Damanhour Journal of Intelligent Systems and Informatics
Volume 1 – Issue 1

An Intelligent Approach to Water Quality Status Classification Using Supervised Machine Learning Algorithms.

Nasser Tamim [1]

[1] Faculty of information Technology and Computer, Sinia University, Egypt

Abstract: The classification of water quality status is the first step towards ensuring safe water for agricultural fields, manufacturing, and daily consumption, including drinking water. Water quality is essential for the survival of humans, animals, and plants. Recently, artificial intelligence techniques, particularly supervised machine learning, have been utilized to develop predictive water quality models. In this paper, we propose a method based on supervised learning that employs a 20-dimensional feature vector along with several supervised machine learning classifiers. Eight classifiers are included in this study: Non-Linear Support Vector Machine (Non-SVM), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), K-Nearest Neighbours (KNN), Decision Tree (DT), Multilayer Perceptron Neural Networks (MLP-NN), AdaBoost, and Random Forest (RF). The 20-dimensional feature vector, which encodes relevant information, is used to train each classifier for binary classification. Additionally, three different cross-validation strategies are employed in the evaluation process. The proposed method is tested using publicly available datasets, and the experimental results—both visual and quantitative—demonstrate the robustness of the approach.

Keywords: Water Quality status, Sustainable development, Supervised Machine Learning.
.

1. Introduction

Water quality is a critical environmental factor influencing public health, biodiversity, and the overall sustainability of ecosystems. The rapid expansion of urbanization, industrial activities, and agricultural practices has led to increased pollution, making it essential to monitor water bodies effectively. Moreover, freshwater is vital for the survival of all living organisms on earth. It popularly represents life — birth, fertility, and food. Even though natural water is abundant worldwide, just 3 % of freshwater is suitable for plant growth and human use [1]. Unfortunately, even this tiny percentage is not without pollutants nor in quality states enough, that is due to the influence of a factory or sewage treatment plant, “point source pollution”. Moreover, widespread sources, like nutrients and pesticides from farming activities and pollutants released by industry into the air, fall back to land and sea, are so-called “diffuse pollution”.

Traditional water quality monitoring methods, often based on manual sampling and laboratory analysis, are time-consuming, costly, and limited in their ability to provide real-time insights. Consequently, there is a growing need for innovative solutions that can automate and expedite water quality assessments.

In recent years, machine learning (ML) techniques have garnered significant attention in environmental monitoring [2], particularly in the classification and prediction of water quality status. Supervised machine learning algorithms, which rely on historical data to train models and predict outcomes, have proven to be effective tools in identifying patterns in water quality parameters and classifying water bodies into distinct quality categories. These techniques have the potential to revolutionize water quality monitoring by enabling timely, accurate, and cost-efficient assessments across large geographical areas.

This study proposes an intelligent approach that leverages supervised machine learning algorithms to classify water quality status based on key water parameters. By utilizing a combination of widely used classifiers, including support vector machines (SVM), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), K-nearest neighbors, decision trees (DT), multilayer perceptron neural networks, ada boast and random forests (RF) classifiers.

This study aims to develop a robust framework capable of accurately determining the status of water bodies in real time. The proposed approach is designed to improve the predictive accuracy of water quality classification, offering valuable insights for environmental managers, policymakers, and researchers seeking to enhance water resource management and protection efforts.

In this paper, we present the methodology, results, and implications of applying these advanced machine-learning techniques to water quality classification, aiming to contribute to the growing field of data-driven environmental monitoring. Figure 1 represents the proposed model for the water quality status classification using the above-mentioned machine learning algorithms.

