

# Comparative Analysis of Machine Learning Algorithms for Groundwater Potability Classification in Jakarta

Diky Arianto Tarihoran<sup>1\*</sup>, Hadi Santoso<sup>2</sup>

<sup>1,2</sup>Faculty of Computer Science, Study Program of Informatics, Universitas Mercu Buana, Indonesia

\*corr-author: hadi.santoso@mercubuana.ac.id

**Abstract** - Groundwater quality is a fundamental aspect of fulfilling clean water needs, particularly in urban areas such as Jakarta, which faces significant supply limitations due to severe contamination from domestic waste, chemical pollutants, industrial activities, and septic tank leakage. This study aims to compare the performance of nine machine learning algorithms in developing a classification model for groundwater feasibility based on physical parameters. Real-time data were collected from three administrative regions in Jakarta using Internet of Things (IoT) sensors, which monitored pH, temperature, total dissolved solids (TDS), and turbidity. Model evaluation involved hyperparameter tuning, cross-validation, feature importance analysis, LIME interpretation, and performance metrics including AUC, accuracy, precision, recall, and F1-score. The results indicate that CatBoost achieved the highest overall performance (AUC: 0.9448, accuracy: 0.9318, F1-score: 0.9209). LightGBM demonstrated competitive results with an F1-score of 0.9211 and AUC of 0.9431, while XGBoost recorded the highest recall at 0.9359. Random Forest and AdaBoost also exhibited consistent performance, with precision of 0.9094 and recall of 0.9327, respectively. In contrast, Support Vector Machine (SVM) yielded the lowest performance (AUC: 0.8860, accuracy: 0.8499). Based on a comprehensive evaluation, CatBoost model is recommended as the most suitable model for IoT-based groundwater quality classification systems.

**Keywords:** Groundwater quality; machine learning; Internet of Things (IoT); hyperparameter tuning

## I. INTRODUCTION

Groundwater serves as the primary source of clean water, particularly in urban areas experiencing increasing pressure due to rapid population growth and environmental degradation [1]. The sustainability of groundwater resources is critically dependent on their quality, especially when intended for consumption, irrigation, and industrial purposes [2]. In Jakarta, a metropolitan city with an estimated population of 11

million as of 2020, the demand for clean water has reached approximately 1.2 billion cubic meters annually. However, the piped water distribution system meets less than half of this demand [3]. Consequently, more than 80% of households depend on groundwater for daily needs such as washing and bathing, highlighting the population's reliance on groundwater quality [4]. Groundwater pollution in Jakarta stems from four primary sources: agricultural runoff, industrial effluents, domestic wastewater, and excessive groundwater extraction, all of which pose significant public health risks [5]. Moreover, all surface water bodies in Jakarta have been reported to be heavily polluted by domestic sewage, industrial waste, agricultural chemicals, solid waste, and seepage from septic tanks [3]. Water quality reflects the environmental disturbances in surrounding ecosystems and is typically evaluated using a range of physicochemical parameters [6].

Various conventional methods have been applied to assess water quality, including manual sampling across multiple sites, laboratory-based analysis, water quality index (WQI) computation, and experimental assessment of critical parameters [1]. However, these traditional techniques are often time-intensive [1,7-8], and involve substantial laboratory costs [6, 8]. In addition, groundwater quality may also be influenced by the type of extraction or pumping technology employed [9]. Given that water serves as a primary medium for disease transmission [5], real-time water quality (WQ) monitoring is essential to mitigate public health risks and to support environmental protection initiatives [10]. To address these challenges, numerous analytical frameworks have been proposed, including multivariate statistical techniques [11-13], explanatory analysis [14], theoretical probability models [15-16], Principal Components Analysis (PCA) [17-18], Hierarchical Cluster Analysis (HCA) [19], and correlation and linear regression analysis [20], with the aim of enhancing the accuracy and efficiency of groundwater potability assessments.

Machine learning (ML), a subfield of artificial intelligence, provides effective solutions to complex problems by leveraging data-driven algorithms that emulate human learning processes to improve system performance and decision-making accuracy [21]. ML techniques have been extensively applied across various domains to automate tasks and extract meaningful insights from large datasets. In the domain of water quality monitoring and assessment, the integration of ML has shown notable advancements in recent years [22]. Several prior studies have investigated the effectiveness of different ML algorithms in classifying groundwater quality, aiming to identify the most suitable models for deployment. For instance, Ref. [23] conducted a comparative analysis of ten classifiers including Extra Trees, Random Forest, XGBoost, and others using a dataset of 2,011 samples with ten physicochemical features. Their methodology incorporated feature selection (UVS, RFE, and feature importance), class imbalance handling (SMOTE), and hyperparameter optimization (Grid Search), with Extra Trees achieving the highest accuracy (89%) and excelling in precision, F1-score, and computational efficiency. In a related study, Ref. [6] compared five classifiers Decision Tree (DT), K-Nearest Neighbors (KNN), Multilayer Perceptron (MLP), Support Vector Machine (SVM), and Random Forest (RF) for water quality classification using the Water Quality Index (WQI) derived from nine parameters, including heavy metals. Despite achieving a maximum accuracy of 88% using RF and SVM, the study was limited by the use of non-real-time, conventionally collected data and small sample sizes (79 records).

