JOIN (Jurnal Online Informatika)
p-ISSN: 2528-1682, e-ISSN: 2527-9165
Volume 10 Number 2 | December 2025: 442-452
DOI: 10.15575/join.v10i2.1623

Random Forest-Based Classification of Greywater Filtration Media for Intelligent Biofiltration Systems

Harun Sujadi[1], Dipa Subandi[2], Nunu Nurdiana[3]
[1,2,3]Department of Informatics, Universitas Majalengka, Indonesia

Article Info

Article history:

Received April 28, 2025
Revised June 13, 2025
Accepted July 13, 2025
Published November 10, 2025

Keywords:

Biofiltration
Classification Model
Greywater
Machine Learning
Random Forest

ABSTRACT

The increasing volume of domestic wastewater, particularly greywater, has raised the demand for intelligent and adaptive treatment systems to support efficient water reuse. This study aims to develop a classification model for filtration media types (physical, chemical, and biological) based on water quality data using the Random Forest algorithm. Initial labeling was conducted using the K-Means Clustering method on a publicly available dataset simulated as greywater, based on ten key water quality parameters relevant to irrigation and environmental standards. Model evaluation demonstrated excellent classification performance, with a macro F1-score reaching 0.97 and consistent results in both 5-fold and 10-fold cross-validation. These findings indicate that the proposed model can be integrated into an IoT-based biofiltration system as an automated classification logic to support adaptive, efficient, and reusable household wastewater treatment in the context of irrigation.

Corresponding Author:

Harun Sujadi,
Informatics Department, Faculty of Engineering, Universitas Majalengka
Jl. Raya K H Abdul Halim No.146, Majalengka Kulon, Kec. Majalengka, Kabupaten Majalengka, Jawa Barat 45418
Email: harunsujadi@unma.ac.id

1. INTRODUCTION

On Urbanization and increased domestic water consumption have put significant pressure on surface water quality, especially in developing countries [1]. One type of domestic wastewater that is often overlooked but has the potential for reuse is greywater, which is non-fecal wastewater derived from bathing, washing, and kitchen activities [2], [3], With proper treatment, greywater can be reused for purposes such as irrigation, in line with the principles of sustainable water management [4].

To be safe for reuse, greywater must go through a filtration process that is capable of handling physical, chemical, and biological contaminants. Intelligent biofiltration systems are emerging as a potential solution through the integration of real-time water quality data from environmental sensors [5], [6], [7]. However, a major challenge in these systems is determining the type of filtration media that best suits the characteristics of the incoming graywater.

Various machine learning (ML) approaches have been used for water quality classification, such as Support Vector Machine (SVM), Decision Tree, and Random Forest [8], [9], [10], [11], [12]. Generally, the focus of previous research has been on general water quality classification or quality index prediction [13], without considering the need for specific classification of filtration media types, which is crucial in the context of data-driven processing.

This study proposes the development of a filtration media type classification model (physical, chemical, biological) using the Random Forest algorithm based on ten key water quality parameters. The dataset used comes from an aquaculture water quality study simulated as graywater, given the similarity

of key parameters such as BOD, DO, ammonia, and turbidity [14]. The selection of parameters refers to the regulation of water quality standards for irrigation and environment such as Regulation of the Minister of Health of the Republic of Indonesia No. 2 Year 2023, Permenkes No. 32 Year 2017, Minister of Environment and Forestry Regulation No. P.68/MENLHK/2016, as well as guidelines from the World Health Organization (WHO) [15], [16], [17], [18]

Initial labeling was performed using K-Means Clustering to form three filtration categories, which were then used as labels in the Random Forest model [19]. Unlike most previous studies which focused on binary water classification or potability prediction, this study introduces a novel approach by classifying greywater into filtration media categories (physical, biological, and chemical) using machine learning. The labels were generated through K-Means clustering and used to train a Random Forest model, offering a data-driven logic for adaptive media selection. This approach provides a foundational step toward intelligent greywater treatment systems that can dynamically respond to real-time water quality conditions The resulting model is an early prototype of automated classification logic that can be integrated into IoT-based intelligent filtration systems, especially for the treatment of household wastewater to be utilized as irrigation water.

2. METHOD

2.1 Dataset and Parameter Characteristics

This research utilizes a public dataset titled Aquaculture - Water Quality Dataset available in the Mendeley Data repository [14]. This dataset was originally developed to evaluate pond water quality in aquaculture systems based on numerical parameters relevant to fish growth. Although not derived from direct field measurements, this dataset is systematically organized and can be used for training machine learning-based classification models.

This dataset was used as the initial simulation data for training the AI model, assuming that the parameters would later be provided by water quality sensors in the real system implementation.

The dataset consists of 4,300 data samples with 14 water quality parameters, namely temperature, turbidity, dissolved oxygen (DO), biological oxygen demand (BOD), carbon dioxide (CO_2), pH, alkalinity, hardness, calcium, ammonia, nitrite, phosphorus, hydrogen sulfide (H_2S), and plankton. All data are organized in numerical format with consistent units of measurement.

Although the original context was aquaculture, a number of parameters in this dataset such as BOD, DO, ammonia, pH, and turbidity are also used in domestic graywater quality assessment. Therefore, this dataset is relevant to use as simulative data to build a filtration media type classification model based on graywater quality [13].