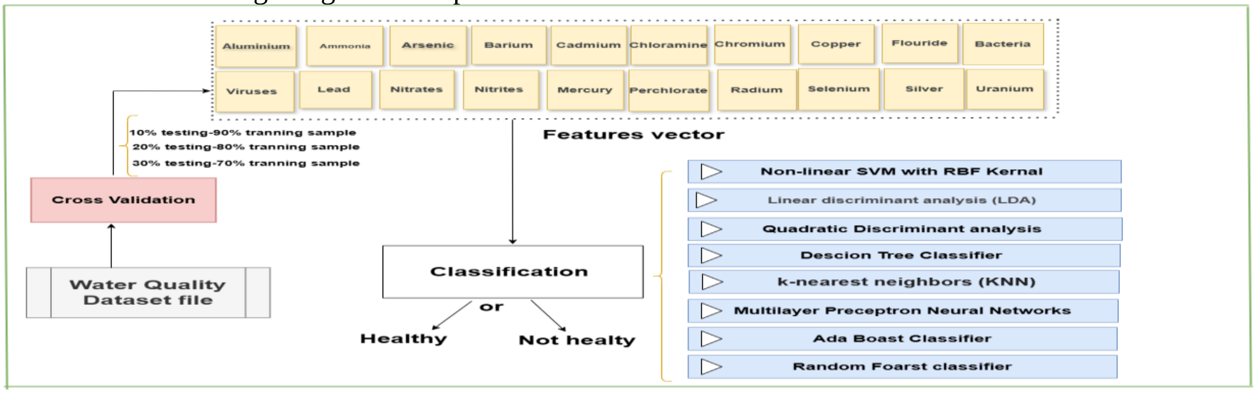

Figure 1. The proposed model for water quality status classification using a band of effective supervised machine learning algorithms.

2. The Research Method

Despite the abundance of applications in the field of artificial intelligence and machine learning, there exists a lack of research related to the classification of water quality status classifications [3]. However several machine learning algorithms have been employed to classify water quality, including supervised and unsupervised learning techniques. These algorithms are trained using labeled data, where the correct classification labels are known. We briefly mentioned some of this work: firstly, related to supervised learning; Decision Trees (DT): used for classifying water quality based on various parameters [4]. A comprehensive framework that integrates Internet of Things (**IoT**) technology and Machine Learning (**ML**) techniques for monitoring and assessing water quality in agricultural settings is proposed. This framework consists of four primary modules: sensing, coordination, data processing, and decision-making. To collect essential water quality data, a series of sensors are deployed along the Rohri Canal and Gajrawah Canal in Nawabshah City. These sensors measure various parameters, including temperature, pH, turbidity, and Total Dissolved Solids (TDS). The data collected is then analyzed using ML algorithms to evaluate the Water Quality Index (WQI) and classify water quality into different categories [5]. Support Vector Machines (SVM): A classification method that works well with high-dimensional data and has been used for classifying water quality status in various studies [6]. K-Nearest Neighbours (k-NN): A simple algorithm that classifies water quality based on the closest training examples in the feature space [7]. A deep learning method - Long short-term memory recurrent neural networks (LSTM RNNs) to produce an intelligent model for drinking water quality classification with principal component analysis (PCA), linear discriminant analysis (LDA), and independent component analysis (ICA) techniques were used for features extraction and data reduction for water quality classification are used in [8].

3. Methodology

This section provides detailed information on the proposed model, data description is described in more detail in Table 1, machine learning models, and evaluation metrics.

3.1 Data set description

The dataset consists of 8000 samples, and each has 20 features. The last one is the label feature that is used to decide if the feature vector is related to healthy or not healthy. So, the dataset is related to the supervised branch where the last feature of the dataset is labeled as healthy or not healthy.

3.2 A supervised machine learning and cross-validation strategies:

A bunch of supervised machine learning algorithms are used to train the proposed models, with three cross-validation techniques strategy, 10-90, 20- 80, 30-70 where 10, 20, and 30 are the percentages of the test sample while the remainder is for the training dataset. In this work and according to the linearly separable check test, we decided to use linear supervised learning algorithms because the linearly separable test declares the non-linearity of datasets as discussed in the next part.

3.3 Model evaluation.

The final stage involved evaluating the performance of the proposed model using the metrics presented in section —— and discussing its relevance and limitations.

