optimistic performance estimates, while cross-validation's more rigorous assessment exposed limited generalization capability. The overfitting gap of 0.238 between training and testing R^2 values, while within the acceptable threshold of 0.25, suggests borderline overfitting tendencies. These performance characteristics underscore the inherent challenge of TDS remote sensing, where dissolved constituents exert minimal direct influence on surface reflectance, necessitating reliance on indirect correlations with optically active co-constituents.

The measured-predicted scatter plot for TDS exhibited greater dispersion than the corresponding TSS plot, with $R^2 = 0.575$ indicating that only 58% of observed TDS variance was explained by spectral features (Figure 6). Predictions displayed moderate adherence to the 1:1 line for TDS concentrations below 400 mg/L, though substantial scatters were evident throughout the concentration range. The RMSE of 85.71 mg/L represents approximately 14% of the mean TDS concentration in the cleaned dataset (estimated mean ≈ 600 mg/L), which, while seemingly reasonable on a percentage basis, translates to considerable absolute uncertainty for low-concentration samples. Several predictions deviated markedly from measured values, with residuals exceeding ±150 mg/L in extreme cases. The cross-validation R^2 of 0.356 ± 0.090 provides a more conservative and realistic estimate of operational performance, suggesting that the model achieves weak-to-moderate predictive power under diverse data conditions. This substantial gap between test and cross-validation performance highlights the model's sensitivity to specific spatial or temporal subsets within the dataset, potentially reflecting variations in the relative composition of dissolved constituents (e.g., organic versus inorganic fractions) across sampling locations or dates.

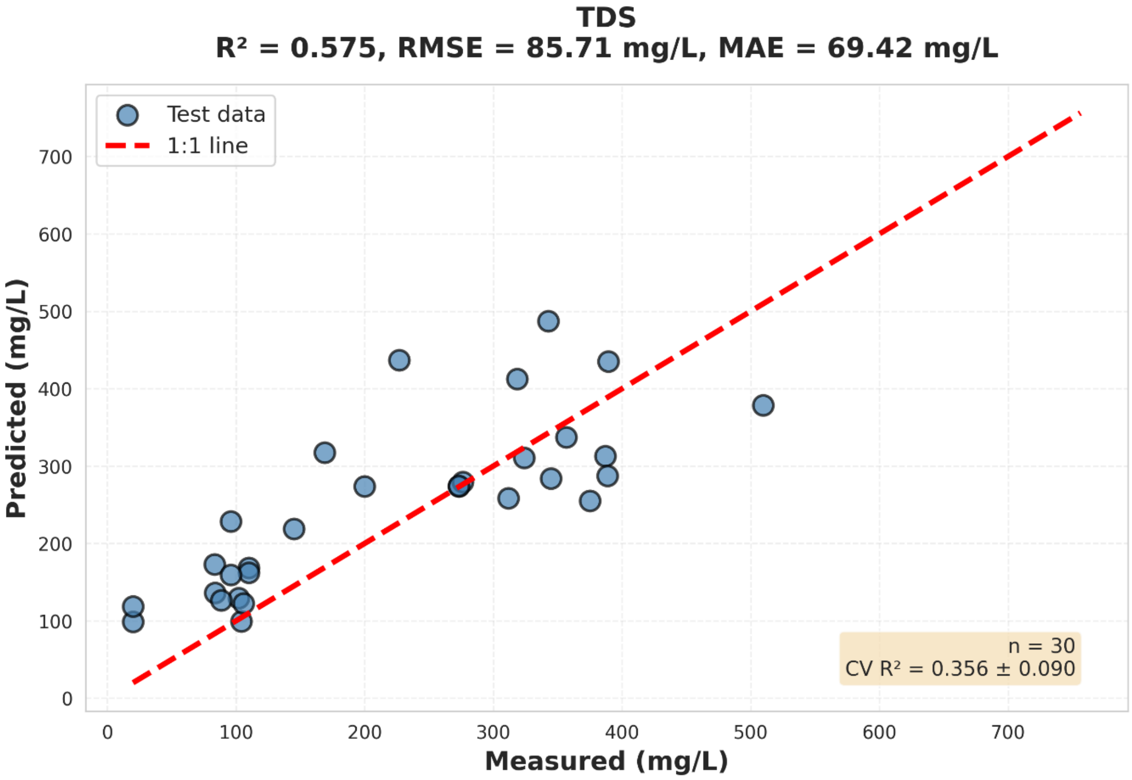
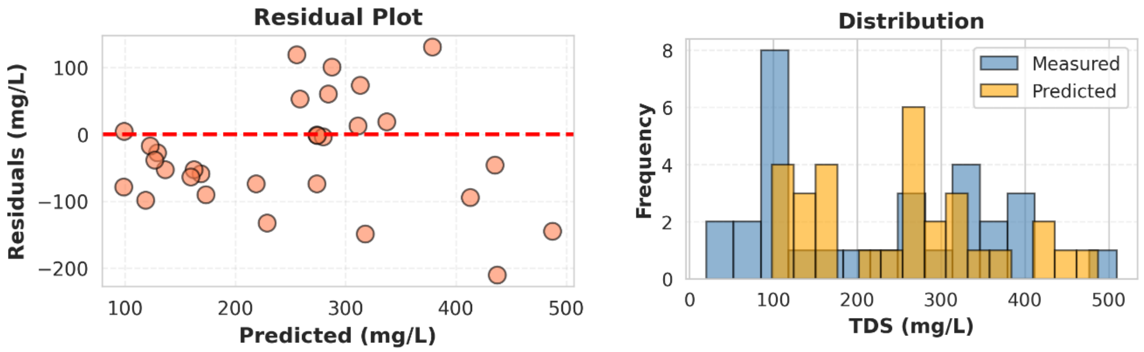