2.2 Preprocessing and Feature Selection

From a total of 14 parameters available in the dataset, 10 parameters were selected for use in training and testing the classification model. The feature selection process was based on three main considerations, namely: (1) consistent data availability across entries, (2) ecological and technical relevance to domestic graywater quality, and (3) regulatory support from applicable environmental quality standards.

The parameters selected include temperature, pH, dissolved oxygen (DO), biological oxygen demand (BOD), ammonia, nitrite, phosphorus, hardness, alkalinity, and turbidity. This selection is based on the provisions listed in the Regulation of the Minister of Health of the Republic of Indonesia No. 2 Year 2023, Permenkes No. 32 Year 2017, Minister of Environment and Forestry Regulation No. P.68/MENLHK/2016, as well as guidelines from the World Health Organization (WHO)[15], [16], [17], [18].

Some parameters such as hardness and alkalinity were considered because they have a high correlation with total dissolved solids (TDS), which directly affects the effectiveness of the filtration media. Previous studies have shown that TDS is strongly correlated with alkalinity and hardness (r > 0.8), as well as conductivity (r = 0.9), suggesting a structural link between water quality parameters. Even in the Ratuwa river analysis, the largest contributions to TDS values came from alkalinity (23.5%) and hardness (19.9%), outweighing other ions [20], [21], [22], [23]. In contrast, parameters such as carbon dioxide, H_2S, plankton, calcium, and nitrite are to some extent not used because they are not the main focus in the filtration process of small-scale domestic graywater, nor are they prioritized in the domestic wastewater quality standard regulations.

p-ISSN: 2528-1682
e-ISSN: 2527-9165

Although a correlation heatmap was also provided (Figure 1) to identify redundancy among features, it was not the main basis for feature selection. Instead, domain knowledge and regulatory standards were prioritized to ensure relevance to greywater filtration contexts. As the correlation values were mostly below 0.5, all selected features were retained to preserve informative diversity.

Table 1. Selected Parameters and Quality Standard Sources

Criteria	units	Source of Quality Standard	Explanation
Temperature	°C	Permenkes No. 2/2023, P.68/MENLHK/2016	Indicator of ambient temperature and biological processes
pH	-	Permenkes No. 2/2023, P.68/MENLHK/2016	Basic parameter of water chemistry balance
Dissolved Oxygen (DO)	mg/L	P.68/MENLHK/2016	Indicates oxygen availability for biological processes
Biochemical Oxygen Demand (BOD)	mg/L	P.68/MENLHK/2016	Measures the load of biodegradable organic pollutants
Ammonia (NH_3)	mg/L	Permenkes No. 2/2023, P.68/MENLHK/2016	Indicates the level of toxic nitrogen contamination
Phosphorus (PO_4^{3-})	mg/L	Permenkes No. 2/2023	Key nutrient causing eutrophication
Nitrite	Mg/L	Permenkes No. 32/2017	Indicator of toxic inorganic nitrogen contaminants
Hardness	mg/L	WHO	Closely related to TDS and scale formation potential
Alkalinity	mg/L	WHO	Indicator of acid neutralizing capacity and projected TDS
Turbidity (CM)	CM	P.68/MENLHK/2016, WHO	Water turbidity based on visual visibility (in centimeters)

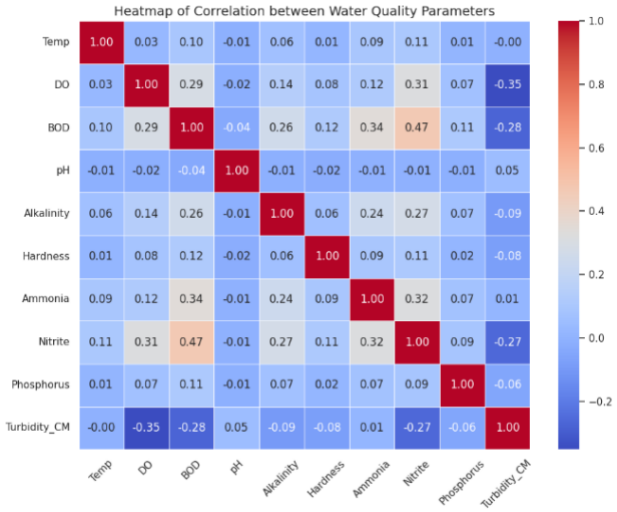

Figure 1. Heatmap of Correlation beetwen Water Quality Parameters

To identify possible redundancies between numerical features and support the input parameter selection process, a correlation analysis between water quality parameters was conducted using Pearson coefficient. The results of the correlation visualization are presented in Figure 1. In general, correlations between parameters were low to moderate (r < 0.5), indicating that each feature contributed relatively unique information to the classification system. The highest correlations were recorded between BOD and Nitrite (r = 0.47) and Ammonia and BOD (r = 0.34), while a negative correlation was found between Turbidity and DO (r = -0.35). No extreme multicollinearity correlations were found between parameters, so all features were considered suitable for inclusion in the classification model training.

2.3 Data Labeling with K-Means Clustering

Since the dataset did not have explicit classification labels, the labeling process was performed using the K-Means Clustering algorithm which has high computational efficiency, fast convergence, and is widely used in clustering environmental data [24].

The number of clusters was set to three (k = 3) to represent the three main categories of filtration media: physical, chemical, and biological. This division refers to the greywater treatment literature that differentiates water quality parameters based on physical (such as turbidity and temperature), chemical (such as pH and BOD), and biological (such as organic residue and coliform) characteristics [3].

