JPPIPA 11(9) (2025)

Jurnal Penelitian Pendidikan IPA

http://jppipa.unram.ac.id/index.php/jppipa/index

Dimensionality Reduction in River Water Quality Classification Using Genetic Algorithm and Correlation-Based Feature Selection

Yudha Riwanto[1*], Fauzia Anis Sekar Ningrum[1]

[1] Department of Informatics, Faculty of Computer Science, Universitas Amikom Yogyakarta, Sleman, Yogyakarta, Indonesia.

Received: June 24, 2025
Revised: August 03, 2025
Accepted: September 25, 2025
Published: September 30, 2025

Corresponding Author:
Yudha Riwanto
yudha.riwanto@amikom.ac.id

DOI: 10.29303/jppipa.v11i9.11863

Abstract: Water quality monitoring is a crucial element in data-driven environmental management. This study aims to identify the most important parameters in river water quality classification through feature selection and machine learning approaches. Eleven physicochemical parameters were used as initial features, and two selection methods were applied: Genetic Algorithm (GA) and Spearman Rank Correlation (RS). Classification was performed using Radial Basis Function Support Vector Machine (RBF-SVM), with performance evaluation based on accuracy, F1 score, and recall. GA testing results identified influential parameters (pH, DHL, DO, BOD, COD, TSS, NO_2^--N), achieving an accuracy of 96.67% and an F1 score of 0.82. RS generated seven different features with an accuracy of 90.00% and an F1 score of 0.67. Both methods revealed five consistently significant features (DHL, BOD, COD, TSS, NO_2^--N), which are the influential features. The model without feature selection, despite producing high accuracy (93.33%), only achieved an F1 score of 0.48, indicating poor recognition of the minority class. These findings confirm that feature selection improves classification efficiency and capability. In conclusion, GA-based feature selection provides the most effective subset for water quality classification and supports the development of intelligent and cost-effective monitoring systems suitable for sensor-based field applications.

Keywords: Feature selection; Genetic algorithm; Spearman rank; Support vector machine; Water quality classification

Introduction

Water is a crucial natural resource containing various physical and chemical substances that can have positive or negative impacts on human health and ecosystems (Saidi et al., 2019). Good water quality is essential for public health and survival, but environmental pollution due to human activities is increasing and becoming a serious threat to water quality. This pollution is often indicated by changes in physical and chemical parameters of water such as pH, BOD, COD, and other contaminants.

Macrobenthos as biological indicators have often been used to evaluate water quality because they are sensitive to changes in the physical and chemical environment (Rosyadi et al., 2020; Nair & Vijaya, 2022). A decrease in the number of macrobenthos is often an early indication of pollution. Household and industrial waste containing hazardous chemicals and excessive organic matter is the main cause of the decline in the quality of river water and other water bodies (Santoso et al., 2021). Manual water quality monitoring is still often carried out, but this method faces challenges in terms of high time and cost as well as limited spatial coverage. Thus, methods using computational technology and

How to Cite:
Riwanto, Y., & Ningrum, F. A. S. (2025). Dimensionality Reduction in River Water Quality Classification Using Genetic Algorithm and Correlation-Based Feature Selection. *Jurnal Penelitian Pendidikan IPA*, *11*(9), 751–758. https://doi.org/10.29303/jppipa.v11i9.11863

machine learning algorithms are efficient choices for fast and precise water quality analysis (Abuzir & Abuzir, 2022; Gai & Guo, 2023; Iswanto et al., 2022; Su et al., 2015).

In predictive modeling, having too many unnecessary or repeated parameters can lower the accuracy of classification and make the process more expensive. Feature selection is important because it helps reduce the number of variables and improves how well the model works in real situations (Iswanto et al., 2022; Putri et al., 2025). Genetic Algorithm (GA) is a good method for finding the best set of features because it uses an approach inspired by evolution (Ileberi et al., 2022; Khatib Sulaiman et al., 2021; Onah et al., 2021). Spearman Rank Correlation is another method that helps understand how strongly different variables are connected. Past research shows that using feature selection with machine learning tools like Support Vector Machine SVM (Chen et al., 2021; Rizwan et al., 2021; Wu & Wang, 2022; Zheng et al., 2024) greatly improves results in tasks related to environmental classification (Onyelowe et al., 2022; Saidi et al., 2019; Zhu et al., 2022).

