OPEN ACCESS
Jurnal Ilmiah Ilmu Terapan Universitas Jambi
Vol. 10 No. 3. June 2026, pp. 1511-1522 DOI. 10.22437/jiituj.v10i3.57126
Research Article

A CLUSTER-LABEL-BASED FRAMEWORK FOR WATER QUALITY RISK PATTERN CLASSIFICATION USING NAIVE BAYES AND RANDOM FOREST

Widya Cholid Wahyudin[1,3,*], Tole Sutikno[2], Rusydi Umar[1]

[1] Department of Informatics, Faculty of Industrial Technology, Ahmad Dahlan University, Yogyakarta, Indonesia
[2] Department of Electrical Engineering, Faculty of Industrial Technology, Ahmad Dahlan University, Yogyakarta, Indonesia
[3] Department of Computer Science, Faculty of Science and Technology, Universitas Muhammadiyah Kudus, Jawa Tengah, Indonesia
Corresponding author email: widyacholidwahyudin@umkudus.ac.id

Article Info
Received: May 13, 2026
Revised: Jun 02, 2026
Accepted: Jun 27, 2026
OnlineVersion: Jun 30, 2026

Abstract
River water quality monitoring requires an analytical approach that can classify risk patterns from field observations, especially when independent official water quality labels are unavailable. Method: This study classified water quality risk patterns in the Bengawan Solo River using Naive Bayes and Random Forest based on K-Means cluster labels. The initial dataset consisted of 1,753 observations with ten attributes. After feature selection and missing-value removal, 1,751 observations with seven features were used, namely temperature, pH, electrical conductivity, total dissolved solids, water color, odor, and weather condition. K-Means with K=2 generated higher-risk and lower-risk pattern labels, which were then used as classification targets. Results: Naive Bayes achieved an accuracy of 0.988604, precision of 0.988880, recall of 0.988339, and F1-score of 0.988577. Random Forest achieved an accuracy of 0.980057, precision of 0.980499, recall of 0.979655, and F1-score of 0.980004. Discussion: The findings indicate that cluster-derived labels can be recognized consistently by supervised models, while Random Forest feature importance shows that total dissolved solids and electrical conductivity are the most dominant parameters. The results should be interpreted as cluster label-based risk pattern classification, not as official pollution status prediction. The novelty of this study lies in integrating cluster-derived labels, supervised classification, and feature-importance analysis for unlabeled water-quality data, while the findings highlight the practical importance of total dissolved solids and electrical conductivity in preliminary water-quality monitoring.

Keywords: Bengawan Solo, K-Means Clustering, Naive Bayes, Random Forest, Water Quality Risk Patterns

© 2026 by the author(s)
This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

INTRODUCTION

Rivers are surface water resources that play an important role in community life, agricultural activities, industry, and ecosystem sustainability. River water quality may change due to anthropogenic activities, domestic waste, industrial discharge, surface runoff, land-use change, and surrounding environmental conditions. Therefore, water quality monitoring needs to be conducted continuously using approaches that can systematically capture changes in physicochemical and environmental parameters. Recent studies show that machine learning has increasingly been used in water resources management because it can support prediction, classification, and interpretation of water quality patterns in complex datasets (Ren et al., 2021; Abraham et al., 2022; Ghobadi & Kang, 2023; Lu & Wang, 2025; Yan et al., 2024; Lokman et al., 2025; Yuan et al., 2025).

Water quality is generally analyzed using parameters such as temperature, pH, electrical conductivity, total dissolved solids, dissolved oxygen, biological oxygen demand, chemical oxygen demand, and other biological indicators (Lencha et al., 2021; Khouri & Al-Mufti, 2022; Pairan et al., 2023; Alharbawee & Mohammed, 2024; Luo et al., 2024). Several studies have shown that machine learning models can be used for water quality index estimation, water quality classification, and reduction of monitoring parameter requirements (J. Xu et al., 2021; A. del Castillo et al., 2022; Khoi et al., 2022; Masood & others, 2023; Shamsuddin & others, 2022; Tran et al., 2022; Firdiani et al., 2024). However, the application of supervised learning to water quality data requires valid and independent class labels. In field datasets without official labels, such as good quality, lightly polluted, or heavily polluted water, direct classification cannot always be applied safely.

Recent water-quality studies also show that model performance is strongly influenced by data preparation, feature characteristics, and the availability of reliable labels. Recent comparative studies consistently report that combining multiple machine learning algorithms can improve prediction accuracy and classification robustness because different algorithms capture nonlinear relationships in water quality datasets differently (Abraham et al., 2022; Mu et al., 2023; Sangwan & Bhardwaj, 2024; Hridoy & others, 2025; Lu & Wang, 2025; Man et al., 2025; Xiao & Shi, 2024; R. Xu et al., 2025). Dritsas & others (2023) demonstrated that data-driven machine learning can support efficient water-quality prediction. Other studies have used machine learning to estimate water quality index and classify river or surface-water conditions using multivariate parameters (Hassan et al., 2021; J. Xu et al., 2021; Khoi et al., 2022; Wu & Wang, 2022). More recent studies further highlight that multi-algorithm comparison is useful because water-quality datasets may contain nonlinear patterns, class imbalance, and parameter interactions that cannot be captured by a single method alone (Ramya & others, 2024; Gao et al., 2025; Hridoy & others, 2025).

