European Research Studies Journal
Volume XXVIII, Issue 3, 2025

pp. 1538-1552

Classification of Water Samples Using a Limited Set of Characteristics for the Economics of Water Resource Management

Submitted 15/08/25, 1st revision 29/08/25, 2nd revision 11/09/25, accepted 30/09/25

Paweł Ziemba[1], Izabela Gago[2], Rafał Różański[3]

Abstract:

Purpose: *The purpose of this article is to develop a framework for reducing water quality assessment parameters while maintaining the precision and accuracy of evaluation.*
Design/Methodology/Approach: *The study was based on machine learning methods, including classification and feature selection techniques. Specifically, data discretization, correlation-based feature selection, and the following classification algorithms were applied: artificial neural network, decision tree, random forest, and support vector machine. The study also utilized two datasets describing water quality based on physical, chemical, and biological parameters.*
Findings: *The conducted research indicates that the correlation-based feature selection filter, combined with other machine learning methods, is an effective tool that enables a substantial reduction in the number of parameters used for water quality assessment without any loss of accuracy or precision.*
Practical Implications: *The study's findings make it possible to optimize the water quality assessment process by reducing the number of required chemical, physical, and biological tests and analyses. This reduction can lower the costs of data collection, analysis, and interpretation, minimize data gaps, and increase monitoring frequency – thereby economically rationalizing water quality testing procedures.*
Originality/Value: *The development of a framework for reducing water quality assessment parameters while maintaining precision and accuracy constitutes a contribution to the field of water quality assessment methods and the economic rationalization of water quality testing.*

Keywords: *Water resources management, water quality, feature selection, classification.*
JEL codes: *Q53, C55, C38.*
Paper type: *Research article.*
Acknowledgement: *This research was co-financed by the Minister of Science under the "Regional Excellence Initiative.*

Regionalna Inicjatywa Doskonałości

Ministerstwo Nauki i Szkolnictwa Wyższego

[1]*University of Szczecin, Poland, e-mail: pawel.ziemba@usz.edu.pl;*
[2]*Maritime University of Szczecin, Poland, e-mail: i.gago@pm.szczecin.pl;*
[3]*Jacob of Paradyz University in Gorzów Wielkopolski, Poland,*
e-mail: rrozanski@ajp.edu.pl;

1. Introduction

Water is essential for all forms of life on Earth – from microorganisms, through plants and animals, to humans. Cells, entire organisms, and even whole ecosystems are fully dependent on the presence of water (Dargaville and Hutmacher, 2022). In the human body, water constitutes approximately 70% of its mass (Komatina, 2004), while on Earth, water covers about 70% of the planet's surface (Dargaville and Hutmacher, 2022).

Water is also a key natural resource, an economic asset, and an indispensable component of sustainable ecological development (Chen *et al.,* 2025). Its importance is evident in various aspects of human life, including agriculture, industry, energy production, and domestic use (Li *et al.,* 2024).

From an economic perspective, it should be noted that water is essential for maintaining the efficient functioning of the economy (Xu *et al.,* 2025) and plays a key role in promoting socio-economic progress (Li *et al.,* 2024). Since the Dublin Conference in 1992, water has been recognized as an economic good, emphasizing its role as a crucial factor of production across nearly all sectors of the economy (Jain and Sekar, 2025).

The use of water, both in industrial and non-industrial sectors, has a significant impact on GDP (Hao *et al.,* 2019). Studies indicate that industrial water consumption is positively correlated with economic development, constituting an important contribution to production processes and the growth of economic activity (Wang *et al.,* 2025). Conversely, water scarcity can constrain economic activity, adversely affecting development (Sun *et al.,* 2023; Wang *et al.,* 2025).

Water shortages are significant not only from an economic perspective but also from environmental and social viewpoints. Insufficient access to clean water for drinking, hygiene, and agriculture can have adverse consequences for human health and food security (Sahoo *et al.,* 2024), as contaminated water can serve as a vector for the spread of diseases (Zhengis *et al.,* 2025).

Moreover, water shortages, which most often affect lower social strata, may lead to conflicts over access to water and exacerbate social tensions (Debsarma and Sahu, 2025; Savelli *et al.,* 2021). From an environmental standpoint, limited water resources result in the overexploitation of available water sources within a given area. This leads to a decline in groundwater levels, a reduction in recharge and discharge rates, and consequently, a deterioration in the quality of available water (Gebremeskel Haile *et al.,* 2019).

Excessive exploitation of water resources, which causes the degradation of aquatic ecosystems, can also harm wildlife and water-dependent plants (Sahoo *et al.,* 2024).

Water is essential for maintaining ecological balance and preserving biodiversity, as it supports the survival of various plant and animal species (Li *et al.,* 2024).