Figure 6: Scatter plot of measured versus predicted TDS values.

Measured TDS values exhibited a unimodal distribution with moderate positive skew, while predicted values showed a similar but slightly compressed distribution, indicating reduced dynamic range in model outputs. The predicted distribution's lower variance reflects the model's tendency to regress extreme values toward the mean, a characteristic amplified by weak predictor-response correlations. Residual analysis revealed random scatter around zero for the majority of predictions, suggesting absence of gross systematic bias in model specification. However, residuals exhibited mild heteroscedasticity, with variance increasing at higher predicted TDS concentrations, similar to the TSS model. Several outlying residuals exceeding ±150 mg/L were identified (Figure 7), corresponding to samples where predicted TDS substantially under- or over-estimated measured values. These extreme residuals may reflect localized compositional anomalies, such as industrial discharge plumes or saline intrusion events, where dissolved constituent mixtures deviate from the baseline spectral-TDS relationships established during training. The absence of pronounced curvature or systematic trends in the residual plot indicates that the linear additive model structure was appropriate, though the substantial residual variance underscores the limited explanatory power of optical remote sensing for TDS estimation in the study area.



(a)

(b)

Figure 7: (a) Residual plot of predicted TDS values and (b) distribution chart of measured and predicted TDS values

Feature importance rankings, derived from XGBoost’s gain metric, revealed that the Red/Blue ratio was the most influential predictor of TDS (0.181), followed by the green band (B3 = 0.089), SWIR1 (B11 = 0.080), and the Blue/Red ratio (0.079) (Figure 8). The prominence of these variables highlights the significance of both visible and shortwave infrared regions in capturing dissolved solid variability, with spectral ratios enhancing sensitivity by reducing atmospheric noise and illumination effects. Blue (B2 = 0.069), SWIR2 (B12 = 0.069), and the red-edge band (B5 = 0.065) contributed moderately, suggesting that both visible and infrared wavelengths carry complementary information about water clarity and dissolved matter. NDWI (0.065) and NIR (B8 = 0.057) also played secondary roles, reinforcing the contribution of green–NIR contrasts in water quality retrieval. By contrast, indices such as NDTI (0.040) and bands in the red-edge region (B7 = 0.040) demonstrated relatively low importance, indicating limited discriminatory capacity for TDS in the study area. Overall, the feature importance hierarchy underscores that while ratio-based indices remain valuable, dissolved solids are better captured when information from both visible and shortwave infrared domains is integrated, reflecting the complex optical interactions of dissolved matter in inland waters.