In contrast, Ref. [1] demonstrated the effectiveness of Internet of Things (IoT) technology in remote water quality monitoring, employing sensors connected via Low-Power Wide Area Network (LPWAN). This system achieved up to 95% measurement accuracy for physical

parameters such as pH, turbidity, TDS, and temperature, surpassing traditional methods which reported accuracy rates of approximately 85%. Building on these findings, the present study utilizes an IoT-based approach for real-time data acquisition, serving as the foundation for training and evaluating ML classification models. This research implements and systematically compares nine supervised learning algorithms Decision Tree, Random Forest, SVM, Extra Trees, Gradient Boosting, LightGBM, AdaBoost, XGBoost, and CatBoost which have been widely applied in environmental and water quality prediction tasks [22, 24]. The novelty of this study lies in its comprehensive benchmarking of ML models using IoT-derived groundwater data in the context of Jakarta, a region for which such applications remain underexplored. In addition to conventional evaluation metrics such as accuracy and AUC-ROC, the study places significant emphasis on model interpretability an often-cited barrier to the practical deployment of ML models [25], the lack of transparency is a key challenge in the deployment of ML models. To address this issue, an Explainable Artificial Intelligence (XAI) approach is employed to understand the contribution of each feature to the model's prediction outcomes. Specifically, the study uses the Local Interpretable Model-Agnostic Explanations (LIME) method to evaluate the significance of individual features in the model's decision-making process, thereby enhancing transparency [26] and reinforcing the model's validity for real-world applications in water quality management [25].

## II. METHOD

This study focuses on the comparative analysis of machine learning algorithms for detecting groundwater potability based on physical parameters. The complete stages of the research process are illustrated in Fig. 1.

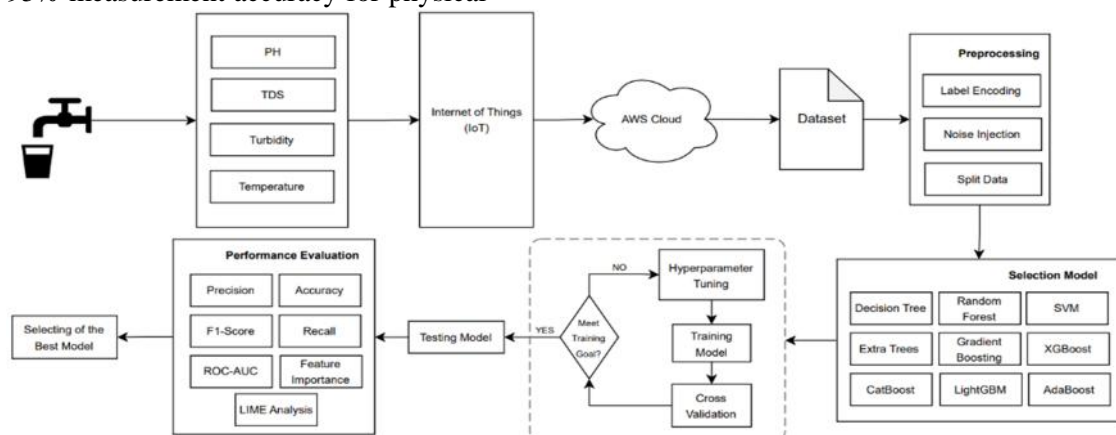


Fig. 1 Research workflow

A. Data Collection

This study utilized primary data obtained directly from groundwater samples collected in three administrative regions of Jakarta Province, namely West Jakarta, East Jakarta, and North Jakarta. The sampling process was conducted using drilling machines or water pumps to ensure that the water originated from underground sources. The collected water was then stored in a dedicated container for further measurement. Subsequently, four IoT-based sensors, namely a pH sensor, temperature sensor, total dissolved solids (TDS) sensor, and turbidity sensor, were placed into the water container to measure the physical parameters of groundwater.

The sensors were connected to an ESP32 microcontroller, which was responsible for data acquisition, preliminary processing, and transmission of measurement results to the AWS IoT Core service via the MQTT protocol. The transmitted data were then directed to AWS DynamoDB, where they were stored in real time in a structured tabular format. These data could subsequently be accessed through an API endpoint developed using Flask, allowing the retrieval of both raw and processed measurements. Finally, the data were exported in CSV format and compiled into a groundwater quality dataset for Jakarta Province.

Data record was generated every 6 seconds, and the data collection duration for each water container ranged between 1 and 2 hours, ensuring sufficient granularity to capture short-term fluctuations in water quality parameters. All collected data were assigned binary labels of “Potable” or “Non-Potable” according to the threshold values established by the Indonesian Ministry of Health Regulation No. 492/Menkes/Per/IV/2010 [27], as presented in Table I. These labels served as the target variable for the classification process in the machine learning model development stage.

B. Data Preprocessing

After the dataset was collected, a preprocessing phase was conducted to ensure data quality and readiness prior to model development. This stage involved three main procedures. First, a noise injection equivalent to 5% of the total dataset was introduced to simulate realistic disturbances commonly arising from sensor inaccuracies or measurement fluctuations. Although noise injection may reduce overall accuracy, prior studies have demonstrated that controlled noise can enhance model generalization and reduce the likelihood of overfitting [28]. Second, the categorical target labels were converted into numerical representations via label encoding, where

“Potable” was assigned a value of 1 and “Non-Potable” a value of 0. Subsequently, the dataset was divided into training and testing subsets using a 70:30 hold-out split, facilitating an unbiased assessment of model performance while reducing the potential for data leakage.

C. Model Development

At this stage, a comparative analysis was conducted on nine machine learning algorithms namely Decision Tree, Gradient Boosting, XGBoost, Random Forest, Extra Trees, Support Vector Machine (SVM), LightGBM, AdaBoost, and CatBoost to determine the most optimal classification model for detecting groundwater potability based on physical parameters. The selection of these algorithms was based on the diversity of modeling approaches, ranging from linear and tree-based models to ensemble learning techniques, thereby representing the performance of various model architectures commonly used in environmental and water quality classification tasks.