Unfortunately, water scarcity is becoming an increasingly severe problem due to climate change, population growth, and the expansion of the global economy (Debaere and Kapral, 2021). Among the factors exacerbating water shortages and deteriorating water quality are urbanization, improper water resource management, and, ultimately, water pollution (Szołdrowska and Smol, 2025).

In fact, water pollution is one of the major causes of water scarcity. Natural contaminants such as pathogens, arsenic, fluoride, and others pose serious health risks, while anthropogenic pollutants further degrade water quality (Olmedo *et al.,* 2025). Clean and safe water, by contrast, is essential for human health, agricultural productivity, and ecological stability (Dharmarathne *et al.,* 2025). Therefore, ensuring adequate access to clean and safe water is of fundamental importance for sustainable development and for the well-being of both people and the planet (Li *et al.,* 2024).

The examination of water samples in terms of quality is a key method for ensuring the availability of clean and safe water. Standard testing methods allow for the assessment of various characteristics and contaminants in water for the purpose of evaluating water quality. Water quality parameters can be classified into:

- physical tests – analyzing water properties detectable by the senses, such as color, turbidity, odor, and taste,
- chemical tests – determining the concentration of mineral and organic substances that affect water quality,
- bacteriological tests – used to identify the presence of bacteria indicative of fecal contamination (Speight, 2020).

However, the large number of parameters to be tested requires substantial time and generates high costs related to data collection, analysis, and interpretation (Misaghi *et al.,* 2017). For practical reasons, analyzing a large number of parameters is not feasible, as such analyses can only be conducted by well-equipped laboratories (Fortes *et al.,* 2023).

Moreover, attempts to monitor numerous water quality parameters on a large scale typically result in data gaps and deficiencies, preventing real-time monitoring (Lesa Chundu *et al.,* 2024). Finally, an excessive number of parameters may lead to the inclusion of interdependent variables (Gikas *et al.,* 2020). Examining dependent parameters is redundant and leads to economically irrational use of laboratory resources. Thus, a research problem emerges—namely, the need to limit the number of water quality testing parameters to the essential minimum without compromising test quality.

Given the aforementioned research problem, the purpose of this article is to develop a framework for reducing water quality assessment parameters while maintaining precision and accuracy of evaluation.

The proposed framework is based on machine learning (ML) methods and algorithms, including classification and feature selection techniques. Achieving this objective simultaneously constitutes a contribution to the field of water quality assessment methods and the economic rationalization of water quality testing.

2. Literature Review

Water quality assessment involves analyzing its composition in terms of physical, chemical, and biological parameters. This process includes sample collection, evaluation of the significance of individual water parameters, and subsequent laboratory and statistical analyses (Jenifel, 2024). It should be noted that conventional water quality assessment methods are time-consuming and require several procedures to achieve the desired results (Karimzadeh Motlagh *et al.,* 2023).

Therefore, in recent years, ML methods have been increasingly applied in the water quality assessment process, offering an alternative to the traditional approach (Yan *et al.,* 2024). ML is considered an appropriate technique for determining relationships between various parameters and response factors. ML models are capable of processing large datasets, identifying complex patterns and dependencies, which translates into more accurate and efficient water quality prediction (Abbasi *et al.,* 2025).

Abbas *et al.* (2024) successfully applied ML models such as Gradient Boosting (GB), Extreme Gradient Boosting (XGB), Decision Trees (DT), Random Forest (RF), K-Nearest Neighbors (kNN), and Support Vector Machine (SVM) to classify water samples in Pakistan into different quality classes based on 12 distinct parameters.

Hassan *et al.* (2021) used ML to classify water quality data from various locations in India. They employed the following classifiers: Multinomial Logistic Regression (MLGR), Bagged Tree Model (BTM), RF, kNN, and SVM, utilizing six water parameters. Similarly, Nasir *et al.* (2022) employed classification models to assess water quality in India. The classifiers investigated included Artificial Neural Network (ANN), Categorical Boosting (CB), XGB, Logistic Regression (LGR), DT, RF, and SVM, with seven water parameters.

Sajib *et al.* (2023) developed an innovative model for assessing groundwater quality in Bangladesh with consideration of heavy metals. The study utilized regression models including ANN, Linear Regression (LR), Gaussian Process Regression (GPR), Ensembles of Trees (ET), Regression Trees (RT), and Support Vector Regression (SVR), using ten water quality parameters.

In a subsequent study, Sajib *et al.* (2024) employed regression models to predict water quality assessments in the Tongi Canal in Bangladesh. The study utilized an extended set of regression models compared to the previous work: ANN, XGB, LR, Step-wise Linear Regression (SLR), GPR, ET, RT, and SVR. Fifteen water parameters were used as regression variables.