Labeling was done based on ten selected parameters, and the clustering results were analyzed through the average value of each parameter in each group. Clusters with high ammonia and BOD levels were associated with biological filtration, while clusters with high ionic levels were associated with chemical filtration, and clusters with high turbidity and low chemical parameters were considered physical filtration.

The use of K-Means in water quality classification has also been proven effective in various studies, both for groundwater, rivers, and industrial effluents [25], [26], [27], [28], [29].

Table 2. Average Value of Water Quality Parameters in Each K-Means Result Cluster.

Parameters	Cluster 0 (Physical)	Cluster 1 (Biological)	Cluster 2 (Chemical)
Temp (°C)	26.229046	25.375313	26.119791
DO (mg/L)	6.161586	4.798653	5.931697
BOD (mg/L)	4.227467	2.451902	4.039728
pH	7.667912	7.719905	7.749116
Alkalinity (mg/L)	114.389701	56.667415	194.772847
Hardness (mg/L)	249.174967	100.093532	64.834258
Ammonia (mg/L)	0.075587	0.029032	0.079713
Nitrite (mg/L)	1.082027	0.364187	1.043787
Phosphorus (mg/L)	1.243793	1.104928	1.315640
Turbidity (cm)	28.358927	44.969636	32.252305

Based on the average values of water quality parameters shown in Table 3, each cluster was classified according to the appropriate filtration type. Cluster 0 was labeled as physical filtration due to its highest hardness and lowest turbidity, indicating a dominance of suspended solid particles. Cluster 1 exhibited the lowest dissolved oxygen (DO) and highest turbidity, suggesting biological contamination that may require microbial treatment, despite relatively low BOD and ammonia levels. Meanwhile, Cluster 2 was categorized as chemical filtration, as it showed elevated levels of alkalinity, nitrite, and phosphorus, along with the lowest hardness, indicating a prevalence of dissolved chemical pollutants.

The interpretation of cluster labels into physical, biological, and chemical categories was based on a synthesis of water quality indicator functions. Cluster 0, characterized by high hardness and low turbidity, reflects the dominance of suspended solids, aligning with typical physical filtration targets [25]. Cluster 1 showed high turbidity and low DO common indicators of organic or microbial contamination therefore classified under biological filtration. Meanwhile, Cluster 2 had elevated concentrations of nitrite, phosphorus, and alkalinity, which are indicative of dissolved chemical pollutants requiring chemical treatment. This interpretive framework is consistent with classifications found in environmental water treatment literature [3], [26].

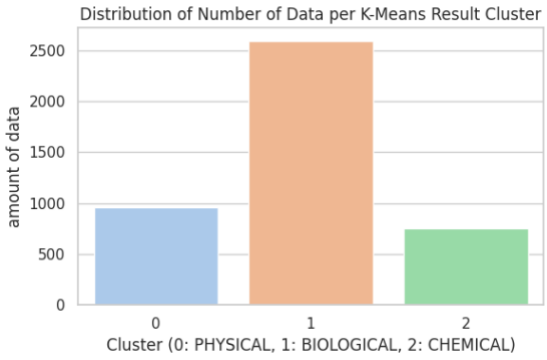

Figure 2a. Distribution Diagram of Number of Data per Cluster

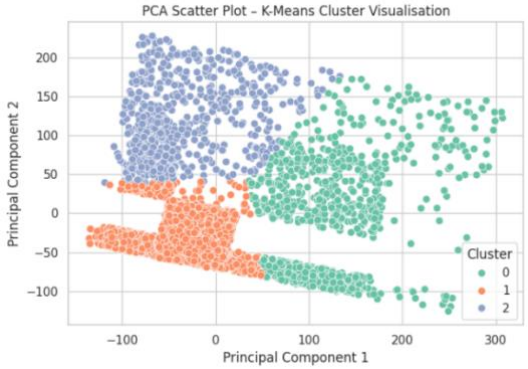

Figure 2b.Scatter Plot (PCA 2D)

These labels were subsequently used as target classes in the Random Forest classification model. Prior to training, the class distribution resulting from K-Means clustering was examined to

p-ISSN: 2528-1682
e-ISSN: 2527-9165

identify potential imbalances. As shown in Figure 2a, the dataset is imbalanced, with 2,590 instances in the biological class, 956 in the physical class, and 754 in the chemical class. To mitigate model bias, the `class_weight='balanced'` parameter was applied during model training.

To further validate the clustering structure, a two-dimensional projection was generated using Principal Component Analysis (PCA), as illustrated in Figure 2b. The three clusters form relatively distinct spatial regions, supporting the coherence of the K-Means result. The silhouette score of 0.4767 and the PCA components accounting for 95.34% of total variance confirm the representativeness and reliability of the clustering.

2.4 Development of the Random Forest Model

After the labeling process with K-Means, a classification model was developed using the Random Forest algorithm due to its ability to handle numerical multivariant data, tolerant of outliers, and stable under unbalanced class distributions [30], [31], [32]. Previous studies have also demonstrated that Random Forest outperforms other machine learning classifiers in managing imbalanced datasets [33] and maintains high predictive accuracy in high-dimensional feature spaces [34]. The model was trained using ten water quality parameters as input features, with the target label being the filtration media category: physical, biological, and chemical.

