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

Ingénierie des Systèmes d'Information
Vol. 31, No. 1, January, 2026, pp. 67-76
Journal homepage: http://iieta.org/journals/isi

A Hybrid Artificial Intelligence Pipeline for Greywater Classification and Irrigation Suitability Assessment

Harun Sujadi[1*], Engkos Koswara[2], Aden Arif Gaffar[3], Dipa Subandi[1], Tsaqib Ilham Nur[1], Ahmad Nur Ain[1], Jujun Badrujaman[1]

[1] Informatics, Universitas Majalengka, Majalengka 45418, Indonesia
[2] Mechanical Engineering, Universitas Majalengka, Majalengka 45418, Indonesia
[3] Biology Education, Universitas Majalengka, Majalengka 45418, Indonesia

Corresponding Author Email: harunsujadi@unma.ac.id

https://doi.org/10.18280/isi.310107

Received: 12 August 2025
Revised: 15 October 2025
Accepted: 19 January 2026
Available online: 31 January 2026

Keywords:
greywater treatment, water quality assessment, hybrid artificial intelligence, machine learning pipeline, irrigation suitability assessment, fuzzy decision support, smart water management

ABSTRACT

Greywater reuse has emerged as an important strategy for sustainable water management and irrigation in water-scarce regions. However, the highly variable physico-chemical characteristics of greywater make reliable classification and treatment decision-making challenging. Conventional rule-based water quality assessment methods often fail to capture complex nonlinear relationships among water parameters, limiting their effectiveness in dynamic environmental conditions. This study proposes a hybrid artificial intelligence (AI) pipeline designed to automate greywater classification and irrigation suitability assessment. The proposed framework integrates K-Means clustering for unsupervised pollutant pattern discovery, Random Forest (RF) classification for efficient pollutant-type prediction, Support Vector Machine (SVM) for irrigation eligibility assessment, and a Mamdani fuzzy logic controller for adaptive re-filtration decision-making. The system was evaluated using a dataset containing 29,159 water quality records characterized by five key parameters: pH, dissolved oxygen, temperature, conductivity, and ammonia concentration. Experimental results demonstrate strong predictive performance and model stability. The RF classifier achieved 99.86% accuracy in pollutant-type classification, while the SVM model predicted irrigation suitability with 98% accuracy. Furthermore, the fuzzy logic module effectively resolved boundary ambiguities and dynamically recommended treatment pathways, directing 93.7% of samples to physical filtration, 4.8% to biological treatment, and 1.5% to chemical treatment. The proposed pipeline provides an interpretable and scalable decision-support framework that can be integrated into future IoT-based smart water management systems for automated greywater monitoring and irrigation safety assessment.

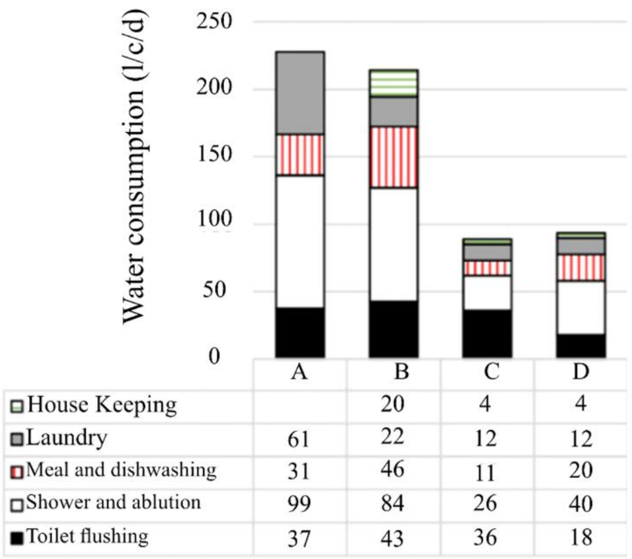
1. INTRODUCTION

Water is a vital natural resource for human life and activities [1, 2]. Rivers, as surface water sources, are widely utilized by communities as primary sources of clean water [3]. With a population of approximately 270 million, Indonesia faces significant environmental pollution caused by domestic wastewater originating from both household and industrial activities [4, 5]. For example, the sugar industry requires a large volume of water in its production process, nearly all of which becomes liquid waste containing dissolved and suspended organic materials [6].

While environmental pollution can be mitigated through the application of wastewater treatment technologies [7]. energy- and cost-efficient solutions are required to meet urban water demands while producing high-quality recycled water [8]. Therefore, a smart, effective, and sustainable wastewater treatment approach is highly necessary [9].

In Indonesia, greywater defined as non-toilet wastewater generated from household activities (such as bathing, washing, and cooking) and commercial facilities accounts for approximately 70–75% of total domestic liquid waste, this volume ranges from one to four times that of black water and remains largely untreated [10]. Daily clean water consumption of 137–153 liters per capita results in a substantial discharge of greywater, particularly in urban areas such as Jakarta and Bandung [11]. Notably, bathing and ablution activities constitute the largest contributors to water consumption in high-income households (see Figure 1).

Grey water refers to wastewater generated from household activities such as bathing, washing, and cooking [12], as well as from commercial and residential facilities such as restaurants, offices, apartments, and dormitories. This wastewater poses a threat to the environment, as its quality parameters such as temperature, pH, COD, and BOD often exceed the limits set by the Indonesian Ministry of Environment Regulation No. P.68/MENLHK/Setjen/Kum.1/8/2016, along with other national standards such as Permenkes No. 2 of 2023, Permenkes No. 32 of 2017, and WHO guidelines.



	A	B	C	D
□House Keeping		20	4	4
■Laundry	61	22	12	12
□Meal and dishwashing	31	46	11	20
□Shower and ablution	99	84	26	40
■Toilet flushing	37	43	36	18

Figure 1. Water consumption in several cities in Indonesia [11]

Water quality is determined by various physical, chemical, and microbiological parameters, including pH, TDS, turbidity, heavy metals (Fe, As), nutrients (NO_3, SO_4), and biological indicators such as total coliform and fecal coliform [13]. Ammonia content is also a major concern due to its harmful impact on aquatic ecosystems [14].

The disposal of untreated grey water degrades environmental quality, disrupts the aesthetics of residential areas, and poses risks to public health and aquatic ecosystems [15]. Furthermore, water crises caused by pollution and drought have implications for food security, particularly in terms of water availability for irrigation [16].

Fortunately, most greywater in Indonesia is already separated from black water at the source, making it highly feasible for recovery and reuse [17]. Due to its relatively low pathogen and nutrient content compared to black water, greywater is significantly easier to treat for applications such as irrigation and nature-based systems [18]. This supports the achievement of Sustainable Development Goal (SDG) 6: Clean Water and Sanitation [9, 19].