Similarly, Asadollah *et al.* (2021) proposed a new ML model for predicting Water Quality Index (WQI) values using nine water parameters. The study compared RT, Extra Tree Regression (ETR), and SVR classifiers. Satish *et al.* (2024) also developed an innovative model for quantitatively assessing water quality in India.

For this purpose, they utilized an ANN meta-model and nine conventional regression models: Adaptive Boosting (ADB), Gradient Boosting (GB), Highest Gradient Boosting (HGB), Light Gradient Boosting Machine (LGBM), XGB, Bagging, Extra Tree Regression (ETR), DT, and RF. Four parameters were used for water quality evaluation. Guo *et al.* (2021) developed an approach for predicting dissolved oxygen (DO) levels in water based on data captured by satellite sensors.

The study employed the following regressors: Multiple Linear Regression (MLR), RF, and SVR. In another study, Guo *et al.* (2021) proposed a new method for monitoring water quality and predicting total phosphorus, total nitrogen, and chemical oxygen demand using remote sensing. To this end, they applied a similar set of regression models, including ANN, RF, and SVR.

Finally, Peng *et al.* (2024) demonstrated that multispectral remote sensing data can be used to predict four water parameters—chlorophyll-a, total phosphorus, total nitrogen, and chemical oxygen demand. Their study employed six different regression models: CB, LGBM, XGB, RF, SVR, and the Spatio-Temporal Ensemble Model (STE).

An analysis of the studies described above clearly indicates that they employed between four and fifteen parameters for water quality prediction. Among the applied ML methods, the most frequently used algorithms were ANN, XGB, various types of RT/DT, RF, and SVM/SVR. Furthermore, an examination of the cited publications reveals the main directions in which ML models have been applied in the field of water quality assessment. These include:

- classification of water into quality classes (Nasir *et al.,* 2022; Abbas *et al.,* 2024; Hassan *et al.,* 2021),
- numerical evaluation of water quality (Satish *et al.,* 2024; Sajib *et al.,* 2023; Sajib *et al.,* 2024; Asadollah *et al.,* 2021),
- prediction of water quality parameters based on remote sensing data (Peng *et al.,* 2024; Guo, Huang, Zhu *et al.,* 2021; Guo, Huang, Chen *et al.,* 2021).

However, there is a noticeable shortage of studies addressing the reduction of the number of parameters used for water quality assessment. This confirms the existence of a research gap in this area and justifies the research objective formulated in the introduction. Although some of the cited studies employed a limited number of water quality parameters, the selection of those parameters was made arbitrarily rather than through a systematic, methodological analysis.

3. Research Methodology

In our study, we used two datasets related to water quality, both obtained from the literature. The first dataset describes the quality of fish pond water and was previously used in a study on deep neural networks (Arabelli *et al.,* 2025). This dataset contains a total of 4,300 water samples divided into three quality classes. The samples are evenly distributed among the classes: 1,500 of 'poor' quality, 1,400 of 'good' quality, and 1,400 of 'excellent' quality.

Each sample is described using 14 parameters (conditional attributes) and one decision attribute indicating the water quality class. The parameters describing water quality include: (1) temperature, (2) turbidity, (3) dissolved oxygen, (4) biochemical oxygen demand, (5) carbon dioxide, (6) pH, (7) alkalinity, (8) hardness, (9) calcium, (10) ammonia, (11) nitrates, (12) phosphorus, (13) hydrogen sulphide, and (14) plankton. The second dataset, in turn, describes coastal water quality using various quantitative and qualitative water indices (Uddin *et al.,* 2023).

This dataset was modified by removing decision attributes that described water quality through different indices. Only the WQM (weighted quadratic mean of WQI) decision attribute was retained. The entire dataset comprises 10,000 water samples divided into three quality classes: 205 of 'marginal' quality, 5,061 of 'fair' quality, and 4,734 of 'good' quality. The dataset includes seven water quality parameters: (1) ammoniacal nitrogen, (2) biological oxygen demand, (3) chlorophyll-a, (4) dissolved inorganic nitrogen, (5) molybdate reactive phosphorus, (6) water temperature, and (7) transparency.

The datasets discussed above served as the foundation for the study on a parameter reduction framework designed to preserve the precision and accuracy of water quality assessment. The study employed classifiers commonly used in water quality evaluation problems: ANN, DT, RF, and SVM. Each classifier was evaluated using several representative metrics: accuracy, precision, F1 score, Cohen's kappa, Area Under the Receiver Operating Characteristic curve (AUROC), and Area Under the Precision-Recall curve (AUPRC) (Mostafaei *et al.,* 2024).