D. Hyperparameter Tuning

Hyperparameter tuning is a fundamental process in machine learning that systematically adjusts model parameters to optimize predictive performance, thereby enhancing accuracy while effectively mitigating the risks of both overfitting and underfitting [29,30]. In this study, hyperparameter optimization was performed individually for each classification algorithm to ensure model-specific tuning and performance consistency.

The tuning process utilized the Randomized Search Cross-Validation (RandomizedSearchCV) method, which is recognized for its computational efficiency compared to exhaustive approaches such as Grid Search particularly in scenarios involving high-dimensional parameter spaces and complex model structures [31]. By randomly sampling from a predefined parameter distribution over a fixed number of iterations, this method effectively balances exploration and efficiency in identifying optimal configurations.

TABLE I  
GROUNDWATER PHYSICAL PARAMETER STANDARDS

Parameter	Permenkes	WHO
Temperature	Air temperature ± 3° C	No temperature limit specified
Turbidity	5 NTU	5 NTU
TDS	500 mg/L	500 mg/L
PH	6.5 – 8.5	6.5-8.5

### E. Model Evaluation

The evaluation stage was conducted to determine the effectiveness of the classification models in predicting groundwater potability using physical parameters collected from IoT devices. Assessments were performed on a reserved test set to ensure the model's ability to generalize to previously unseen data. Performance was quantified through standard metrics such as accuracy, precision, recall, F1-score, and ROC-

## III. RESULT AND DISCUSSION

### A. Data Collecting

Data collection was conducted in two stages to comprehensively represent groundwater quality within the study area. The first stage involved direct measurements from household groundwater sources (pre-filtration) using Internet of Things (IoT) sensors that continuously recorded four physical parameters in real time: pH, total dissolved solids (TDS), turbidity, and temperature. The second stage involved measurements taken after the water passed through a basic filtration unit that mimicked typical household water treatment systems (post-filtration). This dual-stage approach was designed to assess the impact of filtration on physical water quality parameters. Data from both stages were combined to produce a comprehensive dataset that captures both real-world conditions and simulated intervention outcomes, thereby enhancing the variability and richness of data patterns for machine learning model training.

The final dataset comprised 2,442 samples collected from three administrative regions of Jakarta: West Jakarta (858 samples), East Jakarta (833 samples), and North Jakarta (750 samples). West Jakarta exhibited a relatively balanced distribution of potable and non-potable samples, East Jakarta was dominated by potable samples, whereas all samples from North Jakarta were categorized as non-potable. This regional disparity highlights significant spatial variation in groundwater quality, which strengthens the dataset's representativeness and supports the development of more robust and generalizable classification models.

### B. Data Preprocessing

The preprocessing stage in this study was designed to ensure data quality and compatibility with machine learning algorithms, while also enhancing model robustness and generalizability. To simulate real-world fluctuations in IoT sensor readings and improve model resilience against noise, 5% artificial noise was injected

AUC. To further enhance transparency and interpretability, feature importance analysis and the Local Interpretable Model-Agnostic Explanations (LIME) technique were applied, supporting reliable application in practical environmental monitoring. This integrated evaluation approach guarantees that model outputs are both statistically rigorous and interpretable, providing actionable insights for decision-making in water quality management.

into the physical parameter values. The binary target labels were numerically encoded, with "Potable" assigned a value of 1 and "Non-Potable" a value of 0, enabling efficient classification and metric computation. The dataset was partitioned into training and testing subsets with a 70:30 ratio using a randomized hold-out strategy to minimize the risk of data leakage and facilitate unbiased model evaluation. This preprocessing procedure was crucial for ensuring stable, accurate, and interpretable classification outcomes.

### C. Hyperparameters Tuning

Hyperparameter tuning is an essential stage in the development of machine learning models, as carefully selecting optimal parameter settings can markedly enhance predictive accuracy, strengthen model stability, and improve generalization capabilities. In this study, hyperparameter optimization was independently performed for each classification algorithm using RandomizedSearchCV with 10 iterations and 3-fold cross-validation. This method was selected due to its computational efficiency in exploring high-dimensional parameter spaces without compromising the quality of model tuning.

The search space was customized according to the specific architecture of each algorithm. For instance, the Decision Tree model was tuned for `max_depth`, `min_samples_split`, and `criterion`, while Random Forest and Extra Trees were optimized for `n_estimators` and `bootstrap`. For the Support Vector Machine (SVM), tuning focused on `C`, `kernel`, and `gamma`. Ensemble-based models such as Gradient Boosting, AdaBoost, XGBoost, LightGBM, and CatBoost were tuned across multiple hyperparameters including `learning_rate`, `max_depth`, `n_estimators`, as well as regularization components such as `subsample`, `colsample_bytree`, and `l2_leaf_reg`. The optimal hyperparameter combinations and their corresponding computation times are presented in Table II. These configurations were subsequently adopted in the final training phase to ensure that the performance evaluation accurately reflects each model's full predictive capability under its best-tuned condition.