A major issue in unlabeled datasets is the risk of constructing non-independent targets. If water quality labels are created from parameter scores and the same scores or their constituent parameters are reused as model features, data leakage may occur. Data leakage can make model performance appear overly optimistic and may not reflect its generalization ability on new data (Kapoor & Narayanan, 2023). Therefore, this study does not use artificial labels such as total scores as direct classification targets. As an alternative, K-Means clustering results can be used as initial pattern labels for datasets without independent labels. In the previous clustering stage, K-Means was used to group Bengawan Solo River water quality data and produced two main patterns, namely a higher-risk water quality pattern and a lower-risk water quality pattern. Clustering is relevant for water quality grouping and heterogeneous environmental modeling (Ardarsa & Surinta, 2021; Schäfer et al., 2022; Tran et al., 2022; Warsito et al., 2022; Zamri et al., 2022; Mu et al., 2023; Novianta et al., 2023; Abadi & others, 2025). Thus, K-Means cluster labels can be used as initial targets to examine whether the generated patterns can be consistently recognized by supervised learning models.

The use of clustering and preprocessing is also supported by recent national informatics studies. Muhammad et al. (2025), with Tole Sutikno as co-author, compared K-Means and KNN imputation for handling missing data, showing that preprocessing strategy is an important stage before computational modeling. Wala et al. (2024), with Rusydi Umar as co-author, applied K-Means clustering to group heart-disease patient data, indicating that K-Means remains relevant for discovering hidden patterns in unlabeled datasets. In addition, Novianta et al. (2023) used K-Means for grouping rivers based on water-quality parameters, which is closely related to the clustering foundation used in this study.

Despite the growing application of machine learning in water quality assessment, most previous studies have focused on prediction or classification tasks using predefined labels, such as Water Quality

Index categories, pollution levels, or expert-defined water quality classes (Musnansyah et al., 2021; Pandit et al., 2023; Armijos-Arcos et al., 2025). These approaches generally assume the availability of valid and independent target labels. However, many routine river-monitoring datasets contain physicochemical observations without corresponding official pollution-status labels, making the direct application of supervised classification methods difficult.

This condition creates a methodological gap in water quality analytics. Although clustering techniques have been widely used to discover hidden patterns in environmental datasets, limited studies have investigated whether cluster-derived water quality patterns can be consistently recognized by supervised learning algorithms while simultaneously providing interpretable information regarding dominant environmental parameters. In addition, studies that explicitly address the risk of data leakage in cluster-based water quality classification remain relatively limited. Therefore, a methodological framework that combines clustering, supervised classification, and parameter interpretation is needed to support more reliable analysis of unlabeled water quality datasets.

This study focuses on classifying water quality risk patterns in the Bengawan Solo River using Naive Bayes and Random Forest based on K-Means cluster labels. Naive Bayes was selected as a comparative model because it applies a simple and efficient probabilistic approach. Random Forest was used because it can handle nonlinear relationships among features and provides feature importance information to identify the parameters that contribute most to distinguishing water quality risk patterns (Tran et al., 2022; P. Del Castillo & others, 2024; Firdiani et al., 2024; Sangwan & Bhardwaj, 2024; Gao et al., 2025; Hridoy & others, 2025)

Random Forest was selected not only as an alternative classification model, but also as a methodological component to support parameter interpretation in cluster-derived water quality risk patterns. Environmental water quality data often contain nonlinear relationships among physicochemical and field-observed variables, making ensemble-based models useful for capturing complex patterns. In this study, Random Forest helps identify dominant parameters through feature importance analysis, so the classification results are not limited to performance comparison but also provide insight into which water quality variables contribute most to distinguishing higher-risk and lower-risk patterns.

Previous works by the authors also provide methodological support for this study. Wahyudin et al. (2025) compared several data-mining models with feature selection for predictive modeling, while Ridwan et al. (2025), with Tole Sutikno and Widya Cholid Wahyudin as co-authors, compared Naive Bayes with other data-mining classifiers for student-graduation prediction. Kristiyowati et al. (2025) applied Naive Bayes in a chatbot classification system, confirming the practical usefulness of probabilistic classification in structured decision tasks. In the environmental monitoring context, Hernanda et al. (2025) showed that computational and IoT-based approaches can support environmental safety control. These studies strengthen the rationale for combining cluster-derived labels, Naive Bayes, Random Forest, and parameter interpretation in the present water quality risk classification task.

Based on the identified research gap, this study aims to develop a cluster-label-based framework for water quality risk classification in river-monitoring datasets that do not possess independently validated pollution-status labels. Specifically, this study seeks to: (1) identify water quality risk patterns through K-Means clustering, (2) evaluate the ability of Naive Bayes and Random Forest to classify the resulting cluster-derived labels, and (3) determine the dominant environmental parameters contributing to water quality risk differentiation through feature importance analysis. The novelty of this study lies in the integration of unsupervised clustering, supervised classification, and feature-importance-based interpretation within a single analytical framework while explicitly addressing data-leakage concerns. Unlike conventional water quality classification studies that rely on predefined quality labels, this research utilizes cluster-derived labels as an intermediate representation of water quality risk patterns and subsequently evaluates their consistency using supervised learning models.