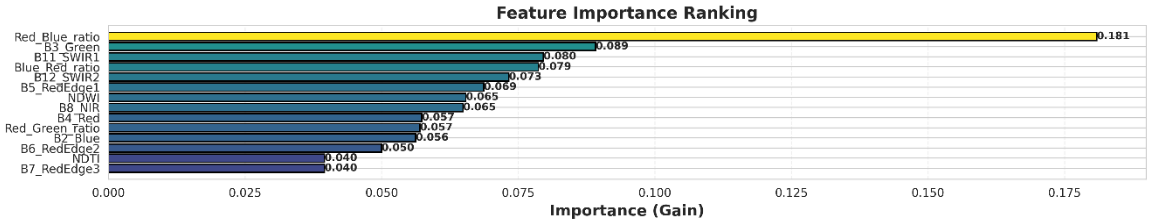


Figure 8: Feature importance ranking derived from XGBoost's gain metric on TDS prediction model.

c. Comparison of TSS and TDS Estimation Model

The comparison of model performance highlights that the XGBoost regression demonstrated stronger predictive capacity for TSS than for TDS. As shown in the left panel of Figure 9, the R² for TSS (0.76) exceeded the "good" threshold (0.70), confirming its reliability in representing suspended sediment variability across different river sections. Meanwhile, the R² for TDS (0.58) fell between the acceptable (0.50) and good (0.70) benchmarks, suggesting moderate performance yet with notable uncertainty, as indicated by the wider confidence range from cross-validation. These results align with previous findings in Porong River, Sidoarjo and Ketapang, South Lampung about TSS estimation with R^2 ranging from 0.7 to 0.9 (Bioresita et al. 2018; Fadel et al. 2023), and TDS value of 0.5321 in Jatiluhur Reservoir (Finita 2021).

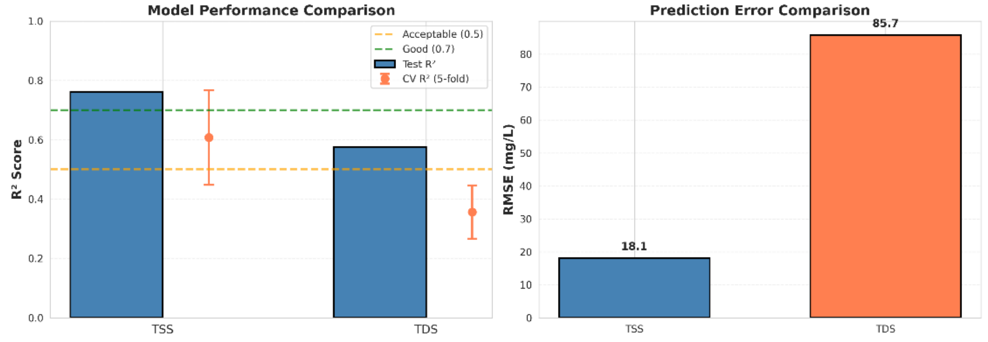


Figure 9: Comparison of XGBoost Model Performance for TSS and TDS Estimation.