This study was done because there is a greater need for smart and affordable water quality monitoring systems. In many areas, it's not possible to check all chemical and physical factors at the same time due to limited resources. By finding the most important factors, the monitoring system can be made simpler without losing accuracy. This allows for creating more affordable and quicker sensor-based systems that can be used for real-time monitoring in the field. Because of this, this study looks at different feature selection methods to find out which key parameters are best at classifying river water quality.

Method

This study aims to identify the most influential attributes in determining the quality of raw water using the feature selection method, then classify water quality based on the selected attributes. The research method applied includes several steps as follows:

Data collection

The data used in this study are raw water quality parameter data consisting of several physical and chemical variables, namely Temperature, pH, DHL, DO, BOD, COD, TSS, NO3N, NO2N, Po4P, and Detergent. In addition, there is a target variable, namely status, which is the result of the classification of raw water quality. This raw data was obtained from the source Jasa Tirta 2.

Data Normalization

Raw data generally has noise and varying feature scales. Therefore, preprocessing steps are taken to handle consistent scale features and avoid the dominance of certain features, Min-Max normalization is performed:

$$X_i^1 \frac{X_1 - MIN(X)}{MAX(X) - MIN(X)} \quad (1)$$

With:
X_1 = original value of feature-*i.*
X_i^1 = value after normalization.

Feature Selection

Feature selection aims to reduce the dimensionality of data by selecting the most relevant features to make the classification model more effective and efficient.

Rank Spearman

Spearman Rank is one method used to find the relationship between two variables by ranking the data and calculating the distance between the variables (Chen et al., 2021; Jurnal et al., 2023; Li et al., 2021; Spearman, 1904). The calculation process includes determining the ranking of the data. The Spearman Rank formula was chosen because this formula compares ordinal data with ratio data (Diamantini et al., 2018; Mohamed et al., 2023; Omar et al., 2022).

Step by Step Workflow

Collect Data: Collect two sets of paired data X={x1,x2,…,xn} and Y={y1,y2,…,yn}, where each xi is paired with a corresponding yi.

Data Ranking

- Rank the data on X and Y individually
- Assign the ranks R(xi) to X and R(yi) to Y, starting at 1 for the smallest value.
- If there is a tie (repeated values), assign an average rank to those values.

Calculate Rank Difference: For each pair, calculate the difference between the ranks xi and yi:

$$d_i=R(x_i)-R(y_i) \quad (2)$$

Square the Difference in Ranks: For each pair, calculate the square of the difference in ranks.

$$d_i^2 = (R(x_i) - R(y_i))^2 \quad (3)$$

Sum of the Differences of Squares: Add up all the squared differences.

$$\sum d_i^2 \quad (4)$$

Apply Spearman's Rank Correlation Formula: Use the formula to calculate the Spearman's Rank correlation coefficient ρ:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \quad (5)$$

Where is:
- n is the number of pairs.
- $\sum d_i^2$ is the sum of the squared rank differences.

Genetic Algorithm

The Genetic Algorithm (GA) is an optimization method inspired by the process of natural selection and genetic mechanisms (Babatunde et al., 2014; Ileberi et al., 2022; Khatib Sulaiman et al., 2021). This method was first introduced by John Holland in 1975 as a solution search approach based on the principles of natural selection and genetics (Putri et al., 2025; Riwanto et al., 2023). GA operates through several main stages that resemble biological evolution. The following describes the workflow and formulas used in the GA process:

Population Initialization: The initial population consists of a number of individuals (potential solutions) represented by chromosomes. These chromosomes are typically encoded as binary strings (0s and 1s), although other formats such as real numbers may be used depending on the problem. Let the initial population consist of *N* individuals.

Initial population = $I_1, I_2,..., I_N$.

Strength Evaluation: Each individual in the population is evaluated using a fitness function to measure how well the solution solves the problem. The objective of GA is to either maximize or minimize this function. Let $f(I_i)$ denote the fitness value of the i-th individual. Then:

Fitness Function = $f(I_i)$

Selection: Individuals with higher fitness are more likely to be selected as parents to produce offspring. Some common selection techniques are:
- Roulette Wheel Selection: individuals are selected in proportion to their fitness values.
- Tournament Selection: A group of individuals are selected at random, and the best of the group is selected.

The probability of selecting an individual with a fitness value in a Roulette Wheel is:

$$P(Ii) = \frac{f(Ii)}{\sum_{j=1}^{N} f(Ij)} \quad (6)$$

Crossover: Crossover is the process of combining two parent individuals to produce new offspring. This occurs with a certain probability called the crossover probability (P_c).

