Water Quality Classification Using SVM And XGBoost Method

Hasriq Izzuan Hasnol Yusri, A'zraa Afhzan Ab Rahim, Siti Lailatul Mohd Hassan, Ili Shairah Abdul Halim, Noor Ezan Abdullah
College of Engineering
Universiti Teknologi MARA
Shah Alam, Selangor
hasriq99@gmail.com, azraa@uitm.edu.my, sitilailatul@uitm.edu.my, shairah@uitm.edu.my, noor_ezan@uitm.edu.my

2022 IEEE 13th Control and System Graduate Research Colloquium (ICSGRC) | 978-1-6654-6806-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICSGRC55096.2022.9845143

***Abstract*— Various pollutants have been endangering water quality over the past decades. As a result, predicting and modeling water quality have become essential to minimizing water pollution. This research has developed a classification algorithm to predict the water quality classification (WQC). The WQC is classified based on the water quality index (WQI) from 7 parameters in a dataset using Support Vector Machine (SVM) and Extreme Gradient Boosting (XGBoost). The results from the proposed model can accurately classify the water quality based on their features. The research outcome demonstrated that the XGBoost model performed better, with an accuracy of 94%, compared to the SVM model, with only a 67% accuracy. Even better, the XGBoost resulted in only 6% misclassification error compared to SVM, which had 33%. On top of that, XGBoost also obtained consistent superior results from 5-fold validation with an average accuracy of 90%, while SVM with an average accuracy of 64%. Considering the enhanced performance, XGBoost is concluded to be better at water quality classification.**

Keywords*— *SVM, XGBoost, water quality, machine learning, classification

I. INTRODUCTION

Water is the most important resource for life, as it is required for the survival of living organisms and humans. From The United Nations world water development report 2017, approximately 80% of the world's wastewater is discharged back into the environment, largely untreated, damaging rivers, lakes, and seas [1]. The widespread problem of water pollution is endangering our health. If nothing is done, the problems will only worsen by 2050, when global freshwater demand is estimated to be one-third more than it is currently [2].

Water quality monitoring through classification has become an important method to control the pollution level. A manual calculation is used to determine the water class when classifying water quality. The manual analysis is inefficient since it takes a long time to complete the operation. As a result, an improved system to classify water quality is necessary [3]. An effort to ensure water quality is the development of a water quality monitoring system.

By utilizing parameters such as the concentration of dissolved oxygen, bacteria levels, pH value, conductivity, Biochemical Oxygen Demand (BOD), Chemical Oxygen Demand (COD), fecal coliform, and total coliform, the system can calculate and classify the pollution of the water to a certain degree [4]. The process involves using an algorithm to calculate the data collected from rivers or lakes.

Support vector machines (SVM) is a kernel approach used to solve classification and regression problems [5]. Hyperplanes are used by SVMs, which were originally developed for binary classification, to construct decision boundaries between data points of different classes. The SVM algorithm is good at generalizing the sample, dealing with nonlinear problems in a simple but highly accurate way, and differentiating data that differs from the norm; therefore, it's ideal for classifying and evaluating water quality data [4][6].

Extreme gradient boosting or XGBoost is a gradient boosting framework from a decision-tree-based ensemble Machine Learning algorithm. XGBoost can be used to solve problems involving regression, classification, ranking, and user-defined prediction. In various machine learning and data mining challenges, the impact of XGBoost has been widely recognized, even for water classification [7].

Water quality needs to be studied comprehensively because of its importance in daily life and its effects on human health. Water quality monitoring offers the objective data needed to make informed decisions about water quality management now and in the future[8]. As a result, water quality must be monitored to ensure that no contaminants exceed levels that are hazardous to human health. Water quality is currently determined through costly and time-consuming lab and statistical analyses. It necessitates sample collection, transportation to labs, and a significant amount of time and calculation, which is ineffective given that water is a highly communicable medium and hazardous if water contaminated with disease-causing waste is not prevented earlier [9]. In this case, an alternative method based on machine learning for the efficient prediction and classification of water quality levels is implemented. We propose the implementation of SVM and XGBoost as an alternative method for water quality classification.