A five-fold data split was applied in the study for cross-validation. This approach is more effective than dividing the data into separate training and test sets, as it allows for better generalization of classification results, making them more robust to overfitting and to the influence of random data partitioning (Li *et al.,* 2026).

The classification process was preceded by preprocessing, during which data discretization was performed. In the discretization process, the values of each parameter were divided into k equal intervals. The number of intervals for each parameter was optimized by minimizing entropy estimated through leave-one-out cross-validation (Mondal *et al.,* 2017).

For each dataset, the study was conducted in two stages. In the first stage, classification was performed using the described preprocessing and 5-fold cross-validation. This procedure yielded reference classification results, which were later compared with the results obtained using the reduced set of water parameters. In the second stage, the preprocessing was extended by applying Correlation-based Feature Selection (CFS) following the discretization step.

CFS is a feature selection method belonging to the filter family, used to identify an optimal subset of conditional attributes based on their correlation values. CFS is based on the assumption that good feature subsets contain conditional attributes that are highly correlated with the class while being uncorrelated with each other (Galar and Kumar, 2017).

Thus, CFS operates in two ways—it eliminates irrelevant features with low correlation to the class and redundant features that are highly correlated with other attributes. In the developed framework, CFS is responsible for the potential reduction in the number of parameters used for water quality assessment. The choice of the CFS filter for this task was motivated by the fact that it is a fully automatic algorithm that does not require the specification of the number of selected features or correlation thresholds (Manouchehri *et al.,* 2024).

Additional arguments in its favor include its computational efficiency, as it does not require extensive resources, and its versatility, since it is independent of the classifier used and applicable to both numerical and nominal parameters (Pham *et al.,* 2019).

Separate classification of the data without and with the use of CFS enabled the assessment of the impact of parameter reduction on the obtained classification results. This approach made it possible to determine whether the reduction of parameters allowed for maintaining the precision and accuracy of water quality assessment.

Consequently, this two-stage procedure constituted the core of the framework and enabled direct verification of the research objective. It should be noted that when referring to accuracy and precision, these terms are not limited to the literal names of specific classification metrics. Rather, they are used in a broader sense, denoting the correctness and consistency of water quality assessment based on the entire set of classification metrics.

4. Results and Discussion

4.1 Research Based on the Quality of Fish Ponds Water Dataset

Tables 1 and 2 present the classification results obtained in the first stage of the study, that is, with preprocessing limited to data discretization and without the use of CFS for reducing water quality parameters. Table 1 contains the confusion matrices produced by each classifier. Their analysis indicates that ANN performed the classification task considerably worse than the other classifiers.

The remaining classifiers achieved a similar level of correct classifications, with SVM correctly classifying samples belonging to the two lower quality classes and misclassifying only those from the highest-quality class. The DT and RF classifiers made errors in predicting each class; however, similar to SVM, the number of misclassifications for the two lower classes was significantly smaller than for the highest class.

These observations are confirmed by Table 2, which presents the metrics of individual classifiers. The metric values demonstrate a clear advantage of the DT, RF, and SVM classifiers over ANN.

Moreover, all metrics indicate that RF achieved the best classification results, followed by DT in second place. The only exception is the precision value, where SVM outperformed DT. Nevertheless, the differences among the three best-performing classifiers are minimal, and, as mentioned earlier, all of them achieved a comparable level of classification quality.

Table 1. *Confusion matrices of individual classifiers obtained without the CFS filter for the first dataset*

ANN				
		Predicted class		
		1	2	3
Actual class	1	**924**	273	203
	2	3	**1011**	386
	3	93	448	**959**

DT				
		Predicted class		
		1	2	3
Actual class	1	**1396**	0	4
	2	1	**1327**	72
	3	159	291	**1050**

RF				
		Predicted class		
		1	2	3
Actual class	1	**1390**	0	10
	2	0	**1371**	29
	3	197	241	**1062**

SVM				
		Predicted class		
		1	2	3
Actual class	1	**1400**	0	0
	2	0	**1400**	0
	3	256	282	**962**

Note: *Classifier: ANN – Artificial Neural Network, DT – Decision Tree, RF – Random Forest, SVM – Support Vector Machine*
Source: *Own study.*

Table 2. *Classification metrics of the classifiers obtained for the classification without the CFS filter for the first dataset*

Classifier	Accuracy / Recall	Precision	F1 score	Cohen's kappa	AUROC	AUPRC
ANN	0.673	0.701	0.678	0.509	0.667	0.565
DT	0.877	0.884	0.873	0.817	0.933	0.869
RF	0.889	0.899	0.884	0.834	0.973	0.946
SVM	0.875	0.895	0.867	0.813	0.907	0.814

Source: *Own study.*

In the second stage of the study, the preprocessing was extended to include the application of CFS. The CFS filter generated a reduced set of water quality parameters containing 12 instead of 14 conditional attributes. The following parameters were removed from the set: (1) temperature and (10) ammonia.