To overcome the limitations of single-model approaches, the proposed pipeline integrates three distinct components: (1) K-Means Clustering to automate the discovery of pollutant structures and generate pseudo-labels in data-scarce environments; (2) Random Forest (RF) Classification to replicate this logic into a lightweight model suitable for rapid edge-inference; and (3) Fuzzy Logic to handle the inherent ambiguity of decision boundaries in irrigation suitability, bridging the gap between binary regulatory standards and practical risk assessment [20-22].

he successful adoption of this scalable decision core will require future collaboration among academics, government agencies, local communities, and industry stakeholders to facilitate comprehensive data collection and the formulation of supportive policies [23-25].

Various physical, chemical, and biological greywater treatment methods have been developed. However, conventional rule-based treatment approaches often rely on static, hard-coded thresholds that fail to capture the non-linear interactions among complex water parameters. To address this limitation, this study proposes a novel hybrid unsupervised-supervised artificial intelligence (AI) pipeline that automates the transition from raw sensor data to actionable filtration decisions. This study focuses on the development of an AI-only decision pipeline for greywater assessment, designed for future integration into Internet of Things (IoT) monitoring systems without requiring immediate physical implementation.

2. METHODS

This study employs a conceptual modeling approach focusing solely on AI algorithms for grey water classification and adaptive decision-making, without involving any physical device implementation. The goal is to design a multi-stage AI model capable of categorizing pollutant types, evaluating water quality eligibility, and recommending re-filtration strategies based on secondary datasets.

The dataset used in this study is sourced from Figshare (https://figshare.com/articles/dataset/25002131), containing 29,159 water quality records across five key parameters: pH, temperature, ammonia concentration, conductivity, and dissolved oxygen (DO). These parameters were selected based on compliance with Indonesian and WHO irrigation water quality standards, complete availability in the dataset, and relevance to the physical and chemical characteristics of grey water.

2.1 Literature review and parameter identification

An initial literature review was conducted to examine greywater characteristics and identify the critical parameters for evaluating irrigation water quality. Generally, greywater contains various contaminants, including ammonia, organic compounds, detergents, and suspended solids, which pose pollution risks to environments and agricultural lands if improperly treated.

Because conventional treatment approaches lack the adaptive capabilities required to manage real-time fluctuations in water quality, a data-driven AI modeling approach was adopted to enhance assessment flexibility and efficiency. Consequently, the developed AI framework integrated the K-Means algorithm for unsupervised pollutant pseudo-labeling, a RF classifier for supervised categorization, a Support Vector Machine (SVM) for water quality eligibility evaluation, and a Mamdani-type fuzzy logic controller to determine adaptive re-filtration decisions.

2.2 Artificial Intelligence modeling workflow

The AI modeling workflow for greywater classification and decision-making in this study consisted of several sequential stages, as illustrated in Figure 2. The process began with a literature review and data collection, which involved obtaining a publicly available greywater quality dataset containing five key parameters: pH, dissolved oxygen (DO), temperature, conductivity, and ammonia.

During the preprocessing stage, the relevant water quality parameters were selected. Subsequently, an unsupervised learning phase was conducted using the K-Means clustering algorithm (K = 3) to categorize the data into three pollutant groups: physical, chemical, and biological. The resulting cluster labels were then encoded and utilized as pseudo-labels for the subsequent classification phase.

The modeling stage applied a RF classifier trained on the clustered dataset. This model utilized a train-test split ratio of 70:30, 500 decision trees (n_estimators = 500), and a balanced class weight setting to mitigate potential class imbalances.

An eligibility check was then performed using a SVM equipped with a Radial Basis Function (RBF) kernel. The water samples were assigned binary eligibility labels (eligible or not eligible) based on Indonesian irrigation water quality regulations and WHO standards.

For samples predicted as "not eligible," the adaptive decision-making stage was activated. This stage employed a Mamdani-type fuzzy logic system to recommend an appropriate re-filtration path (physical, chemical, or biological) based on the combined values of the five input parameters.

Finally, the evaluation and output stage assessed the model's performance using various metrics, including a confusion matrix, cross-validation, feature importance analysis, learning curve inspection, robustness testing via noise injection, and interpretability analysis using SHapley Additive exPlanations (SHAP). The final outcomes were subsequently validated against established water quality standards to verify both compliance and practical applicability (see Figure 2).