Table 1. The statistical characteristics of the Water Quality Status dataset

	Aluminum	Ammonia	Arsenic	Barium	Cadmium
Count	*7996*	*7996*	*7996*	*7996*	*7996*
Mean	*0.6663957*	*14.278*	*0.1614*	*1.567*	*0.0428*
Std	*1.2653*	*8.8798032*	*0.252632*	*1.2162*	*0.0360*
25%	*0.04*	*6.5775*	*0.03*	*0.56*	*0.008*
50%	*0.07*	*14.13*	*0.05*	*1.19*	*0.023*
75%	*0.28*	*22.1325*	*0.1*	*2.4825*	*0.07*
Max	*5.05*	*29.84*	*1.05*	*4.94*	*0.13*
	Chloramine	**Chromium**	**Copper**	**Fluoride**	**Bacteria**
Count	*7996*	*7996*	*7996*	*7996*	*7996*
Mean	*2.1775*	*0.2472*	*0.8059*	*0.771646*	*0.3197*
Std	*2.56720*	*0.27066*	*0.6535.*	*0.435423*	*0.3294.*
25%	*0.01*	*0.05*	*0.09*	*0.4075*	*0*
50%	*0.53*	*0.09*	*0.75*	*0.77*	*0.22*
75%	*4.24*	*0.44*	*1.32*	*1.16*	*0.61*
Max	*8.68*	*0.9*	*2*	*1.5*	*1*
	Viruses	**Lead**	**Nitrates**	**Nitrites**	**Mercury**
Count	*7996*	*7996*	*7996*	*7996*	*7996*
Mean	*0.32870*	*0.09943*	*9.81925*	*1.329846*	*0.005193*
Std	*0.37811*	*0.05816*	*5.541977*	*0.573271*	*0.002967*
25%	*0.002*	*0.048*	*5*	*1*	*0.003*
50%	*0.1*	*0.08*	*9.3*	*1.3*	*0.004*
75%	*0.7*	*0.151*	*14.61*	*1.76*	*0.008*
Max	*1*	*0.2*	*19.83*	*2.93*	*0.01*
	Perchlorate	**Radium**	**Selenium**	**Silver**	**Uranium**
Count	*7996*	*7996*	*7996*	*7996*	*7996*
Mean	*16.465266*	*2.920106*	*0.04968*	*0.14718*	*0.04467*
Std	*17.68827*	*2.328051*	*0.028773*	*0.143569*	*0.020696*
25%	*2.17*	*0.82*	*0.02*	*0.04*	*0.03*
50%	*7.75*	*1.47*	*0.04*	*0.08*	*0.04*
75%	*24.75*	*4.05*	*0.07*	*0.14*	*0.06*
Max	*60.01*	*7.99*	*0.1*	*0.65*	*0.09*

In this work, a set of data created from imaginary data on water quality in urban environments is used to quantify our proposed model. It consists of 20 features or elements including their toxic concentration. Fig. 2 shows the dangerous dosage for each element that is not allowed to increase this concentration in the human daily intake of freshwater that is used for drinking. The dataset consists of 8000 samples with 21 features : (Aluminum, Ammonia, arsenic, Barium, Cadmium, Chloramine, Chromium, Copper, Fluoride, Bacteria, Viruses, Nitrites, Mercury, Perchlorate, Radium, Selenium, Silver, and Uranium). The final feature is the decision feature. If the water is as healthy as freshwater for human use or not healthy, the proposed model will act as a binary Classification. Table 2 illustrates the data illustrate the dataset`s features with the maximum and dangerous concentrations per Milliliters must be avoided for human life in their daily intake of fresh water.

3.4 Linearly separable check

To decide which supervised algorithm is the match fit for classification, a linearity separable test is done, where Linear Separable implies the existence of a hyperplane separating the two classes. We used a support virtual machine with a linear kernel of a sklearn for this purpose and the results identified that the dataset cannot separate linearly, therefore the dataset under this study is completely non-linearly separable. Figure 2 shows the output points that are not possible to draw a line that can separate the red and blue points from each other [2].

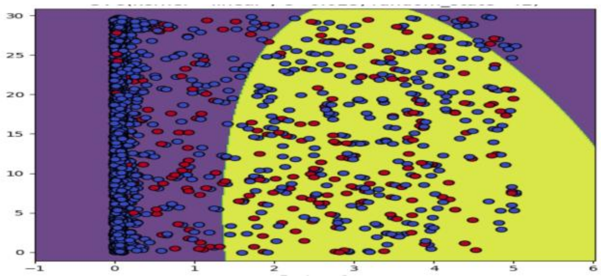

Figure 2. The output for the linearity test is clear that the hyperplane of the tester fails to separate the two different points that represent the two classes of water quality

3.5 Cross-validation strategies

Cross-validation is the most powerful statistical technique used in machine learning to evaluate the model's performance by splitting the data into two subsets, where the first part is used for training the dataset and the other for testing. Therefore, cross-validation is crucial for fine-tuning model parameters and guaranteeing high predictive accuracy. Table 3 shows the number of samples for both training and testing at the different cross-validation strategies that are used. Various cross-validation methods can be used, but the most suitable ones for the proposed model according to the characteristics of the used dataset are 10- 90, 20-80, and 30-70 cross-validation strategies. In the 10-90 cross-validation strategy, the dataset is divided into two subsets, 10% for testing and 90% for training. The same is true for the selected cross-validation techniques 20%, 30% for testing samples, and 80%, 70% for training samples.