The prediction error comparison further reinforces this distinction, where the RMSE for TSS (18.1 mg/L) was substantially lower than that for TDS (85.7 mg/L) (Figure 9). The markedly higher error in TDS predictions reflects the limitations of optical satellite imagery in resolving dissolved constituents, which are often influenced by chemical and hydrological factors not easily captured by reflectance features. Sediment-related parameters are more directly linked to surface reflectance signals, whereas dissolved solids are less optically active and thus more difficult to capture through satellite-based predictors (Adjovu 2023; Al-Fartusi 2023). Nevertheless, the model still provides useful approximations for TDS that can complement traditional field monitoring, particularly in identifying broader spatial and temporal trends. Overall, these findings underscore the suitability of machine learning combined with Sentinel-2 data for robust TSS estimation. However, they also highlight the need for methodological improvements, such as integrating multi-sensor datasets or hybrid modeling approaches, to enhance the predictive accuracy for TDS.

Conclusion and Recommendation

This study demonstrated the potential of integrating Sentinel-2 satellite imagery with XGBoost regression for estimating river water quality parameters, specifically TSS and TDS. The model for TSS achieved strong predictive performance with $R^2 = 0.761$, RMSE = 18.08 mg/L, and MAE = 15.21 mg/L, indicating its robustness in capturing suspended sediment variability across both spatial and temporal scales. In contrast, the estimation of TDS yielded lower accuracy, with $R^2 = 0.575$, RMSE = 85.71 mg/L, and MAE = 69.42 mg/L, although the results still provide informative insights into dissolved solid dynamics. These outcomes confirm that machine learning combined with remote sensing can reduce dependence on resource-intensive field monitoring while enabling cost-effective and spatially continuous assessments of river water quality.

Despite these promising results, the study faced a minor limitation due to temporal discrepancies between in-situ sampling and Sentinel-2 acquisitions, which occurred only within a few hours and may have introduced a degree of uncertainty, particularly in TDS estimation. Addressing such discrepancies through closer synchronization of field campaigns with satellite overpasses or incorporating multi-sensor observations could further improve predictive performance. In addition, expanding the dataset with more extensive temporal and spatial coverage would allow the models to capture greater variability and thus enhance their generalizability. Future research could also explore the integration of multi-sensor satellite data to complement the spectral limitations of Sentinel-2, as well as the application of alternative machine learning algorithms or ensemble approaches to identify the most suitable predictive model. Overall, the findings validate the applicability of satellite-based machine learning models in advancing large-scale river water quality monitoring, highlighting their potential to strengthen management strategies and contribute to sustainable water resource governance.

References

Adjovu, G. E., Stephen, H., James, D., & Ahmad, S. (2023). Measurement of total dissolved solids and total suspended solids in water systems: A review of the issues, conventional, and remote sensing techniques. *Remote Sensing, 15*(14), 1938. https://doi.org/10.3390/rs15071938

Al Fadel, A., Ningsih, N., Ellis, U., & Ulqodry, T. Z. (2023). *Pola sebaran total suspended solid (TSS) menggunakan citra Sentinel 2-A di perairan Ketapang, Lampung Selatan* (Undergraduate thesis). Sriwijaya University.

Al-Fartusi, A. J., Malik, M. I., & Abduljabbar, H. M. (2023). Utilizing Spectral Indices to Estimate Total Dissolved Solids in Water Body Northwest Arabian Gulf. Ilmu Kelautan: Indonesian Journal of Marine Sciences, 28(3), 217-224.

American Public Health Association. (2017). *Standard methods for the examination of water and wastewater* (23rd ed.). APHA.

Ananta, M. F., Nariswana, R., Supriyadi, D. F., Al Fauzan, H. T. G., & Nugroho, R. S. (2024). Smart water quality monitoring with IoT and AI: An XGBoost approach in household water feasibility. *Journal of Scientech Research and Development, 6*(2), 1092–1099.

Apip, Sagala, S. A. H., & Pingping, L. (2015). Overview of Jakarta water related environmental challenges. *Water and Urban Initiative Working Paper Series, 4,* 1–5.

Bilotta, G. S., & Brazier, R. E. (2008). Understanding the influence of suspended solids on water quality and aquatic biota. *Water Research, 42*(12), 2849–2861.