II. LITERATURE REVIEW

In a recent research by Bouamar and Ladjal [10], SVM was compared against Artificial Neural Network (ANN) for water quality classification. The research was conducted with

two classified states, whether the water was drinkable or not. SVM performed better with an error of 4% compared to the ANN model with a 7% error, owing to SVM's characteristic of being more robust against noise [10]. Another research from China that used SVM for water quality monitoring system found that the outcome of the SVM classification is superior with 99.2% accuracy, compared to Logistic regression Multi-layered perceptron (MLP) and Radial Basis Function (RBF) [6]. In [3], SVM water quality classification competed against K-Nearest Neighbour (KNN), a method for classifying objects based on the learning data closest to the object. The study found that SVM performed better with a 92.40% accuracy. In comparison, KNN only achieved 71.28% accuracy by using 10-fold cross-validation on 120 datasets. SVM obtained the highest accuracy from the linear kernel, and KNN's highest accuracy is when K=7 [3].

As mentioned in the Introduction section, another classification method considered in this research is the XGBoost. A study conducted by researchers from Turkey found that XGBoost performed better at water quality classification for accuracy, precision, and recall with 95%, 96%, and 96%, respectively, against LogitBoost, RF, and AdaBoost [11]. Another study on water quality classification using a tree-based ensemble model, XGBoost, is on par with the LightGBM classification method with 85% accuracy classification compared to other tree-based learning models such as Classification tree, Random Forest, and CatBoost [12]. In [13], XGBoost performed moderately with 88% accuracy compared to CatBoost, which had the highest accuracy of 94% but still performed better than SVM, with only 80% accuracy.

From the literature reviewed, SVM and XGBoost have been applied to water quality classification; however, their execution is non-trivial. Issues such as the constraint of limited water quality parameters [14] and the algorithm's robustness when dealing with noises [15] are still actively investigated. Thus, water quality classification using the SVM and XGBoost should be given more attention.

III. METHODOLOGY

This section discusses the classification algorithms of SVM and XGBoost, whose performances are evaluated in the water quality classification. The methodology includes the project workflow, method description, and evaluation performance criteria, illustrated in Figure 1.

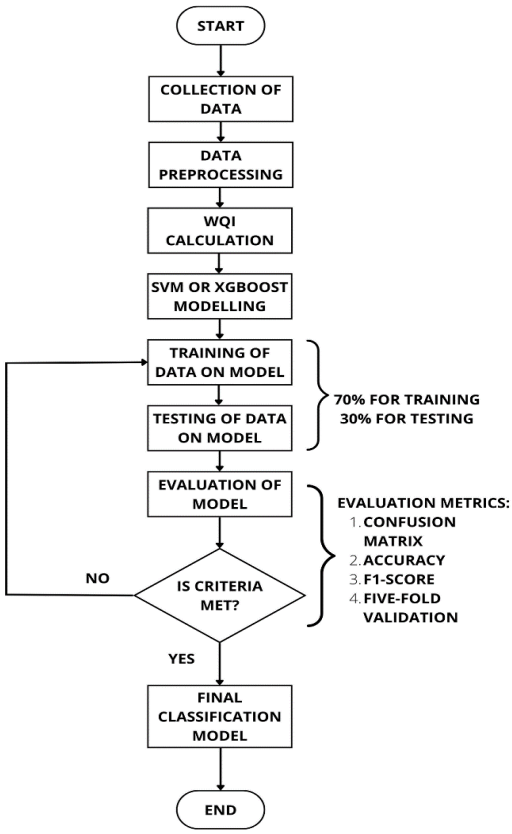

Figure 1. Flowchart of methodology

A. Dataset

The dataset used for this research was obtained from the Kaggle website, sourced from an Indian Government Website [16]. Data for historical water quality in a few locations in India has been combined and cleaned. The data were recorded in the range from 2003 to 2014. The data is appropriate for the current research project since the characteristics required to construct the water quality index are available in this dataset. A water quality classification can be derived from the water quality index. The classification model may employ a large amount of data, up to 2000 samples, for classification training and prediction.

