# The Classification of Water Quality using Machine Learning Approach

Hritwik Ghosh[1], Mahatir Ahmed Tusher[2], Irfan Sadiq Rahat[3], Syed Khasim[4,] Sachi Nandan Mohanty[5]

[1,2,3,4,5] School of Computer Science and Engineering (SCOPE), VIT-AP University, Amaravati, Andhra Pradesh

[1]hritwik.21bce8973@vitapstudent.ac.in, [2]ahmed.21bce8971@vitapstudent.ac.in, [3]rahat.21bce9985@vitapstudent.ac.in, [4]profkhasim@gmail.com
[5]sachinandan09@gmail.com

## Abstract

For protecting environment and human health water quality monitoring is very essential. In recent years Artificial Intelligence has paved the way to bring a great improvement in the classification and prediction of water quality. Here, for developing a reliable approach to forecast water quality and to to distinguish between Potable and Non-Potable water, we have used various machine learning models in this study. The studied machine learning classifiers and their stacking ensemble models included Logistic Regression, Gaussian Naive Bayes, Bernoulli Naive Bayes, Support Vector Machine(SVM), Kth Nearest Neighbours(KNN), X Gradient Boosting, and Random Forest. In this study, our used dataset has 3277 samples and it's a historical dataset which was collected for over 9 years from various places of Andhra Pradesh, India. It has been taken from Andhra Pradesh Pollution Control Board (APPCB). We have used precision-recall curves to evaluate the performance of the various classifiers. Among our used models, Random Forest provided the highest accuracy with a percentage of 78.96, where SVM provided us the least amount of accuracy with a percentage of 68.29.

**Keywords** Water quality, Machine learning, Logistic Regression, Gaussian Naivd Bayes, Bernoulli Naive Bayes, Support Vector Machine(SVM), Kth Nearest Neighbours(KNN), X Gradient Boosting, Random Forest.

## 1 Introduction

For drinking and domestic use, food production or recreational purposes safe and readily available water is essentially important. Despite having a massive amount of water, the amount of drinkable water is still inadequate [1]. In India, around 70% of surface water is not fit for consumption [2]. In case of poverty reduction better management of water resources, improved water supply and sanitation can favour a country's economic growth.

Transmission of contagious disease, such as cholera, diarrhoea, dysentery, hepatitis A, typhoid, and polio etc are directly linked the contaminated water and poor sanitation. Generally, after collecting water samples from various sources, manual water sampling and lab analysis of water quality can not be efficient. At the same time, it can be time consuming and prodigal. Therefore, the use of intelligent systems are increasing exponentially to monitor water quality especially when we need real time data.[3][4]

Machine learning is a subset of artificial intelligent which teaches a system to automatically learn and improve from the experience without the manual interference[5] . The procedures that are used in machine learning are trained to capture trends and pursuant to they update themselves [6], [7]. In water studies, Machine learning paves the way to assess, classify and predict water quality indicators. For instance, we can successfully simulate hydro-logical processes subject to the accessibility of bountiful sets data. However, using this water potability dataset of Vijayawada, Rand Pradesh [8] and applying machine learning to it, we are going to distinguish between Potable and Non-Potable water using some parameters such as pH value, chloramines, sulfate, conductivity, organic carbon, hardness, solids, conductivity, Trihalomethanes, turbidity, potability.

# 2 Literary Survey

For classifying the water quality of Chao Phraya river, (RivSillberg et al.) have developed a ML-based technique integrating attribute-realization (AR) and support vector machine (SVM)[9] . Attribute realization has identified the most significant elements of improving the quality of river, using the linear function. The most availing characteristics were TCB, NH3-N, FCB, DO, BOD, Sal and DO in the assortment, with contributed values in the range of 0.80–0.98. Using SVM linear approach, they got the best classification results, which had an F1-score average of 0.84 ,accuracy of 0.94, a recall average of 0.84 and a precision average of 0.84. With 0.86–0.95 accuracy AR-SVM was a strong and effective method for identifying the quality of the river water. They could find it when they applied to -6 parameters. To simulate the quality index of Akaki river water Yilma et al used ANN [10]. They calculated the index using 12 WQ indicators from 27 different wet & dry season sample locations.Almost all projected findings figured low WQ. The exception was one upstream location. The ANN model was corroborated and trained using the number of hidden layer neurons (5, 10, 15, 20, and 25), hidden layers (2–20), learning functions and transfer training. It was performed through 12 inputs. In their research, ANN with fifteen hidden neurons and eight hidden layers predicted the water quality index with an accuracy of 0.93.