Binding, C. E., Greenberg, T. A., Jerome, J. H., Bukata, R. P., & Letourneau, G. (2010). An assessment of MERIS algal products during an intense bloom in Lake of the Woods. *Journal of Plankton Research, 33*(5), 793–806. https://doi.org/10.1093/plankt/fbq133

Bioresita, F., Firdaus, H. S., Pribadi, C. B., Hariyanto, T., & Puissant, A. (2018). The use of Sentinel-2 imagery for total suspended solids (TSS) estimation in Porong River, Sidoarjo. *Elipsoida: Jurnal Geodesi dan Geomatika, 1*(1). https://doi.org/10.14710/elipsoida.2018.2726

Bui, D. T., Khosravi, K., & Tiefenbacher, J. P. (2020). Improving prediction of water quality indices using novel hybrid machine-learning algorithms. *Science of the Total Environment, 721,* 137612. https://doi.org/10.1016/j.scitotenv.2020.137612

Butler, B. A., & Ford, R. G. (2018). Evaluating relationships between total dissolved solids (TDS) and total suspended solids (TSS) in a mining-influenced watershed. *Mine Water and the Environment, 37,* 18–30. https://doi.org/10.1007/s10230-017-0474-4

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). https://doi.org/10.1145/2939672.2939785

Dekker, A. G., Vos, R. J., & Peters, S. W. M. (2002). Analytical algorithms for lake water TSM estimation for retrospective analyses of TM and SPOT sensor data. *International Journal of Remote Sensing, 23*(1), 15–35. https://doi.org/10.1080/01431160010006917

DKI Jakarta Environmental Agency. (2024). *Monitoring of river water quality in DKI Jakarta.* Jakarta: DKI Jakarta Environmental Agency. https://lingkunganhidup.jakarta.go.id/publikasi/laporan-kualitas-air.

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., Meygret, A., Spoto, F., Sy, O., Marchese, F., & Bargellini, P. (2012). Sentinel-2: ESA's optical high-resolution mission for GMES operational services. *Remote Sensing of Environment, 120,* 25–36. https://doi.org/10.1016/j.rse.2011.11.026

Effendi H. 2003. Telaah Kualitas Air bagi Pengelolaan Sumber Daya dan Lingkungan Perairan. Yogyakarta (ID): Kanisius.

Finita, R. (2021). *Pemetaan estimasi distribusi total zat padat terlarut menggunakan citra penginderaan jauh Landsat 8 OLI di Waduk Jatiluhur, Jawa Barat* (Diploma thesis). Universitas Gadjah Mada.

Gholizadeh, M. H., Melesse, A. M., & Reddi, L. (2016). A comprehensive review on water quality parameters estimation using remote sensing techniques. *Sensors, 16*(8), 1298. https://doi.org/10.3390/s16081298

Ngamile, S., Madonsela, S., & Kganyago, M. (2025). Trends in remote sensing of water quality parameters in inland water bodies: A systematic review. *Environmental Informatics and Remote Sensing, 13.* https://doi.org/10.3389/fenvs.2025.1549301

United Nations. (2023). *SDG Summit 2023.* https://www.un.org/en/conferences/SDGSummit2023

Vanhellemont, Q., & Ruddick, K. (2018). Atmospheric correction of Sentinel-2 and Landsat-8 over coastal and inland waters. *Remote Sensing of Environment, 216,* 586–597. https://doi.org/10.1016/j.rse.2018.07.015

Virdis, S., Xue, W. M., Winijkul, E., Nitivattananon, V., & Punpukdee, P. (2022). Remote sensing of tropical riverine water quality using Sentinel-2 MSI and field observations. *Ecological Indicators, 144,* 109472. https://doi.org/10.1016/j.ecolind.2022.109472

Zheng, G., & DiGiacomo, P. M. (2017). Remote sensing of chlorophyll-a in coastal waters based on the light absorption coefficient of phytoplankton. *Remote Sensing of Environment, 201,* 331–341. https://doi.org/10.1016/j.rse.2017.09.008