B. Data Preprocessing

Before using the data for training, data need to go through data pre-processing, which refers to identifying and correcting errors in the dataset that may negatively impact a predictive model [4]. Common errors in a dataset include missing columns and duplicated rows. The data pre-processing phase is critical in data analysis to improve data quality. Using raw data from the dataset for classifications may generate incorrect results; hence, data cleaning is critical [4]. The water quality index (WQI) was calculated using the dataset's most important parameters, which are Dissolved oxygen (DO), pH, conductivity, biological oxygen demand (BOD), nitrate, fecal coliform, and total coliform. The WQI scores were then used to categorize the water samples. This research uses the weighted arithmetic water quality index method to calculate the WQI. WQI is chosen because it

incorporates various water quality criteria into a mathematical equation that measures the health of a water body and describes the acceptability of surface and groundwater sources for human use [17].

C. Water quality index

WQI is one of the most effective instruments for expressing water quality because it provides a simple, consistent unit of measurement for describing water quality. This renders it an important aspect in surface water evaluation and management due to its sensitivity to the presence of contaminants that affect water quality[18], [19], [20]. WQI is a rating system that calculates the combined impact of individual water quality parameters on the overall water quality. Seven important parameters were chosen for the study to calculate the water quality index. The weighted arithmetic method was chosen to calculate the WQI. The benefit of this approach is that it is slightly better than other methods, such as Canadian methods, in calculating WQI [18]. The World Health Organization's (WHO) drinking water quality standards were used to determine the WQI [21]. Before obtaining the WQI, several steps must be completed using the arithmetic index method in the following steps, as suggested in [22].

Calculation of sub-index of quality rating (q_n)

Let (n) be the number of water quality parameters and q_n be the quality rating or sub-index corresponding to the n^{th} parameter. Eq. 1 is used to compute the value of q_n.

$$q_n = 100\left[\frac{V_n - V_{i0}}{S_n - V_{i0}}\right] \quad (1)$$

Such that: q_n = rating quality for the nth water quality parameter; V_n = estimated value of the nth parameter at a given determinant; V_{i0} = ideal value of the nth parameter in pure water; S_n = standard permissible value of the nth parameter.

For all other parameters, except for pH, which is 7.0, and dissolved oxygen = 14 mg/L, all ideal (V_{i0}) values are taken as zero (0) for drinking water [21].

Unit weight (W_n)

For various water quality parameters, the calculation of unit weight (W_n) is inversely related to the suggested standards S_n for the relevant parameters in Table 2, as shown in Eq. 2

$$W_n = K/S_n \quad (2)$$

Such that: W_n= unit weight for the nth parameter; S_n = standard permissible value of the nth parameter; K = constant for proportionality.

WQI

The formula for WQI can be seen in Eq. 3:

$$WQI = \sum q_n W_n / \sum W_n \quad (3)$$

The overall water quality index was produced by linearly aggregating the standards of quality.

Assessment of WQI

The water quality has been classified into four groups, as shown in Table 1, in which the water quality classes were based using the weighted arithmetic method.

Table 1. Range of class for water quality classification

WQI range	Water quality	Class value
0 - 25	Excellent	3
26 - 50	Good	2
51 - 75	Poor	1
Greater than 75	Very Poor	0

By following the weighted arithmetic method, the WQI can be calculated. It includes the estimation of the unit weight assigned to each parameter. They are then converted to a common scale, illustrated in Table 2, representing the water quality standards.

Table 2. Water quality parameter and its standard, ideal and relative unit value

No.	Parameter	Standard value recommended	Ideal value	Relative weight
1	Dissolved oxygen (DO) mg/L	10	14	0.2213
2	pH	8.5	7.0	0.2604
3	Conductivity µS/cm	1000	0	0.0022
4	Biological oxygen demand (BOD) mg/L	5	0	0.4426
5	Nitrate (NI) mg/L	45	0	0.0492
6	Fecal coliform (FC)	100	0	0.0221
7	Total coliform (TC)	1000	0	0.0022

D. Machine learning

Support Vector Machine (SVM)

SVMs use hyperplanes to define decision boundaries between data points of different classes, originally developed for binary classification problems [4]. The hyperplanes are the decision functions that distinguish between positive and negative data and have marked the maximum margins, illustrated in Figure 2. Because of its low sensitivity to feature space dimensions, SVM is considered a reliable classifier for the Hughes effect; classification using SVM has minimal impact on the outcome [23].