TABLE II  
COMPARISON OF HYPERPARAMETER TUNING

No	Method	Hyperparameters	Best Parameters	Training Time (s)
1	Decision Tree	max_depth: [3, 5, 7, 10, None], min_samples_split: [2, 5, 10], min_samples_leaf: [1, 2, 4], criterion: ['gini', 'entropy']	max_depth: 3, min_samples_split: 2, min_samples_leaf: 1, criterion: 'gini'	5.95
2	Random Forest	n_estimators: [50, 100, 200], max_depth: [None, 5, 10, 20], min_samples_split: [2, 5, 10], min_samples_leaf: [1, 2, 4], bootstrap: [True, False]	n_estimators: 50, max_depth: None, min_samples_split: 10, min_samples_leaf: 2, bootstrap: True	3.03
3	Extra Trees	n_estimators: [50, 100, 200], max_depth: [None, 5, 10], min_samples_split: [2, 5], min_samples_leaf: [1, 2]	n_estimators: 200, max_depth: 10, min_samples_split: 5, min_samples_leaf: 2,	2.96
4	SVM	C: [0.1, 10], kernel: ['rbf', 'sigmoid'], gamma: ['scale', 'auto'] + [0.1, 1]	C: 10, kernel: 'rbf', gamma: 1	8.39
5	Gradient Boosting	n_estimators: [50, 100, 200], learning_rate: [0.01, 0.1, 0.2], max_depth: [3, 5, 7], min_samples_split: [2, 5]	n_estimators: 200, learning_rate: 0.01, max_depth: 3, min_samples_split: 5	8.58
6	AdaBoost	n_estimators: [50, 100, 200], learning_rate: [0.01, 0.1, 1.0], algorithm: ['SAMME', 'SAMME.R']	n_estimators: 200, learning_rate: 1.0, algorithm: 'SAMME'	4.45
7	XGBoost	n_estimators: [50, 100, 200], max_depth: [3, 5, 7], learning_rate: [0.01, 0.1, 0.2], subsample: [0.6, 0.8, 1.0], colsample_bytree: [0.6, 0.8, 1.0]	n_estimators: 200, max_depth: 7, learning_rate: 0.01, subsample: 0.6, colsample_bytree: 0.6	1.36
8	LightGBM	n_estimators: [50, 100, 200], max_depth: [3, 5, 7, -1], learning_rate: [0.01, 0.1, 0.2], num_leaves: [31, 63, 127], subsample: [0.6, 0.8, 1.0]	n_estimators: 50, max_depth: 3, learning_rate: 0.1, num_leaves: 31, subsample: 0.6	3.59
9	CatBoost	iterations: [50, 100, 200], depth: [3, 5, 7], learning_rate: [0.01, 0.1, 0.2], l2_leaf_reg: [1, 3, 5]	iterations: 50, depth: 5, learning_rate: 0.1, l2_leaf_reg: 1	3.79

*D. Model Evaluation*

Model evaluation was performed on nine different machine learning techniques to classify the drinkability of groundwater using four key physical indicators: pH, total dissolved solids (TDS), turbidity, and temperature. The methods tested comprised Decision Tree, Extra Trees, Support Vector Machine (SVM), Gradient

Boosting, Random Forest, XGBoost, LightGBM, AdaBoost, and CatBoost. Their effectiveness was benchmarked using multiple criteria, including accuracy, precision, recall, F1-score, AUC values, as well as assessments of training efficiency and overfitting risk. To enhance interpretability, results were further illustrated through confusion matrices and ROC curve visualizations.

As shown in the confusion matrix results (Fig. 2), all models exhibited generally high performance in distinguishing between "potable" and "non-potable" groundwater samples. Among them, CatBoost, LightGBM, AdaBoost, Extra Trees, and Random Forest achieved the highest identical test accuracy of 0.9318. This was closely followed by XGBoost (0.9304), Gradient Boosting (0.9291), and Decision Tree (0.9263). Conversely, the Support Vector Machine (SVM) model recorded the lowest accuracy of 0.8499, indicating limited effectiveness in modeling the complex and non-linear relationships present in the groundwater dataset. These findings underscore the critical importance of selecting models that align with the underlying data structure. Ensemble-based models, particularly CatBoost and LightGBM, demonstrated superior capability in capturing nonlinear interactions and delivering robust classification outcomes under diverse groundwater quality conditions.

The ROC curve in Fig. 3 provides a more detailed illustration of the models' discriminative capabilities. The ROC curve reflects how well a model can distinguish between two classes. The CatBoost model achieved the highest AUC score of 0.9448, indicating its superior ability to differentiate between "Potable" and "Non-Potable" water classes. This was followed by LightGBM (AUC = 0.9431) and AdaBoost (AUC = 0.9378). In contrast, Support Vector Machine (SVM) recorded the lowest AUC score of 0.8806, reflecting its limited effectiveness in discriminating between the two classes. As detailed in Table III, the CatBoost model not

only led in AUC but also recorded high precision (0.904) and recall (0.9327), yielding a robust F1-score of 0.9209. Additionally, the model maintained a high true positive rate (TPR = 0.9327) while minimizing the false positive rate (FPR = 0.0689), suggesting reliable identification of potable water samples with minimal misclassification of non-potable instances. These results affirm CatBoost's strength in balancing sensitivity and specificity, making it highly suitable for practical groundwater potability classification tasks.

One key aspect examined is the overfit gap defined as the accuracy difference between training and testing sets. CatBoost exhibited the lowest overfit gap (0.0214), suggesting minimal overfitting and strong generalization. With the highest AUC (0.9448), test accuracy (0.9318), and the fastest training time (3.8 s), CatBoost emerged as the top-performing model. Although XGBoost and LightGBM offered competitive results, CatBoost demonstrated superior classification stability and sharper probabilistic predictions. The CatBoost model's ability to generate realistic and interpretable probability estimates provides strong support for data-driven decision-making.

In addition to strong metrics, CatBoost model effectively handled non-linear data, missing values, and sensor noise, making it robust for real-world conditions. Its ability to produce interpretable probability estimates also enhances transparency in data-driven decision making. These advantages position CatBoost model as the most reliable model for integration into IoT-based groundwater quality monitoring systems.