The main contributions of this study are to classify K-Means cluster labels as water quality risk patterns using Naive Bayes and Random Forest, compare the performance of both models based on accuracy, precision, recall, and F1-score, and identify dominant parameters using Random Forest feature importance. The study also maintains a data leakage prevention principle by avoiding score-based artificial labels as classification targets. Recent studies have demonstrated that Explainable Artificial Intelligence (XAI), particularly SHAP-based interpretation, has become increasingly important for improving transparency, parameter interpretation, and decision support in water quality prediction models (Mallick et al., 2022; Ezzat et al., 2023; Madni et al., 2023; R. Aldrees & others, 2024; Makumbura et al., 2024; Nallakaruppan et al., 2024; Choudhary et al., 2025).

This research therefore contributes to the literature by bridging an unsupervised clustering stage and a supervised classification stage in a dataset that does not have official water quality class labels. It also prepares a methodological basis for future explainable artificial intelligence, as current studies increasingly emphasize model transparency, interpretability, and parameter level explanation in water-quality assessment (A. Aldrees & others, 2024; Makumbura et al., 2024; Nallakaruppan et al., 2024; Choudhary et al., 2025).

RESEARCH METHOD

This study used a supervised learning approach to classify water quality risk patterns in the Bengawan Solo River based on K-Means cluster labels. The design was quantitative and computational, involving feature preprocessing, cluster label generation, supervised classification, performance evaluation, and interpretation of dominant parameters. Because the classification target was generated from K-Means cluster labels rather than from external regulatory water quality classes, this study was designed as cluster label-based classification and internal consistency validation. Therefore, the classification models were used to examine whether the risk patterns produced by K-Means could be consistently learned by supervised algorithms, not to claim an independent prediction of official pollution status.

This study used secondary water quality monitoring data obtained from the Bengawan Solo River Basin Authority (Balai Besar Wilayah Sungai Bengawan Solo, BBWS Bengawan Solo). The dataset consisted of 1,753 observations representing routine river-water monitoring records. Because all available observations that met the completeness criteria were included in the analysis, a total sampling approach was applied. The original dataset contained measurements of water temperature, pH, electrical conductivity, total dissolved solids, water color, water odor, weather condition, local water level, and river discharge. In this study, seven parameters directly related to water quality characteristics were selected as modeling features, namely water temperature, pH, electrical conductivity, total dissolved solids, water color, water odor, and weather condition. The dataset was subsequently processed through cleaning, feature selection, missing-value handling, clustering, and classification procedures.

Table 1. Water Quality Variables Used in the Study

No.	Variable	Description
1	Water Temperature	Water temperature (°C)
2	pH	Acidity level of water
3	Electrical Conductivity	Electrical conductivity of water (µS/cm)
4	Total Dissolved Solids	Dissolved solids concentration (mg/L)
5	Water Color	Encoded water-color observation
6	Water Odor	Encoded water-odor observation
7	Weather Condition	Encoded weather-condition observation

The final dataset used in this study consisted of 1,751 observations after data cleaning and missing-value removal. This sample size is considered adequate for machine-learning-based classification because it provides sufficient observations for training and testing processes while maintaining representative variability across the selected water-quality parameters. The dataset consisted of 1,753 observations with ten attributes, namely ID, Water Temperature, pH, Electrical Conductivity (EC), Total Dissolved Solids (TDS), Water Color, Water Odor, Weather Condition, Local Water Level, and Manual River Discharge. Seven main parameters were used as modeling features: water temperature, pH, electrical conductivity, total dissolved solids, water color, odor, and weather condition. The preprocessing stage included missing-value checking, relevant feature selection, numerical format validation, and removal of observations with missing values in the selected features (Ji et al., 2022; Yu et al., 2022). After missing values were removed, 1,751 observations were used in the modeling process. Categorical variables such as water color, odor, and weather condition were represented in numerical form through categorical encoding.

This preprocessing strategy was designed to reduce the possibility of biased modeling caused by incomplete observations and inconsistent feature formats. Missing-value handling and feature preparation are critical in water-quality machine learning because classification models are sensitive to the consistency of input variables (Bouchelouche et al., 2021; Kapoor & Narayanan, 2023; Muhammad et

al., 2025). In this study, the preprocessing stage was kept transparent by excluding identifier attributes and retaining only the features directly related to physicochemical and environmental characteristics.

Table 2. Research Features

No.	Feature	Description
1	Water Temperature	Water temperature in degrees Celsius
2	pH	Acidity level of water
3	EC	Electrical conductivity of water
4	TDS	Total dissolved solids
5	Water Color	Numerical representation of water color
6	Water Odor	Numerical representation of water odor
7	Weather Condition	Numerical representation of weather condition

Because the dataset did not contain independent water quality labels, K-Means cluster labels were used as classification targets. Cluster labels were generated using K-Means with K=2. The use of K=2 refers to the previous clustering evaluation that showed the best configuration based on Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Score. Before applying K-Means, all features were standardized using Z-score normalization to ensure comparable scales among features, as shown in Equation (1).

Table 3. K-Means Cluster Evaluation Results

K	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Score
2	0.187638	1.926465	456.873808
3	0.160465	1.972924	354.929149
4	0.158596	2.069388	314.704099
5	0.145326	2.050436	271.311654

The evaluation summary shows that K=2 produced the highest Silhouette Score (0.187638), the lowest Davies-Bouldin Index (1.926465), and the highest Calinski-Harabasz Score (456.873808). Therefore, K=2 was selected as the most representative configuration for generating cluster-derived classification targets.

$$z = \frac{x - \mu}{\sigma} \quad \ldots (1)$$

After cluster labels were generated, the dataset was divided into training and testing data using a train-test split with 80% training data and 20% testing data. The split was performed using a stratified scheme to preserve the distribution of cluster labels in both subsets. Naive Bayes was used as a comparative classification model because it is based on a simple and efficient probabilistic approach. Random Forest was used as the second classification model because it can handle nonlinear relationships among features and is relatively stable against data variation. Random Forest was also used to obtain feature importance values.