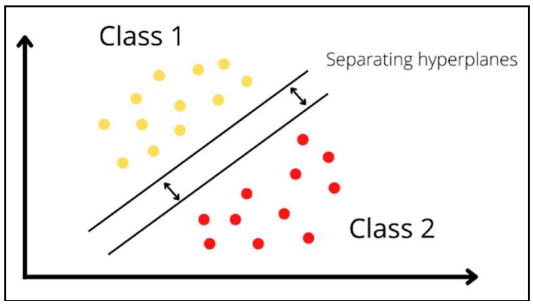

Figure 2. Hyperplane illustration for SVM

In this research, the radial basis function (RBF) is used as the kernel function of SVM. The most important parameter to consider during the modeling of SVM is C, the penalty coefficient (cost), error tolerance, and gamma in the RBF function. A higher C value contributes to lower tolerance, and the more likely overfitting occurs. The lower the value of C, the more easily it can be fitted. C values that are too large or too small reduce the model's generalizability[23].

Suppose that the dataset is represented by $x_i = \{x_1 x_2 \dots x_n\}$ for the input and the corresponding space as $y_i \epsilon \{-1, 1\}$the SVM needs to construct a hyperplane to classify the data into two classes. Therefore, the formula for the SVM using RBF is shown in Eq. 4, which $k(x, y)$ denotes the function expression and gamma used to give the curvature weight of the hyperplane.

$$k(x, y) = \exp(-gamma \times ||x - y||^2) \quad (4)$$

Extreme gradient boosting (XGBoost)

XGBoost is an algorithm that generates many shallow decision trees, and combining all trees results in a high prediction accuracy [24]. The decision trees created by the XGBoost algorithm are not only used to minimize an objective function by accounting for the loss function, but they also protect the tree from overfitting by using a regularization process [11].

Data scientists prefer XGBoost because of its fast out-of-core compute execution [13].

E. Evaluation metrics

Water quality classification performance is evaluated using a confusion matrix. Since this study has four classes of water classification, a confusion matrix for multi-class classification is employed to portray the true classes of the data samples. The four classes; are Class 3 as Excellent, Class 2 as Good, Class 1 as Poor, and Class 0 as Very Poor.

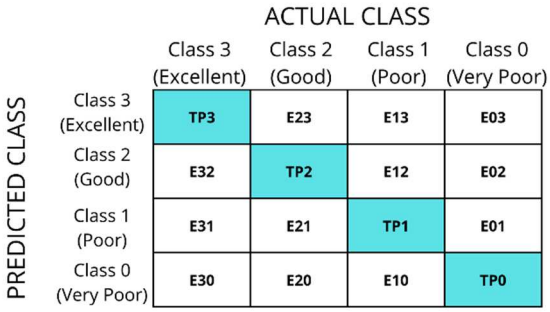

Figure 3. Confusion matrix for multi-class classification of water quality

As shown in Figure 3, the TP3 represents the number of True Positive samples in class 3, where the number of samples is correctly classified from class 3, and E32 is the samples for negative in class 3, where samples from class 3 are misclassified as class 2. The False Negative in class 3 can be expressed as Eq. 5

$$FN3 = E32 + E31 + E30 \quad (5)$$

By obtaining the values for True Positive (TP), False Negative (FN), True Negative (TN), and False positive (FP), other matrices such as accuracy and F1-score can be calculated using Eq. 6 and Eq. 7, respectively.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

$$F1 - Score = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \quad (7)$$

Another metric for evaluating the classification model is cross-validation using the k-fold method. The k-fold cross-validation is conducted to lessen the overfitting issue and random sampling biases [25]. The k-fold algorithm was used in the current research to analyze the performance of the models that divided the data sample into five subclasses. These test subclasses are then utilized to determine the final algorithms' accuracy, which is expressed as a mean accuracy attained by the five models in five validation rounds.

IV. Result and Discussion

A. Classification analysis

The result of water quality classification between SVM and XGboost is covered in this section. A total of 1893 data were split into two parts, one for training the model and the other for testing. A split of 70% was used for the training and 30% for the testing validation.

The SVM model's accuracy is very low, with only 47% for the training dataset and 48% for the test dataset. The most likely factor which causes the SVM to have a poor accuracy model is the selection of C and gamma parameters. C and gamma were tuned in to increase accuracy until the most suitable parameter are achieved [26], which is C = 1000 and gamma = 0.0001.

