Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15540
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Machine Learning Analysis of Jakarta Bay Water Quality: Comparing Models

Aura Savira[1)], **Andrianingsih**[2)*]
[1)2)]Sistem Informasi, Fakultas Teknologi Komunikasi dan Informatika, Universitas Nasional
[1)]aurasavira2022@student.unas.ac.id, [2)]andrianingsih@civitas.unas.ac.id

Submitted :Nov 11, 2025 | **Accepted** : Dec 3, 2025 | **Published** : Jan 02, 2026

Abstract: Jakarta Bay experiences persistent anthropogenic pressures that produce spatially heterogeneous water-quality conditions. This study develops a regulation-aligned, explainable classification framework using a 2024 in-situ dataset collected at 53 stations across two sampling periods (March and August). After preprocessing—including unit harmonization, outlier screening, missing-value imputation, and treatment of below-detection-limit measurements—the dataset yielded 104 complete samples classified into Good (n=46), Lightly Polluted (n=28), and Moderately Polluted (n=34) categories based on KEPMEN LH No. 51/2004. Three ensemble algorithms (LightGBM, CatBoost, and Random Forest) were evaluated using stratified cross-validation to maintain class balance and prevent spatial leakage. CatBoost achieved the best overall performance (Accuracy = 0.8338; F1 = 0.8257), followed by Random Forest, while LightGBM showed the highest variability across folds. Class-level metrics indicate that CatBoost produced the most balanced predictions, particularly for the borderline Lightly Polluted class. SHAP analysis identified turbidity/TSS, nutrients, dissolved oxygen, salinity, and spatial gradients as dominant predictors, enabling transparent interpretation of model decisions. The resulting framework provides a reproducible and operationally deployable approach for rapid screening, hotspot detection, and decision support in Jakarta Bay's water-quality management.

Keywords: Jakarta Bay; Water quality classification; LightGBM; CatBoost; Explainable AI (SHAP)

INTRODUCTION

Jakarta Bay is a semi-enclosed tropical embayment that receives the outputs of a megacity: domestic wastewater, industrial effluents, port activities, and storm-driven runoff collectively imprint spatially variable signals on water quality (Edward & Kusnadi, 2023). In this context, managers require concise yet reliable means to translate monitoring results into actionable condition categories. This study uses a 2024 monitoring dataset comprising two periods—March (Period 1) and August (Period 2)—covering 53 sites across three location types (Bay, Estuary Mouth, and Island). The variables span physical, chemical, heavy-metal, biological, and water-quality index parameters, providing a representative multiparameter view for operational assessment. With a compact temporal scope but broad spatial coverage, the study focuses on site/period classification rather than long-term trend modeling.

Although several prior studies attempt to analyze Jakarta Bay water quality, most focus on single parameters, satellite-derived indicators, or descriptive assessments that are not aligned with national regulatory classes. Existing works lack (i) multi-parameter in-situ classification aligned with KEPMEN LH No. 51/2004, (ii) head-to-head comparison of modern boosting models under consistent evaluation, and (iii) explainable AI analysis that links predictions to environmental processes. These gaps highlight the need for a regulation-based, interpretable, and operationally deployable classification model for Jakarta Bay.

The challenge is intensified by the nonlinear and heterogeneous nature of environmental data, where interactions between nutrients, physical properties, and trace metals often deviate from simple linear patterns. Prior comparative machine-learning studies have demonstrated that tree-based gradient boosting—such as Gradient Boosting Machine and LightGBM—tends to outperform conventional learners when handling mixed-type tabular datasets with complex feature interactions (Hindarto, 2024). Similarly, broader supervised-learning evaluations show that algorithmic performance varies substantially across domains and requires rigorous, reproducible benchmarking to ensure reliability (Hindarto & Santoso, 2022). These methodological insights reinforce the

*name of corresponding author

importance of using comparative ensemble frameworks for Jakarta Bay, where classification decisions have operational consequences for monitoring, enforcement, and coastal management.

Despite routine monitoring, Jakarta Bay lacks a transparent and reproducible framework capable of translating multiparameter measurements into official regulatory categories. The specific problems addressed in this study are:
(1) how to classify water quality into Good, Lightly Polluted, and Moderately Polluted categories using modern machine-learning ensembles;
(2) which environmental parameters most strongly influence class assignments; and
(3) which ensemble algorithm offers the best balance between accuracy, class-level stability, and interpretability.

The objective of this study is to develop, evaluate, and interpret a multi-class classification framework for Jakarta Bay water quality based on KEPMEN LH No. 51/2004. It is hypothesized that CatBoost, due to its ordered boosting and categorical stability, will outperform LightGBM and Random Forest in class-level performance, particularly for borderline classes such as Lightly Polluted. This study also aims to produce an interpretable model capable of supporting operational decision-making through transparent identification of key environmental drivers.