The data was stratified into training and test subsets with a ratio of 60:40. The model was trained using 500 decision trees without advanced hyperparameter tuning, as the main focus was to evaluate the feasibility of classification based on the water quality data. Initial evaluation was done based on accuracy, precision, recall, and f1-score metrics, as well as confusion matrix visualization. Full results are presented in the Results and Discussion section.

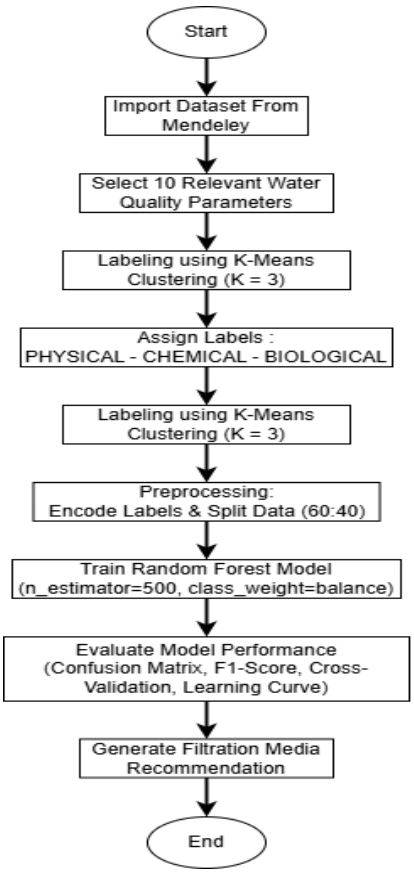

Figure 3. Research Process Flow Chart

Figure 3 presents the flow of the research process, starting from data preprocessing, feature selection, K-Means labeling, to the construction and evaluation of the Random Forest classification

model. The diagram illustrates the systematic stages carried out in the development of a classification model for graywater filtration media types.

This research is focused on building and preliminary validation of the data-based classification model, without direct implementation of the IoT system. Therefore, the resulting model serves as the foundation of classification logic that can be integrated into a water quality sensor-based intelligent biofiltration system at a later stage of development.

3. RESULT AND DISCUSSION

This section presents the results of a computational evaluation of a Random Forest algorithm-based greywater filtration media type classification model, which distinguishes between physical, biological, and chemical filtration based on ten water quality parameters. The evaluation was conducted using accuracy, precision, recall, and f1-score metrics, and supported by cross-validation and learning curve visualization to assess the stability and generalizability of the model. The model is intended as an automatic classification logic in data-driven intelligent biofiltration systems.

3.1 Classification Model Evaluation

Model evaluation was conducted by stratifying the dataset into training and test data (60:40 ratio), in order to maintain proportional distribution between classes. The Random Forest model was trained using ten water quality parameters to predict the filtration media labels: physical, biological, and chemical, resulting from K-Means labeling.

Table 3. Classification Report Model Random Forest (60:40)

Class	Precision	Recall	F1-score	Support
Biological	0.99	0.99	0.99	1036
Physical	0.97	0.98	0.97	382
Chemical	0.96	0.95	0.95	302
Accuracy	-	-	0.98	1720
Macro Avg	0.97	0.97	0.97	1720
Weighted Avg	0.98	0.98	0.98	1720

The classification results on the test dataset are shown in Table 3, demonstrating a macro F1-score of 0.97. The model exhibited balanced predictive performance across all classes, with the Biological class achieving the highest F1-score of 0.99. This indicates that biologically contaminated samples—characterized by features such as high turbidity and low dissolved oxygen—were well distinguished. The Physical and Chemical classes attained F1-scores of 0.97 and 0.95 respectively, with minor misclassifications to other categories as seen in the confusion matrix. Overall, the model achieved an accuracy of 98%, confirming its robustness for filtration media classification tasks.

Prediction results are visualized in the confusion matrix (Figure 4a), which shows a strong concentration of correct predictions along the main diagonal. These findings confirm that the model is capable of consistently distinguishing between the three filtration classes without significant misclassification.

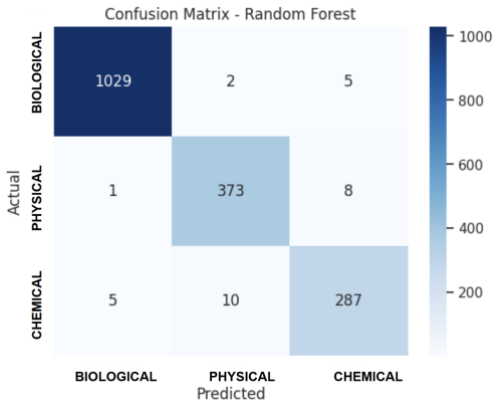

Figure 4a. Confusion Matrix of Random Forest Model

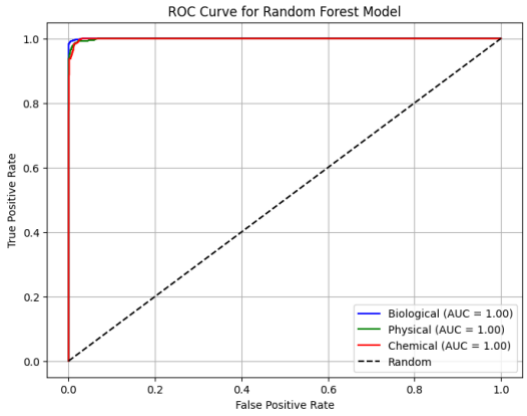

Figure 4b. ROC Curve of Random Forest Model

p-ISSN: 2528-1682
e-ISSN: 2527-9165

In addition to the standard evaluation metrics, the Receiver Operating Characteristic (ROC) curve was utilized to assess the model's ability to distinguish between the three filtration media categories. As illustrated in Figure 4b, the Random Forest model achieved an Area Under the Curve (AUC) score of 1.00 for each class biological, physical, and chemical. This indicates that the classifier was able to perfectly separate positive and negative instances across all threshold levels. The ROC analysis provides a more robust performance assessment, especially in scenarios involving imbalanced data and pseudo-labels derived from K-Means clustering. These results strengthen the model's reliability and demonstrate its potential for integration into real-time classification systems for intelligent biofiltration applications.