The SVM model is retrained using the optimized SVM parameters that have been obtained. The SVM classifier is prepared to identify fresh samples in the testing phase after the training phase. The accuracy improved for the SVM model with a score of 67% for the accuracy in the training dataset, accuracy prediction, and F1-score, then 81% for the accuracy test dataset.

On the other hand, the XGBoost classifier outperformed SVM in terms of accuracy with an accuracy score of 1 for the accuracy train dataset and 0.94 for the accuracy test dataset, accuracy predictions, and F1-Score. Figure 4 illustrates the performance of each model.

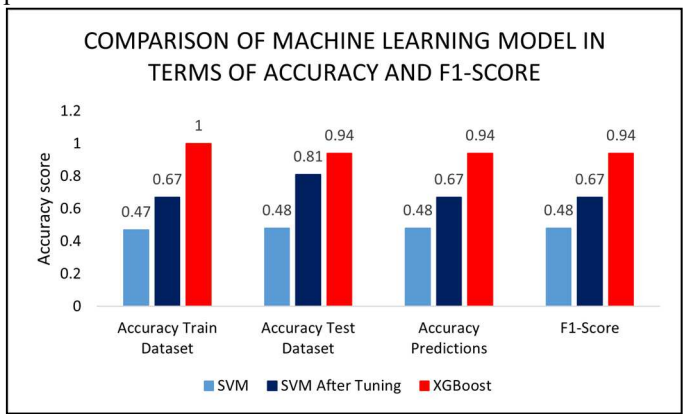

Figure 4. Comparison of machine learning model in terms of accuracy

In terms of using the K-fold validation method, five-fold were used in determining the accuracy of the models. The tuned SVM model was used as the final model, resulting in a 65% accuracy. In contrast, XGBoost had a mean accuracy score of 90% for the five-fold validation. Even using a different dataset for each fold, XGBoost proved superior to the SVM method, illustrated in Table 3.

Table 3. Number of folds with accuracy score

Number of fold	SVM Accuracy	XGBoost
1	0.66	0.92
2	0.68	0.88
3	0.61	0.90
4	0.58	0.90
5	0.67	0.94
Mean	0.64	0.90

From Fig. 5 and Fig. 6, the confusion matrix for SVM and XGBoost were constructed, and it is observed that XGBoost correctly classified most of the test data, while SVM resulted in a higher misclassification.