LITERATURE REVIEW

Recent water-quality studies increasingly adopt ensemble machine-learning models to handle heterogeneous environmental data. Random Forest and XGBoost remain widely used, while LightGBM and CatBoost provide faster computation and strong performance on mixed numerical–categorical inputs. Comparative WQI optimization studies consistently show that boosting models reduce index uncertainty and outperform traditional statistical approaches (Uddin et al., 2023). Recent methodological developments also emphasize Explainable AI (XAI)—including SHAP and Permutation Importance—to connect model outputs with environmental mechanisms and improve transparency for regulatory applications (Makumbura et al., 2024).

Alongside ML developments, remote sensing combined with ML has enhanced spatial monitoring of coastal parameters such as turbidity, TSS, and chlorophyll-a using Sentinel-2, Landsat, and MODIS (Bai et al., 2024). In Jakarta Bay, Sentinel-2/GEE studies map NDTI and TSS to identify spatial heterogeneity (Ardyan, 2025), yet these works generally focus on single-parameter estimation and lack systematic integration with multi-parameter in-situ data or multi-class regulatory categories. Nationally, water-quality assessment for marine and coastal waters is guided by KEPMEN LH No. 51/2004, which defines thresholds for Good, Lightly Polluted, and Moderately Polluted waters. However, most Jakarta Bay studies do not incorporate these regulatory classes into ML-based classification frameworks.

Comparative ensemble-learning research shows that LightGBM and CatBoost often outperform traditional ensembles, particularly with heterogeneous input features (Nishat et al., 2025). Nonetheless, many such studies are based on non-tropical or inland datasets, raising questions about their applicability in tropical, complex coastal systems like Jakarta Bay. Furthermore, XAI tools are often used merely for visualization and are not translated into operational rules or monitoring guidelines.

Local Indonesian work (Candra & Andrianingsih, 2025) demonstrates that ensemble ML integrated with spatial platforms such as QGIS can enhance public-sector decision-support workflows. Although this study focuses on demographic indicators rather than environmental data, it highlights the value of combining ML predictions with spatial visualization. However, it does not involve boosting algorithms, environmental parameters, or national water-quality standards—underscoring a methodological gap relevant to Jakarta Bay.

Across the literature, several limitations remain unresolved:
(1) lack of head-to-head comparison between LightGBM, CatBoost, and Random Forest for multi-class, regulation-based coastal water-quality classification;
(2) limited integration of KEPMEN LH No. 51/2004 in ML studies;
(3) inconsistent use of XAI for deriving actionable monitoring insights; and
(4) limited validation of ML models for tropical, high-variability coastal environments.

To address these gaps, this study develops a regulation-aligned, multi-parameter classification framework using CatBoost, LightGBM, and Random Forest for 2024 Jakarta Bay in-situ data. It incorporates SHAP and Permutation Importance to identify environmental drivers behind model predictions and provides spatial visualization outputs to support operational monitoring and decision-making.

Table 1. Summary of Previous Studies and Identified Research Gaps

Study	Data Type	Methods	Key Findings	Limitations / Gap
Uddin et al. (2023)	WQI multiparameter	CatBoost, LightGBM, SVM	Boosting improves predictive stability	No coastal focus; no regulatory classes

Makumbura et al. (2024)	Water-quality datasets	SHAP, XAI	Improves interpretability	No operational thresholding
Bai et al. (2024)	Remote sensing (Sentinel-2)	ML regression	Maps turbidity & TSS spatially	Single parameter; no multi-class
Ardyan (2025)	Jakarta Bay satellite data	NDTI/TSS mapping	Shows spatial heterogeneity	No ML classification; not KEPMEN-based
Candra & Andrianingsih (2025)	DKI Jakarta socio-demographic	RF + QGIS	Strong ML–GIS integration	Not water-quality; no boosting/XAI
Present Study	In-situ multiparameter (2024)	LightGBM, CatBoost, RF + XAI	Regulation-based, interpretable multi-class classification	Fills all above gaps

METHOD

This research method was designed to reliably and clearly address the need for water quality classification in Jakarta Bay. We utilized 2024 in-situ data (two periods; 53 locations) and formulated the task as a multi-class classification using three tree-based ensemble algorithms—LightGBM, CatBoost, and Random Forest. Key contributions include handling values below the detection limit (BDL) with sensitivity auditing, structured imputation and outlier control, consistent feature scaling/coding, and location-based Stratified Group K-Fold validation to prevent spatial leakage. Decision-threshold optimization and probability calibration were also applied to balance false alarms and missed detections. Model performance was evaluated using accuracy, precision, recall, and F1-score and enriched with an interpretability layer (Permutation Importance/SHAP) to reveal the dominant environmental drivers behind classification outputs.