The use of two different classifiers was intended to compare a simple probabilistic model with an ensemble-based tree model. Naive Bayes is computationally efficient and can perform well when class patterns are separable, whereas Random Forest is useful for modeling nonlinear relationships and obtaining feature-importance values (Firdiani et al., 2024; Ramya & others, 2024; Sangwan & Bhardwaj, 2024). This combination enables the study not only to compare predictive performance but also to interpret the parameters that shape the classification results.

$$P(C \mid X) = \frac{P(X \mid C)P(C)}{P(X)} \ldots (2)$$

$$\hat{C} = \arg\max_{C}[P(C) \times \Pi\, P(x_i \mid C)] \ldots (3)$$

$$\hat{y} = mode\{h_{1(x)}, h_{2(x)}, \ldots, h_{T(x)}\} \ldots (4)$$

The data analysis process was conducted in several stages. First, data preprocessing was performed through missing-value checking, feature selection, categorical encoding, and Z-score standardization. Second, K-Means clustering was applied to generate cluster-derived labels representing water quality risk patterns. Third, the resulting cluster labels were used as target variables for supervised classification using Naive Bayes and Random Forest algorithms. To evaluate the classification

performance, the dataset was divided into training and testing subsets using an 80:20 stratified train-test split. Model performance was assessed using accuracy, precision, recall, and F1-score. In addition, Random Forest feature importance analysis was employed to identify the relative contribution of each parameter to the classification results. All analyses were performed using Python and relevant machine-learning libraries.

Model evaluation was performed using accuracy, precision, recall, and F1-score. These metrics were used to assess the ability of Naive Bayes and Random Forest to classify water quality risk pattern labels. The evaluation formulas are shown in Equations (5) to (8).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \dots (5)$$

$$Precision = \frac{TP}{TP + FP} \dots (6)$$

$$Recall = \frac{TP}{TP + FN} \dots (7)$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \dots (8)$$

In these equations, TP represents true positive, TN represents true negative, FP represents false positive, and FN represents false negative.

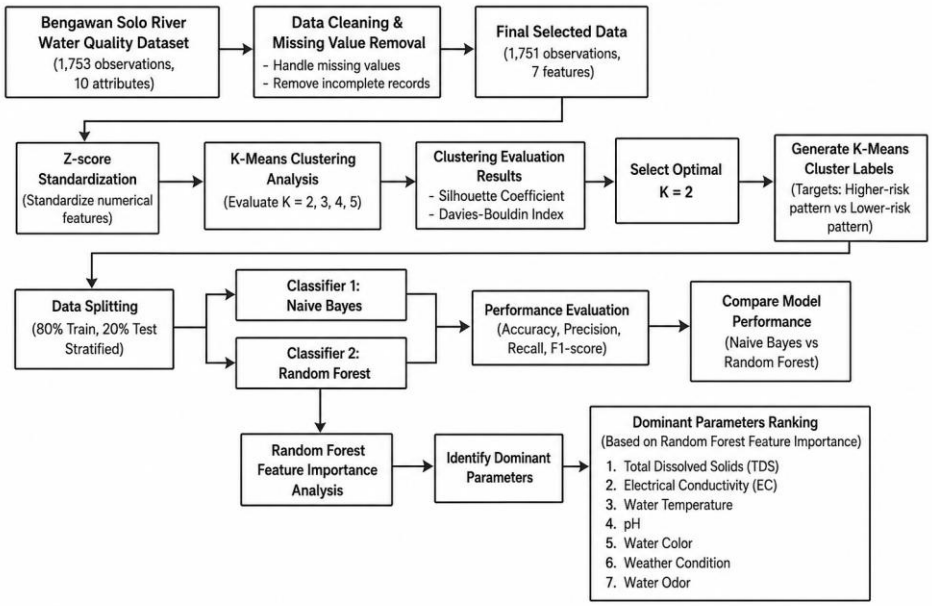

Figure 1. Research Workflow for Water Quality Risk Pattern Classification Based on K-Means Cluster Labels

RESULTS AND DISCUSSION

Cluster Label Formation

The initial dataset consisted of 1,753 observations with ten attributes. After feature selection and missing-value removal, 1,751 observations with seven main features were used in the modeling process. K-Means with K=2 was applied to generate cluster labels as classification targets. The cluster label distribution is shown in Table 4.

Table 4. Distribution of K-Means Cluster Labels

Cluster	Number of Observations	Percentage	Interpretation
0	840	47.97%	Higher-risk pattern
1	911	52.03%	Lower-risk pattern

Based on the average parameter characteristics, Cluster 0 had an electrical conductivity value of 558.7875 and a TDS value of 345.6404, while Cluster 1 had an electrical conductivity value of 284.8330

and a TDS value of 164.5026. Therefore, Cluster 0 was interpreted as a higher-risk pattern and Cluster 1 as a lower-risk pattern.