In Malaysia, to calculate the water quality of Kinta River, Gazzaz et al. [11] designed a feed-forward, three layer, fully connected neural ANN model and the IoT . In Iran, to predict water quality index in 47 springs and wells, Sakizadeh [12] used 3 different artificial neural network algorithms, such as, an ensemble of artificial neural network, ANN with early stopping and artificial neural network with Bayesian regularization. One of the earliest works in this field is (Mahapatra et al., 2015). In this study, the authors applied various machine learning algorithms, such as K-Nearest Neighbors (KNN), Decision Tree (DT), Random Forest (RF), and Artificial Neural Network (ANN), to classify water quality in India based on parameters such as pH, temperature, dissolved oxygen, etc. The results showed that the ANN algorithm performed better than the other algorithms in terms of accuracy and efficiency[13].In [14] ANN has been used to analyse the comparative risk on prediction of Diabetes Mellitus. In [15], Internet of Things with cloud-based clinical decision support system has been used to predict and observe Chronic Kidney Disease where Deep neural network classifier predicted Chronic Kidney Disease with an accuracy of 98.25%, and later it has been enhanced to 99.25 by Particle Swarm Optimization method.

Another study that used machine learning algorithms for water quality monitoring is (Patil et al., 2019) . In this work, the authors employed KNN, SVM, and RF algorithms to monitor water quality in India based on parameters such as pH, TDS, and EC. The results showed that the RF algorithm performed better than the other algorithms in terms of accuracy and efficiency [16]. In (Nakamura et al.,2020) , the authors proposed the development of machine learning models for water quality classification in Asian countries. The study also discussed the challenges and limitations of using machine learning for water quality classification, such as the need for large amounts of labeled data and the difficulty in interpreting the results of complex algorithms [17]. K-Nearest Neighbors (KNN) is another machine learning technique that has been used in water quality classification. In a study by (Jiang et al., 2019), a KNN model was trained to classify water quality into three categories: safe, caution, and danger. The model was trained using water quality data from the Yangtze River, China. The results showed that the KNN model achieved an accuracy of 93.6% [18].

Naive Bayes is a probabilistic machine learning technique that has been used in water quality classification. In a study by (Liu et al., 2020), a Naive Bayes model was trained to classify water quality into two categories: safe and caution. The model was trained using water quality data from Lake Taihu, China. The results showed that the Naive Bayes model achieved an accuracy of 89.3% [19]. In this study (Shamsuddin et al.,2016), they aimed to supervise the performance of ML models for multiclass classification in the Langat River Basin water quality assessment. Among SVM, DT and ANN, Support Vector Machine performed the best [20].

# 3 Proposed Methodology

We have taken prior actions to prepare the data as input before employing ML models (Logistic Regression, Gaussian Naivd Bayes, Bernoulli Naive Bayes, Support Vector Machine, Kth Nearest Neighbors, X Gradient Boosting, Random Forest). We have divided the data into two sets, training and testing sets to train our 7 machine learning models and assess the performance. In addition, we have cleaned the dataset by removing inexact values and replacing empty cells with the median of the dataset's input variables. The following figure (Fig.1) has demonstrated the framework of the proposed system.