This reduced dataset was then subjected to classification, analogously to the first stage. The classification results are presented in Tables 3 and 4. Table 3 contains the confusion matrices, while Table 4 presents the classifier metrics.

Table 3. *Confusion matrices of individual classifiers obtained with the CFS filter for the first dataset*

ANN				
		Predicted class		
		1	2	3
Actual class	1	**1359**	0	41
	2	0	**1350**	50
	3	91	87	**1322**

DT				
		Predicted class		
		1	2	3
Actual class	1	**1390**	0	10
	2	1	**1333**	66
	3	188	289	**1023**

RF				
		Predicted class		
		1	2	3
Actual class	1	**1370**	0	30
	2	0	**1346**	54
	3	186	235	**1079**

SVM				
		Predicted class		
		1	2	3
Actual class	1	**1400**	0	0
	2	0	**1400**	0
	3	246	280	**974**

Source: *Own study.*

Table 4. *Classification metrics of the classifiers obtained for the classification with the CFS filter for the first dataset*

Classifier	Accuracy / Recall	Precision	F1 score	Cohen's kappa	AUROC	AUPRC
ANN	0.937	0.937	0.937	0.906	0.971	0.947
DT	0.871	0.879	0.866	0.807	0.927	0.860
RF	0.883	0.887	0.878	0.824	0.959	0.914
SVM	0.878	0.897	0.870	0.817	0.909	0.817

Source: *Own study.*

An analysis of Table 3 reveals a marked improvement in the performance of the ANN classifier compared with the classification results obtained without the use of CFS. For the remaining classifiers, only minimal differences are observed relative to the reference classification results obtained in the first stage.

Furthermore, an examination of Table 4 and its comparison with Table 2 indicate that the application of CFS significantly improved the performance of the ANN classifier. The results of the SVM classifier also improved slightly (by no more than 0.004), while those of the DT and RF classifiers showed a minor decrease (by no more than 0.01).

However, in the case of SVM, DT, and RF, these differences are negligible compared to the reference classification, allowing the conclusion that the classification performance of these algorithms remained effectively unchanged relative to the baseline results.

4.2 Research Based on the Coastal Water Quality Dataset

Tables 5 and 6 present the classification results obtained without the use of CFS for reducing water quality parameters. Table 5 shows the confusion matrices produced by each classifier. An analysis of Table 5 reveals that the ANN and SVM classifiers struggle to handle the substantial imbalance in the number of samples belonging to the lowest water quality class.

As a result, these classifiers fail to correctly classify any samples as having ‘marginal’ water quality, with all samples from this class being incorrectly assigned to the ‘fair’ water quality category. The RF and DT classifiers perform only slightly better in classifying samples from the ‘marginal’ class, as the majority of these samples are still misclassified as ‘fair’ quality.

These misclassifications of the least represented water quality class have a major impact on the classification metrics summarized in Table 6. For the ANN and SVM classifiers, accuracy and precision metrics cannot be computed, since the denominators of their respective formulas contain zero due to the absence of correctly classified samples in the lowest water quality class. The remaining classifiers, DT and RF, achieve somewhat better results, with DT showing a slight advantage over RF.

Table 5. *Confusion matrices of individual classifiers obtained without the CFS filter for the second dataset*

ANN				
		Predicted class		
Actual class		1	2	3
	1	**0**	205	0
	2	0	**4197**	864

DT				
		Predicted class		
Actual class		1	2	3
	1	**19**	186	0
	2	4	**4180**	877

	3	0	531	**4203**

	3	0	352	**4382**

RF				
		Predicted class		
		1	2	3
Actual class	1	**44**	161	0
	2	39	**4188**	834
	3	0	543	**4191**

SVM				
		Predicted class		
		1	2	3
Actual class	1	**0**	205	0
	2	0	**3880**	1181
	3	0	352	**4382**

Source: *Own study.*

Table 6. *Classification metrics of the classifiers obtained for the classification without the CFS filter for the second dataset*

Classifier	Accuracy / Recall	Precision	F1 score	Cohen's kappa	AUROC	AUPRC
ANN	0.840	-	-	0.687	0.920	0.901
DT	0.858	0.860	0.851	0.723	0.922	0.894
RF	0.842	0.839	0.839	0.693	0.918	0.896
SVM	0.826	-	-	0.661	0.831	0.761

Source: *Own study.*

The application of CFS reduced the feature set to five water quality parameters by eliminating (4) dissolved inorganic nitrogen and (6) water temperature. The classification results obtained from this reduced dataset are presented in Tables 7 and 8. Table 7 contains the confusion matrices, while Table 8 presents the classifier metrics.