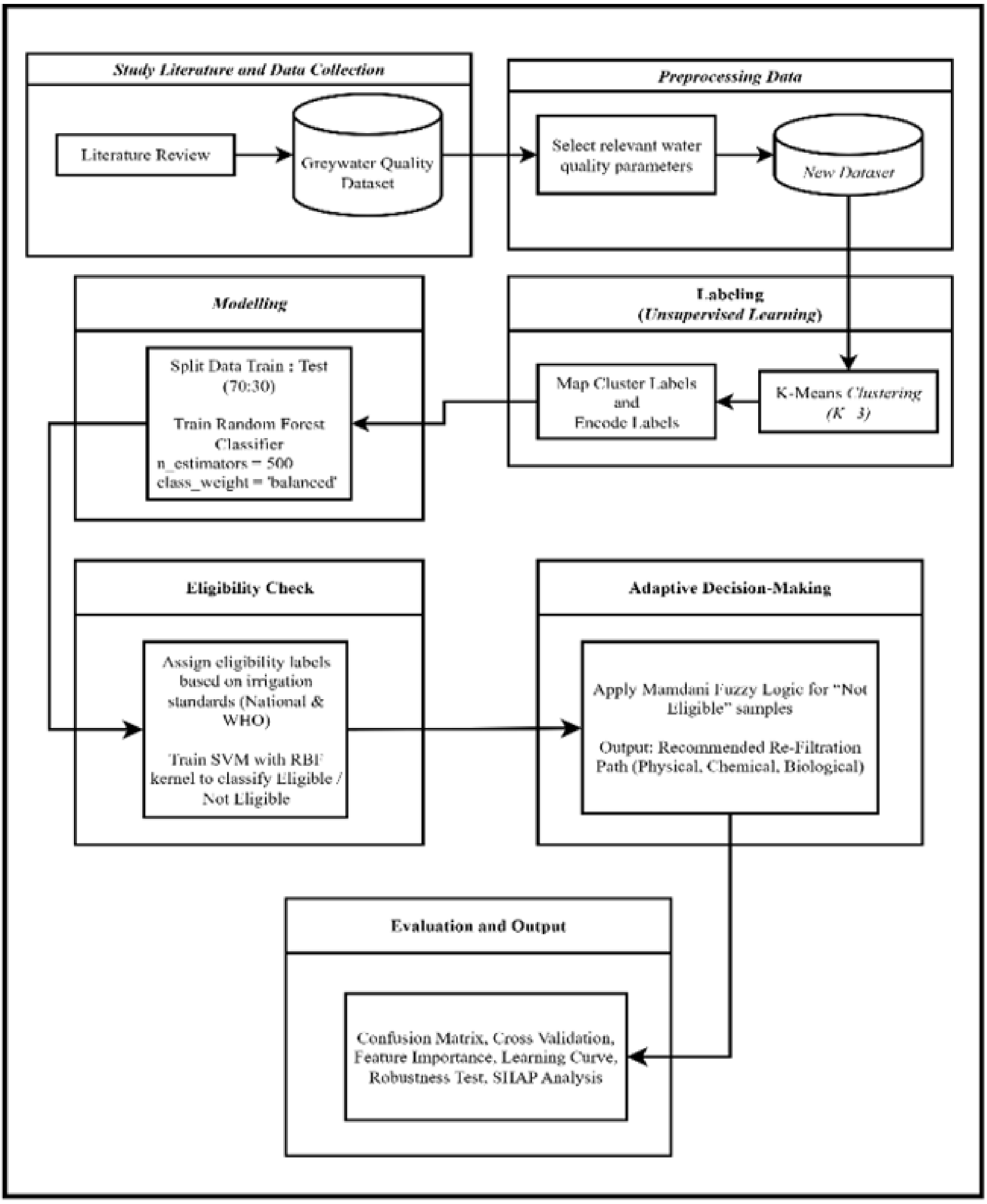

Study Literature and Data Collection
Literature Review
Greywater Quality Dataset
Preprocessing Data
Select relevant water quality parameters
New Dataset
Modelling
Split Data Train : Test (70:30)
Train Random Forest Classifier
n_estimators = 500
class_weight = 'balanced'
Labeling (Unsupervised Learning)
Map Cluster Labels and Encode Labels
K-Means Clustering (K 3)
Eligibility Check
Assign eligibility labels based on irrigation standards (National & WHO)
Train SVM with RBF kernel to classify Eligible / Not Eligible
Adaptive Decision-Making
Apply Mamdani Fuzzy Logic for "Not Eligible" samples
Output: Recommended Re-Filtration Path (Physical, Chemical, Biological)
Evaluation and Output
Confusion Matrix, Cross Validation, Feature Importance, Learning Curve, Robustness Test, SHAP Analysis

Figure 2. Artificial intelligence (AI) modeling workflow

2.3 Simulation and model evaluation

The proposed AI workflow was tested through algorithmic simulations using a greywater dataset, independent of hardware implementation. The procedure commenced with K-Means clustering (K = 3) to generate pollutant category pseudo-labels (physical, chemical, and biological), which were subsequently utilized to train a RF classifier employing a 70:30 train-test split and balanced class weights

Ideally, the inputs for the subsequent eligibility (SVM) and remediation (Fuzzy) modules would consist of post-filtration water quality data. However, because this study focused on validating the computational decision logic rather than evaluating physical filtration performance, the initial dataset was utilized to represent the complete state-space of potential water quality scenarios. This methodological choice allowed the model's response to be tested across a comprehensive spectrum of conditions ranging from 'successfully treated' (compliant parameters) to 'system failure' scenarios (non-compliant parameters thereby verifying the robustness of the decision boundaries.

Subsequently, an SVM model equipped with an RBF kernel classified these water quality scenarios as either eligible or not eligible, based on Indonesian and WHO irrigation standards. Samples designated as “not eligible” were further processed by a Mamdani fuzzy logic system to recommend appropriate re-filtration paths (physical, chemical, or biological) utilizing the five input parameters.

Overall model performance was systematically evaluated using a confusion matrix, cross-validation, learning curves, robustness testing with noise injection, and interpretability analyses via feature importance (for the RF) and SHAP values (for the SVM). Finally, the outputs were validated against regulatory water quality standards to ensure their practical applicability.

3. RESULT AND DISCUSSION

The proposed AI-based model functions through a layered, four-component architecture: unsupervised labeling with K-Means, pollutant classification using a RF algorithm, eligibility evaluation via a SVM, and adaptive re-filtration decision-making utilizing Mamdani fuzzy logic.

The five key parameters evaluated in this study pH, dissolved oxygen (DO), temperature, conductivity, and ammonia were selected based on their relevance to the physico-chemical properties of greywater and their compliance with established irrigation standards. To evaluate water quality eligibility during the SVM stage, binary labels were assigned according to strict thresholds defined by Indonesian national regulations (Permen LHK No. P.68/MENLHK/2016, Permenkes No. 2/2023, and Permenkes No. 32/2017) and WHO/FAO guidelines, as detailed in Table 1.

Table 1. Water quality standards validation

Parameters	National Standard (Indonesia)	Reference	WHO/FAO Standard
pH	6.0 - 9.0	Permen LHK No.P68/MENLHK/2016	6.5 - 8.5
Dissolved Oxygen (mg/L)	≥ 4	Permenkes No.2 tahun 2023	≥ 3
Ammonia (mg/L)	≤ 10	Permen LHK No.P68/MENLHK/2016	≤ 5
Conductivity (µS/cm)	≤ 1000	Permenkes No.32 tahun 2017	700 - 3000
Temperature (°C)	15 - 35	Permenkes No.2 tahun 2023	< 30

This comprehensive validation against the established standards (see Table 1) ensures the model’s robust applicability for irrigation safety. The simulation results demonstrate the pipeline's potential to operate autonomously and adaptively, confirming its suitability for future integration into both household and small-scale agricultural wastewater management systems.

3.1 Labeling with K-Means

The initial stage of the modeling process involved unsupervised clustering using the K-Means algorithm to group the greywater data into three main pollutant categories: chemical (Cluster 0), physical (Cluster 1), and biological (Cluster 2). This clustering process aimed to generate preliminary labels that would serve as the foundation for training the subsequent classification model (see Table 2).