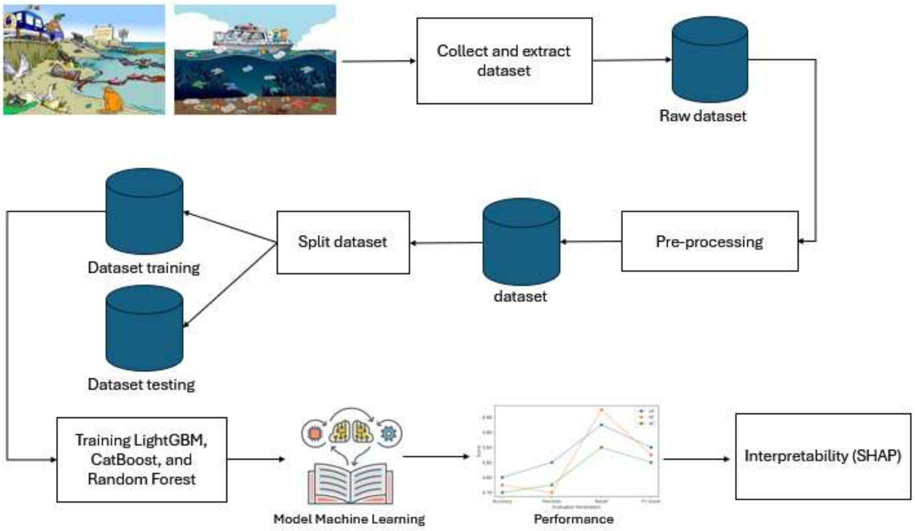

Figure 1. Proposed methodology comparison with model machine learning

Figure 1 illustrates the complete methodological workflow of this study. The process begins with the acquisition and documentation of the raw in-situ dataset from DLH DKI Jakarta. The next stage includes preprocessing and feature engineering, where below-detection-limit values, unit inconsistencies, missing entries, and outliers are systematically handled. The cleaned dataset is then consolidated into wide format and divided into training and testing sets using station-based grouping to reduce spatial leakage. Model development includes hyperparameter tuning and Stratified Group K-Fold validation for each ensemble algorithm. Final evaluation is performed using standard classification metrics and confusion matrices, followed by explainability analysis to identify environmental drivers influencing class predictions.

Dataset

This study uses in-situ water-quality monitoring data collected by the Dinas Lingkungan Hidup Provinsi DKI Jakarta (DLH DKI) during two sampling periods (March and August) in 2024 across 53 monitoring locations classified into three area types: bay, estuary, and island. The dataset is provided in a "long" format in which each row represents a single parameter measurement at a specific station and sampling time. It contains physical, chemical, biological, and heavy-metal parameters, along with station metadata such as coordinates, sampling period, measurement units, and quality flags including below-detection-limit (BDL) indicators.

The raw dataset was obtained through the official open-data portal Satu Data Jakarta under the title *"Data Kualitas Air Teluk dan Laut Jakarta 2024"* (https://data.jakarta.go.id). All values were laboratory-verified by DLH Jakarta's marine water-quality analysis division. Before preprocessing, the dataset consisted of more than 1,000 individual parameter-level measurements.

Table 2 illustrates an excerpt of the raw long-format dataset used for analysis.

Table 2. Sample Dataset Bay and Sea Water Quality Data

Location	Month	Location type	Parameter	Measurement	Label
Pulau Pari	8	Island	Total Phenols	< 0.0008	Good
Pulau Pari	8	Island	PAHs	< 0.0001	Good
Pulau Pari	8	Island	PCBs	< 0.001	Good
Pulau Pari	8	Island	Surfactants	< 0.01	Good
Pulau Semak	8	Island	pH	7.85	Lightly Polluted
Pulau Semak	8	Island	Salinity	30.4	Lightly Polluted

Following the dataset acquisition, the raw long-format file was transformed into a wide analytical format through a series of preprocessing and feature-engineering procedures. All parameter units were aligned to DLH laboratory standards, and below-detection-limit (BDL) values (e.g., "<0.001") were converted using the standard LOD/√2 substitution, with sensitivity checks ensuring that this transformation did not distort class boundaries. Missing values were imputed using median values for numerical parameters and mode assignment for categorical metadata, while outliers were addressed through domain-based thresholds and winsorization at the 1st and 99th percentiles. The resulting dataset was pivoted so that each row represented a unique station-period sample; this process produced 106 expected samples (53 stations × 2 periods), of which 104 remained complete and ready for model development.

Water-quality classes were assigned in accordance with KEPMEN LH No. 51/2004, which defines regulatory thresholds for marine and coastal waters. Samples were categorized into Good, Lightly Polluted, or Moderately Polluted based on multi-parameter criteria including TSS, DO, pH, nutrients, salinity, ammonia, and trace metals. This ensured full consistency with Indonesia's operational classification system used by DLH DKI Jakarta.

All analysis was performed using Python 3.10 in Google Colab Pro equipped with an NVIDIA T4 GPU and 25 GB RAM. Data processing relied on pandas and numpy, model training on scikit-learn, LightGBM, CatBoost, and RandomForest modules, visualization on matplotlib and seaborn, and interpretability on the SHAP framework. The computational environment ensured reproducibility of all model results.

Hyperparameter tuning was conducted using Grid Search in combination with Stratified Group K-Fold validation, where monitoring stations served as grouping units to prevent spatial leakage between folds. LightGBM parameters tuned included num_leaves, max_depth, learning_rate, feature_fraction, bagging_fraction, and λ-based regularization. CatBoost tuning involved depth, iterations, learning_rate, l2_leaf_reg, border_count, and class_weights, while Random Forest tuning covered n_estimators, max_depth, min_samples_split, min_samples_leaf, and max_features. Because the final dataset showed moderate imbalance (46 Good, 28 Lightly Polluted, 34 Moderately Polluted samples), class-weight adjustments were applied across all models to ensure proportional learning.