Table 7. *Confusion matrices of individual classifiers obtained with the CFS filter for the second dataset*

ANN				
		Predicted class		
		1	2	3
Actual class	1	**0**	205	0
	2	0	**4208**	853
	3	0	406	**4328**

DT				
		Predicted class		
		1	2	3
Actual class	1	**19**	186	0
	2	4	**4178**	879
	3	0	349	**4385**

RF				
		Predicted class		
		1	2	3
Actual class	1	**59**	146	0
	2	53	**4176**	832
	3	0	438	**4296**

SVM				
		Predicted class		
		1	2	3
Actual class	1	**0**	205	0
	2	0	**3880**	1181
	3	0	352	**4382**

Source: *Own study.*

An analysis of Table 7 indicates that little has changed with respect to the classification of the least represented class. The ANN and SVM classifiers still fail to correctly classify any samples belonging to the 'marginal' class. In this regard, the reduced set of parameters led only to a modest improvement in the performance of the RF classifier.

Table 8. *Classification metrics of the classifiers obtained for the classification with the CFS filter for the second dataset*

Classifier	Accuracy / Recall	Precision	F1 score	Cohen's kappa	AUROC	AUPRC
ANN	**0.854**	-	-	**0.713**	**0.929**	**0.910**
DT	0.858	0.860	0.851	0.723	0.922	0.894
RF	**0.853**	**0.851**	**0.850**	**0.715**	**0.926**	**0.905**
SVM	0.826	-	-	0.661	0.831	0.761

Source: *Own study.*

The remaining classification results are very similar to the reference results, and in the case of SVM, they are identical. These observations are confirmed by the analysis of Table 8. For SVM, the metric values are exactly the same as in the reference classification, while for the DT classifier, the changes are so minor that results reported with a precision of three decimal places cannot capture them. A slight improvement in all metric values, however, is observed for the ANN and RF classifiers.

5. Conclusions

The research results clearly indicate that reducing the number of water quality parameters by eliminating those identified by the CFS filter allows the level of accuracy and precision of assessment to be maintained. In each dataset, CFS identified two parameters for removal. Their elimination generally resulted in classification outcomes comparable to those obtained with the full dataset, with only minimal differences in results (SVM, DT, and RF in the first dataset; SVM and DT in the second dataset).

Moreover, for some classifiers, the removal of the indicated parameters positively affected classification performance (ANN in the first dataset; ANN and RF in the second dataset). This demonstrates that the developed framework employing CFS is an effective tool for efficiently reducing the number of parameters used in water quality assessment without compromising accuracy or precision.

The findings make it possible to optimize the water quality assessment process by reducing the number of required chemical, physical, and biological tests and analyses. This, in turn, can lower the costs of data collection, analysis, and interpretation, reduce the occurrence of data gaps, and increase monitoring frequency – thereby economically rationalizing water quality testing procedures.

Naturally, the study has certain limitations. Among them, the most significant is the use of only two datasets. Another limitation concerns the scope of the tested ML techniques, as the study employed only one type of discretization, a single feature selection method, and four classification methods. These limitations indicate directions for future research on the parameter reduction framework for water quality assessment. To enable broader generalization of the obtained results, the study should be repeated using a larger number of datasets and a wider range of discretization, feature selection, and classification techniques in order to validate the current findings.

References:

Abbas, F., Cai, Z., Shoaib, M., Iqbal, J., Ismail, M., Arifullah, Alrefaei, A.F., Albeshr, M.F. 2024. Machine Learning Models for Water Quality Prediction: A Comprehensive Analysis and Uncertainty Assessment in Mirpurkhas, Sindh, Pakistan. Water, 16(7), 941.

Abbasi, F., Kazemi, A., Badeenezhad, A., Moazamfard, M., Armand, R., Mohammadpour, A. 2025. Assessing the impact of reverse osmosis plant operations on water quality index improvement through machine learning approaches and health risk assessment. Results in Engineering, 25, 104363.

Arabelli, R., Bernatin, T., Veeramsetty, V. 2025. Water quality assessment for aquaculture using deep neural network. Desalination and Water Treatment, 321, 101016.

Asadollah, S.B.H.S., Sharafati, A., Motta, D., Yaseen, Z.M. 2021. River water quality index prediction and uncertainty analysis: A comparative study of machine learning models. Journal of Environmental Chemical Engineering, 9(1), 104599.

Chen, Y., He, Z., Niu, X., Huang, D. 2025. How water resource management policies shape the coupled coordination development of the water-energy-food nexus: Evidence from the dual pathways of taxation and property rights. Journal of Environmental Management, 381, 125311.