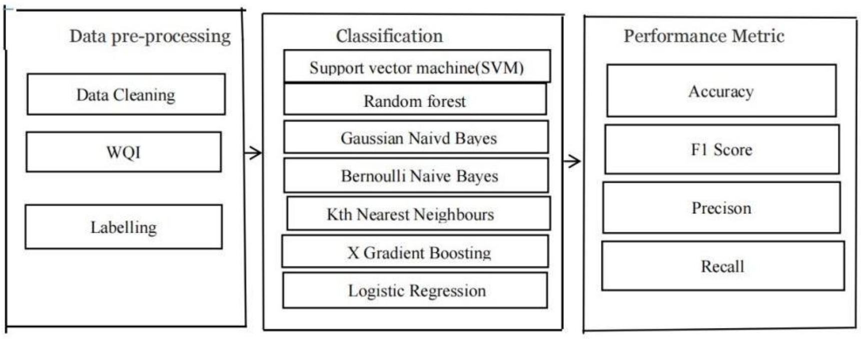

**Fig.1** Methodology of the proposed system.

## 3.1 Data Description

Our dataset contains water quality metrics for 3277 different water bodies and 10 features, such as pH value, chloramines, sulfate, conductivity, organic carbon, hardness, solids, conductivity, Trihalomethanes, turbidity, potability. At first we have imported all the necessary libraries which have been used to train the machine learning models or visualize the data. Then using a Pandas's function read_csv(), we have loaded the data set. Then we have performed Exploratory Data Analysis. In Exploratory Data Analysis, firstly have checked the shape of the data set. Then after handling the null values we have checked the value counts of our target feature Potability. After that, we were able to visualize the potability using a countplot function of seaborn (Fig.2). Then, we have displayed the entire dataset using the hist method (Fig.3). One can explore the dataset through it. We can say, there is no correlation between any feature. Because ,in the visualization of the correlation of all the features using a heat map function of seaborn, we found no relation(Fig.4). So, we can conclude that, the dimension cannot be reduced.