The unified definitions of mathematical symbols used across the gradient-boosting formulas are maintained for consistency, including x_i(feature vector), y_i(true label), $\hat{y}_i$(predicted output), $g_i$and h_i(first- and second-order gradients), G_L, G_R, H_L, H_R(aggregated gradients and Hessians), and regularization parameters $\lambda$and γ. These notations apply to the gradient-boosting formulations in LightGBM and CatBoost as well as impurity-based criteria in Random Forest.

This fully preprocessed and regulation-aligned dataset was then used for the development and evaluation of the three ensemble learning models described in the next section.

Model Descriptions

Three ensemble-learning algorithms—LightGBM, CatBoost, and Random Forest—were used to model the multi-class coastal water-quality classification task. These models were selected due to their robustness in handling heterogeneous tabular features, nonlinearity, and mixed numerical–categorical structures commonly found in environmental datasets. Their mathematical foundations and key mechanisms are summarized below.

LightGBM

LightGBM is a gradient-boosting framework based on decision trees, where a sequence of weak learners is added to minimize the loss function iteratively. Its efficiency comes from three main mechanisms: (i) leaf-wise tree growth, which expands nodes that provide the highest gain; (ii) Gradient-based One-Side Sampling (GOSS), which retains samples with large gradients to accelerate computation; and (iii) Exclusive Feature Bundling (EFB), which reduces sparsity by grouping mutually exclusive features.

LightGBM follows the additive boosting formulation:

$$\hat{y}_i^{(T)} = \sum_{t=1}^{T} f_t(x_i), f_t \in F \qquad (1)$$

The optimal leaf value for a tree node is calculated as:

$$w_j^{\backslash *} = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \qquad (2)$$

Split selection is based on maximizing gain:

$$\text{Gain} = \frac{1}{2}\left(\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\right) \qquad (3)$$

LightGBM was tuned using parameters such as *num_leaves*, *max_depth*, *learning_rate*, *feature_fraction*, *bagging_fraction*, and regularization coefficients λ. The definitions of $g_i, h_i, G_L, G_R, H_L, H_R$ follow the unified symbol list provided earlier.

CatBoost

CatBoost is a gradient-boosting algorithm optimized for categorical features. It employs two key innovations: **ordered boosting**, which eliminates prediction shift when generating target statistics for categories, and **oblivious/symmetric trees**, where the same splitting criterion is applied at each depth, resulting in stable and efficient tree structures.

The objective function for boosting is expressed as:

$$L^{(t)} = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \qquad (4)$$

with log-loss:

$$\ell(y, \hat{y}) = -[y\log p + (1 - y)\log(1 - p)], p = \sigma(\hat{y}) \qquad (5)$$

Ordered target encoding for categories is computed as:

$$TE_i = \frac{\sum_{j:\sigma(j)<\sigma(i)} y_j + aP}{N_{<i} + a} \qquad (6)$$

The split gain used to choose tree nodes is:

$$\text{Gain} = \frac{1}{2}\left(\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\right) - \gamma \qquad (7)$$

CatBoost tuning included *depth*, *iterations*, *learning_rate*, *l2_leaf_reg*, *border_count*, and *class_weights*. Its integration with SHAP allows model interpretability consistent with XAI standards for environmental decision-making.

Random Forest

Random Forest is a bagging-based ensemble that constructs multiple decision trees using bootstrap samples and random feature subsampling (mtry). These mechanisms reduce correlation among trees and lower variance, improving generalization. Each tree predicts independently, and the final output is obtained through voting (classification) or averaging (regression):

$$\hat{y}(x) = \text{mode}\{h_b(x)\}_{b=1}^{B} \qquad (8)$$

$$\hat{y}(x) = \frac{1}{B}\sum_{b=1}^{B} h_b(x) \qquad (9)$$

Random Forest uses Gini impurity to evaluate split quality:

$$\text{Gini} = 1 - \sum_{k=1}^{K} p_k^2 \qquad (10)$$

and supports internal performance estimation through out-of-bag (OOB) error:

$$Err_{OOB} = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, \hat{y}_i^{OOB}) \qquad (11)$$

Model tuning used parameters such as *n_estimators*, *max_depth*, *min_samples_split*, *min_samples_leaf*, and *max_features*. Random Forest provides permutation-based importance values that complement SHAP results from boosting models.

RESULT

The comparative evaluation of the three ensemble models—LightGBM, CatBoost, and Random Forest—was conducted using the cleaned Jakarta Bay water-quality dataset collected in 2024, comprising 53 monitoring sites across two sampling periods (March and August). The dataset included physical, chemical, biological, and heavy-metal parameters representing complex environmental conditions across bay, estuarine, and island locations. All models were trained and validated using stratified 5-fold cross-validation to maintain class balance among the categories Good, Lightly Polluted, and Moderately Polluted. Model performance was assessed using accuracy, precision, recall, and F1-score metrics, as presented in Table 3. This evaluation ensures a fair and consistent comparison of model performance across all classes and parameters.