Example: If the crossover point is at the third position, parts of the chromosome after that point are exchanged between the parents.

C= P_c X (two randomly selected individuals)

Mutation: introduces random changes to genes within a chromosome to maintain genetic diversity and prevent premature convergence. Mutation occurs with a certain probability called the mutation probability (P_m). Example: Mutating the second bit of the chromosome *1101010* results in *1001010*.

μ=P_m×(random change in a gene)

Population Replacement: A new population is formed after crossover and mutation. The process of selection, crossover, and mutation is repeated until a certain number of iterations or generations are reached or the best solution is found. Some of the best individuals from the previous generation are usually also retained to ensure that the best solution persists.

Termination, The algorithm is terminated if some conditions are met, such as reaching the maximum number of generations or if the best solution does not evolve over several iterations.

Classification

After feature selection using Genetic Algorithm and Spearman Rank Correlation, and the intersection of both, the next stage is to classify the quality of raw water using several machine learning algorithms. The goal is to measure how well the selected features can distinguish water quality classes.

Is a supervised machine learning method commonly employed for classification tasks because of its robust theoretical principles and efficiency in high-dimensional environments. SVM operates by locating the best hyperplane that divides data points from various classes with the largest margin (Awalullaili et al., 2023; Gai & Guo, 2023; Restiani & Purwadi, 2024). The margin is characterized as the space between the hyperplane and the closest data points from each class, referred to as support vectors.

In this research, the Radial Basis Function (RBF) kernel is employed, mapping input data into a higher-

dimensional space to address non-linear connections. The RBF kernel function is characterized as:

$$K(X_i, X_j) = EXP\left(-\gamma \left|\left|X_i - X_j\right|\right|^2\right) \quad (7)$$

where γ is a kernel parameter that determines the extent of a single training example's influence. A more compact γ indicates a broader impact, whereas a greater γ suggests a more limited one. The objective of SVM optimization is expressed as:

$$\underset{\boldsymbol{w}, b, \xi}{MIN} \ \frac{1}{2}||W||^2 + C\sum_{i=1}^{n} \xi_i \quad (8)$$

Subject to:

$$y_i\,(\boldsymbol{W}.\emptyset(X_i) + b) \geq 1 - \xi_i, \xi_i \geq 0 \quad (9)$$

Where: w is the weight vector, b is the bias term, ξ_i are slack variables for soft margin classification, C is the regularization parameter that controls the trade-off between maximizing the margin and minimizing the classification error.

In this study, SVM is utilized to assess the effects of feature selection through Genetic Algorithm (GA) and Spearman Rank (SR) (Andriani & Wihartiko, 2024; Awalullaili et al., 2023; Sakaa et al., 2022). Classification is conducted utilizing: Complete feature list (excluding selection), Features selected by GA, SR-chosen characteristics, and Intersected characteristics (GA ∩ SR). Performance metrics like accuracy, F1-score, and recall are calculated for every scenario to evaluate classification effectiveness, particularly when dealing with imbalanced class distribution (Koranga et al., 2021; Razaque et al., 2021). Employing SVM with an RBF kernel guarantees the capability to represent intricate patterns in water quality information.

Result and Discussion

This study aims to identify the most relevant parameters in water quality classification to select significant features. This study begins the water quality classification step by filtering input parameters from eleven initial physicochemical parameters, namely: Temperature, pH, DHL, DO, BOD, COD, TSS, NO_3^--N, NO_2N, $PO_4^{3}P$, and Detergent. Feature selection is carried out using two complementary approaches: genetic algorithm (GA) as an evolutionary method and Spearman Rank (RS) as a statistical method that focuses on correlation. In addition, the truncation of both methods is used to find the most consistent and prominent features. The GA method produces seven optimally selected features based on fitness performance for the classification function, namely: pH, DHL, DO, BOD, COD, TSS, and NO_2N. In contrast, RS finds seven characteristics that are significantly related to water quality, namely: Temperature, DHL, BOD, COD, TSS, NO_3N, and NO_2N. Of the two, five features emerged consistently: DHL, BOD, COD, TSS, and NO_2N. These features were then utilized as a description of the slice parameters for the third scenario.