TABLE III  
COMPARISON OF MODEL EVALUATION METRIX RESULTS

No	Model	AUC	Accuracy	Precision	Recall	F1-Score	FPR	TNR	Overfit Gap
1	Decision Tree	0.9278	0.9263	0.9031	0.9263	0.9146	0.0736	0.9264	0.0286
2	Random Forest	0.9361	<b>0.9318</b>	<b>0.9094</b>	0.9327	0.9209	<b>0.0689</b>	<b>0.9311</b>	0.0249
3	Extra Trees	0.9316	<b>0.9318</b>	0.9068	<b>0.9359</b>	<b>0.9211</b>	0.0713	0.9287	0.0225
4	SVM	0.8860	0.8499	0.8117	0.8429	0.8270	0.1449	0.8551	0.1466
5	Gradient Boosting	0.9286	0.9291	0.9088	0.9263	0.9175	<b>0.0689</b>	<b>0.9311</b>	0.0253
6	AdaBoost	0.9378	<b>0.9318</b>	<b>0.9094</b>	0.9327	0.9209	<b>0.0689</b>	<b>0.9311</b>	0.0225
7	XGBoost	0.9357	0.9304	0.9040	<b>0.9359</b>	0.9197	0.0736	0.9264	0.0251
8	LightGBM	0.9431	<b>0.9318</b>	0.9068	<b>0.9359</b>	<b>0.9211</b>	0.0713	0.9287	0.0220
9	CatBoost	<b>0.9448</b>	<b>0.9318</b>	<b>0.9094</b>	0.9327	0.9209	<b>0.0689</b>	<b>0.9311</b>	<b>0.0214</b>