3.6 Classification

Based on training data, the Classification algorithm is a Supervised Learning technique used to categorize new observations. In classification, a program uses the dataset or observations provided to learn how to categorize new observations into various classes or groups. In the proposed model, a feature vector is characterized by a vector in a 20-D feature space for each observation.

$$F(x, y) = \quad f_1(x, y), \ \dots , \quad f_{20}(x, y) \qquad (1)$$

A classification procedure assigns each candidate feature vector to one of two classes: ”Healthy” or “Non-Healthy” once its representation is known. To select a suitable classifier, it's important to analyze the distribution of the training set data in the feature space[9]. Determining the most appropriate supervised machine learning algorithm can be a challenging task Because of this complexity, experimentation is often the best approach for identifying the algorithm that can deliver satisfactory performance. After conducting thorough analyses and balancing speed, performance, accuracy, and complexity across several classifiers, we found that this dilemma was effectively resolved by a combination of supervised machine learning algorithms. These algorithms include linear discriminant analysis, quadratic discriminant analysis, decision trees, k-nearest neighbors, non-linear support vector machines, multilayer perceptron neural networks, Ada-Boost, and random forest classifiers. These classifiers have demonstrated high performance on the datasets under study.

Table 2. Dangerous Concentrations of Various Substances per ml/litter must not exceed this level.

Feature	Aluminum	Ammonia	Arsenic	Barium	Cadmium
Concentration	2.8	32.5	0.01	2	0.005
Feature	Chloramine	Chromium	Copper	Fluoride	Bacteria
Concentration	4	0.1	0.3	1.5	0
Feature	Viruses	Lead	Nitrites	Nitrites	Mercury
Concentration	0	0.015	10	1	0.002
/Feature	Perchlorate	Radium	Selenium	Silver	Uranium
Concentration	56	5	0.5	0.1	0.3

Table 3. The output for the linearity test is clear that the hyperplane of the tester fails to separate the two different points that represent the two classes of water quality.

Cross Validation	Training Sample	Test Sample
10-90 Strategy	7196	800
20-80 Strategy	6396	1600
30-70 Strategy	5597	2399

4 Results and Discussion

This section presents the results and discussion of the proposed machine learning models for classifying the water quality status into two distinct categories: healthy and not-healthy We have tested the proposed system by experimenting extensively on water quality status datasets, originating from publicly available datasets. Our experiments were run on a Colab platform including VMs with a standard system memory profile and using Python 3. The eight classifiers considered in the present work, Non-linear SVM, Linear Discriminant Analysis, Quadratic Discriminant Analysis, Decision Tree, k-nearest neighbor, Multilayer Perceptron neural networks, AdaBoost, and Random Forest, were tested separately to classify the water quality status dataset. A 2 × 2 confusion matrix is typically used to evaluate the performance of a classification model, especially in binary classification tasks. It summarizes the results of the model's predictions by comparing them to the actual outcomes (true labels). Here's the structure of a 2 × 2 confusion matrix:

1 True Positive (TP). The model correctly predicted the positive class.

2 False Negative (FN): The model incorrectly predicted the negative class when the actual class was positive.

3 False Positive (FP): The model incorrectly predicted the positive class when the actual class was negative.

4 True Negative (TN): The model correctly predicted the negative

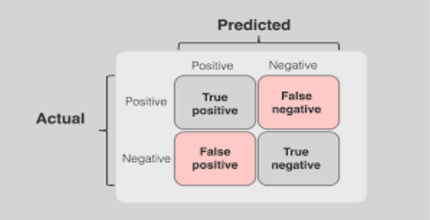

Figure 4. The confusion matrix for the classification outputs.

To quantify the classification results, we use the following standard metrics: accuracy (***Acc***), Precision (***Pr***), and Recall (**Re**).

$$\boldsymbol{Pr} = TP/ TP+FP, \qquad (2)$$

$$\boldsymbol{Re} = TP / TP+FN, \qquad (3)$$

$$\boldsymbol{Acc}=TP+TN/TP+FP+TN+FN. \qquad (5)$$

Tables 4, 5, and 6 show the accuracy, precision, and recall results over the eight different supervised machine learning that deployed in the for the Quality Water Status for our proposed model.