The superiority of GA in this study is consistent with findings by Saidi et al. (2019) who showed that GA-based feature selection improves classification efficiency in environmental datasets. Similarly, Awalullaili et al., (2023) also demonstrated that GA-SVM outperforms traditional methods in handling complex medical data classification, supporting our observation that GA provides better feature subsets for improving predictive accuracy. On the other hand, the relatively lower performance of Spearman Rank selection aligns with Wu et al. (2022), who reported that correlation-based feature selection is effective but limited when feature interactions are nonlinear. Furthermore, Zhu et al. (2022) highlighted that combining statistical and evolutionary feature selection approaches can yield robust results, which is reflected in our intersection experiment where consistent features (DHL, BOD, COD, TSS, and NO_2N) maintained reliable performance despite reduced dimensionality.

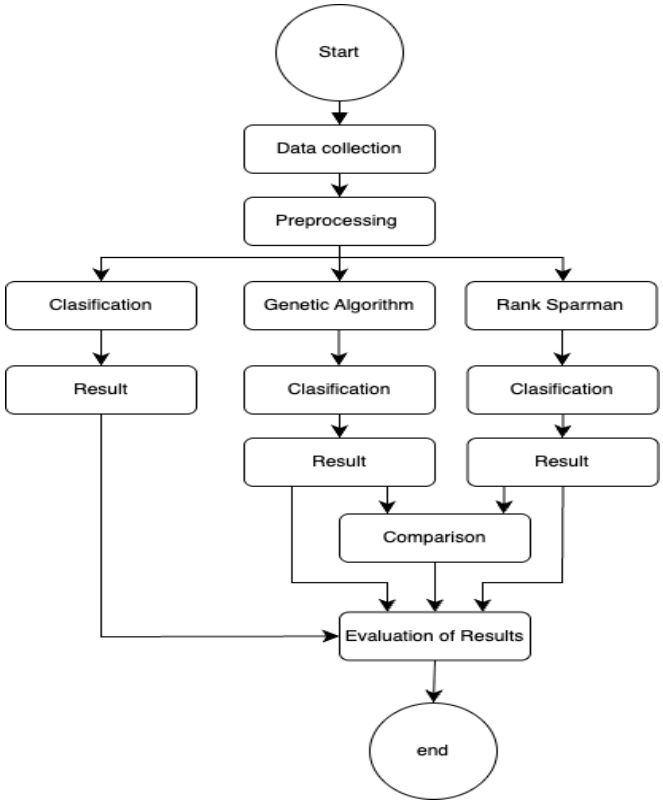

Figure 1. Flowchart diagram

Classification Performance Evaluation

Performance evaluation is performed using the Radial Basis Function Support Vector Machine (RBF-SVM) algorithm, measured on three main metrics: accuracy, F1-score, and recall. Table 1 presents a summary of the evaluation results from four scenarios: no feature selection, GA selection, RS selection, and GA ∩ RS slice features.

Table 1. Scenario Results

Scenario	Selected Features	Accuracy	F1	Recall
1	Temperature, pH, DHL, DO, BOD, COD, TSS, NO_3^--N, NO_2N, PO_4^3P, and Detergent	0.9333	0.4828	0.5000
2	pH, DHL, DO, BOD, COD, TSS, NO_2^--N	0.9667	0.8246	0.7500
3	Temperature, DHL, BOD, COD, TSS, NO_3^--N, NO_2^--N	0.9000	0.6727	0.7143
4	DHL, BOD, COD, TSS, NO_2^--N	0.9000	0.6727	0.7143

To further evaluate the distribution of predictions between classes, a confusion matrix is used for each scenario. The following is the confusion matrix for the existing scenarios.