Dargaville, B.L., Hutmacher, D.W. 2022. Water as the often neglected medium at the interface between materials and biology. Nature Communications, 13(1), 4222.

Debaere, P., Kapral, A. 2021. The potential of the private sector in combating water scarcity: The economics. Water Security, 13, 100090.

Debsarma, C., Sahu, P. 2025. Enhancement of urban drought resilience using green infrastructural strategies. In: Handbook of Nature-Based Drought Solutions. Elsevier, 231-252.

Dharmarathne, G., Abekoon, A.M.S.R., Bogahawaththa, M., Alawatugoda, J., Meddage, D.P.P. 2025. A review of machine learning and internet-of-things on the water quality assessment: Methods, applications and future trends. Results in Engineering, 26, 105182.

Fortes, A.C.C., Barrocas, P.R.G., Kligerman, D.C. 2023. Water quality indices: Construction, potential, and limitations. Ecological Indicators, 157, 111187.

Galar, D., Kumar, U. 2017. Preprocessing and Features. In eMaintenance. Essential Electronic Tools for Efficiency, 129-177.

Gebremeskel Haile, G., Tang, Q., Sun, S., Huang, Z., Zhang, X., Liu, X. 2019. Droughts in East Africa: Causes, impacts and resilience. Earth-Science Reviews, 193, 146-161.

Gikas, G.D., Sylaios, G.K., Tsihrintzis, V.A., Konstantinou, I.K., Albanis, T., Boskidis, I. 2020. Comparative evaluation of river chemical status based on WFD methodology and CCME water quality index. Science of The Total Environment, 745, 140849.

Guo, H., Huang, J.J., Chen, B., Guo, X., Singh, V.P. 2021. A machine learning-based strategy for estimating non-optically active water quality parameters using Sentinel-2 imagery. International Journal of Remote Sensing, 42(5), 1841-1866.

Guo, H., Huang, J.J., Zhu, X., Wang, B., Tian, S., Xu, W., Mai, Y. 2021. A generalized machine learning approach for dissolved oxygen estimation at multiple spatiotemporal scales using remote sensing. Environmental Pollution, 288, 117734.

Hao, Y., Hu, X., Chen, H. 2019. On the relationship between water use and economic growth in China: New evidence from simultaneous equation model analysis. Journal of Cleaner Production, 235, 953-965.

Hassan, Md. Mehedi, Hassan, Md. Mahedi, Akter, L., Rahman, Md.M., Zaman, S., Hasib, K.Md., Jahan, N., Smrity, R.N., Farhana, J., Raihan, M., Mollick, S. 2021. Efficient Prediction of Water Quality Index (WQI) Using Machine Learning Algorithms. Human-Centric Intelligent Systems, 1(3), 86-97.

Jain, S., Sekar, A. 2025. Unveiling the flow: A comprehensive exploration of virtual water trade globally and in India. In: Water Sustainability and Hydrological Extremes. Elsevier, 63-74.

Jenifel, G.M. 2024. Secure water quality prediction system using machine learning and blockchain technologies. Journal of Environmental Management, 350, 119357.

Karimzadeh Motlagh, Z., Derakhshani, R., Sayadi, M.H. 2023. Groundwater vulnerability assessment in central Iran: Integration of GIS-based DRASTIC model and a machine learning approach. Groundwater for Sustainable Development, 23, 101037.

Komatina, M.M. 2004. Other Natural Factors. In: Developments in Earth and Environmental Sciences. Elsevier, 236-262.

Lesa Chundu, M., Banda, K., Sichingabula, H.M., Nyambe, I.A. 2024. Integrated water quality assessment of open water bodies using empirical equations and remote sensing techniques in Bangweulu Wetland lakes, Zambia. Journal of Great Lakes Research, 50(6), 102451.

Li, S., Yang, H., Zhang, X., Sun, M., Liu, Y., Song, Y. 2026. Rapid prediction of heat transfer characteristics in regenerative cooling channels of ram/scramjet using machine learning. Aerospace Science and Technology, 168, 111088.

Li, Z., Wang, G., Lin, D., Mashhadi, A. 2024. Hybrid approach for accurate water demand prediction using socio-economic and climatic factors with ELM optimization. Heliyon, 10(3), e25028.

Manouchehri, L., Zinati, Z., Nazari, L. 2024. Population-Specific gene expression profiles in prostate cancer: insights from Weighted Gene Co-expression Network Analysis (WGCNA). World Journal of Surgical Oncology, 22(1), 177.