To better understand the contribution of each input parameter, feature importance analysis was conducted based on the Random Forest model. As illustrated in Figure 5, the most influential features were Hardness and Alkalinity, followed by Ammonia, Nitrite, and BOD. These top features are commonly associated with dissolved ionic content and organic pollution in greywater, supporting their relevance in identifying suitable filtration media. Parameters such as DO, Turbidity, and pH were found to be less dominant, indicating that the model relies more on chemical characteristics than physical ones when performing classification. The Random Forest model was trained using class_weight='balanced' and n_estimators=500, which were selected to handle class imbalance and ensure robust performance across categories.

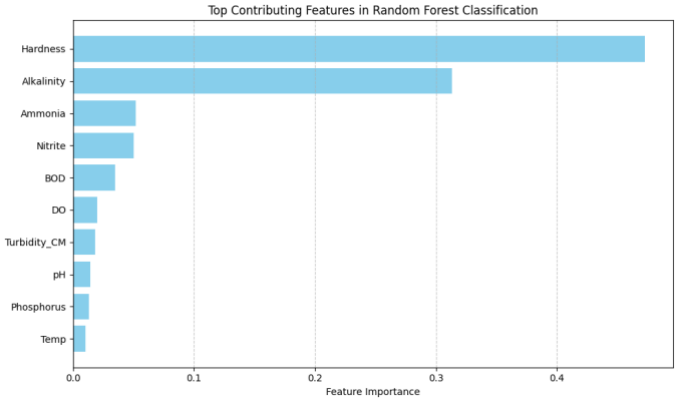

Figure 5. Feature Importance in Random Forest Model

3.2 Comparative Evaluation with Support Vector Machine

To validate the robustness of the Random Forest classifier, a comparative evaluation was performed using a Support Vector Machine (SVM) model. Both models were trained and tested on the same stratified dataset using ten water quality parameters.

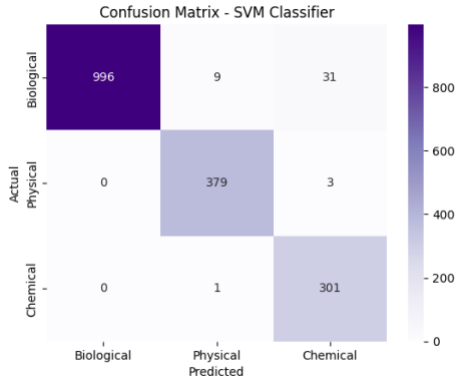

Figure 6a. Confusion Matrix of the SVM Classifier

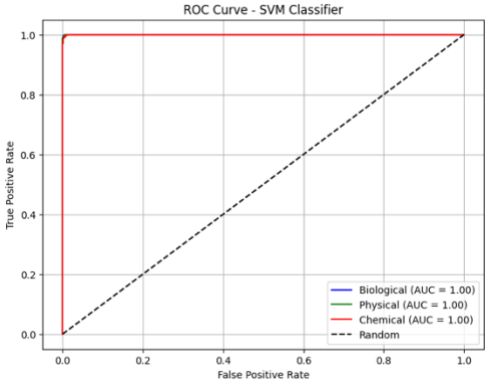

Figure 6b. ROC Curve of the SVM Classifier

The SVM achieved an accuracy of 98%, a macro F1-score of 0.97, and ROC-AUC scores of 1.00 across all classes, comparable to the performance of the Random Forest. However, the confusion matrix showed that Random Forest produced slightly more balanced predictions, particularly in the majority class of biological filtration. These findings support the reliability of the Random Forest model while positioning SVM as a valid baseline for greywater filtration media classification tasks.

3.3 Model Validation Using Cross-Validation

To evaluate the consistency of the model's performance, cross-validation was carried out using two schemes: 5-Fold and 10-Fold.

Table 4. Cross-Validation Results (Macro F1-Score)

Validation Scheme	F1 Macro Avg	Standard Deviation
5-Fold CV	0.9773	0.0090
10-Fold CV	0.9785	0.0124

The results are presented in Table 4. In the 5-Fold scheme, the model achieved an average macro f1-score of 0.9773 with a standard deviation of 0.0090, while in the 10-Fold scheme, the average macro f1-score slightly increased to 0.9785 with a standard deviation of 0.0124. The low deviation values indicate stable model performance across different subsets of data. These results suggest that the model does not suffer from overfitting and demonstrates strong generalization capability, making it suitable as a classification logic in a smart biofiltration system based on water quality data.