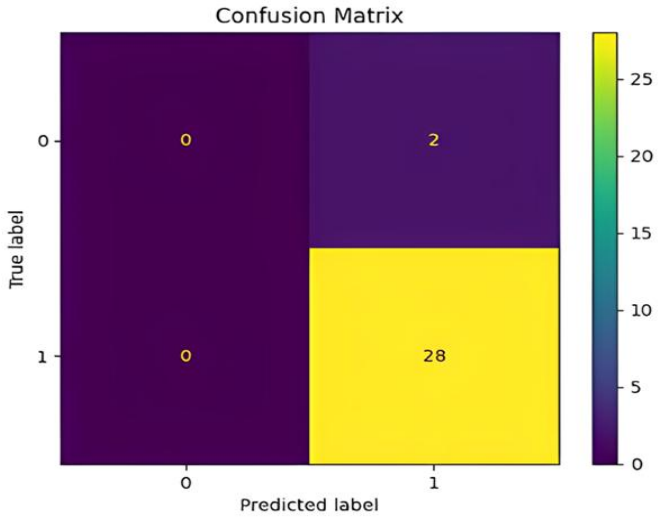

Figure 2. Confusion matrix without feature selection

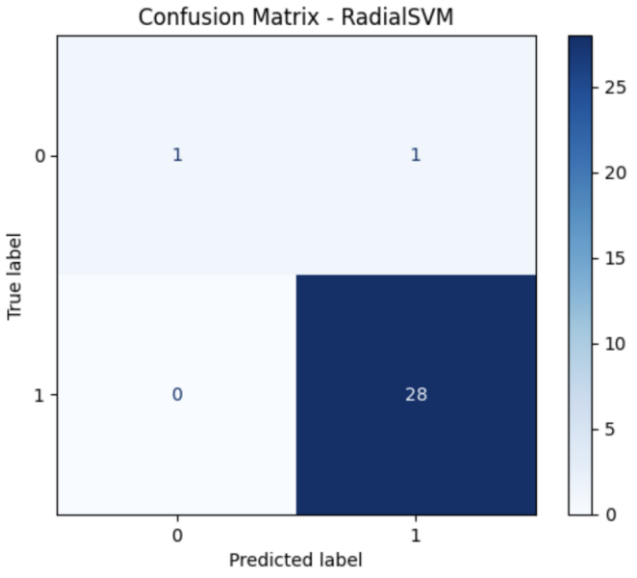

Figure 3. Confusion matrix genetic algorithm

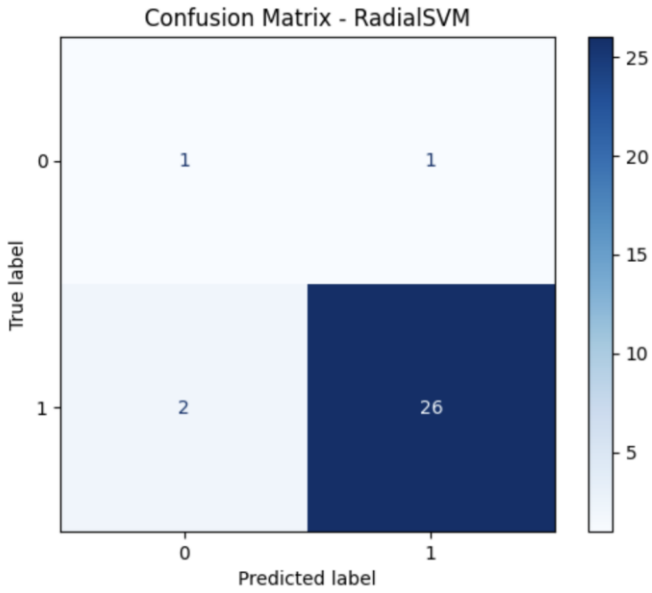

Figure 4. Confusion matrix rank spearman

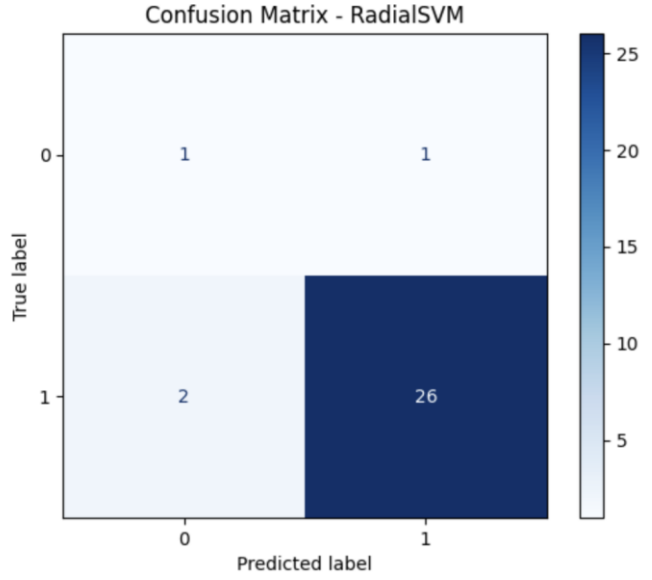

Figure 5. Confusion matrix genetic algorithm and rank spearman

The results show that the model without feature selection tends to experience overfitting, as seen from the high accuracy (93.33%) but low F1-score (0.48),

indicating an imbalance in prediction for the minority class. In contrast, feature selection using GA showed the best overall performance with significant improvements in F1-score (0.82) and recall (0.75), indicating a better ability to correctly recognize the target class.