Table 3. Class Distribution

Class	Count	Percentage (%)
Good	46	42.59
Lightly Polluted	28	25.93
Moderately Polluted	34	31.48

Table 2 illustrates the distribution of samples across the three regulatory water-quality categories defined by KEPMEN LH No. 51/2004. The dataset shows a moderately imbalanced structure, with the Good class representing the largest proportion, followed by Moderately Polluted and Lightly Polluted samples. This imbalance highlights the need for stratified validation and class-weight adjustment in the modeling stage to ensure that minority classes are learned adequately by the ensemble algorithms.

Table 4. Cross-Validated Ensemble Model Performance (Mean ± Std, 5-Fold Stratified K-Fold)

Model	Mean Accuracy	Std Accuracy	Mean Precision	Std Precision	Mean Recall	Std Recall	Mean F1	Std F1
Random Forest	0.7870	0.0464	0.7956	0.0564	0.7870	0.0464	0.7712	0.0459
LightGBM	0.7784	0.0824	0.8028	0.0855	0.7784	0.0824	0.7772	0.0763
CatBoost	**0.8338**	**0.0458**	**0.8434**	**0.0555**	**0.8338**	**0.0458**	**0.8257**	**0.0439**

Table 4 summarizes the cross-validated performance of the three ensemble models using 5-Fold Stratified K-Fold. CatBoost achieved the highest mean accuracy (0.8338) and mean F1-score (0.8257), indicating strong and stable predictive capability across folds, supported by the lowest standard deviations among all models. Random Forest showed moderate performance (mean F1 = 0.7712), while LightGBM exhibited the lowest stability, reflected in its highest standard deviations (Std Accuracy = 0.0824). These results highlight CatBoost as the most robust and consistent model for Jakarta Bay's multi-parameter water-quality classification.

Table 5. Per-Class Performance for All Models

Model	Class	Precision	Recall	F1
RandomForest	Good	0.8077	0.9130	0.8571
RandomForest	Lightly Polluted	0.7059	0.4286	0.5333
RandomForest	Moderately Polluted	0.7949	0.9118	0.8493
LightGBM	Good	0.8696	0.8696	0.8696
LightGBM	Lightly Polluted	0.6071	0.6071	0.6071

LightGBM	Moderately Polluted	0.7941	0.7941	0.7941
CatBoost	Good	**0.8958**	**0.9348**	**0.9149**
CatBoost	Lightly Polluted	**0.8000**	0.5714	0.6667
CatBoost	Moderately Polluted	0.7750	**0.9118**	**0.8378**

Table 5 presents the per-class precision, recall, and F1-score for all models. CatBoost shows the most balanced performance, achieving high precision and recall in the Good and Moderately Polluted classes and improving recall for the Lightly Polluted class compared to Random Forest and LightGBM. Random Forest performs well for the majority classes but struggles to correctly identify Lightly Polluted samples, while LightGBM maintains moderate scores across all categories. These patterns are consistent with the aggregated metrics shown in Table 3 and further emphasize CatBoost's superior stability across all water-quality categories.

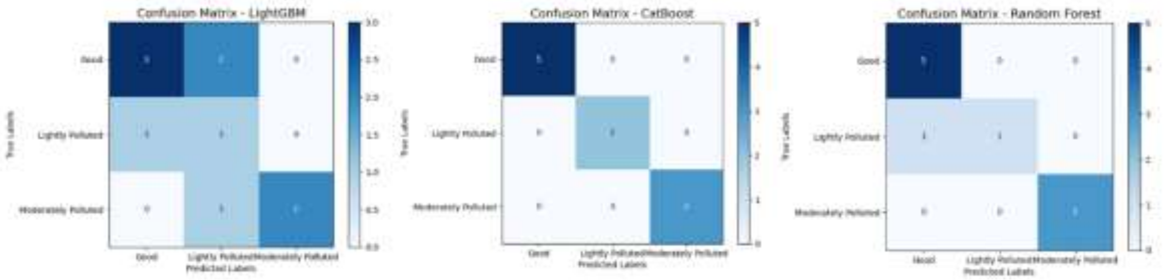

Figure 2. Confusion Matrices of LightGBM, CatBoost, and Random Forest Models

Figure 2 illustrates the confusion matrices of the three ensemble models used in this study. The LightGBM model correctly classified most Good samples but produced several misclassifications in the Lightly Polluted category, indicating moderate sensitivity to borderline water-quality conditions. At the class level, Random Forest achieved perfect recall (1.00) for Good and Moderately Polluted samples but showed reduced recall (0.50) for Lightly Polluted. LightGBM demonstrated relatively consistent performance across classes but with lower precision for Lightly Polluted (0.57). CatBoost produced the most stable results, maintaining recall above 0.65 for all categories and achieving the highest precision in Good (0.90) and Moderately Polluted (0.86) samples. These visual and numerical results are consistent with the metrics in Table 2, confirming that CatBoost provides the best balance between accuracy and stability among the evaluated ensemble algorithms.