3.4 Learning Curve Analysis and Model Generalization

To assess the potential overfitting behavior and model generalization, a learning curve analysis was conducted based on varying training sizes (Figure 7a). The training F1-score remained at 1.00, as typically expected in Random Forest models. Meanwhile, the validation curve stabilized around 0.978 with lower fluctuation as the data size increased. The narrower confidence bounds further indicated consistent cross-validation performance. These results suggest that the relabeling process, informed by K-Means clustering, significantly contributed to improving model reliability and reducing labeling-induced variance.

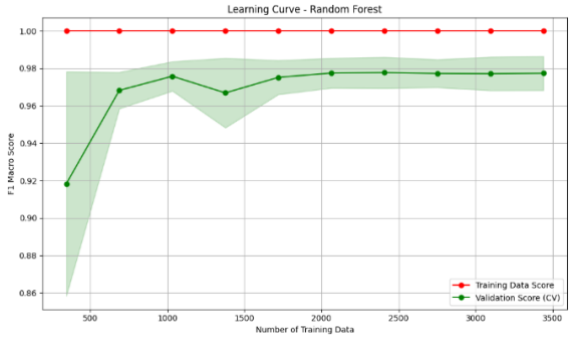

Figure 7a. Learning Curve for Evaluating Overfitting in the Random Forest Model

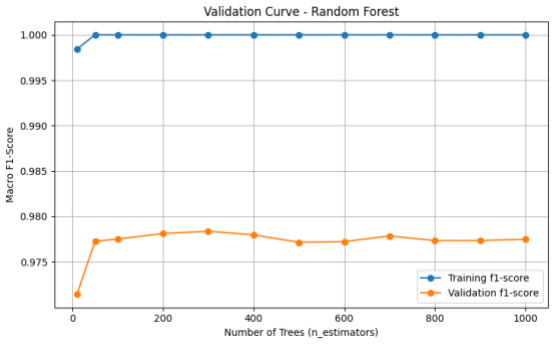

Figure 7b. Validation Curve of the Random Forest Model

As a complement, Figure 7b shows the validation curve across varying numbers of decision trees (n_estimators) ranging from 10 to 1000. The results indicate that model performance on validation data stabilized early, from approximately 100 estimators, maintaining an F1-score around 0.978. No significant decrease was observed even at higher estimator counts, suggesting that the increased model complexity did not lead to overfitting. Therefore, choosing n_estimators within the range of 300 to 500 is considered optimal to balance accuracy and computational efficiency.

3.5 Comparison with Previous Studies

Various previous studies have demonstrated the effectiveness of the Random Forest algorithm in water quality classification, albeit with different focuses. Some studies performed binary classification

of water potability with high accuracy [5], predicted BOD values with 81.21% accuracy using DO, temperature, and pH as input features [19] or categorized groundwater quality into three classes using Conditional Inference Tree with 93.75% accuracy and 6.25% OOB error [8]. Other works employed PCA and Multi-Layer Perceptron (MLP) to assess water consumption feasibility [6], Other works employed PCA and Multi-Layer Perceptron (MLP) to assess water consumption feasibility [13], . However, these approaches primarily focused on general water quality status and not on filtration media selection based on data.

Another study using Random Forest with 10 water quality parameters also reported perfect results across all evaluation metrics [35]. However, the classification was limited to binary potable vs. non-potable status, without any linkage to filtration media or treatment system design.

As a distinctive contribution, this study proposes a classification approach for filtration media types (physical, chemical, biological) based on numerical water quality data simulated as greywater, utilizing an open dataset from Mendeley Data [14]. Media labels were derived using K-Means Clustering (k = 3), based on ten relevant water quality parameters. The classification model was subsequently developed using Random Forest with class_weight adjustment to address class imbalance, without applying synthetic oversampling methods such as SMOTE.

Evaluation was conducted comprehensively through a confusion matrix, classification report, 5-fold and 10-fold cross-validation, as well as learning and validation curve analyses. With this approach, the proposed model contributes as a data-driven classification logic that can be integrated into future smart biofiltration systems based on water quality sensor data.

4. CONCLUSION

This study successfully developed and evaluated a classification model for greywater filtration media using the Random Forest algorithm, incorporating ten water quality parameters as input features. The classification labels were derived through an unsupervised K-Means clustering approach, which grouped the data into three main filtration categories: physical, chemical, and biological.

The model demonstrated high performance with a macro F1-score consistently ranging from 0.97 to 0.98 on test data, supported by robust cross-validation results from 5-Fold and 10-Fold schemes. Learning curve and validation curve analyses confirmed the model's ability to generalize well without signs of overfitting, even as data volume and estimator count increased.

The dominant characteristics of each filtration type were successfully identified, indicating that the model can recognize relevant patterns in water quality data for each category. As a foundational step, this data-driven classification approach offers a promising basis for intelligent decision-making in smart biofiltration systems. The findings of this study are expected to contribute to the development of machine learning-based domestic wastewater treatment systems integrated with real-time water quality sensors. However, as the dataset used in this study was derived from aquaculture simulations rather than actual household greywater, further validation using real greywater samples is necessary to confirm the model's applicability under real-world conditions.

In future work, additional classification models and optimization strategies may be explored to enhance performance and scalability. Integrating the proposed model with real-time IoT-based water quality monitoring systems also presents a valuable direction for applied research.

REFERENCES

[1] M. Strokal *et al.*, "Urbanization: an increasing source of multiple pollutants to rivers in the 21st century," *npj Urban Sustainability*, vol. 1, no. 1, 2021, doi: 10.1038/s42949-021-00026-w.