Misaghi, F., Delgosha, F., Razzaghmanesh, M., Myers, B. 2017. Introducing a water quality index for assessing water for irrigation purposes: A case study of the Ghezel Ozan River. Science of The Total Environment, 589, 107-116.

Mondal, D., Kole, D.K., Roy, K. 2017. Gradation of yellow mosaic virus disease of okra and bitter gourd based on entropy based binning and Naive Bayes classifier after identification of leaves. Computers and Electronics in Agriculture, 142, 485-493.

Mostafaei, S.H., Tanha, J., Sharafkhaneh, A. 2024. A novel deep learning model based on transformer and cross modality attention for classification of sleep stages. Journal of Biomedical Informatics, 157, 104689.

Nasir, N., Kansal, A., Alshaltone, O., Barneih, F., Sameer, M., Shanableh, A., Al-Shamma'a, A. 2022. Water quality classification using machine learning algorithms. Journal of Water Process Engineering, 48, 102920.

Olmedo, S.I., Valeggia, C.R., Palavecino, C., Pérez-Escamilla, R. 2025. "I Never Don't Have Water Because I Collect Rainwater": Domains of Water Insecurity and Their Sociocultural Correlates in an Indigenous Community of Northern Argentina. Current Developments in Nutrition, 9(9), 107519.
Peng, C., Xie, Z., Jin, X. 2024. Using Ensemble Learning for Remote Sensing Inversion of Water Quality Parameters in Poyang Lake. Sustainability, 16(8), 3355.
Pham, L.T.H., Vo, T.Q., Dang, T.D., Nguyen, U.T.N. 2019. Monitoring mangrove association changes in the Can Gio biosphere reserve and implications for management. Remote Sensing Applications: Society and Environment, 13, 298-305.
Sahoo, S.K., Sahu, A.K., Baitharu, I. 2024. Environmental and health risks assessment of chemical pollutants in drinking water and wastewaters. In: Emerging Innovative Trends in the Application of Biological Processes for Industrial Wastewater Treatment. Elsevier, 291-311.
Sajib, A.M., Diganta, M.T.M., Moniruzzaman, Md., Rahman, A., Dabrowski, T., Uddin, M.G., Olbert, A.I. 2024. Assessing water quality of an ecologically critical urban canal incorporating machine learning approaches. Ecological Informatics, 80.
Sajib, A.M., Diganta, M.T.M., Rahman, A., Dabrowski, T., Olbert, A.I., Uddin, M.G. 2023. Developing a novel tool for assessing the groundwater incorporating water quality index and machine learning approach. Groundwater for Sustainable Development, 23, 101049.
Satish, N., Anmala, J., Rajitha, K., Varma, M.R.R. 2024. A stacking ANN ensemble model of ML models for stream water quality prediction of Godavari River Basin, India. Ecological Informatics, 80, 102500.
Savelli, E., Rusca, M., Cloke, H., Di Baldassarre, G. 2021. Don't blame the rain: Social power and the 2015-2017 drought in Cape Town. Journal of Hydrology, 594.
Speight, J.G. 2020. Remediation technologies. In: Natural Water Remediation. Chemistry and Technology.
Sun, D., Shi, M., Wei, J., Chen, Z. 2023. Economic contribution and rebound effect of industrial water: The case of the Yangtze River Delta. Water Resources and Economics, 42, 100222.
Szołdrowska, D., Smol, M. 2025. The current state of water resources in Poland - possibilities of water reuse and management by the circular economy. Desalination and Water Treatment, 323, 101287.
Uddin, M.G., Nash, S., Rahman, A., Olbert, A.I. 2023. Performance analysis of the water quality index model for predicting water state using machine learning techniques. Process Safety and Environmental Protection, 169, 808-828.
Wang, C., Shuai, C., Chen, X., Sun, J., Zhao, B. 2025. Revealing the heterogeneity in the relationship between water consumption and economic output of enterprises across sectors and regions in China. Journal of Hydrology, 660, 133481.
Xu, Y., Yang, L., Sun, K., Zhu, J. 2025. Synergistic security relationships and risk measurement of water resources-social economy-ecological environment in Beijing-Tianjin-Hebei region. Ecological Indicators, 175, 113512.
Yan, Y., Zhang, Y., Yang, S., Wei, D., Zhang, J., Li, Q., Yao, R., Wu, X., Wang, Y. 2024. Optimized groundwater quality evaluation using unsupervised machine learning, game theory and Monte-Carlo simulation. Journal of Environmental Management, 371, 122902.
Zhengis, A., Yergesheva, A., Amrenova, Y., Abutalip, M., Nuraje, N. 2025. Zwitterionic SBMA-based nanofiltration membranes: Synthesis techniques, antifouling mechanisms, and future perspectives. Polymer, 337, 128944.