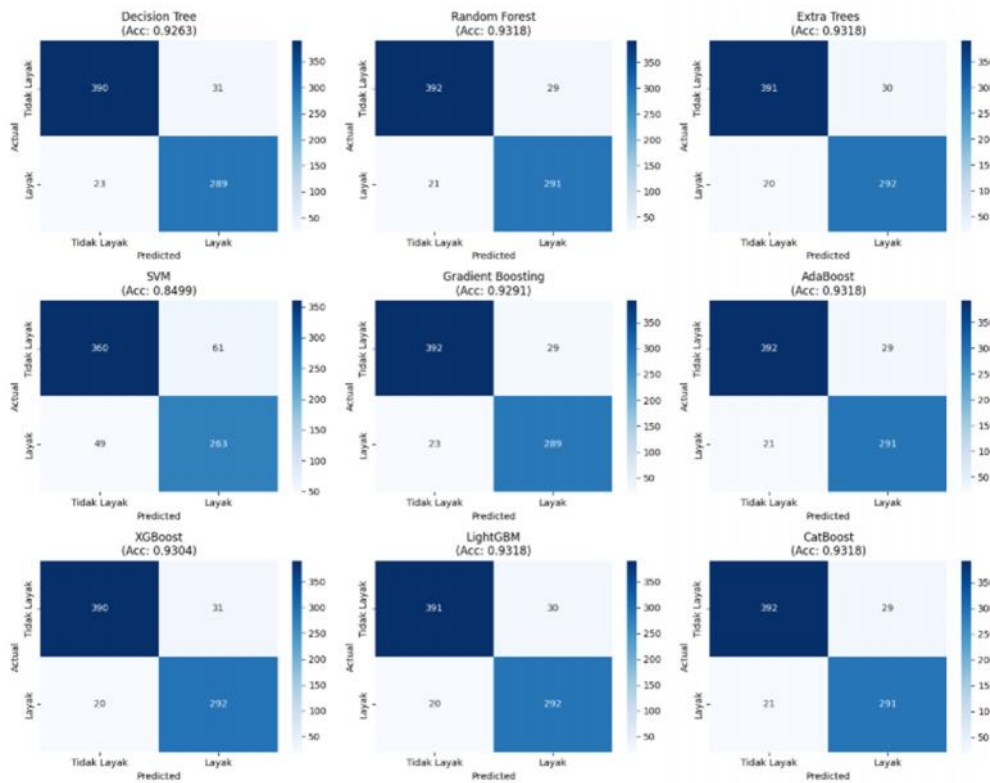


Fig. 2 Comparison of confusion matrix

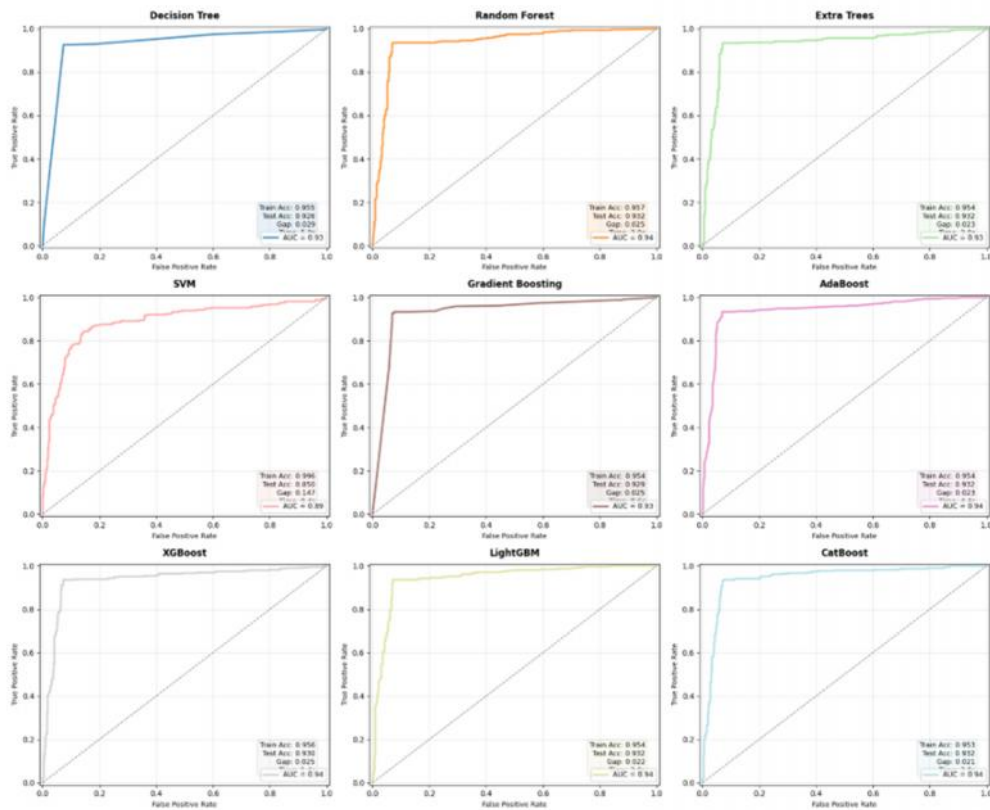


Fig. 3 Comparison of ROC-AUC curves

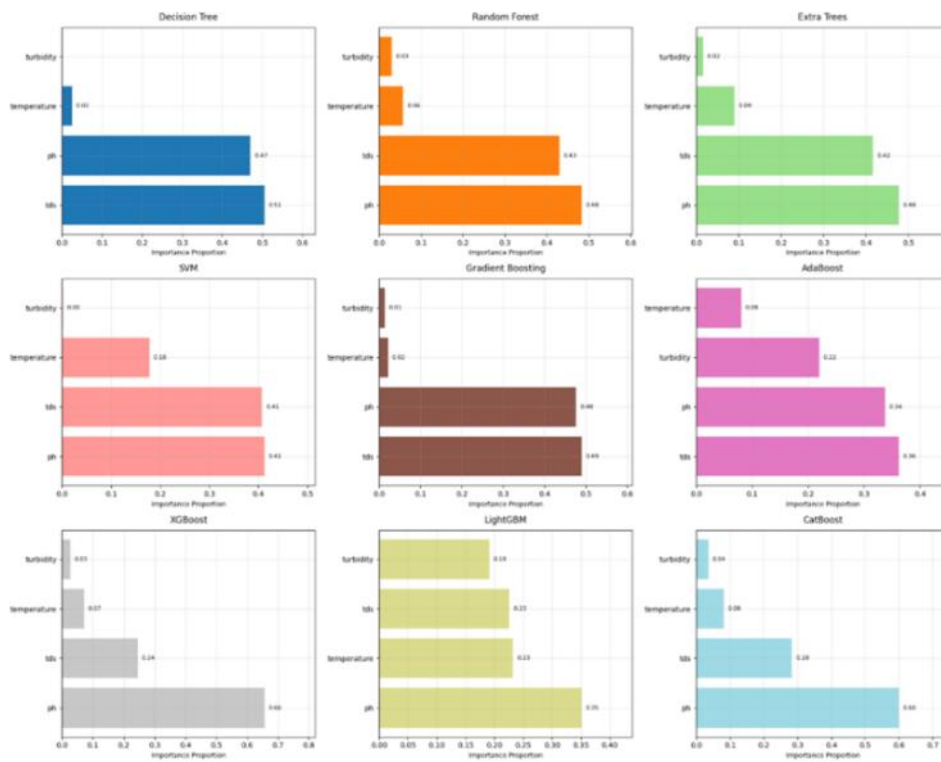


Fig. 4 Comparison of feature importance across models



Fig. 5 LIME-based model comparison

As shown in Fig. 4, all models consistently identified pH and TDS as the most influential parameters in predicting drinking water potability. In the CatBoost model, pH had the highest feature importance (0.60), followed by TDS (0.28), with turbidity and temperature contributing less. Similar patterns were observed in LightGBM, Gradient Boosting, and Random Forest, where pH remained the dominant factor (importance > 0.3). The LIME analysis in Fig. 5 further illustrated this by explaining the prediction for a sample instance (Instance 229), highlighting that TDS < 300 and pH between 7.4 - 8.3 were key indicators of potable water, while temperatures above 30 °C negatively affected classification outcomes. The SVM model demonstrated a distinctive pattern in feature contributions, where temperature and pH exerted significant negative effects, and only turbidity showed a positive influence. This result suggests SVM's limited capacity to capture non-linear relationships and its tendency toward overfitting within this dataset. Conversely, the CatBoost model showed balanced and interpretable behavior. LIME analysis confirmed that all four features contributed proportionally and aligned logically with prediction directions pH and TDS positively supported the classification of potable water, while temperature and turbidity contributed negatively.

These consistent patterns underscore CatBoost's strengths in both predictive accuracy and interpretability. By integrating feature importance and LIME-based analysis, the evaluation extended beyond standard metrics (e.g., accuracy, AUC), emphasizing decision transparency an essential requirement for drinking water quality assessment systems. These findings reaffirm the dominant influence of pH and TDS, while validating CatBoost as the most reliable model in terms of both performance and explainability.

#### IV. CONCLUSION

This study conducted a comprehensive comparative analysis of nine machine learning algorithms for classifying groundwater potability using real-time data from IoT-based sensors that measured four key physical parameters: pH, total dissolved solids (TDS), turbidity, and temperature across three administrative regions in Jakarta. The preprocessing workflow incorporated controlled noise injection, label encoding, and stratified train-test splitting, followed by hyperparameter optimization using RandomizedSearchCV with 3-fold cross-validation. Among the evaluated models, CatBoost consistently demonstrated superior performance, achieving the highest test accuracy (93.18%), the highest

AUC score (0.9448), and the lowest overfit gap, indicating strong generalization capabilities. Interpretability analysis using feature importance and LIME further highlighted the dominance of pH and TDS in influencing potability classification and confirmed that CatBoost model produced transparent and domain-consistent predictions. These results affirm CatBoost model as the most suitable model for real-time groundwater quality classification within IoT-based systems. The integration of predictive performance, robustness, and interpretability supports its application in environmental monitoring and public health, particularly in developing explainable and trustworthy decision-support tools.