Table 5. Cluster Label Characteristics

Cluster	Average Electrical Conductivity	Average TDS	Interpretation
0	558.7875	345.6404	Higher-risk pattern
1	284.8330	164.5026	Lower-risk pattern

Classification Performance

The K-Means cluster labels were used as classification targets to compare the performance of Naive Bayes and Random Forest. The evaluation results based on accuracy, precision, recall, and F1-score are shown in Table 6 and Figure 2. Naive Bayes achieved an accuracy of 0.988604 and an F1-score of 0.988577, while Random Forest achieved an accuracy of 0.980057 and an F1-score of 0.980004. These values indicate the classification performance of both models in recognizing the cluster-derived labels.

Table 6. Classification Performance Comparison

Model	Accuracy	Precision	Recall	F1-score
Naive Bayes	0.988604	0.988880	0.988339	0.988577
Random Forest	0.980057	0.980499	0.979655	0.980004

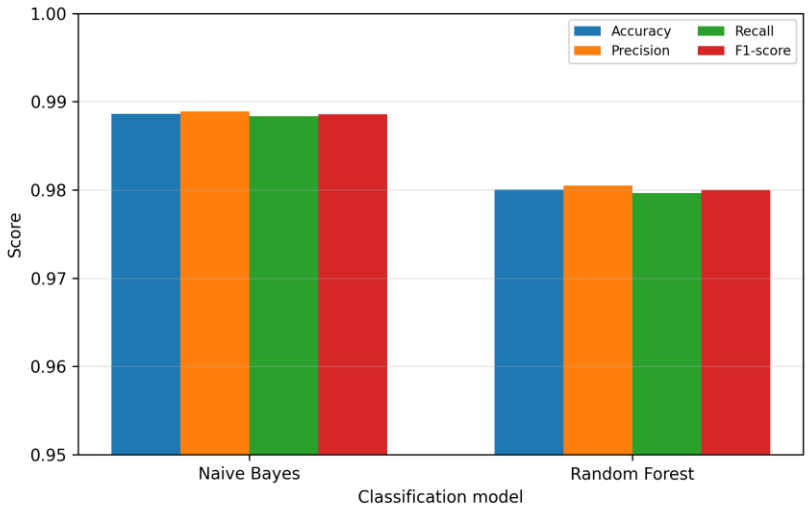

Figure 2. Classification Performance Comparison of Naive Bayes and Random Forest

Dominant Parameters Based on Random Forest

Random Forest was used to obtain feature importance and identify the parameters that contributed most to distinguishing water quality risk patterns. The feature importance results are shown in Table 7 and Figure 3. TDS had the highest contribution with an importance value of 0.509529, followed by electrical conductivity with an importance value of 0.456166. Other parameters, including temperature, pH, water color, weather condition, and odor, had much smaller contributions.

Table 7. Random Forest Feature Importance

Rank	Parameter	Importance
1	TDS	0.509529
2	EC	0.456166
3	Water Temperature	0.011738
4	pH	0.010366
5	Water Color	0.005468
6	Weather Condition	0.004339
7	Water Odor	0.002393

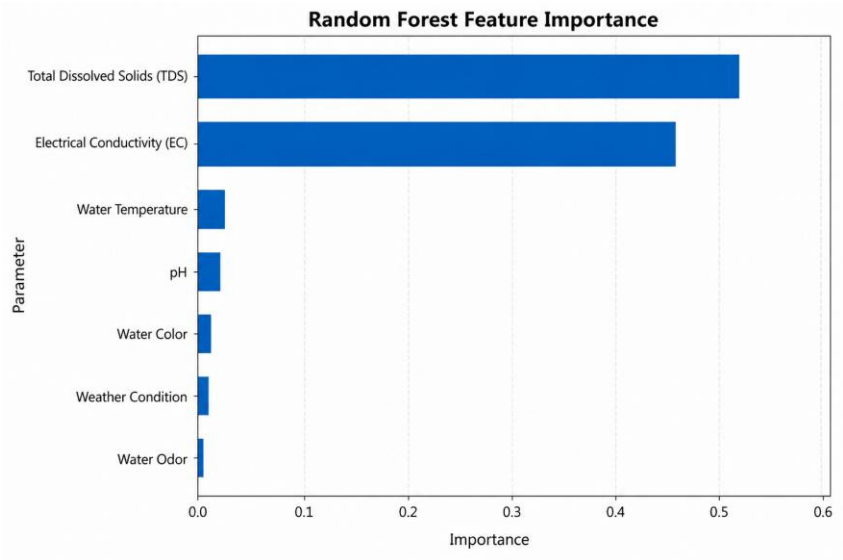

Figure 3. Random Forest Feature Importance

The discussion section interprets the results presented in the previous section and positions the findings within the methodological design of cluster label-based classification. Therefore, this section focuses on the meaning of the high classification performance, the role of Random Forest feature importance, and the limitations of using cluster-derived labels rather than official pollution-status categories. The results show that K-Means cluster labels can be used as classification targets for water quality risk patterns in the Bengawan Solo River. The relatively balanced label distribution, with 840 observations in Cluster 0 and 911 observations in Cluster 1, indicates that the classification models were not trained on highly imbalanced targets. This condition supports more stable evaluation of Naive Bayes and Random Forest.

Although the classification results show high performance, they must be interpreted within the methodological design of this study. The labels used as classification targets were generated from the same set of selected features through K-Means clustering. Therefore, the high accuracy values should not be interpreted as evidence that the models can independently determine official water-pollution status. Instead, the results indicate that the cluster boundaries formed by K-Means are sufficiently learnable by supervised algorithms. In this context, Naive Bayes and Random Forest function as internal consistency validation models and as cluster-label-based classifiers.