Interestingly, the application of the slice feature (only five parameters) can still maintain the model performance on par with the RS method, even though the number of features is smaller. This confirms that selecting appropriate features is more important than the number of features in improving the effectiveness of classification.

Comparative Analysis of Selection Methods

The GA approach relies on heuristic exploration of the solution space based on classification performance, so that selected features directly improve prediction results. On the other hand, the RS method emphasizes linear statistical relationships, without considering the complexity of feature interactions in the classification context. Although RS is easier and faster, GA proves to be more appropriate for the ultimate goal of prediction. The feature intersection of both methods has strategic value, as it brings together statistically meaningful and predictively maximal features. This offers a compromise between efficiency and effectiveness, especially for the application of sensor-based real-time monitoring systems.

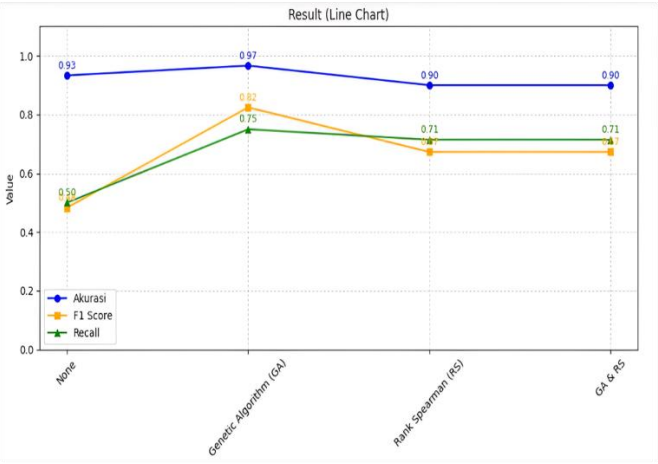

Figure 6. Result diagram

Implications of Findings

The main results of this study indicate that model efficiency can be significantly improved by performing feature selection without compromising prediction accuracy. Three to five key parameters such as DHL, BOD, COD, TSS, and NO_2N were shown to be sufficient to produce reliable water quality classification. This provides significant opportunities for the implementation of cost-effective and efficient sensor-based automated monitoring systems, targeting only important parameters.

Conclusion

This study shows that choosing the right features makes machine learning work better for classifying water quality. Out of eleven possible physical and chemical factors, the Genetic Algorithm (GA) picked the best seven features: pH, DHL, DO, BOD, COD, TSS, and NO_2^--N. These features gave the highest accuracy of 96.67% and an F1-score of 0.82. Spearman Rank (RS) found seven different features, but they performed less well. When both methods were compared, five features—DHL, BOD, COD, TSS, and NO_2^--N—were consistently important and helped classify water quality reliably. Models that didn't use feature selection often made mistakes and had lower F1-scores. GA was better at recognizing rare categories and made predictions more balanced. The study shows that picking the right features is better than using all possible ones. This has real-world value for creating cheaper, sensor-based systems that monitor water quality in real time.

Acknowledgments

Gratitude is expressed to Amikom University Yogyakarta and the Institute for Research and Community Service who have provided the facilities and resources needed to carry out this research. This support means a lot to us in exploring the topic of Water Quality Feature Selection in Recognizing River Water Quality Standard Status Patterns, as well as all parties involved in this research until the completion of this research.

Author Contributions

Yudha Riwanto is the first author to conduct a literature review on previous research, data collection, idea research, testing, analysis, and implementation, while Fauzia Sekar Anis Ningrum as the second author provides suggestions and input on the research concept.

Funding

This research was funded by Dikti through the Beginner Lecturer Research Grant (PDP), grant number 107/E5/PG.02.00.PL/2024 and APC was funded by the same funds.

Conflicts of Interest

The authors declare no conflict of interest.