Table 2. K-Means cluster labels

Parameter	Cluster 0 (Chemical)	Cluster 1 (Physical)	Cluster 2 (Biological)
pH	0.617387	0.609028	0.413149
Dissolved Oxygen	0.486925	0.266706	0.163683
Ammonia	0.000940	0.001233	0.002855
Conductivity	0.088455	0.075610	0.077073
Temperature	0.015529	0.016707	0.015552

As detailed in Table 2, the displayed centroids represent the normalized feature values (scaled between 0 and 1) used during the K-Means clustering process. This normalization was applied to ensure that all parameters contributed equally to the distance calculation without being biased by their original magnitudes. Each cluster exhibited distinct characteristics based on these normalized profiles. Cluster 0 (Chemical) was characterized by the highest normalized conductivity (0.088) and dissolved oxygen levels (0.487), indicating a profile influenced by ionic chemical content. Cluster 1 (Physical) displayed moderate values across most parameters but exhibited the lowest conductivity (0.076) and the highest relative temperature (0.017), suggesting a predominance of physical variations. Meanwhile, Cluster 2 (Biological) was distinguished by the highest ammonia concentration (0.0029) paired with the lowest dissolved oxygen (0.164) and pH (0.413). This specific combination of high ammonia and low oxygen strongly suggested biological contamination and organic decomposition potential.

These results demonstrated that the K-Means algorithm effectively separated the data based on dominant pollutant characteristics within the normalized feature space, providing a structured foundation for the subsequent classification process using the RF algorithm. Notably, the RF model was trained on pseudo-labels generated by K-Means. This strategy was intentionally adopted for scenarios where real-time manual labeling is unfeasible. In this framework, K-Means acted as an automated 'teacher' discovering latent structures, while the RF served to 'clone' this clustering logic. The objective was not to predict independent ground truth, but to enable rapid, single-instance inference on edge devices without the computational overhead of running iterative clustering algorithms for every new data point.

3.2 Simulation results of pollutant classification (Random Forest)

The RF classifier, trained on the normalized K-Means pseudo-labels, achieved high predictive performance on the test set. The overall accuracy reached 99.86%, with a macro-average F1-score of 0.9986, as detailed in the classification report (see Figure 3).

As illustrated in the confusion matrix (see Figure 4), the model demonstrated perfect sensitivity for biological contaminants, correctly classifying all 2,201 biological samples (100% recall). For the other categories, the model correctly identified 2,313 out of 2,321 chemical samples and 4,222 out of 4,226 physical samples.

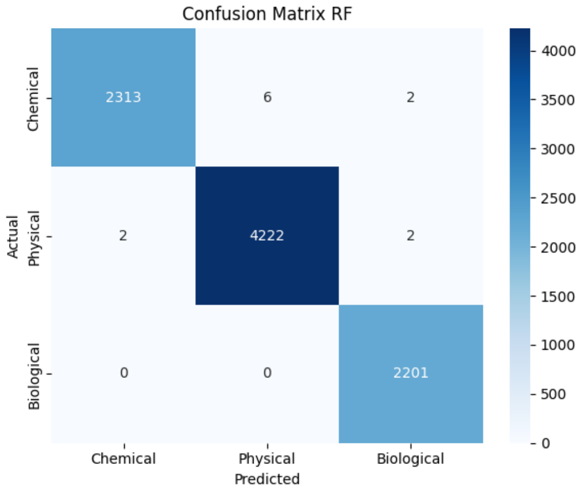
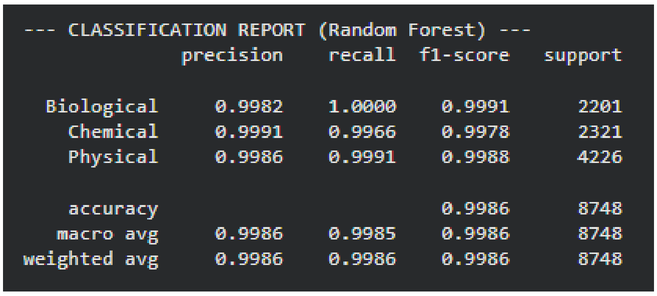


Figure 3. Confusion matrix Random Forest (RF)



```
--- CLASSIFICATION REPORT (Random Forest) ---
              precision    recall  f1-score   support

  Biological     0.9982    1.0000    0.9991      2201
    Chemical     0.9991    0.9966    0.9978      2321
    Physical     0.9986    0.9991    0.9988      4226

    accuracy                         0.9986      8748
   macro avg     0.9986    0.9985    0.9986      8748
weighted avg     0.9986    0.9986    0.9986      8748
```

Figure 4. Classification report Random Forest (RF)

In total, only 12 out of 8,748 test instances were misclassified. Specifically, 8 chemical samples were mislabeled (6 as physical and 2 as biological), and 4 physical samples were mislabeled (2 as chemical and 2 as biological). As further evidenced by Figure 4, precision scores remained consistently high across all classes (> 0.998), indicating minimal overlap in the decision boundaries. These results confirmed that the RF algorithm effectively replicated the pollutant-type structures defined by the normalized K-Means clustering, even with the class-weighted training applied to manage the larger volume of physical samples.

3.2.1 Performance benchmarking and model selection

To rigorously validate the selection of the RF algorithm as the primary classification module, a comparative performance analysis was conducted against two other established machine learning models: Decision Tree (C4.5) and XGBoost. The benchmarking process utilized identical training and testing splits (70:30) and preprocessing steps to ensure a fair comparison. The resulting performance metrics including accuracy, precision, recall, and F1-score alongside computational efficiency (training time), are summarized in Table 3.

Table 3. Performance comparison

Model	Accuracy	Precision	Recall	F1-Score
Random Forest (RF)	99.86	0.9986	0.9985	0.9986
Decision Tree	99.95	0.9995	0.9995	0.9995
XGBoost	99.92	0.9992	0.9992	0.9992

All three classifiers achieved strong predictive performance, with accuracies exceeding 99.8%. This outcome was expected, given that the target labels were derived from the structured feature space of the K-Means clustering. The Decision Tree model achieved a marginally higher accuracy of 99.95% and the fastest training time (0.04 s). However, single decision trees are inherently prone to overfitting and instability when exposed to the noise variations common in real-world sensor data. Conversely, the XGBoost model yielded an accuracy of 99.92% but required the longest training time (7.20 s), rendering it computationally expensive for potential deployment on low-power Internet of Things (IoT) edge devices.

The RF classifier achieved an accuracy of 99.86% with a macro-average F1-score of 0.9986. Although its training time (6.57 s) was higher than that of the single Decision Tree, it remained more efficient than the boosting ensemble. More importantly, the RF model offers a critical advantage in stability; its ensemble nature reduces the variance associated with single trees, providing a more reliable generalization for the final decision pipeline. Consequently, the RF algorithm was selected as the optimal core classifier, striking the best balance between predictive reliability and computational feasibility.