Figure 3. Feature Importance (RF, LGBM, CatBoost)

Figure 3 compares feature importance across the three ensemble models. Nutrient indicators (ortho-phosphate, nitrate), turbidity, and BOD5 consistently appear as the strongest predictors, while spatial gradients captured by latitude also contribute substantially. In contrast, several heavy-metal parameters show low importance. The agreement across models confirms that the primary drivers of water-quality classification are nutrient enrichment and sediment-related pressure along the coastal zone.

*name of corresponding author

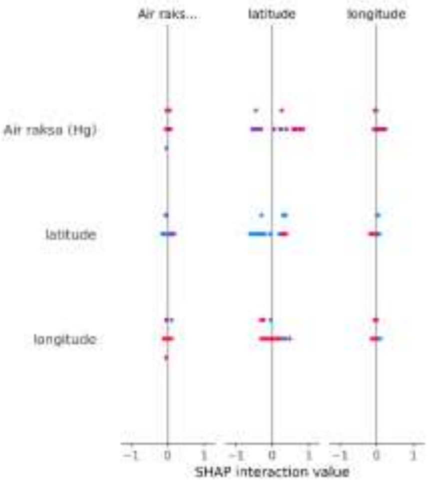

Figure 4. SHAP Summary Plot

Figure 4 presents the SHAP summary plot, which illustrates both the direction and magnitude of each feature's contribution to the CatBoost model predictions. The plot shows that latitude and longitude exhibit clear spatial influence: higher latitude values (corresponding to offshore island stations) contribute positively toward the *Good* class, while lower latitude values (stations located near river mouths and coastal inputs) shift predictions toward the *Lightly Polluted* and *Moderately Polluted* classes. In contrast, the Air raksa (Hg) parameter displays minimal variation in SHAP values and does not show a strong contribution pattern, indicating a relatively limited role in determining the final classification. Overall, the SHAP visualization confirms that the CatBoost model captures underlying spatial gradients while remaining primarily driven by dominant chemical and nutrient-related parameters in assessing coastal water quality.

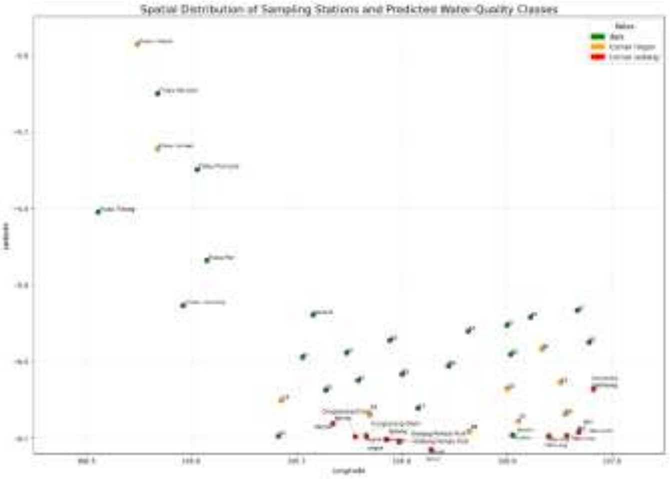

Figure 5. Spatial Spatial Distribution of Sampling Stations and Predicted Water-Quality Classes

Figure 5 illustrates the spatial distribution of the 53 sampling stations across Jakarta Bay with their predicted water-quality classes (Good, Lightly Polluted, Moderately Polluted). The northern offshore island stations predominantly fall within the Good category, while stations located near coastal areas and river mouths display higher concentrations of Lightly Polluted and Moderately Polluted conditions. This spatial pattern aligns with known land-based inputs, where anthropogenic discharge and riverine flows contribute to localized degradation toward the shoreline.

Overall, the experimental results demonstrate that CatBoost delivers the most consistent and accurate performance across all evaluation settings, supported by its superior class-level metrics, clearer confusion-matrix separation, and more stable behavior under cross-validated testing. Feature-importance and SHAP analyses jointly confirm that nutrient-related parameters (such as ortho-phosphate and nitrate), sediment indicators (turbidity and TSS), and spatial gradients (latitude and longitude) are the dominant drivers influencing model predictions. The spatial visualization further verifies the environmental plausibility of these findings, showing that pollution levels intensify in coastal and river-influenced areas while offshore island stations remain comparatively cleaner. These combined results provide a robust foundation for interpreting the environmental mechanisms behind the predicted water-quality classes, which are further explored in the Discussion section.

DISCUSSIONS

The findings of this study highlight several important environmental and methodological implications for coastal water-quality assessment in Jakarta Bay. The superior performance of CatBoost—evident from its consistent cross-validated metrics, stronger class-level recall, and more coherent confusion-matrix patterns—suggests that boosting-based ensemble models are well suited for heterogeneous, non-linear ecological datasets. This result aligns with recent literature indicating that gradient-boosting ensembles outperform traditional classifiers when dealing with mixed numerical–categorical environmental variables and moderate class imbalance (Uddin et al., 2023; Makumbura et al., 2024). The relatively weaker performance of LightGBM, particularly in identifying the Lightly Polluted class, further supports prior studies showing that model sensitivity to borderline conditions varies substantially across boosting architectures.