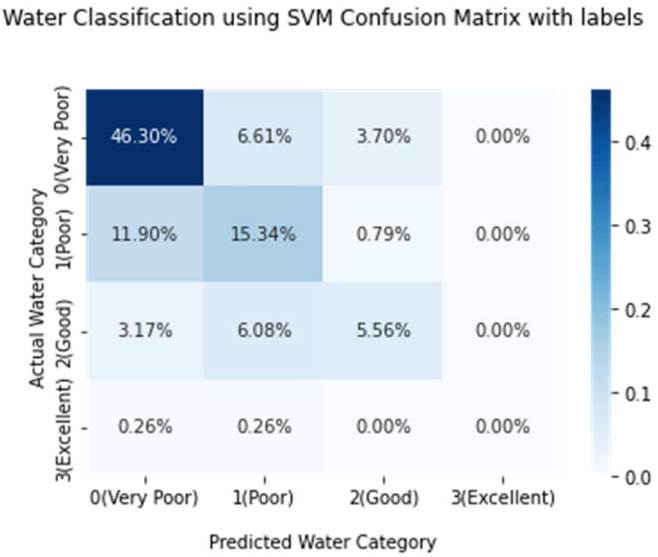

Figure 5. Confusion matrix for classification of SVM algorithm

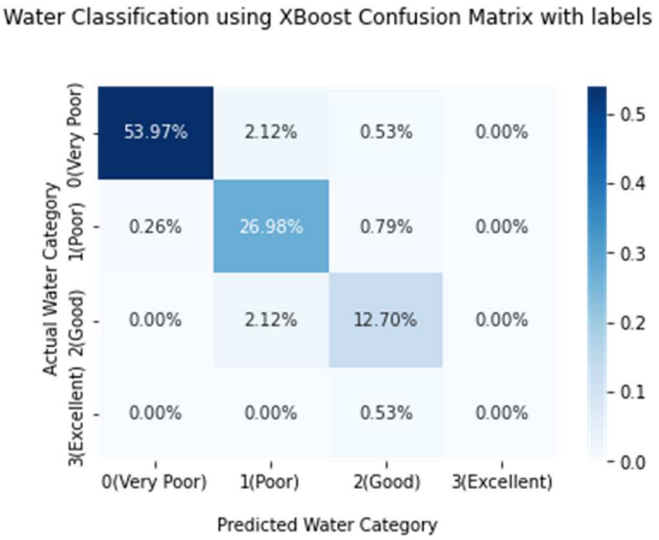

Figure 6. Confusion matrix for classification of XGBoost algorithm

V. CONCLUSION

Water quality is important in determining whether the water source is qualified for consumption. WQI is essential to classify whether the water is safe for consumption. Rather than requiring expensive and complex analysis to test the water quality, this research uses two machine learning algorithms, SVM and XGBoost, to predict water quality using readily available water quality parameters. The parameters employed for the classification algorithm are dissolved oxygen, pH, conductivity, biological oxygen demand, nitrate, fecal coliform, and total coliform. The outcome showed that XGBoost outperformed the SVM algorithm even after the parameters had been tuned. The results showed that SVM resulted in more misclassification of data than XGBoost. Despite the achievement of this research, improvements, such as applying more parameters to the model or using even more advanced

deep learning algorithms, can be applied in future research to improve water quality classification.

ACKNOWLEDGEMENT

The authors would like to thank the College of Engineering, Universiti Teknologi MARA, Shah Alam for providing necessary financial support.

REFERENCES

[1] World Water Assessment Programme (United Nations), *Wastewater : the untapped resource : the United Nations world water development report 2017*.

[2] P. Burek *et al.*, "The Water Futures and Solutions Initiative of IIASA," 2016.

[3] A. Danades, D. Pratama, D. Anggraini, and D. Anggriani, "Comparison of accuracy level K-Nearest Neighbor algorithm and support vector machine algorithm in classification water quality status," in *Proceedings of the 2016 6th International Conference on System Engineering and Technology, ICSET 2016*, Feb. 2017, pp. 137–141. DOI: 10.1109/FIT.2016.7857553.

[4] K. P. Singh, N. Basant, and S. Gupta, "Support vector machines in water quality management," *Analytica Chimica Acta*, vol. 703, no. 2, pp. 152–162, Oct. 2011, DOI: 10.1016/j.aca.2011.07.027.

[5] T. Eitrich and B. Lang, "Efficient optimization of support vector machine learning parameters for unbalanced datasets," *Journal of Computational and Applied Mathematics*, vol. 196, no. 2, pp. 425–436, Nov. 2006, DOI: 10.1016/j.cam.2005.09.009.

[6] Z. Pang and K. Jia, "Designing and accomplishing a multiple water quality monitoring system based on SVM," in *Proceedings - 2013 9th International Conference on Intelligent Information Hiding and Multimedia Signal Processing, IIH-MSP 2013*, 2013, pp. 121–124. DOI: 10.1109/IIH-MSP.2013.39.

[7] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, vol. 13-17-August-2016, pp. 785–794. DOI: 10.1145/2939672.2939785.

[8] D. N. Myers, "Why monitor water quality?" [Online]. Available: https://www.epa.gov/assessing

[9] "Artificial Neural Network Modeling of the Water Quality Index Using Land Use Areas as Predictors".

[10] M. Bouamar and M. Ladjal, "Evaluation of the performances of ANN and SVM techniques used in water quality classification."

[11] F. Hassanbaki Garabaghi, "Performance Evaluation of Machine Learning Models with Ensemble Learning approach in Classification of Water Quality Indices Based on Different Subset of Features," 2021, DOI: 10.21203/rs.3.rs-876980/v1.

[12] L. Li *et al.*, "Interpretable tree-based ensemble model for predicting beach water quality," *Water Research*, vol. 211, Mar. 2022, DOI: 10.1016/j.watres.2022.118078.

[13] N. Nasir *et al.*, "Water quality classification using machine learning algorithms," *Journal of Water Process Engineering*, vol. 48, p. 102920, Aug. 2022, DOI: 10.1016/j.jwpe.2022.102920.