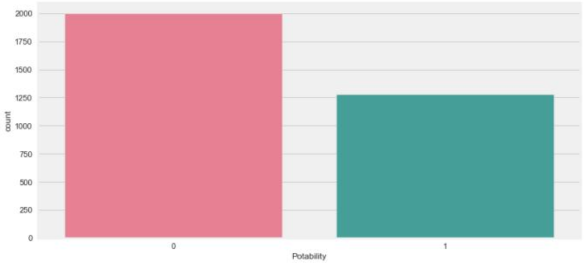

**Fig.2** Visualization of the potability

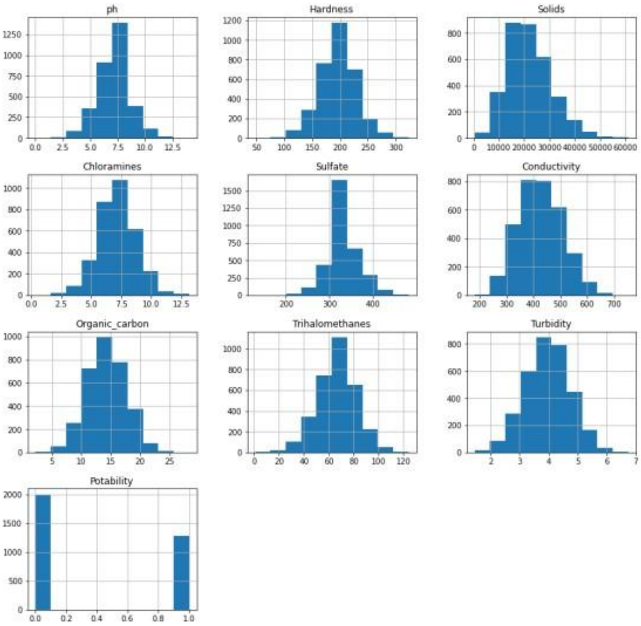

**Fig.3** Visualization of the entire dataset using hist method.

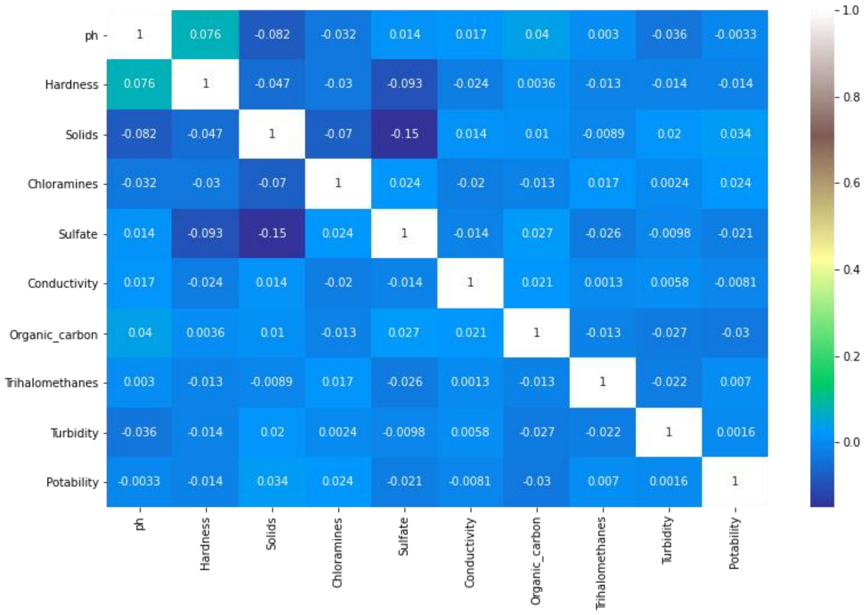

**Fig.4** Correlation heatmap

## 3.2 Data Preparation

In data preparation, we have divided the data into two categories, dependent and independent features. Except potability, all our features are independent. Then using train_test_split function we have split the dataset into testing and training. After that, using the data set ( X_train, Y_train ), we trained the model and defined the decision tree classifier model.  Using the test data set (X_test), we tested the model. Finally using the accuracy score, we have evaluated the model , confusion matrix and classification report. The techniques of evaluation take two parameters; the first one is the predicted data and the second one is actual data. Here we are demonstrating a graph of models vs accuracy (Fig.5)

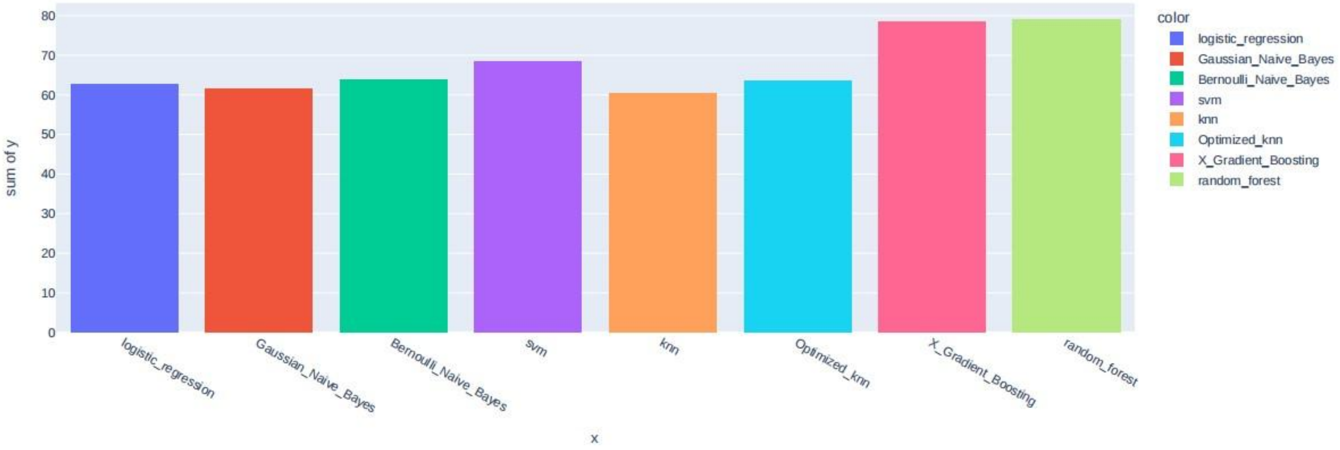

**Fig.5** Models v/s Accuracy

# 4 Experimental Analysis

Using the boxplot function, we can find outliers where these are contained by the solid feature (**Fig.**9). But these outliers cannot be removed. Point to be noted, the water will be safe to drink if the outliers are removed from the solid feature. Basically these outliers in solid features make the water impure. Presence of high amount of solid particles make the water unsafe to drink. But we cannot remove this outlier to train the model.