References

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

Andriani, S., & Wihartiko, D. (2024). Comparison of Genetic Algorithm Optimization with Support

Vector Machine (SVM) for Weather Forecast Introduction. *Journal of Applied Science and Advanced Technology Journal Homepage*. https://doi.org/10.24853/JASAT.6.3.83-90

Awalullaili, F. O., Ispriyanti, D., & Widiharih, T. (2023). Klasifikasi Penyakit Hipertensi Menggunakan Metode SVM Grid Search dan SVM Genetic Algorithm (GA). *Jurnal Gaussian*, *11*(4), 488–498. https://doi.org/10.14710/j.gauss.11.4.488-498

Babatunde, O. H., Armstrong, L., Leng, J., & Diepeveen, D. (2014). Available here This Journal Article is posted at Research Online. *International Journal of Electronics Communication and Computer Engineering*, *5*(4), 899–905. Retrieved from https://ro.ecu.edu.au/ecuworkspost2013

Chen, B., Mu, X., Chen, P., Wang, B., Choi, J., Park, H., & Yang, H. (2021). Machine learning-based inversion of water quality parameters in typical reach of the urban river by UAV multispectral data. *Ecological Indicators*, *133*. https://doi.org/10.1016/j.ecolind.2021.108434

Diamantini, E., Lutz, S. R., Mallucci, S., Majone, B., Merz, R., & Bellin, A. (2018). Driver detection of water quality trends in three large European river basins. *Science of the Total Environment*, *612*, 49–62. https://doi.org/10.1016/j.scitotenv.2017.08.172

Gai, R., & Guo, Z. (2023). A water quality assessment method based on an improved grey relational analysis and particle swarm optimization multi-classification support vector machine. *Frontiers in Plant Science*, *14*. https://doi.org/10.3389/fpls.2023.1099668

Ileberi, E., Sun, Y., & Wang, Z. (2022). A machine learning based credit card fraud detection using the GA algorithm for feature selection. *Journal of Big Data*, *9*(1). https://doi.org/10.1186/s40537-022-00573-8

Iswanto, I., Tulus, T., & Poltak, P. (2022). Comparison Of Feature Selection To Performance Improvement Of K-Nearest Neighbor Algorithm In Data Classification. *Jurnal Teknik Informatika (Jutif)*, *3*(6), 1709–1716. https://doi.org/10.20884/1.jutif.2022.3.6.471

Khatib Sulaiman, J., Riau Taslim, P., Toresa, D., Jollyta, D., Suryani, D., & Sabna, E. (2021). Optimasi K-Means dengan Algoritma Genetika untuk Target Pemanfaat Air Bersih. *Indonesian Journal of Computer Science Attribution-ShareAlike*, *4*(1), 1. Retrieved from https://repository.uir.ac.id/22410/

Koranga, M., Pant, P., Pant, D., Bhatt, A. K., Pant, R. P., Ram, M., & Kumar, T. (2021). SVM Model to Predict the Water Quality Based on Physicochemical Parameters. *International Journal of Mathematical, Engineering and Management Sciences*, *6*(2), 645–659. https://doi.org/10.33889/IJMEMS.2021.6.2.040

Li, K., Huang, G., & Baetz, B. (2021). Development of a Wilks feature importance method with improved variable rankings for supporting hydrological inference and modelling. *Hydrology and Earth System Sciences*, *25*(9), 4947–4966. https://doi.org/10.5194/hess-25-4947-2021

Mohamed, S. A., Metwaly, M. M., Metwalli, M. R., AbdelRahman, M. A. E., & Badreldin, N. (2023). Integrating Active and Passive Remote Sensing Data for Mapping Soil Salinity Using Machine Learning and Feature Selection Approaches in Arid Regions. *Remote Sensing*, *15*(7). https://doi.org/10.3390/rs15071751

Nair, J. P., & Vijaya, M. S. (2022). River Water Quality Prediction and index classification using Machine Learning. *Journal of Physics: Conference Series*, *2325*(1). https://doi.org/10.1088/1742-6596/2325/1/012011

Omar, N., Aly, H., & Little, T. (2022). Optimized Feature Selection Based on a Least-Redundant and Highest-Relevant Framework for a Solar Irradiance Forecasting Model. *IEEE Access*, *10*, 48643–48659. https://doi.org/10.1109/ACCESS.2022.3171230

Onah, J. O., Abdulhamid, S. M., Abdullahi, M., Hassan, I. H., & Al-Ghusham, A. (2021). Genetic Algorithm based feature selection and Naïve Bayes for anomaly detection in fog computing environment. *Machine Learning with Applications*, *6*, 100156. https://doi.org/10.1016/j.mlwa.2021.100156

Onyelowe, K. C., Gnananandarao, T., & Ebid, A. M. (2022). Estimation of the erodibility of treated unsaturated lateritic soil using support vector machine-polynomial and -radial basis function and random forest regression techniques. *Cleaner Materials*, *3*. https://doi.org/10.1016/j.clema.2021.100039