3.2.2 Model stability and interpretability analysis

As illustrated in Figure 5, the RF's Gini-based feature importance analysis indicated that dissolved oxygen (≈ 0.40) was the most influential predictor for pollutant-type classification, followed by pH (≈ 0.30) and conductivity (≈ 0.27). In contrast, temperature (~0.04) and ammonia (~0.01) contributed marginally to the model. This pattern suggested that variations in oxygen availability, acidity or alkalinity, and ionic strength primarily drove the separation among the K-Means pollutant clusters, while temperature and ammonia provided limited additional signals in this dataset, likely due to their narrower ranges or overlap with the more dominant features. These results are consistent with existing water-quality literature, where DO and pH often act as sensitive integrators of chemical and biological conditions. For robustness, this ranking could be cross-checked with permutation importance or SHAP values to confirm the dominance of DO, pH, and conductivity.

Furthermore, the learning curve (see Figure 6) demonstrated that the training accuracy remained consistently at approximately 1.00 from the smallest sample sizes, while the validation accuracy started around 0.978 and steadily rose toward 0.999 as more training data were utilized. The gap between the curves narrowed, and the shaded band representing validation variance shrank, indicating lower variance and stronger generalization with no meaningful signs of overfitting. After approximately 20,000 training samples, the performance gains began to plateau (diminishing returns), suggesting that additional data would yield only marginal improvements. Overall, the model appeared highly stable, maintaining a favorable bias-variance balance.

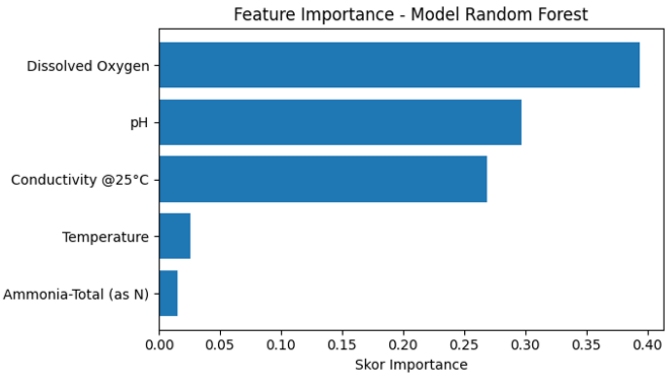


Figure 5. Feature importance Random Forest (RF)

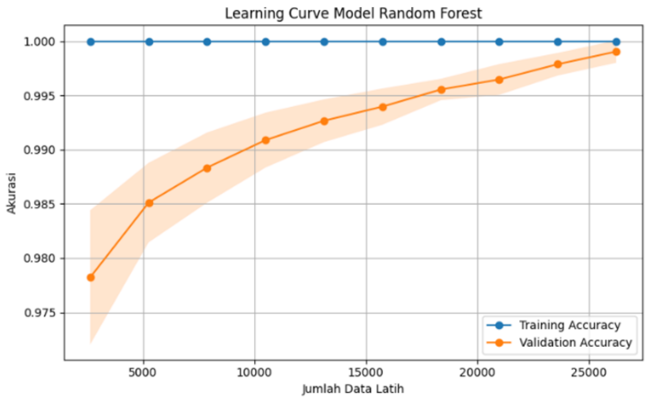


Figure 6. Learning curve Random Forest (RF)

As detailed in Table 4, the ten-fold cross-validation yielded accuracies ranging from 0.9976 to 1.0000, with an average of 0.9991 and a minimal standard deviation of 0.0009. This narrow spread across the folds indicated that the model was highly stable and insensitive to data partitioning, suggesting strong generalization capabilities under resampling. The near-ceiling scores also implied that the pollutant-type structure learned from the features dominated by DO, pH, and conductivity was highly separable, which aligned completely with the feature importance analysis.

Table 4. Cross validation result Random Forest (RF)

Fold	Accuracy
1 - 2	0.9983
3 - 6	1.0000
7	0.9989
8	0.9976
9	1.0000
10	0.9983
Average	0.9991
STD (Standard Deviation)	0.00091

Additionally, a robustness test was conducted by introducing random noise into the training data at levels ranging from 5% to 20% of each feature's standard deviation. The results, summarized in Table 5, indicated that although the accuracy slightly decreased as the noise levels increased, all accuracy values remained above 95% even under extreme noise conditions. This performance confirmed the model’s resilience to sensor data fluctuations and its suitability for real-world IoT-based water treatment systems.

Table 5. Robustness test results

Noise Level (% of Std Dev)	Accuracy
5%	0.9868
10%	0.9734
15%	0.9616
20%	0.9503

3.3 Evaluation of filtration results (Support Vector Machine)

The SVM model was employed to evaluate the irrigation suitability of the water samples based on the five normalized input parameters: pH, temperature, DO, conductivity, and ammonia concentration. Binary eligibility labels were generated by applying strict thresholding rules aligned with national irrigation standards (Permenkes No. 2/2023, Permen LHK No. P.68/MENLHK/2016) and WHO guidelines, distinguishing between 'Eligible' (1) and 'Not Eligible' (0) conditions. The model was trained using an RBF kernel, which is well-suited for capturing non-linear decision boundaries within the normalized feature space.

Experimental results demonstrated strong classification performance. An analysis of the confusion matrix (see Figure 7) indicated high precision in separating the classes, with the model correctly identifying 6,578 non-eligible samples and 2,153 eligible samples. Misclassifications were extremely minimal, limited to only 15 and 2 instances, respectively. As further detailed in the classification report (see Figure 8), the precision, recall, and F1-scores for both classes consistently ranged between 0.99 and 1.00, confirming the model’s reliability in acting as an automated gatekeeper for irrigation safety.