[2] A. Mahmoudi, S. A. Mousavi, and P. Darvishi, "Greywater as a sustainable source for development of green roofs: Characteristics, treatment technologies, reuse, case studies and future developments," *J Environ Manage*, vol. 295, no. June, p. 112991, 2021, doi: 10.1016/j.jenvman.2021.112991.

[3] O. M. Ikumapayi, O. T. Laseinde, and E. T. Akinlabi, "An overview of sustainable greywater treatment processes," *E3S Web of Conferences*, vol. 552, pp. 1–16, 2024, doi: 10.1051/e3sconf/202455201047.

[4] I. I. S. Shamsuddin, Z. Othman, and N. S. Sani, "Water Quality Index Classification Based on Machine Learning: A Case from the Langat River Basin Model," *Water (Switzerland)*, vol. 14, no. 19, 2022, doi: 10.3390/w14192939.

[5] V. Kukartsev, V. Orlov, E. Semenova, and A. Rozhkova, "Optimizing water quality classification using random forest and machine learning," *BIO Web Conf*, vol. 130, pp. 1–10, 2024, doi: 10.1051/bioconf/202413003007.

[6] S. Y. Abuzir and Y. S. Abuzir, "Machine learning for water quality classification," *Water Quality Research Journal*, vol. 57, no. 3, pp. 152–164, 2022, doi: 10.2166/wqrj.2022.004.
[7] H. Sujadi, Nunu Nurdiana, and Reyna Indra Maulana, "Pengembangan Sistem Smart Village Berbasis Internet of Things untuk Meningkatkan Kualitas Hidup di Desa," *Journal of Applied Computer Science and Technology*, vol. 4, no. 2, pp. 141–146, 2023, doi: 10.52158/jacost.v4i2.474.
[8] S. V. S. Ganga Devi, "Random forest advice for water quality prediction in the regions of Kadapa District," *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, no. 6 Special Issue 4, pp. 1464–1466, 2019, doi: 10.35940/ijitee.F1298.0486S419.
[9] N. Maulidah, M. Maulidah, R. Supriyadi, H. Nalatissifa, S. Diantika, and A. Fauzi, "Prediksi Kualitas Air Menggunakan Metode Random Forest, Decision Tree, Dan Gradient Boosting," *Jurnal Khatulistiwa Informatika*, vol. 12, no. 1, pp. 1–6, 2024, doi: 10.31294/jki.v12i1.16004.
[10] K. Abdi, A. Warjaya, I. Muthmainnah, and P. H. Pahutar, "Penerapan Algoritma Random Forest dalam Prediksi Kelayakan Air Minum," *Jurnal Ilmu Komputer dan Informatika*, vol. 3, no. 2, pp. 81–88, 2024, doi: 10.54082/jiki.81.
[11] D. Hartanti and A. I. Pradana, "Komparasi Algoritma Machine Learning dalam Identifikasi Kualitas Air," *SMARTICS Journal*, vol. 9, no. 1, pp. 1–6, 2023, [Online]. Available: https://doi.org/10.21067/smartics.v9i1.8113
[12] N. H. A. Malek, W. F. W. Yaacob, S. A. M. Nasir, and N. Shaadan, "Prediction of Water Quality Classification of the Kelantan River Basin, Malaysia, Using Machine Learning Techniques," *Water (Switzerland)*, vol. 14, no. 7, 2022, doi: 10.3390/w14071067.
[13] X. Xu, T. Lai, S. Jahan, F. Farid, and A. Bello, "A Machine Learning Predictive Model to Detect Water Quality and Pollution," *Future Internet*, vol. 14, no. 11, pp. 1–14, 2022, doi: 10.3390/fi14110324.
[14] T. Veeramsetty, Venkataramana; Arabelli, Rajeshwarrao; Bernatin, "Aquaculture - Water Quality Dataset," 2024. doi: 10.17632/y78ty2g293.1.
[15] K. L. H. dan K. R. Indonesia, *Peraturan Menteri Lingkungan Hidup dan Kehutanan Republik Indonesia Nomor P.68/Menlhk-Setjen/2016 tentang Baku Mutu Air Limbah Domestik*. Indonesia, 2016.
[16] K. K. R. Indonesia, *Peraturan Menteri Kesehatan Republik Indonesia Nomor 2 Tahun 2023 tentang Standar Kegiatan Usaha dan/atau Produk pada Penyelenggaraan Perizinan Berusaha Berbasis Risiko Sektor Kesehatan*, no. 55. Indonesia: Kementerian Kesehatan Republik Indonesia, 2023.
[17] World Health Organization (WHO), *WHO Guidelines for the Safe Use of Wastewater, Excreta and Greywater: Volume 2 – Wastewater Use in Agriculture*, Third Edit. Geneva: World Health Organization (WHO), 2006.
[18] Menteri Kesehatan Republik Indonesia, *Peraturan Menteri Kesehatan Republik Indonesia Nomor 32 Tahun 2017 Tentang Standar Baku Mutu Kesehatan Lingkungan Dan Persyaratan Kesehatan Air Untuk Keperluan Higiene Sanitasi, Kolam Renang, Solus Per Aqua dan Pemandian Umum*. Indonesia: Kementerian Kesehatan Republik Indonesia, 2017, pp. 1–20. [Online]. Available: https://peraturan.bpk.go.id/Home/Details/112721/permenkes-no-32-tahun-2017