This interpretation is important because using cluster-derived labels differs from using externally validated water-quality classes. The present study confirms the consistency of the generated risk patterns, while external validation using complete laboratory measurements, expert assessment, or regulatory water-quality standards remains necessary in future research. Thus, the contribution of this study is positioned as a data driven risk pattern classification framework rather than as a final regulatory pollution-classification system. Naive Bayes produced slightly higher performance than Random Forest. This indicates that the cluster-derived patterns could be recognized well by a simple probabilistic model. Nevertheless, Random Forest still provided an important contribution because it produced feature importance values that support parameter interpretation. Thus, Naive Bayes can be positioned as an efficient classification model, while Random Forest can be positioned as an interpretative model in this study.

The role of Random Forest in this study is therefore complementary to Naive Bayes. Although Naive Bayes achieved slightly higher classification performance, Random Forest provides an additional advantage by explaining the relative contribution of each input feature. This capability is important for environmental data analysis because the model can indicate which parameters should receive more attention in preliminary monitoring. In this dataset, the feature importance results show that TDS and electrical conductivity were the dominant variables, indicating that Random Forest helped clarify the parameter structure behind the cluster-derived risk patterns.

From a practical perspective, these findings provide useful implications for preliminary river-water monitoring. The identification of total dissolved solids and electrical conductivity as dominant parameters suggests that these variables may serve as priority indicators for distinguishing higher-risk

and lower-risk water quality patterns. Consequently, monitoring programs with limited resources may benefit from prioritizing these parameters during routine observations and early risk assessment activities.

The finding that TDS and electrical conductivity were the dominant parameters is consistent with the cluster characteristics. The cluster with higher TDS and electrical conductivity values was interpreted as the higher-risk water quality pattern. This indicates that dissolved solids and ion concentration in water are important indicators for distinguishing water quality patterns in the dataset. The dominance of TDS and electrical conductivity is also consistent with previous studies that used physicochemical parameters to explain water quality variation and classification outcomes (J. Hu & others, 2024; P. Hu & Feng, 2024; C. Xu et al., 2023). Several studies show that dissolved solids, conductivity-related variables, and other ion related parameters often play major roles in water-quality modeling because they reflect the concentration of dissolved substances in water (Madni et al., 2023; Nallakaruppan et al., 2024; Ramya & others, 2024; Hridoy & others, 2025; Yuan et al., 2025). In the present dataset, the higher-risk cluster was characterized by higher average TDS and electrical conductivity, which explains why both variables also emerged as the dominant Random Forest features.

This study does not claim the classification results as official pollution status. Instead, it classifies risk patterns based on K-Means cluster labels. This approach is more appropriate for field datasets that do not yet have independent water quality labels. In addition, it avoids score-based label construction that may cause data leakage if the scores or their constituent parameters are reused in model training. Methodologically, this study strengthens the clustering stage by showing that the generated patterns can be consistently recognized by supervised learning models. The results may serve as a basis for further research, particularly the use of explainable artificial intelligence such as SHAP to explain parameter contributions in greater depth.

Compared with studies that directly classify official WQI categories, this study uses a more cautious strategy because independent water quality labels are not yet available. The approach is consistent with the recommendation that machine-learning experiments should clearly separate feature construction, target definition, and model evaluation to reduce leakage risk and improve reproducibility (Kapoor & Narayanan, 2023). Therefore, the findings should be viewed as a data driven risk pattern classification framework that can be further validated using more complete laboratory parameters and regulatory standards.

This study has several limitations. First, the classification targets were derived from K-Means clustering rather than from independently validated water-quality classes. Second, the dataset was limited to physicochemical and field-observed parameters available in the monitoring records and did not include laboratory-based indicators such as dissolved oxygen, biological oxygen demand, or chemical oxygen demand. Therefore, the results should be interpreted within the context of cluster-derived risk patterns rather than official pollution-status assessment. Future studies are recommended to incorporate laboratory-based water-quality parameters, regulatory water-quality classifications, and explainable artificial intelligence approaches such as SHAP to improve model transparency and provide deeper interpretation of parameter contributions (P. Del Castillo & others, 2024; Yuan et al., 2025). Additional validation using datasets from other river systems may also strengthen the generalizability of the proposed framework.

CONCLUSION

This study demonstrated that K-Means cluster labels can be effectively used as classification targets for identifying water quality risk patterns in unlabeled river-monitoring datasets. The results showed that both Naive Bayes and Random Forest were able to consistently recognize the cluster-derived patterns, indicating that the generated risk categories possessed meaningful internal structure. In addition, Random Forest feature importance analysis identified total dissolved solids and electrical conductivity as the dominant parameters contributing to water quality risk differentiation. The novelty of this study lies in the integration of cluster-derived labels, supervised classification, and feature-importance analysis within a data-leakage-aware framework for water quality assessment. Practically, the findings provide a basis for preliminary river-water monitoring by highlighting key parameters that can support early risk identification. However, the results should be interpreted as cluster-based risk pattern classification rather than official pollution-status prediction. Future research is recommended to incorporate laboratory-based water-quality indicators, regulatory classifications, and explainable artificial intelligence techniques to further improve model interpretability and external validation.