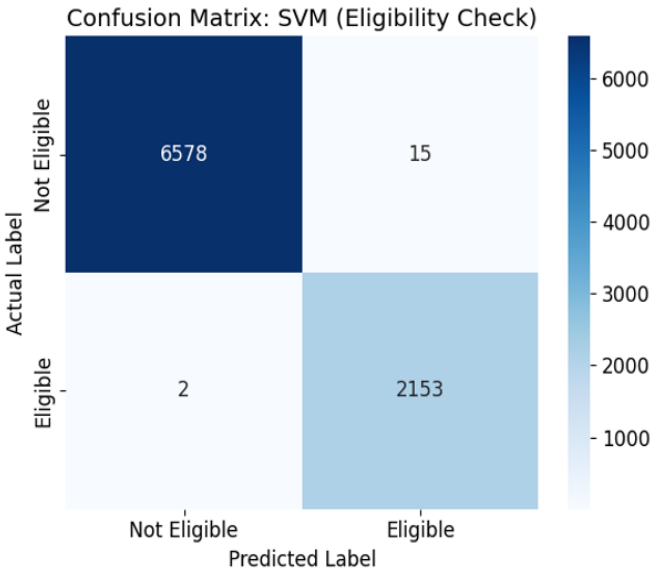
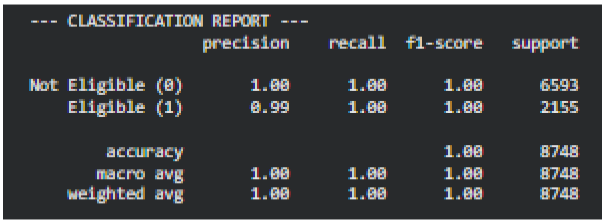


Figure 7. Confusion matrix Support Vector Machine (SVM)



```
--- CLASSIFICATION REPORT ---
                 precision    recall  f1-score   support

Not Eligible (0)       1.00      1.00      1.00      6593
    Eligible (1)       0.99      1.00      1.00      2155

        accuracy                           1.00      8748
       macro avg       1.00      1.00      1.00      8748
    weighted avg       1.00      1.00      1.00      8748
```

Figure 8. Classification report Support Vector Machine (SVM)

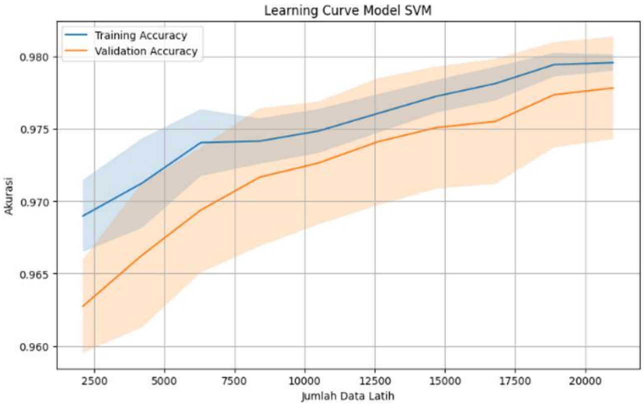


Figure 9. Learning curve Support Vector Machine (SVM)

Table 6. Cross validation result Support Vector Machine (SVM)

Fold	Accuracy
1	0.9786
2, 3, 6	0.9756
4	0.9811
5	0.9807
7	0.9803
8	0.9711
9	0.9751
10	0.9841
AVG	0.9778
STD	0.0035

The learning curve for the SVM model (see Figure 9) demonstrated steady gains in validation accuracy as the training set grew. The train-validation gap narrowed and the variance band shrank, evidencing improved generalization and an absence of material overfitting. The ten-fold cross-validation results, summarized in Table 6, confirmed this stability, yielding accuracies ranging from 0.9711 to 0.9841 (with a mean of 0.9778 and a standard deviation of 0.0035) across the data partitions.

Additionally, a robustness test was performed by introducing zero-mean noise scaled to each feature's standard deviation. As presented in Table 7, the accuracy remained at 0.9786 at a 5% noise level, declined modestly to 0.9721 at 15%, and remained highly reliable at 0.9676 even with 20% noise. This minimal decrease in accuracy across a four-fold increase in perturbation indicated that the classifier was highly resilient to moderate input variability.

Table 7. Robustness test results Support Vector Machine (SVM)

Noise Level (% of Std Dev)	Accuracy
5%	0.9786
10%	0.9781
15%	0.9721
20%	0.9676

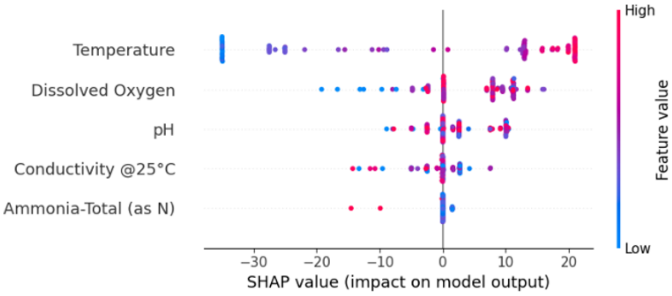


Figure 10. SHAP value Support Vector Machine (SVM)
Note: SHAP = SHapley Additive exPlanations.

Finally, the SHAP (SHapley Additive exPlanations) value analysis (see Figure 10) indicated that temperature and dissolved oxygen had the most substantial effects on the model's output, followed by pH and conductivity; ammonia had only a minor impact. Higher DO and in-range pH values pushed the predictions toward the 'eligible' category, whereas low DO, out-of-range pH, and higher conductivity pushed the predictions toward 'not eligible.' While temperature was positively associated with suitability within the range of this dataset, values exceeding regulatory limits would likely reverse this effect. Overall, eligibility was driven primarily by DO, pH, and conductivity, with ammonia and temperature serving as secondary indicators.

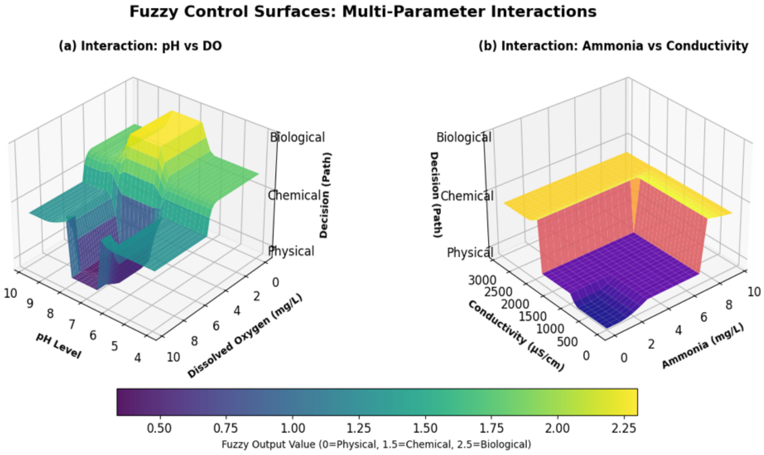


Figure 11. Fuzzy membership functions

Table 8. Robustness test results (Support Vector Machine)