[19] R. Khare, A. A. Khurshid, A. Jain, and S. Shukla, "Predictive Sensor for Biological Oxygen Demand in water using Active Learning based Random Forest Algorithm," *NeuroQuantology*, vol. 20, no. 9, pp. 1983–1988, 2022, doi: 10.48047/nq.2022.20.9.nq44230.
[20] V. Kothari, S. Vij, S. K. Sharma, and N. Gupta, "Correlation of various water quality parameters and water quality index of districts of Uttarakhand," *Environmental and Sustainability Indicators*, vol. 9, no. December 2020, p. 100093, 2021, doi: 10.1016/j.indic.2020.100093.
[21] A. K. Shrestha and N. Basnet, "The Correlation and Regression Analysis of Physicochemical Parameters of River Water for the Evaluation of Percentage Contribution to Electrical Conductivity," *J Chem*, vol. 2018, 2018, doi: 10.1155/2018/8369613.
[22] A. Maulizar, M. Masykur, and J. Supardi, "ANALISIS pH, TDS, TOTAL HARDNESS, ALKALINITY, DAN SILICA PADA BOILER FEET WATER DI PT. SOCFINDO PERKEBUNAN KELAPA SAWIT DI SEUNAGAN," *Jurnal Mekanova: Mekanikal, Inovasi dan Teknologi*, vol. 8, no. 1, p. 129, 2022, doi: 10.35308/jmkn.v8i1.5630.
[23] K. Khofifah and M. Utami, "Analysis of total dissolved solid (TDS) and total suspended solid (TSS) levels in liquid waste from sugar cane industry," *Indonesian Journal of Chemical Research*, vol. 7, no. 1, pp. 43–49, 2022.
[24] M. I. Mashur and Y. Salim, "Analisis Performa Metode Cluster K-Means pada Dataset Ocular Disease Recognition," *Indonesian Journal of Data and Science*, vol. 3, no. 1, pp. 35–46, 2022, doi: 10.56705/ijodas.v3i1.47.
[25] H. Zou, Z. Zou, and X. Wang, "An enhanced K-Means algorithm for water quality analysis of the haihe river in China," *Int J Environ Res Public Health*, vol. 12, no. 11, pp. 14400–14413, 2015, doi: 10.3390/ijerph121114400.
[26] A. E. M. Celestino, D. A. M. Cruz, E. M. O. Sánchez, F. G. Reyes, and D. V. Soto, "Groundwater quality assessment: An improved approach to K-means clustering, principal component analysis and spatial analysis: A case study," *Water (Switzerland)*, vol. 10, no. 4, pp. 1–21, 2018, doi: 10.3390/w10040437.
[27] M. A. R. Hamed, "Application of surface water quality classification models using principal components analysis and cluster analysis," *Water and Energy International*, vol. 62r, no. 1, pp. 54–62, 2019, doi: 10.4236/gep.2019.76003.
[28] N. Laili, N. S. Indrasti, and D. Wahyudi, "Design of sustainable coffee processing wastewater treatment system using K-means clustering algorithm," *IOP Conf Ser Earth Environ Sci*, vol. 1063, no. 1, 2022, doi: 10.1088/1755-1315/1063/1/012032.
[29] K. Kaur, K. Singh, S. K. Gupta, and A. K. Tiwary, "Prediction of water quality and LULC analysis using machine learning and geospatial techniques," *Water Pract Technol*, vol. 20, no. 1, pp. 275–294, 2025, doi: 10.2166/wpt.2024.310.
[30] P. S. Varma and V. Anand, "Random Forest Learning Based Indoor Localization as an IoT Service for Smart Buildings," *Wirel Pers Commun*, vol. 117, no. 4, pp. 3209–3227, 2021, doi: 10.1007/s11277-020-07977-w.
[31] R. Tyasnurita and S. W. Hapsari, "Identification of Chronic Kidney Disease Using Naive Bayes, Adaboost, and Random Forest Learning Methods," *JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer)*, vol. 6, no. 1, pp. 115–120, 2020, doi: 10.33480/jitk.v6i1.1403.
[32] V. R. Danestiara and N. N. Azkiya, "PREDICTION OF INHIBITOR BINDING AFFINITY AND MOLECULAR INTERACTIONS IN MPRO DENGUE USING MACHINE LEARNING," vol. 10, no. 3, pp. 461–468, 2025, doi: 10.33480/jitk.v10i3.5994.PREDICTION.

[33] K. N. Khikmah, B. Sartono, B. Susetyo, and G. A. Dito, "Performance Comparative Study of Machine Learning Classification Algorithms for Food Insecurity Experience by Households in West Java," *Jurnal Online Informatika*, vol. 9, no. 1, pp. 128–137, 2024, doi: 10.15575/join.v9i1.1012.

[34] J. Asian, M. Dholah Rosita, and T. Mantoro, "Sentiment Analysis for the Brazilian Anesthesiologist Using Multi-Layer Perceptron Classifier and Random Forest Methods," *Jurnal Online Informatika*, vol. 7, no. 1, p. 132, 2022, doi: 10.15575/join.v7i1.900.

[35] S. M. Alomani, N. I. Alhawiti, and A. Alhakamy, "Prediction of Quality of Water According to a Random Forest Classifier," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 6, pp. 892–899, 2022, doi: 10.14569/IJACSA.2022.01306105.