#### REFERENCES

- [1] P. Jayaraman, K. K. Nagarajan, P. Partheeban, and V. Krishnamurthy, "Critical review on water quality analysis using IoT and machine learning models," *Int. J. Inf. Manag. Data Insights*, vol. 4, no. 1, p. 100210, 2024, doi: 10.1016/j.jjime.2023.100210.
- [2] R. Haggerty, J. Sun, H. Yu, and Y. Li, "Application of machine learning in groundwater quality modeling - A comprehensive review," *Water Res.*, vol. 233, no. February, p. 119745, 2023, doi: 10.1016/j.watres.2023.119745.
- [3] N. Ardhanie, D. Daniel, P. Purwanto, and K. Kismartini, "Jakarta water supply provision strategy based on supply and demand analysis," *H2Open J.*, vol. 5, no. 2, pp. 221–233, 2022, doi: 10.2166/h2oj.2022.076.
- [4] D. P. J. D. L. Hidup, "Laporan Akhir Pemantauan Kualitas Air Tanah Provinsi DKI Jakarta tahun 2022," pp. 78–79, 2019.
- [5] Y. I. Memon, S. A. Shah, S. H. Mahesar, S. H. Memon, and M. K. Jatoti, "Statistical analysis and physicochemical characteristics of groundwater quality parameters: a case study," *Int. J. Environ. Anal. Chem.*, vol. 103, no. 10, pp. 2270–2291, 2023, doi: 10.1080/03067319.2021.1890064.
- [6] M. Akhlaq, A. Ellahi, R. Niaz, M. Khan, S. S. Sammen, and M. Scholz, "Comparative Analysis of Machine Learning Algorithms for Water Quality Prediction," *Tellus A Dyn. Meteorol. Oceanogr.*, vol. 76, no. 1, pp. 177–192, 2024, doi: 10.16993/tellusa.4069.
- [7] E. Dritsas and M. Trigka, "Efficient Data-Driven Machine Learning Models for Water Quality Prediction," *Computation*, vol. 11, no. 2, 2023, doi: 10.3390/computation11020016.
- [8] M. Y. Vishnoi and S. V. Taral, "IoT-Enabled Machine Learning for Water Quality Monitoring," *Int. J. Adv. Comp. Theory Engg.*, vol. 14, no. 1, pp. 123–134, 2025, [Online]. Available:

- <https://journals.mriindia.com/index.php/ijacte/article/view/338>
- [9] A. A. Suleiman, A. K. Yousafzai, and M. Zubair, "Comparative Analysis of Machine Learning and Deep Learning Models for Groundwater Potability Classification †," *Eng. Proc.*, vol. 56, no. 1, 2023, doi: 10.3390/ASEC2023-15506.
- [10] S. Kaddoura, "Evaluation of Machine Learning Algorithm on Drinking Water Quality for Better Sustainability," *Sustain.*, vol. 14, no. 18, 2022, doi: 10.3390/su141811478.
- [11] E. O. Thomas, "Evaluation of groundwater quality using multivariate, parametric and non-parametric statistics, and GWQI in Ibadan, Nigeria," *Water Sci.*, vol. 37, no. 1, pp. 117–130, 2023, doi: 10.1080/23570008.2023.2221493.
- [12] B. Hama, A. Alodah, F. Bouaicha, M. F. Bekkouche, A. Barkat, and E. E. Hussein, "Hydrochemical assessment of groundwater using multivariate statistical methods and water quality indices (WQIs)," *Appl. Water Sci.*, vol. 14, no. 2, pp. 1–18, 2024, doi: 10.1007/s13201-023-02084-0.
- [13] N. Q. Pham and G. T. Nguyen, "Evaluating Groundwater Quality Using Multivariate Statistical Analysis and Groundwater Quality Index," *Civ. Eng. J.*, vol. 10, no. 3, pp. 699–713, 2024, doi: 10.28991/CEJ-2024-010-03-03.
- [14] A. Suleiman, A. Ibrahim, and U. Abdullahi, "Statistical Explanatory Assessment of Groundwater Quality in Gwale LGA, Kano State, Northwest Nigeria," *Hydrospatial Anal.*, vol. 4, no. 1, pp. 1–13, 2020, doi: 10.21523/gcj3.2020040101.
- [15] A. A. Suleiman, A. Ibrahim, U. A. Abdullahi, and S. A. Suleiman, "Assessment of probability distributions of groundwater quality data in Gwale area, north-western Nigeria," *Ann. Optim. Theory Pract.*, vol. 3, no. 1, pp. 37–46, 2020, doi: 10.22121/aotp.2020.243381.1039.
- [16] A. Ibrahim, A. A. Suleiman, U. A. Abdullahi, and S. A. Suleiman, "Monitoring Groundwater Quality using Probability Distribution in Gwale, Kano state, Nigeria," *J. Stat. Model. Anal.*, vol. 3, no. 2, pp. 95–108, 2021, doi: 10.22452/josma.vol3no2.6.
- [17] S. Arıman, N. G. Soydan-Oksal, N. Beden, and H. Ahmadzai, "Assessment of Groundwater Quality through Hydrochemistry Using Principal Components Analysis (PCA) and Water Quality Index (WQI) in Kızılırmak Delta, Turkey," *Water*, vol. 16, no. 11, p. 1570, May 2024, doi: 10.3390/w16111570.
- [18] A. J. F. Gadelha, C. O. da Rocha, J. G. V. Neto, and M. A. Gomes, "Multivariate statistical analysis of physicochemical parameters of groundwater quality using PCA and HCA techniques," *Eclat. Quim.*, vol. 48, no. 4, pp. 37–47, 2023, doi: 10.26850/1678-4618eqj.v48.4.2023.p37-47.
- [19] P. Hartanto, R. F. Lubis, B. Y. C. S. S. Syah Alam, Y. A. Sendjaja, I. Ismawan, T. Y. W. M. Iskandarsyah, and H. Hendarmawan, "Multivariate Data Analysis to Assess Groundwater Hydrochemical Characterization in Rawadanau Basin, Banten Indonesia," *Rud. Geol. Naft. Zb.*, vol. 39, no. 1, pp. 141–154, 2024, doi: 10.17794/rgn.2024.1.12.
- [20] A. A. Suleiman, U. A. Abdullahi, A. Suleiman, S. A. Suleiman, and H. U. Abubakar, "Correlation and Regression Model for Physicochemical Quality of Groundwater in the Jaen District of Kano State, Nigeria," *J. Stat. Model. Anal.*, vol. 4, no. 1, pp. 14–24, 2022, doi: 10.22452/josma.vol4no1.2.
- [21] N. Muniroh and E. Agus Priatno, "PENERAPAN ALGORITMA K-NN PADA MACHINE LEARNING UNTUK KLASIFIKASI KUALITAS AIR BUDIDAYA AKUAPONIK BERBASIS IoT," *J. Teknol. dan Bisnis*, vol. 4, no. 2, pp. 73–86, 2022, doi: 10.37087/jtb.v4i2.87.
- [22] C. N. Ihsan, N. Agustina, M. Naseer, H. Gusdevi, J. F. Rusdi, A. Hadhiwibowo, and F. Abdullah, "Comparison of Machine Learning Algorithms in Detecting Tea Leaf Diseases," *J. RESTI (Rekayasa Sist. dan Teknol. Informatika)*, vol. 8, no. 1, pp. 135–141, 2024, doi: 10.29207/resti.v8i1.5587.
- [23] R. G. de Luna, K. L. Enriquez, E. S. A. Molino, V. C. Magnaye, J. M. R. Dalguntas, A. A. L. Pucyutan, D. Y. D. Umali, R. A. L. Reaño, A. J. M. Lizardo, and J. C. Solis, "A Comparative Study of Machine Learning Techniques for Water Potability Classification," *IEEE Reg. 10 Annu. Int. Conf. Proceedings/TENCON*, pp. 1345–1350, 2023, doi: 10.1109/TENCON58879.2023.10322335.
- [24] T. Z. Jasman, M. A. Fadhlullah, A. L. Pratama, and R. Rismayani, "Analisis Algoritma Gradient Boosting, Adaboost dan Catboost dalam Klasifikasi Kualitas Air," *J. Tek. Inform. dan Sist. Inf.*, vol. 8, no. 2, pp. 392–402, 2022, doi: 10.28932/jutisi.v8i2.4906.
- [25] J. Patel, C. Amipara, T. A. Ahanger, K. Ladhva, R. K. Gupta, H. O. Alsaab, Y. S. Althobaiti, and R. Ratna, "A Machine Learning-Based Water Potability Prediction Model by Using Synthetic Minority Oversampling Technique and Explainable AI," *Comput. Intell. Neurosci.*, vol. 2022, 2022, doi: 10.1155/2022/9283293.
- [26] A. M. Salih, Z. Raisi-Estabragh, I. B. Galazzo, P. Radeva, S. E. Petersen, K. Lekadir, and G. Menegaz, "A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME," *Adv. Intell. Syst.*, vol. 2400304, pp. 1–8, 2024, doi: 10.1002/aisy.202400304.
- [27] E. P. Umar, A. Nawir, H. M. Pakka, J. Jamaluddin, N. S. Tappa, and W. Joemsittiprasert, "Analysis of Shallow Groundwater Quality as Consumable Water in Maros Baru District Aquifer Systems, South Sulawesi, Indonesia," *Int. J. Hydrol. Environ. Sustain.*, vol. 1, no. 1, pp. 33–40, 2022, doi: 10.58524/ijhes.v1i1.55.
- [28] A. Orvieto, H. Kersting, F. Proske, F. Bach, and A. Lucchi, "Anticorrelated Noise Injection for Improved Generalization," *Proc. Mach. Learn. Res.*, vol. 162, pp.

- 17094–17116, 2022, doi: <https://doi.org/10.48550/arXiv.2202.02831>.
- [29] A. Larasati, S. Surono, A. Thobirin, and D. A. Dewi, “Performance Analysis of Resampling Techniques for Overcoming Data Imbalance in Multiclass Classification,” *JUITA J. Inform.*, vol. 13, no. 1, pp. 57–66, Mar. 2025, doi: 10.30595/juita.v13i1.25270.
- [30] J. A. Ilemobayo, O. Durodola, O. Alade, O. J. Awotunde, A. T. Olanrewaju, O. Falana, A. Ogungbire, A. Osinuga, D. Ogunbiyi, A. Ifeanyi, I. E. Odezuligbo, and O. E. Edu, “Hyperparameter Tuning in Machine Learning: A Comprehensive Review,” *J. Eng. Res. Reports*, vol. 26, no. 6, pp. 388–395, 2024, doi: 10.9734/jerr/2024/v26i61188.
- [31] T. Elansari, M. Ouanan, and H. Bourray, “Mixed Radial Basis Function Neural Network Training Using Genetic Algorithm,” *Neural Process. Lett.*, vol. 55, no. 8, pp. 10569–10587, 2023, doi: 10.1007/s11063-023-11339-5.