From an environmental perspective, the dominance of ortho-phosphate, nitrate, turbidity, and TSS in the feature-importance and SHAP analyses is consistent with established pollution drivers in Jakarta Bay. Previous research has shown that land-based nutrient inflows from major rivers such as Citarum, Angke, and Ciliwung contribute to eutrophication and elevated chlorophyll-a concentrations (Damar & Hesse, 2019; Ladwig et al., 2016). The elevated TSS and turbidity levels near coastal and estuarine stations reflect sediment resuspension driven by river discharge, port activity, and seasonal monsoon-driven mixing. These patterns explain why the spatial map generated in this study shows higher pollution probabilities near river mouths and anthropogenic hotspots, while offshore island stations—such as Pulau Pari and Pulau Semak Daun—exhibit more stable “Good” conditions. The SHAP summary plot further confirms the presence of spatial gradients, with lower latitudes (coastal/urban areas) contributing more strongly to polluted classes, matching satellite-based TSS and NDTI findings reported by regional studies (Bai et al., 2024; Ardyan, 2025).

Despite the strong predictive performance, several limitations must be acknowledged. First, the dataset contains only two sampling periods (March and August 2024), which restricts the ability to capture full monsoonal and inter-seasonal variability. Coastal systems in Indonesia exhibit pronounced seasonal patterns, and the absence of data from transitional monsoon periods may limit generalizability. Second, the moderate spatial clustering of stations—particularly around river mouths—raises the possibility of spatial autocorrelation, meaning that nearby stations may share similar feature signatures and thus inflate model performance. Third, some parameters exhibit skewness or high BDL frequency, which may reduce the stability of certain predictors despite the implemented preprocessing steps. Future work should incorporate multi-year monitoring data, more spatially dispersed sampling, and spatial cross-validation to mitigate autocorrelation effects.

The findings offer several actionable implications for environmental agencies. Instead of broad recommendations, this study identifies specific parameter combinations that can trigger early-warning signals: elevated turbidity coupled with high ortho-phosphate or nitrate consistently pushes predictions toward polluted classes, even before DO declines. These patterns can be integrated into DLH Jakarta’s rapid-assessment workflow by prioritizing sites such as river mouths (e.g., Muara Angke, Muara Kamal) during high-discharge periods. Furthermore, the strong spatial gradient detected by SHAP suggests that fixed monitoring stations can be supplemented with mobile or drone-based sampling along the coast to capture short-term pollutant transport. For island conservation zones, routine in-situ monitoring of chlorophyll-a and turbidity should be emphasized to detect early signs of eutrophication.

Overall, this study demonstrates that interpretable ensemble learning provides not only accurate classification results but also meaningful ecological insights into pollution processes in Jakarta Bay. By connecting model behavior with environmental mechanisms and spatial drivers, the approach strengthens the transparency and operational value of machine-learning tools for coastal water-quality management.

CONCLUSION

This study developed an explainable ensemble-learning framework for classifying Jakarta Bay water quality into three regulatory categories based on KEPMEN LH No. 51/2004. Using in-situ measurements from 53 monitoring stations across two 2024 sampling periods, the models demonstrated that tree-based ensembles are effective for multi-parameter coastal classification. CatBoost achieved the highest performance with mean accuracy 0.8338 and mean F1-score 0.8257, followed by Random Forest (F1 = 0.7712), while LightGBM showed lower stability (F1 = 0.7772). Class-level diagnostics further revealed that CatBoost provided the most balanced results across all categories, particularly improving recall for the Lightly Polluted class compared with the other models.

Explainability analysis using feature importance and SHAP confirmed that turbidity/TSS, nitrate, ortho-phosphate, BOD5, ammonia, dissolved oxygen, and spatial variables (latitude/longitude) were the dominant drivers shaping model decisions. These findings correspond with known ecological gradients in Jakarta Bay, where island stations generally exhibit better water quality, while estuary-mouth locations—affected by river discharge—tend to fall into polluted categories.

Several limitations should be acknowledged. The dataset covers only two sampling periods, limiting seasonal interpretation; the number of stations, although operationally realistic, constrains the robustness of spatial generalization; and potential spatial autocorrelation may influence model independence despite grouped cross-validation. Future studies should incorporate longer multi-year time series, additional sensors (e.g., chlorophyll-a or turbidity from Sentinel-2), and spatial modeling approaches to reduce autocorrelation effects.

In practical applications, the proposed framework can support DLH Jakarta and coastal managers in rapid screening, prioritizing intervention zones, designing verification sampling, and communicating data-driven assessments aligned with national regulatory classes. Overall, the results demonstrate that explainable boosting ensembles—particularly CatBoost—offer a reliable, interpretable, and deployment-ready solution for operational coastal water-quality classification in Jakarta Bay.