ACKNOWLEDGMENTS

The authors would like to thank all parties who supported data provision and the preparation of this research.

AUTHOR CONTRIBUTIONS

Conceptualization, W.C.W., T.S., and R.U.; methodology, W.C.W.; software, W.C.W.; validation, T.S. and R.U.; formal analysis, W.C.W.; investigation, W.C.W.; data curation, W.C.W.; writing - original draft preparation, W.C.W.; writing - review and editing, W.C.W., T.S., and R.U.; visualization, W.C.W.; supervision, T.S. and R.U.; project administration, W.C.W.

CONFLICTS OF INTEREST

The author(s) declare no conflict of interest.

USE OF ARTIFICIAL INTELLIGENCE (AI)-ASSISTED TECHNOLOGY

The authors declare that no artificial intelligence (AI) tools were used in the generation, analysis, or writing of this manuscript. All aspects of the research, including data collection, interpretation, and manuscript preparation, were carried out entirely by the authors without the assistance of AI-based technologies.

REFERENCES

Abadi, A. M., & others. (2025). Determining river water quality in the special region of yogyakarta using k-means and fuzzy c-means algorithms. *TEM Journal*.

Al-Adhaileh, M. H., & Alsaade, F. W. (2021). Modelling and prediction of water quality by using artificial intelligence. *Sustainability*, *13*(8), 4259. https://doi.org/10.3390/su13084259.

Aldrees, A., & others. (2024). Evaluation of water quality indexes with novel machine learning and SHAP-based interpretation. *Environmental Research*, *238*, 118269. https://doi.org/10.1016/j.envres.2024.118269.

Aldrees, R., & others. (2024). Explainable AI for environmental risk modeling. *Environmental Informatics*.

Choudhary, R., Kumar, A., C., P., Naik, M. M., Choudhury, M., & Khan, N. A. (2025). Predicting water quality index using stacked ensemble regression and shap based explainable artificial intelligence. *Scientific Reports*, *15*, 31139.

del Castillo, A., Yebra-Montes, C., Verduzco Garibay, M., de Anda, J., Garcia-Gonzalez, A., & Gradilla-Hernández, M. S. (2022). Simple prediction of an ecosystem-specific water quality index and the water quality classification of a highly polluted river through supervised machine learning. *Water*, *14*(8), 1235. https://doi.org/10.3390/w14081235.

Dritsas, E., & others. (2023). Efficient data-driven machine learning models for water quality prediction. *Computers*, *12*(2), 16. https://doi.org/10.3390/computers12020016.

Ezzat, D., Soliman, M., Ahmed, E., & Hassanien, A. E. (2023). An optimized explainable artificial intelligence approach for sustainable clean water. *Environment, Development and Sustainability*, *26*(10), 25899–25919. https://doi.org/10.1007/s10668-023-03712-0.

Firdiani, F., Mandala, S., Adiwijaya, & Abdullah, A. H. (2024). WaQuPs: a ros-integrated ensemble learning model for precise water quality prediction. *Applied Sciences*, *14*(1), 262. https://doi.org/10.3390/app14010262.

Gao, M., Yun, Q., Zhang, Q., Liu, Y., & Zhang, Z. (2025). Surface water quality classification and prediction model based on multiple machine learning algorithms. *Scientific Reports*, *15*, 39907. https://doi.org/10.1038/s41598-025-23775-5.

Ghobadi, F., & Kang, D. (2023). Application of machine learning in water resources management: a systematic literature review. *Water*, *15*(4), 620. https://doi.org/10.3390/w15040620.

Hassan, M. M., Hassan, M. M., Akter, L., Rahman, M. M., Zaman, S., Hasib, K. M., Jahan, N., & others. (2021). Efficient prediction of water quality index (WQI) using machine learning algorithms. *Human-Centric Intelligent Systems*, *1*, 86–97.

Hernanda, T., Nugroho, S. S. P., Izzati, T. I., Nisa, F. K., Wahyudin, W. C., & Nuriyatman, E. (2025). Iot-based legal policy in co₂ emission safety control to support green transportation. *Jurnal Ilmiah Ilmu Terapan Universitas Jambi*, *9*(4), 1434–1443. https://doi.org/10.22437/jiituj.v9i4.48755.

Hridoy, M. A. A. M., & others. (2025). Advanced machine learning models for accurate water quality

classification and wqi prediction: implications for aquatic disease risk management. *Science of the Total Environment*, *1008*, 180965. https://doi.org/10.1016/j.scitotenv.2025.180965.

Hridoy, Md. A. A. M., Shawkat, A. I., Bordin, C., Acharjee, M. R., Masood, A., Baki, A. O., & Al Mamun, Md. A. (2025). Advanced machine learning models for accurate water quality classification and wqi prediction: implications for aquatic disease risk management. *Science of The Total Environment*, *1008*, 180965. https://doi.org/10.1016/j.scitotenv.2025.180965.

Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, *4*(9), 100804. https://doi.org/10.1016/j.patter.2023.100804.

Khoi, D. N., Quan, P. T., Linh, D. T., Nhi, N. T. T., & Thuy, N. T. (2022). Using machine learning models for predicting the water quality index in the la buong river, Vietnam. *Water*, *14*(10), 1552. https://doi.org/10.3390/w14101552.