Table 4. The accuracy of the classification performance under the three different cross-validation strategies.

Classifier	10-90 cross-validation	20–80 cross-validation	30–70 cross-validation
Non-Linear SVM (None-SVM)	**0.9325**	**0.9375**	**0.9399**
Linear Discriminant Analysis (LDA)	**0.8937**	**0.8912**	**0.8932**
Quadratic Discriminant Analysis (QDA)	**0.8725**	**0.8562**	**0.8670**
K-Nearest Neighbours (KNN)	**0.8812**	**0.8725**	**0.8716**
Decision Tree (DT)	**0.9475**	**0.9500**	**0.9508**
Multilayer perceptron neural networks (MLP-NN)	**0.9139**	**0.9068**	**0.9003**
Ada Boost	**0.9225**	**0.9287**	**0.9271**
Random Forest (RF)	**0.9576**	**0.9887**	**0.9513**

Table 5. The Precision of the classification performance under the three different cross-validation strategies

Classifier	10-90 cross-validation	20–80cross-validation	30–70cross-validation
Non-Linear SVM (None-SVM)	***0.9419***	***0.9476***	***0.9492***
Linear Discriminant Analysis (LDA)	***0.9309***	***0.9277***	***0.9284***
Quadratic Discriminant Analysis (QDA)	***0.9496***	***0.9500***	***0.9560***
K-Nearest Neighbours (KNN)	***0.9028***	***0.9003***	***0.8993***
Decision Tree (DT)	***0.9537***	***0.9613***	***0.9588***
Multilayer perceptron neural networks (MLP-NN)	***0.9210***	***0.9128***	***0.9075***
Ada Boost	***0.9342***	***0.9429***	***0.9383***
Random Forest (RF)	***0.9576***	***0.9622***	***0.9651***

Table 6. The recall of the classification performance under the three different cross-validation strategies

Classifier	10-90 cross-validation	20–80cross-validation	30–70cross-validation
Non-Linear SVM (None-SVM)	***0.9844***	***0.9838***	***0.9849***
Linear Discriminant Analysis (LDA)	***0.9506***	***0.9513***	***0.9529***
Quadratic Discriminant Analysis (QDA)	***0.9040***	***0.8843***	***0.9808***
K-Nearest Neighbours (KNN)	***0.9028***	***0.9003***	***0.8993***
Decision Tree (DT)	***0.9887***	***0.9830***	***0.9868***
Multilayer perceptron neural networks (MLP-NN)	***0.9873***	***0.9949***	***0.9882***
Ada Boost	***0.9816***	***0.9788***	***0.9821***
Random Forest (RF)	***0.9887***	***0.9887***	***0.9905***

Table 7. The classification results as a confusion matrix containing: the number of correctly classified as a *TP* and *TN* beside the misclassification as an *FP* and *FN* for each given classifier.

Classifier-SVM / Cross-validation	***TP***	***TN***	***FP***	***FN***
10–90	*698*	*11*	*43*	*48*
20–80	*1395*	*23*	*77*	*105*
30–70	*2093*	*32*	*112*	*162*
Classifier-LDA / Cross-validation	***TP***	***TN***	***FP***	***FN***
10–90	*674*	*25*	*50*	*41*
20–80	*1394*	*69*	*105*	*77*
30–70	*2025*	*100*	*156*	*118*
Classifier-QDA / Cross-validation	***TP***	***TN***	***FP***	***FN***
10–90	*641*	*68*	*34*	*57*
20–80	*1254*	*164*	*66*	*116*
30–70	*1893*	*232*	*87*	*187*
Classifier-kNN / Cross-validation	***TP***	***TN***	***FP***	***FN***
10–90	*686*	*21*	*74*	*17*
20–80	*1365*	*53*	*151*	*31*
30–70	*2046*	*79*	*229*	*45*
Classifier-DT / Cross-validation	***TP***	***TN***	***FP***	***FN***
10–90	*701*	*8*	*34*	*57*
20–80	*1394*	*24*	*56*	*126*
30–70	*2097*	*28*	*90*	*184*
Classifier-MLPNN / Cross-validation	***TP***	***TN***	***FP***	***FN***
10–90	*700*	*9*	*60*	*31*
20–80	*1403*	*15*	*134*	*48*
30–70	*2100*	*25*	*214*	*60*
Classifier-Ada boost / Cross-validation.	***TP***	***TN***	***FP***	***FN***
10–90	*696*	*13*	*49*	*42*
20–80	*1388*	*30*	*84*	*98*
30–70	*2087*	*38*	*137*	*137*
Classifier-RF / Cross-validation	***TP***	***TN***	***FP***	***FN***
10–90	*701*	*8*	*28*	*63*
20–80	*1402*	*16*	*47*	*135*
30–70	*2103*	*22*	*84*	*190*