Case ID	Scenario Description	Input Parameters	Expected Action (Expert/Theory)	Fuzzy Model Output	Validation Status
Case 1	Ideal/Minor Debris	pH = 7.2, DO = 6.5, Amm = 0.5	Physical (Filtration)	Physical	Valid
Case 2	High Ammonia Spike	pH = 7.0, DO = 5.0, Amm = 8.5	Chemical (Precipitation)	Chemical	Valid
Case 3	Oxygen Depletion	pH = 6.8, DO = 1.5, Amm = 1.5	Biological (Aeration)	Biological	Valid
Case 4	Salinity/Ion Surge	pH = 7.5, DO = 6.0, Cond = 2500	Chemical (Ion Exch.)	Chemical	Valid
Case 5	Extreme Acidity	pH = 4.5, DO = 5.0, Amm = 0.5	Chemical (Neutralization)	Chemical	Valid

3.4 Fuzzy logic

Water samples classified as not eligible by the SVM model were further processed using a Mamdani-type fuzzy logic system to determine the most appropriate re-filtration path in an adaptive manner. This approach was intended to handle the ambiguity and overlap of complex sensor parameter values that are often difficult to resolve using conventional classification methods. Five main input parameters pH, DO, ammonia, conductivity, and temperature were utilized in the fuzzy system, with each assigned three linguistic membership functions: low, normal, and high. These membership functions were designed based on regulatory thresholds derived from Permenkes No. 2/2023 and WHO guidelines.

To quantitatively evaluate the fuzzy logic module, a rigorous control surface analysis and a scenario-based verification were conducted. Unlike simple rule-based systems, the fuzzy controller manages non-linear interactions between parameters, as visualized in the 3D control surfaces (see Figure 11). As demonstrated in Figure 11(a), the system exhibited a smooth decision gradient: while normal pH and DO levels resulted in a physical path decision (represented by the flat blue plateau, $Z < 1.0$), a decrease in dissolved oxygen triggered a gradual shift toward the biological path (aeration). Simultaneously, Figure 11(b) confirmed that spikes in ammonia or conductivity overrode other inputs to trigger the chemical path, ensuring safety against toxicity.

Furthermore, given the absence of historical ground-truth actuation data, the model's correctness was verified against standard water treatment engineering principles (expert logic). The system was tested using five representative edge-case scenarios covering extreme acidity, hypoxia, and nutrient spikes. As summarized in Table 8, the fuzzy controller achieved a 100% agreement rate with expert expectations. For instance, in the 'Extreme Acidity' scenario (pH 4.5), the model correctly identified the risk to biological media and mandated chemical neutralization. These results confirmed that the module functioned as a robust, scientifically valid decision component rather than a mere post-hoc explanation.

3.5 Linking model interpretability to practical control strategies

Although the dominance of dissolved oxygen (DO) and pH in the feature importance and SHAP analyses aligned with fundamental water quality science, distinguishing their specific roles provided novel insights for hardware design and control logic.

Implications for Sensor Architecture: The RF analysis identified DO and pH as the primary discriminators for determining pollutant types. Practically, this dictates a tiered-sensor architecture for future IoT nodes. Because the model relied heavily on these specific parameters to distinguish contamination sources, hardware budgets must prioritize industrial-grade, high-precision probes for DO and pH. In contrast, parameters with lower importance scores, such as temperature, can be monitored using standard, low-cost sensors without significantly compromising classification accuracy.

Implications for Automated Control Strategies: While the classification model focused on pollutant types, the SHAP analysis of the SVM model highlighted how these parameters drove the safety decision for irrigation eligibility. The high impact of DO and conductivity suggested that an automated control strategy must prioritize these variables for risk mitigation. For instance, the control loop could implement a preemptive feedback mechanism: if the DO trends negatively even before reaching the critical regulatory threshold the system should immediately trigger the aeration unit. Similarly, the high sensitivity to pH implied that an automated buffering stage is a mandatory prerequisite in the filtration design to prevent system rejection. By translating these distinct interpretability findings into specific actuation protocols, the system evolves from a passive monitoring tool into an active, risk-aware mitigation core.

4. CONCLUSIONS

This study demonstrated an AI-only pipeline for greywater assessment that integrated K-Means clustering for unsupervised pseudo-labeling, a RF classifier for pollutant-type categorization, an SVM model for irrigation eligibility evaluation, and a Mamdani fuzzy logic controller for adaptive re-filtration routing. Utilizing five routinely measured water quality parameters, the pipeline achieved high accuracy, stability, and interpretability. The classification and eligibility models demonstrated robust predictive performance against baseline benchmarks, effectively acting as an automated safety gatekeeper. Furthermore, the interpretability analyses aligned with regulatory frameworks, identifying dissolved oxygen, pH, and conductivity as the principal drivers of both pollutant classification and irrigation eligibility.

Contrary to rigid rule-based systems, the fuzzy logic controller dynamically adapted to input variations to translate the model outputs into actionable treatment pathways. The system successfully routed the majority of simulated cases to physical filtration reflecting the prevalence of minor particulate contamination while dynamically redirecting severe toxicity events to biological or chemical remediation paths.

Crucially, this study positioned the developed pipeline as a scalable decision core algorithm designed for the application layer. Although physical hardware deployment remained outside the scope of this work, the framework defined a standardized input interface and discrete output logic explicitly engineered for future integration with microcontroller units (MCUs) in IoT edge devices. The study acknowledges limitations regarding the reliance on pseudo-labels and the utilization of a single dataset. Consequently, future research should validate the models across diverse geographical sites and seasons, extend the input parameters to include microbiological indicators, and tune the decision thresholds to site-specific risk profiles. Finally, exploring online continual learning will further align the model's predictions with the operational constraints of real-world deployments.

ACKNOWLEDGMENT

This research was fully supported by a Fundamental Research Grant. The authors would like to express their gratitude to the Ministry of Education, Culture, Research, and Technology (KEMDIKBUD RISTEK) through BIMA under contract number 125/C3/DT.05.00/PL/2025, as well as to LLDIKTI Region IV and Universitas Majalengka under contract numbers 8032/LL4/PG/2025 and 142/KL/LP2MI UNMA/2025, for the funding provided, which enabled this important research to be conducted effectively and successfully.