REFERENCES

Ardyan, P. A. N. (2025). Water Quality Analysis Using NDTI and TSS Parameters Based on Sentinel Image Data in Jakarta Bay Waters. *Maritime Park: Journal of Maritime Technology and Society*, *4*(June), 103–109. https://doi.org/10.62012/mp.vi.43831

Bai, Y., Xu, Z., Lan, W., Peng, X., Deng, Y., Chen, Z., Xu, H., Wang, Z., Xu, H., Chen, X., & Cheng, J. (2024). Predicting Coastal Water Quality with Machine Learning, a Case Study of Beibu Gulf, China. *Water (Switzerland)*, *16*(16), 1–20. https://doi.org/10.3390/w16162253

Chen, B., Chen, Y., & Chen, H. (2024). An Interpretable CatBoost Model Guided by Spectral Morphological Features for the Inversion of Coastal Water Quality Parameters. *Water*, *16*(24), 3615. https://doi.org/10.3390/w16243615

Ding, F., Hao, S., Zhang, W., Jiang, M., Chen, L., Yuan, H., Wang, N., Li, W., & Xie, X. (2025). Using multiple machine learning algorithms to optimize the water quality index model and their applicability. *Ecological Indicators*, *172*, 113299. https://doi.org/https://doi.org/10.1016/j.ecolind.2025.113299

Edward, E., & Kusnadi, A. (2023). Review on monitoring of water quality of the Jakarta Bay, Indonesia. *E3S Web of Conferences*, *454*, 1–18. https://doi.org/10.1051/e3sconf/202345402003

Frincu, R. M. (2025). Artificial intelligence in water quality monitoring: A review of water quality assessment applications. *Water Quality Research Journal*, *60*(1), 164–176. https://doi.org/10.2166/wqrj.2024.049

Gharehbaghi, A., Heddam, S., Mehdizadeh, S., & Kim, S. (2025). Development of interpretable intelligent frameworks for estimating river water turbidity. *Engineering Applications of Computational Fluid Mechanics*, *19*(1). https://doi.org/10.1080/19942060.2025.2511886

Hindarto, D. (2022). Perbandingan Kinerja Akurasi Klasifikasi K-NN, NB dan DT pada APK Android. *JATISI (Jurnal Teknik Informatika Dan Sistem Informasi)*, *9*(1), 486–503. https://doi.org/10.35957/jatisi.v9i1.1542

Hindarto, D. (2024). Case Study: Gradient Boosting Machine vs Light GBM in Potential Landslide Detection. *Journal of Computer Networks, Architecture and High Performance Computing*, *6*(1), 169–178. https://doi.org/10.47709/cnahpc.v6i1.3374

Hindarto, D., & Santoso, H. (2022). PERFORMANCE COMPARISON OF SUPERVISED LEARNING USING NON-NEURAL NETWORK AND NEURAL NETWORK. *Janapati*, *11*, 49–62.

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

Makumbura, R. K., Mampitiya, L., Rathnayake, N., Meddage, D. P. P., Henna, S., Dang, T. L., Hoshino, Y., & Rathnayake, U. (2024). Advancing water quality assessment and prediction using machine learning models, coupled with explainable artificial intelligence (XAI) techniques like shapley additive explanations (SHAP) for interpreting the black-box nature. *Results in Engineering*, *23*, 102831. https://doi.org/10.1016/j.rineng.2024.102831

Nishat, M. H., Khan, Md. H. R. B., Ahmed, T., Hossain, S. N., Ahsan, A., El-Sergany, M. M., Shafiquzzaman, Md., Imteaz, M. A., & Alresheedi, M. T. (2025). Comparative analysis of machine learning models for predicting water quality index in Dhaka's rivers of Bangladesh. *Environmental Sciences Europe*, *37*(1), 31. https://doi.org/10.1186/s12302-025-01078-w

Setiawan, S. (2025). Trends and gaps in ai-driven predictive models for coastal water quality: A bibliometric study. *BIO Web of Conferences*, *188*, 04004. https://doi.org/10.1051/bioconf/202518804004

Shah, F. U., Khan, A. U., Khan, A. W., Ullah, B., Khan, M. R., & Javed, I. (2024). Comparative analysis of ensemble learning algorithms in water quality prediction. *Journal of Hydroinformatics*, *26*(12), 3041–3059. https://doi.org/10.2166/hydro.2024.071

Singh, P., Hasija, T., Bharany, S., Naeem, H. N. T., Rao, B. C., Hussen, S., & Rehman, A. U. (2025). An ensemble-driven machine learning framework for enhanced water quality classification. *Discover Sustainability*, *6*(1), 552. https://doi.org/10.1007/s43621-025-01467-4

Uddin, M. G., Nash, S., Rahman, A., & Olbert, A. I. (2023). Performance analysis of the water quality index model for predicting water state using machine learning techniques. *Process Safety and Environmental Protection*, *169*, 808–828. https://doi.org/10.1016/j.psep.2022.11.073