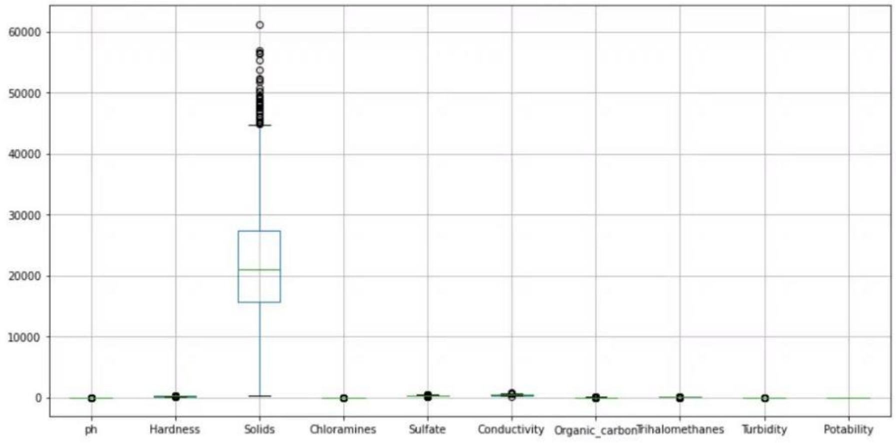

**Fig.9** Outliers in the solid feature.

Now we are going to see the performances of our used machine learning models (Logistic Regression, Gaussian Naive Bayes, Bernoulli Naive Bayes, Support Vector Machine, Kth Nearest Neighbours, X Gradient Boosting, Random Forest.) through performance metrics, Precision, recall, f1 score and accuracy (**Table.1** to **Table.8**)

| **Model Name** | **Accuracy** | **Precision** | **Recall** | **F1_Score** |
|---|---|---|---|---|
| Logistic Regression | 0.63 | 0.63 | 1.00 | 0.77 |
| Gaussian Naïve Bayes | 0.61 | 0.65 | 0.85 | 0.74 |
| Bernoulli Naïve Bayes | 0.64 | 0.68 | 0.81 | 0.73 |
| Support Vector Machine | 0.68 | 0.68 | 0.93 | 0.89 |
| K Nearest Neighbours | 0.63 | 0.68 | 0.69 | 0.73 |
| X Gradient Boosting | 0.78 | 0.81 | 0.86 | 0.83 |
| Random Forest | 0.79 | 0.81 | 0.88 | 0.84 |

**Table.1** performances of our used machine learning models

## 4.1 Logistic Regression

Logistic regression is one of the regression algorithms of supervised learning. We use it to predict or calculate the probability of a binary (yes/no) event occurring. Its function is a simple S-shaped curve that is generally used to convert data in binary expression (0 or 1).

$$h\Theta(x) = 1/1 + e - (\beta o + \beta 1X) \qquad (1)$$

Here, in our study, 0 represents "non-potability"and 1 represents "potability". We are going to demonstrate its performance metrics.

| | Precision | Recall | F1 Score | Support |
|---|---|---|---|---|
| 0 | 0.63 | 1.00 | 0.77 | 412 |
| 1 | 0.00 | 0.00 | 0.00 | 244 |
| Macro avg | 0.31 | 0.50 | 0.39 | 656 |
| Weighted avg | 0.39 | 0.63 | 0.48 | 656 |

**Table.2** Performance Metrics of Logistic Regression

## 4.2 Gaussian Naive Bayes

Naïve Bayes algorithm is one of the supervised learning algorithms based on Bayes theorem. Generally we use it for solving classification problems. Basically it's used in text classification which includes a high-dimensional training dataset. Gaussian Naive Bayes is an extension of naive Bayes. The following expression is its mathematical expression.

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)} \qquad (2)$$

| | Precision | Recall | F1 Score | Support |
|---|---|---|---|---|
| 0 | 0.65 | 0.85 | 0.74 | 412 |
| 1 | 0.46 | 0.22 | 0.30 | 246 |
| Macro avg | 0.56 | 0.53 | 0.52 | 656 |
| Weighted avg | 0.58 | 0.62 | 0.57 | 656 |

**Table.3** Performance Metrics of Gaussian Naive Bayes

## 4.3 Bernoulli Naive Bayes

Naive Bayes is one of the supervised machine learning algorithms. Based on numerous attributes it can predict the probability of different classes. It indicates the likelihood of occurrence of an event. We also know it as conditional probability. Bernoulli Naive Bayes is an extension of naive Bayes

| | Precision | Recall | F1 Score | Support |
|---|---|---|---|---|
| 0 | 0.68 | 0.80 | 0.73 | 412 |
| 1 | 0.52 | 0.37 | 0.43 | 244 |
| Macro avg | 0.60 | 0.58 | 0.58 | 656 |
| Weighted avg | 0.62 | 0.64 | 0.62 | 656 |