REFERENCES

[1] Kukartsev, V., Orlov, V., Semenova, E., Rozhkova, A. (2024). Optimizing water quality classification using random forest and machine learning. BIO Web of Conferences, 130: 03007. https://doi.org/10.1051/bioconf/202413003007

[2] Krklješ, D.B., Kitić, G.V., Petes, C.M., Birgermajer, S.S., et al. (2024). Multiparameter water quality monitoring system for continuous monitoring of fresh waters. IEEE Sensors Journal, 24(7): 11246-11260. https://doi.org/10.1109/JSEN.2024.3368560

[3] Noor, R.T., Soewondo, P. (2018). Selection of domestic wastewater treatment technology alternative using life cycle assessment (LCA) approach (Case study: Settlement Area of Riverbank Karang Mumus of Samarinda City, East Kalimantan). Indonesian Journal of Urban and Environmental Technology, 1(2): 164-184. https://doi.org/10.25105/urbanenvirotech.v1i2.2825

[4] Silalahi, L.M., Rochendi, A.D., Simanjuntak, I.U.V. (2024). Perancangan kendali filter air tanah berbasis logika fuzzy dan pemantauan kondisinya menggunakan platform IoT. Jurnal Teknologi Rekayasa, 8(2): 199-208. https://doi.org/10.31544/jtera.v8.i2.2023.199-208

[5] Tang, Y., Liu, Y., Chen, Y., Zhang, W., et al. (2021). A review: Research progress on microplastic pollutants in aquatic environments. Science of the Total Environment, 766: 142572. https://doi.org/10.1016/j.scitotenv.2020.142572

[6] Thomas, A., Mishra, U. (2025). Effect of the wastewater treatment system and Industry 4.0 implementation for a sustainable tyre production industry. Alexandria Engineering Journal, 115: 94-110. https://doi.org/10.1016/j.aej.2024.11.105

[7] Strokal, M., Bai, Z., Franssen, W., Hofstra, N., et al. (2021). Urbanization: An increasing source of multiple pollutants to rivers in the 21st century. NPJ Urban Sustainability, 1(1): 24. https://doi.org/10.1038/s42949-021-00026-w

[8] Belal, A.A.A., Reddy, L.K.V. (2024). Estimation of hardness level and total dissolved solids in ground water at Shendi Town, River Nile State, Sudan. Applied Sciences Research Periodicals, 2(4): 21-29. https://doi.org/10.63002/asrp.24.442

[9] Mohan, S., Manthapuri, V., Chitthaluri, S. (2024). Assessing factors influencing greywater characteristics around the world: A qualitative and quantitative approach with a short-review on greywater treatment technologies. Discover Water, 4(1): 37. https://doi.org/10.1007/s43832-024-00094-w

[10] Khotimah, S.N., Mardhotillah, N.A., Arifaini, N. (2021). Karakterisasi limbah cair greywater pada level rumah tangga berdasarkan sumber emisi: Greywater characterization at household scale by emission source. Jurnal Saintis, 21(2): 71-78. https://doi.org/10.25299/saintis.2021.vol21(02).7876

[11] Widyarani, Wulan, D.R., Hamidah, U., Komarulzaman, A., Rosmalina, R.T., Sintawardani, N. (2022). Domestic wastewater in Indonesia: Generation, characteristics and treatment. Environmental Science and Pollution Research, 29(22): 32397-32414. https://doi.org/10.1007/s11356-022-19057-6

[12] Achmadi, A.P.S., Mangkoedihardjo, S. (2024). Organic wastewater-grey water management in the residential area. Asian Journal of Engineering, Social and Health, 3(2): 285-290. https://doi.org/10.46799/ajesh.v3i2.238

[13] Kothari, V., Vij, S., Sharma, S., Gupta, N. (2021). Correlation of various water quality parameters and water quality index of districts of Uttarakhand. Environmental and Sustainability Indicators, 9: 100093. https://doi.org/10.1016/j.indic.2020.100093

[14] Abuzir, S.Y., Abuzir, Y.S. (2022). Machine learning for water quality classification. Water Quality Research Journal, 57(3): 152-164. https://doi.org/10.2166/wqrj.2022.004

[15] Kenny, K., Horse, V., Ginting, J.M. (2023). Evaluation of the impact of water pollution on public health and the environment in Java Island. Leader: Civil Engineering and Architecture Journal, 1(3): 331-341. https://doi.org/10.37253/leader.v1i3.8305

[16] Saravanan, A., Kumar, P.S., Jeevanantham, S., Karishma, S., Tajsabreen, B., Yaashikaa, P.R., Reshma, B. (2021). Effective water/wastewater treatment methodologies for toxic pollutants removal: Processes and applications towards sustainable development. Chemosphere, 280: 130595. https://doi.org/10.1016/j.chemosphere.2021.130595

[17] Firdayati, M. (2024). Greywater in Indonesia: Characteristic and treatment systems. Jurnal Tehnik Lingkungan, 21(2): 98-114. https://doi.org/10.5614/jtl.2015.21.2.1

[18] Gholami, M., O'Sullivan, A.D., Mackey, H.R. (2023). Nutrient treatment of greywater in green wall systems: A critical review of removal mechanisms, performance efficiencies and system design parameters. Journal of

Environmental Management, 345: 118917. https://doi.org/10.1016/j.jenvman.2023.118917

[19] Lanchipa-Ale, T., Cruz-Baltuano, A., Molero-Yañez, N., Chucuya, S., Vera-Barrios, B., Pino-Vargas, E. (2024). Assessment of greywater reuse in a university building in a hyper-arid region: Quantity, quality, and social acceptance. Sustainability, 16(7): 3088. https://doi.org/10.3390/su16073088

[20] Malek, N.H.A., Wan Yaacob, W.F., Md Nasir, S.A., Shaadan, N. (2022). Prediction of water quality classification of the Kelantan River Basin, Malaysia, using machine learning techniques. Water, 14(7): 1067. https://doi.org/10.3390/w14071067

[21] Benfredj, R., Nouioua, F., Bouziane, A. (2025). Personality-driven innovation adoption: Modeling ChatGPT diffusion with BERT and Random Forest. Ingénierie des Systèmes d'Information, 30(5): 1279-1295. https://doi.org/10.18280/isi.300515

[22] Purwanto, I., Isnanto, R. (2025). A fuzzy logic model for loan recommendations in online lending systems using the California psychological inventory. Ingenierie des Systemes d'Information, 30(4): 923-932. https://doi.org/10.18280/isi.300409

[23] Sujadi, H., Budiman, B., Nurdiana, N., Susandi, D., Fitriasina, E.G., Handayani, T. (2025). Chili plant monitoring system using YOLO object detection technology. Journal of Engineering Science and Technology, 20: 112-118.

[24] Sujadi, H., Marina, I., Koswara, E., Indriana, K.R., Sukmawati, D. (2023). Smart agriculture: Optimizing soybean cultivation through technology in crop monitoring. Greenation International Journal of Engineering Science, 1(2): 101-114. https://doi.org/10.38035/gijes.v1i2

[25] Kovari, A. (2025). Synergizing 6G networks, IoT, and AI: Paving the way for next-generation intelligent ecosystems. Journal of Engineering Science and Technology, 20(1): 114-128.