Kristiyowati, L. A. S., Hana, F. M., & Wahyudin, W. C. (2025). Penerapan algoritma naive bayes pada sistem chatbot persewaan kos [Application of the naive Bayes algorithm to a boarding house rental chatbot system]. *Sainteks*, *22*(1), 53–62. https://doi.org/10.30595/sainteks.v22i1.25962.

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.

Madni, H. A., Umer, M., & others. (2023). Water-quality prediction based on h2o automl and explainable ai techniques. *Water*, *15*(3), 475.

Makumbura, R. K., Mampitiya, L. I., Rathnayake, N., & Meddage, D. P. P. (2024). Advancing water quality assessment and prediction using machine learning models, coupled with explainable artificial intelligence techniques like shap for interpreting the black-box nature. *Results in Engineering*, *23*, 102831. https://doi.org/10.1016/j.rineng.2024.102831.

Mallick, J., Pandey, S., Alsubih, M., & Singh, R. (2022). Integrated remote sensing and machine learning approach for monitoring and predicting river water quality. *Environmental Monitoring and Assessment*, *194*(10), 1–20. https://doi.org/10.1007/s10661-022-10045-0.

Man, G., Yun, Q., Qilin, Z., Yuyong, L., & Zhuoshi, Z. (2025). Surface water quality classification and prediction model based on multiple machine learning algorithms. *Scientific Reports*, *15*(1). https://doi.org/10.1038/s41598-025-23775-5.

Masood, A., & others. (2023). a machine learning-based framework for water quality index estimation in the southern bug river. *Water*, *15*(20), 3543. https://doi.org/10.3390/w15203543.

Muhammad, M., Sutikno, T., & Riadi, I. (2025). A Comparative study of k-means and knn imputation for handling missing data in scholarship applicant datasets. *JUITA: Jurnal Informatika*, *13*(3). https://doi.org/10.30595/juita.v13i3.26502.

Nallakaruppan, M. K., Gangadevi, E., Shri, M. L., Balusamy, B., Bhattacharya, S., & Selvarajan, S. (2024). Reliable water quality prediction and parametric analysis using explainable ai models. *Scientific Reports*, *14*, 7520. https://doi.org/10.1038/s41598-024-56775-y.

Novianta, M. A., Syafrudin, S., & Warsito, B. (2023). K-Means clustering for grouping rivers in diy based on water quality parameters. *JUITA: Jurnal Informatika*, *11*(1). https://doi.org/10.30595/juita.v11i1.16986.

Ramya, S., & others. (2024). Comprehensive analysis of multiple classifiers for assessment of water quality index. *Water-Energy Nexus*.

Ridwan, A., Sutikno, T., Riyadi, I., & Wahyudin, W. C. (2025). On-Time student graduation prediction modeling: a comparative analysis of naive bayes algorithm and other data mining classifications. *JOINCS (Journal of Informatics, Network, and Computer Science)*, *8*(2).

Sangwan, V., & Bhardwaj, R. (2024). Machine learning framework for predicting water quality classification. *Water Practice and Technology*, *19*(11), 4499–4521. https://doi.org/10.2166/wpt.2024.259.

Schäfer, K., & others. (2022). Machine learning applications for river water quality prediction. *Environmental Modelling & Software*.

Shamsuddin, I. I. S., & others. (2022). Prediction of water quality classification of the langat river basin, Malaysia, using machine learning. *Water*, *14*(19), 2939. https://doi.org/10.3390/w14192939.

Tran, Y. B., Arias-Rodriguez, L. F., & Huang, J. (2022). Predicting high-frequency nutrient dynamics in the Danube River with surrogate models using sensors and Random Forest. *Frontiers in Water*, *4*, 894548. https://doi.org/10.3389/frwa.2022.894548.

Wahyudin, W. C., Sutikno, T., Umar, R., & Ridwan, A. (2025). Comparison of data mining model

performance in heart disease detection with feature selection application. *JOINCS (Journal of Informatics, Network, and Computer Science)*, *8*(1), 87–93.

Wala, J., Herman, H., & Umar, R. (2024). Implementasi k-means clustering pada pengelompokan pasien penyakit jantung [Implementation of k-means clustering in grouping heart disease patients]. *JISKA (Jurnal Informatika Sunan Kalijaga)*, *9*(3), 205–216. https://doi.org/10.14421/jiska.2024.9.3.205-216.

Wu, J., & Wang, Z. (2022). A hybrid model for water quality prediction based on an artificial neural network, wavelet transform, and long short-term memory. *Water*, *14*(4), 610. https://doi.org/10.3390/w14040610.

Xu, J., Xu, Z., Kuang, J., Lin, C., Xiao, L., Huang, X., & Zhang, Y. (2021). An Alternative to Laboratory Testing: Random Forest-Based Water Quality Prediction Framework for Inland and Nearshore Water Bodies. *Water*, *13*(22), 3262. https://doi.org/10.3390/w13223262.

Yan, X., Zhang, T., Du, W., Meng, Q., Xu, X., & Zhao, X. (2024). A comprehensive review of machine learning for water quality prediction over the past five years. *Journal of Marine Science and Engineering*, *12*(1), 159. https://doi.org/10.3390/jmse12010159.

Yuan, P., Li, H., Yi, X., Wang, J., Ning, C., Xu, X., & Nong, X. (2025). Optimizing water quality index using machine learning: A six-year comparative study in riverine and reservoir systems. *Scientific Reports*, *15*, 33919. https://doi.org/10.1038/s41598-025-10187-8.