**Table.4** Performance Metrics of Bernoulli Naive Bayes

## 4.4 Support Vector Machine (SVM)

Creating the best line or decision boundary which can separate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. SVM can be used as both classifier and predictor.

| | Precision | Recall | F1 Score | Support |
|---|---|---|---|---|
| 0 | 0.68 | 0.93 | 0.79 | 412 |
| 1 | 0.69 | 0.27 | 0.38 | 244 |
| Macro avg | 0.69 | 0.60 | 0.59 | 656 |
| Weighted avg | 0.69 | 0.68 | 0.64 | 656 |

**Table.5** Performance Metrics of Support Vector Machine

## 4.5 Kth Nearest Neighbours(KNN)

Among the simplest machine learning algorithms used for classification is K-Nearest Neighbors. Data points are classified based on their neighbors' classifications. New cases are classified based on similar characteristics based on the stored cases.

| | Precision | Recall | F1 Score | Support |
|---|---|---|---|---|
| 0 | 0.68 | 0.71 | 0.69 | 412 |
| 1 | 0.47 | 0.43 | 0.45 | 244 |
| Macro avg | 0.57 | 0.57 | 0.57 | 656 |
| Weighted avg | 0.60 | 0.61 | 0.60 | 656 |

**Table 6** Performance Metrics of KNN

## 4.6 X Gradient Boosting

Gradient boost machines (GBM) are one of the most effective and mention-worthy algorithms of supervised machine learning. a, and X Gradient Boost is one of the implementations of GBM. In addition, it can also be used to solve regression and classification related problems.

| | Precision | Recall | F1 Score | Support |
|---|---|---|---|---|
| 0 | 0.81 | 0.86 | 0.83 | 412 |
| 1 | 0.74 | 0.65 | 0.69 | 244 |
| Macro avg | 0.77 | 0.76 | 0.76 | 656 |
| Weighted avg | 0.78 | 0.79 | 0.78 | 656 |

**Table.7** Performance Metrics of X Gradient Boosting

### 4.7 Random Forest

Random Forest is one of the well performing algorithms of popular supervised machine learning . We can use it for both, regression problems and classification. Random Forest is based on the notion of ensemble learning.

| Precision | | Recall | F1 Score | Support |
|---|---|---|---|---|
| 0 | 0.80 | 0.88 | 0.84 | 412 |
| 1 | 0.76 | 0.64 | 0.69 | 244 |
| Macro avg | 0.78 | 0.76 | 0.77 | 656 |
| Weighted avg | 0.79 | 0.79 | 0.79 | 656 |

**Table.8** Performance Metrics of Random Forest

## Conclusion

The Bureau of Indian Standards (BIS) declared the upper limit of total dissolved solids (TDS) levels in water is 500 ppm. For having values of TDS on an average of 40 times as much as the upper limit o the safe drinking water, the solid level seem to contain some descripancy. Equal number of basic and acidic pH level water samples are contained in our data. Very less correlation coefficients has been noticed between the features. Among our used models, Random Forest provided the highest accuracy with a percentage of 78.96, where SVM provided us the least amount of accuracy with a percentage of 68.29. But in case of training the model, XGBoost and Random forest performed the best. Both of the models gave us f1 score (Balanced with recall and precision) around 76%.

## References

[1] A.S. Brar, Consumer behaviour and perception for efficient water use in urban Punjab[Online]. Available: http://shodhganga.inflibnet.ac.in:8080/jspui/handle/10603/8807 (2011), Accessed 2nd Sep 2021

[2] https://www.weforum.org/agenda/2019/10/water-pollution-in-india-data-tech-solution/

[3] B. O'Flynn, F. Regan, A. Lawlor, J. Wallace, J. Torres, C. O'Mathuna Experiences and recommendations in deploying a real-time, water quality monitoring system Meas. Sci. Technol., 21 (12) (Oct. 2010), Article 124004, 10.1088/0957-0233/21/12/124004