[14] D. Dezfooli, S. M. Hosseini-Moghari, K. Ebrahimi, and S. Araghinejad, "Classification of water quality status based on minimum quality parameters: application of machine learning techniques," *Modeling Earth Systems and Environment*, vol. 4, no. 1, pp. 311–324, Apr. 2018, DOI: 10.1007/s40808-017-0406-9.

[15] T. H. H. Aldhyani, M. Al-Yaari, H. Alkahtani, and M. Maashi, "Water Quality Prediction Using Artificial Intelligence Algorithms," *Applied Bionics and Biomechanics*, vol. 2020, 2020, DOI: 10.1155/2020/6659314.

[16] "Indian water quality data | Kaggle." https://www.kaggle.com/datasets/anbarivan/indian-water-quality-data (accessed Jul. 01, 2022).

[17] S. C. Dendukuri, D. S. Chandra, M. V. S. Raju, and S. S. Asadi, "Estimation of Water Quality Index By Weighted Arithmetic Water Quality Index Method: A Model Study," *International Journal of Civil Engineering and Technology*, vol. 8, no. 4, pp. 1215–1222, 2017, [Online]. Available: http://www.iaeme.com/IJCIET/index.asp1215http://www.iaeme.com/IJCIET/issues.asp?JType=IJCIET&VType=8&IType=4http://www.iaeme.com/IJCIET/issues.asp?JType=IJCIET&VType=8&IType=4http://www.iaeme.com/IJCIET/index.asp1216

[18] F. M. Kizar, "A comparison between weighted arithmetic and Canadian methods for a drinking water quality index at selected locations in shatt al-kufa," in *IOP Conference Series: Materials Science and Engineering*, Nov. 2018, vol. 433, no. 1. DOI: 10.1088/1757-899X/433/1/012026.

[19] M. D. Noori, "Comparative analysis of weighted arithmetic and CCME Water Quality Index estimation methods, accuracy and representation," in *IOP Conference Series: Materials Science and Engineering*, Mar. 2020, vol. 737, no. 1. DOI: 10.1088/1757-899X/737/1/012174.

[20] C. K. Ojukwu, G. O. C. Okeah, and P. C. Mmom, "A Comparative Analysis of the Weighted Arithmetic and Canadian Council of Ministers of the Environment Water Quality Indices for Water Sources in Ohaozara, Ebonyi State, Nigeria," *International Journal of Engineering Research & Technology (IJERT) www.ijert.org*, vol. 10, 2021, [Online]. Available: www.ijert.org

[21] "Fourth edition incorporating the first and second addenda Guidelines for drinking-water quality."

[22] S. Bouslah, L. Djemili, and L. Houichi, "Water quality index assessment of Koudiat Medouar Reservoir, northeast Algeria using weighted arithmetic index method," *Journal of Water and Land Development*, vol. 35, no. 1, pp. 221–228, Dec. 2017, DOI: 10.1515/jwld-2017-0087.

[23] S. Feng, J. Zhao, T. Liu, H. Zhang, Z. Zhang, and X. Guo, "Crop Type Identification and Mapping Using Machine Learning Algorithms and Sentinel-2 Time Series Data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 9, pp. 3295–3306, Sep. 2019, DOI: 10.1109/JSTARS.2019.2922469.

[24] S. Singha, S. Pasupuleti, S. S. Singha, R. Singh, and S. Kumar, "Prediction of groundwater quality using efficient machine learning technique," *Chemosphere*, vol. 276, Aug. 2021, DOI: 10.1016/j.chemosphere.2021.130265.

[25] M. I. Shah, M. F. Javed, and T. Abunama, "Proposed formulation of surface water quality and modelling using gene expression, machine learning, and regression techniques," *Environmental Science and Pollution Research*, vol. 28, no. 11, pp. 13202–13220, Mar. 2021, DOI: 10.1007/s11356-020-11490-9.

[26] M. Y. Cho and T. T. Hoang, "Feature Selection and Parameters Optimization of SVM Using Particle Swarm Optimization for Fault Classification in Power Distribution Systems," *Computational Intelligence and Neuroscience*, vol. 2017, 2017, DOI: 10.1155/2017/4135465.