Putri, A. S., Suhartanto, E., & Andawayanti, U. (2025). Validation of NRECA Parameters for Rainfall-to-Discharge Modeling in the Rejoso Watershed. *Jurnal Penelitian Pendidikan IPA*, *11*(5), 1081–1088. https://doi.org/10.29303/jppipa.v11i5.11107

Razaque, A., Ben Haj Frej, M., Almi'ani, M., Alotaibi, M., & Alotaibi, B. (2021). Improved support vector machine enabled radial basis function and linear variants for remote sensing image classification. *Sensors*, *21*(13). https://doi.org/10.3390/s21134431

Restiani, Y., & Purwadi, J. (2024). Support Vector Machine for Classification: A Mathematical and Scientific Approach in Data Analysis. *Jurnal Penelitian Pendidikan IPA*, *10*(11), 9896–9903. https://doi.org/10.29303/jppipa.v10i11.8122

Riwanto, Y., Nuruzzaman, M. T., Uyun, S., & Sugiantoro, B. (2023). Data Search Process Optimization using Brute Force and Genetic Algorithm Hybrid Method. *IJID (International Journal on Informatics for Development)*, *11*(2), 222–231. https://doi.org/10.14421/ijid.2022.3743

Rizwan, A., Iqbal, N., Ahmad, R., & Kim, D. H. (2021). Wr-svm model based on the margin radius approach for solving the minimum enclosing ball problem in support vector machine classification. *Applied Sciences (Switzerland)*, *11*(10). https://doi.org/10.3390/app11104657

Rosyadi, H. I., & Ali, M. (2020). Biomonitoring makrozoobentos sebagai indikator kualitas air sungai. *Envirotek: Jurnal Ilmiah Teknik Lingkungan, 12*(1), 11-18. https://doi.org/10.33005/envirotek.v12i1.43

Saidi, R., Bouaguel, W., & Essoussi, N. (2019). Hybrid feature selection method based on the genetic algorithm and pearson correlation coefficient. In *Studies in Computational Intelligence* (Vol. 801, pp. 3–24). Springer Verlag. https://doi.org/10.1007/978-3-030-02357-7_1

Sakaa, B., Elbeltagi, A., Boudibi, S., Chaffaï, H., Islam, A. R. M. T., Kulimushi, L. C., & Wong, Y. J. (2022). Water quality index modeling using random forest and improved SMO algorithm for support vector machine in Saf-Saf river basin. *Environmental Science and Pollution Research*, *29*(32), 48491–48508. https://doi.org/10.1007/s11356-022-18644-x

Santoso, T., Sutanto, A., & Achyani, A. (2021). Keanekaragaman Makrozoobentos Sebagai Bioindikator Kualitas Air Di Danau Asam Suoh Lampung Barat. *Bioedukasi (Jurnal Pendidikan Biologi), 12*(2), 213-220. http://dx.doi.org/10.24127/bioedukasi.v12i2.4450

Spearman, C. (1904). The Proof and Measurement of Association between Two Things. *American Journal of Psychology*, *15*, 45-58 https://psycnet.apa.org/doi/10.1037/11491-005

Su, J., Wang, X., Zhao, S., Chen, B., Li, C., & Yang, Z. (2015). A Structurally Simplified Hybrid Model of Genetic Algorithm and Support Vector Machine for Prediction of Chlorophyll a in Reservoirs. *Water (Switzerland)*, *7*(4), 1610–1627. https://doi.org/10.3390/W7041610

Wu, J., & Wang, Z. (2022). A Hybrid Model for Water Quality Prediction Based on an Artificial Neural Network, Wavelet Transform, and Long Short-Term Memory. *Water (Switzerland)*, *14*(4). https://doi.org/10.3390/w14040610

Zheng, Z., Jiang, Y., Zhang, Q., Zhong, Y., & Wang, L. (2024). A Feature Selection Method Based on Relief Feature Ranking with Recursive Feature Elimination for the Inversion of Urban River Water Quality Parameters Using Multispectral Imagery from an Unmanned Aerial Vehicle. *Water (Switzerland)*, *16*(7). https://doi.org/10.3390/w16071029

Zhu, M., Wang, J., Yang, X., Zhang, Y., Zhang, L., Ren, H., & Ye, L. (2022). A review of the application of machine learning in water quality evaluation. *Eco-Environment & Health, 1*(2), 107-116. https://doi.org/10.1016/j.eehl.2022.06.001