According to the quantitative results of the classification using the aforementioned supervised machine learning classifiers, the accuracy, precision, and recall metrics are presented in Tables 4, 5, and 6. The overall performance of the eight classifiers utilized in this study is impressive, particularly considering that the dataset was not processed in any way. To further demonstrate the strong performance of the proposed method, we calculated the confusion matrix for each classifier's output. This matrix details the number of test samples categorized as true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), as outlined in Table 7.

5. Conclusion

In conclusion, this study reveals the possibility of machine learning techniques for practical water quality classification. By leveraging algorithms such as Non-Linear SVM (None-SVM), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), K-Nearest Neighbours (KNN), Decision Tree (DT), Multilayer perceptron neural networks (MLP-NN), Ada Boost, and Random Forest (RF). We have demonstrated that machine learning can provide proper and efficient predictions of water quality parameters. The results indicate that these models can effectively classify water quality based on key variables, offering a promising tool for real-time monitoring and decision-making. Forthcoming research should focus on improving model interpretability, developing the dataset to include diverse environmental conditions, and exploring the integration of sensor networks for continuous monitoring. Overall, this approach facilitates more accessible and scalable water quality assessment, supporting the sustainable management of water resources not only locally but also globally. Machine learning algorithms are proving to be effective tools for classifying water quality status, offering faster, more accurate, and automated methods for monitoring and assessment. However, challenges such as data quality, model generalization, and class imbalance need to be addressed for better performance. The future of water quality classification lies in integrating multiple data sources, improving model robustness, and leveraging advanced ML models like deep learning to further enhance accuracy and scalability.

References

[1] S. Y. Abuzir and Y. S. Abuzir, "Machine learning for water quality classification," *Water Quality Research Journal*, vol. 57, pp. 152-164, 2022.

[2] N. Tamim, M. Elshrkawey, G. Abdel Azim, and H. Nassar, "Retinal blood vessel segmentation using hybrid features and multi-layer perceptron neural networks," *Symmetry*, vol. 12, p. 894, 2020.

[3] M. Zhu, J. Wang, X. Yang, Y. Zhang, L. Zhang, H. Ren, *et al.*, "A review of the application of machine learning in water quality evaluation," *Eco-Environment & Health*, vol. 1, pp. 107-116, 2022.

[4] F. Aydin, R. Durgut, M. Mustu, and B. Demir, "Prediction of wear performance of ZK60/CeO2 composites using machine learning models," *Tribology International*, vol. 177, p. 107945, 2023.

[5] M. A. Rahu, M. M. Shaikh, S. Karim, S. A. Soomro, D. Hussain, and S. M. Ali, "Water Quality Monitoring and Assessment for Efficient Water Resource Management through Internet of Things and Machine Learning Approaches for Agricultural Irrigation," *Water Resources Management*, pp. 1-42, 2024.

[6] M. Mustapha, M. Zineddine, E. Kaufman, L. Friedman, M. Gmira, K. U. Majikumna, *et al.*, "A hybrid machine learning approach for imbalanced irrigation water quality classification," *Desalination and Water Treatment*, vol. 321, p. 100910, 2025.

[7] M. V. S. Kotwal, S. Pati, and J. Patil, "Review On Ai And Iot Based Integrated Smart Water Management And Distribution System," *Educational Administration: Theory and Practice*, vol. 30, pp. 594-605, 2024.

[8] S. Dilmi and M. Ladjal, "A novel approach for water quality classification based on the integration of deep learning and feature extraction techniques," *Chemometrics and Intelligent Laboratory Systems*, vol. 214, p. 104329, 2021.

[9] N. Tamim, M. Elshrkawey, and H. Nassar, "Accurate diagnosis of diabetic retinopathy and glaucoma using retinal fundus images based on hybrid features and genetic algorithm," *Applied Sciences*, vol. 11, p. 6178, 2021.