[4] N. Kedia, Water quality monitoring for rural areas- a Sensor Cloud based economical project. 2015 1st International Conference on Next Generation Computing Technologies (NGCT) (Sep. 2015), pp. 50-54, 10.1109/NGCT.2015.7375081

[5] A.Y. Sun, B.R. Scanlon, How can big data and machine learning benefit environment and water management: a survey of methods, applications, and future directions. Environ. Res. Lett., 14 (7) (Jul. 2019), Article 073001, 10.1088/1748-9326/ab1b7d

[6] M. Bagheri, A. Akbari, S.A. Mirbagheri, Advanced control of membrane fouling in filtration systems using artificial intelligence and machine learning techniques: a critical review. Process Saf. Environ. Prot., 123 (Mar. 2019), pp. 229-252, 10.1016/j.psep.2019.01.013

[7] F. Hassanpour, S. Sharifazari, K. Ahmadaali, S. Mohammadi, Z. Sheikhalipour, Development of the FCM-SVR hybrid model for estimating the suspended sediment load. KSCE J. Civ. Eng., 23 (6) (Jun. 2019), pp. 2514-2523, 10.1007/s12205-019-1693-7

[8] Link of dataset: https://drive.google.com/file/d/1YpthGvFexSjMgOLtd4Hps-fA42pmO7SP/view?usp=drivesdk

[9] C.V. Sillberg, P. Kullavanijaya, O. Chavalparit, Water quality classification by integration of attributerealization and support vector machine for the chao phraya river, Journal of Ecological Engineering 22. (2021), 70–86.

[10] M. Yilma, Z. Kiflie, A. Windsperger, N. Gessese, Application of arti- ficial neural network in water quality index prediction: a case study in little Akaki River, Addis Ababa, Ethiopia, Modeling Earth Systems and Environment 4 (2018), 175–187.

[11] N.M. Gazzaz, M.K. Yusoff, A.Z. Aris, H. Juahir, M.F. Ramli, Artificial neural network modeling of the water quality index for Kinta River (Malaysia) using water quality variables as predictors. Mar. Pollut. Bull., 64 (11) (Nov. 2012), pp. 2409-2420, 10.1016/j.marpolbul.2012.08.005

[12] M. Sakizadeh, Artificial intelligence for the prediction of water quality index in groundwater systems. Model. Earth Syst. Environ., 2 (1) (Mar. 2016), p. 8, 10.1007/s40808-015-0063-9

[13] K. S. Mahapatra, S. K. Sahoo, Assessment of water quality using machine learning algorithms: A case study in India(2015).

[14] A. Swain, S. N. Mohanty and A. C. Das, "Comparative risk analysis on prediction of Diabetes Mellitus using machine learning approach," 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), Chennai, India, 2016, pp. 3312-3317, doi: 10.1109/ICEEOT.2016.7755319

[15] Lakshmanaprabu S.K., Sachi Nandan Mohanty, Sheeba Rani S., Sujatha Krishnamoorthy, Uthayakumar J., K. Shankar, Online clinical decision support system using optimal deep neural networks, Applied Soft Computing, Volume 81, 2019, 105487, ISSN 1568-4946, https://doi.org/10.1016/j.asoc.2019.105487

[16] S. B. Patil, S. M. Pawar, N. R. Shinde, Water Quality Monitoring using Machine Learning Algorithms: A Study in India (2019).

[17] T. Nakamura, Y. Uchida and T. Kimura, Development of machine learning models for water quality classification in Asian countries (2020).

[18] Jiang, Y., Chen, J., Yen, K., Xu, J. (2019) Ectopically Expressed IL-34 Can Efficiently Induce Macrophage Migration to the Liver in Zebrafish. Zebrafish. 16(2):165-170.

[19] Huiyuan Liu, Jun Xia, Lei Zou, Ran Huo, Comprehensive quantitative evaluation of the water resource carrying capacity in Wuhan City based on the “human–water–city” framework: Past, present and future, Journal of Cleaner Production. Volume 366,2022, ISSN 0959-6526, https://doi.org/10.1016/j.jclepro.2022.132847.

[20] Shamsuddin, I.I.S.; Othman, Z.; Sani, N.S. Water Quality Index Classification Based on Machine Learning: A Case from the Langat River Basin Model. Water 2022, 14, 2939. https://doi.org/10.3390/ w14192939