Water Quality Prediction & Classification Based on Machine Learning Technique

Chavana Sateesh[1], T.N.R Kumar[2]
[1]*M.Tech, Department of Computer Science, Msrit, Bangalore,*
[2] *Associate Professor, Department of Computer Science, Msrit, Bangalore*

Abstract-One of the major problems the globe has faced in recent decades is estimating the quality of the water supply. This study offers a more precise classification and prediction model for water quality. Our everyday lives depend greatly on the quality of urban water. Urban water quality forecasting aids in reducing water pollution and safeguarding public health. However, estimating the quality of urban water is difficult since urban water quality fluctuates nonlinearly and depends on a variety of variables, including weather, water usage patterns, and land uses. Using the water quality data and water hydraulic data supplied by existing monitor stations and a range of data sources we saw in, we used a data-driven approach to predict the water quality over the following few hours in this work.

The city, including the weather, pipe networks, road network design, and points of interest (POIs). By conducting comprehensive experiments based on the literature, we first determine the key elements that have a significant impact on urban water quality. There are many machine learning algorithms for categorization, but selecting the right one is a crucial challenge. The proposed system experiment and study the machine learning algorithms to determine the optimal algorithm for the water quality monitoring system. In this experimental investigation, a real dataset is utilized to classify and predict the water quality in order to compare the effectiveness of the various classification methods.

Index Terms: Machine Learning, Support vector Machine, Decision Tree

INTRODUCTION

Water quality analysis is a complex topic due to the different factors that influence it. This concept is inextricably linked to the various purposes for which water is used. Different needs necessitate different standards. There is a lot of study being done on water quality prediction. Water quality is normally determined by a set of physical and chemical parameters that are closely related to the water's intended usage. The acceptable and unacceptable values for each variable must then be established. Water that meets the predetermined parameters for a specific application is considered appropriate for that application. If the water does not fulfil these requirements, it must be treated before it may be used. Water quality can be assessed using a variety of physical and chemical properties. As a result, studying the behaviour of each individual variable independently is not possible in practice to accurately describe water quality on a spatial or temporal basis. The more challenging method is to combine the values of a group of physical and chemical variables into a single value. A quality value function (usually linear) represented the equivalence between the variable and its quality level was included in the index for each variable. These functions were created using direct measurements of a substance's concentration or the value of a physical variable derived from water sample studies. The major goal of this project is to examine how machine learning algorithms may be used to predict water quality. The proposed model is evaluated by different datasets that and the results proved that the proposed model outperforms constant models where it achieved an accuracy of 88.16%, 98.67%, and 97.63%, respectively. Furthermore, the proposed method outperformed the previous models. In addition to accuracy, three other measurements, precision, recall, and F1-score were used.

LITERATURE REVIEW

For example, Barzegar et al. (2020), applied a CNN-LSTM amalgam model to predict two water quality variables, named Dissolved Oxygen (DO) and chlorophyll-a. Results indicated that the CNN-LSTM amalgam model outperformed both the individual CNN and LSTM model and the machine learning

models such as SVR, Decision Tree. Oladipo et al. (2021), compared two statistical methods, including Fuzzy Logic Inference (FLI) and WQI methods, for evaluating the water quality in the Ikare community, Nigeria. They found moderate and poor water quality conditions using FLI and WQI methods, respectively. They also found that the FLI method is superior to the WQI method because of the relationship between measured values and WQI standard values. For the estimation of dissolved oxygen in aquaculture, Li et al. (2018), suggested a synthetic model by combining Sparse- autoencoder and long short-term memory networks (LSTM). Although both CNN-LSTM and Sparse-autoencoder-LSTM models showed excellent performance since they predicted only DO and chlorophyll, it may be challenging to deal with more water quality variables using such models. In another research, Asadollah et al. (2021), applied an ensemble machine learning method called Extra Tree Regression (ETR) which combines multiple week learners such as decision tree to predict WQI values in Tsuen River, Hong Kong. They applied the ETR method on ten water quality variables. Results indicated that the ETR method achieved 98% prediction accuracy, which outperformed the other state-of-the-art models such as support vector regression and decision tree. Further, Hameed et al. (2017), developed two neural artificial network techniques: a radial basis function neural network (RBFNN) and a backpropagation neural network (BNN) to predict the WQI in the tropical region of Malaysia. The WQI was measured using sub-indices equations in this study (Agamuthu and Victor, 2011). In both RBFNN and BNN strategies, the training is faster, but the prediction takes a long time, making the model slow. Bui et al. (2020), proposed a hybrid machine learning algorithm by combining the random tree and bagging (BA-RT) technique. The BA-RT method achieved 94% prediction accuracy using a 10–fold cross-validation technique, outdoing 15 standalone and hybrid algorithms. A more comprehensive study into the application of machine learning methods for modeling river water quality was performed by Rajaee et al. (2020), where they reviewed a total of 51 articles published from 2000 to 2016. According to this study, artificial neural networks and wavelet-neural networks were the most widely used methods for predicting water quality. Furthermore, Samsudin et al. (2019), developed an artificial neural network. For this study, the most significant water quality parameters were found through a spatially discriminant analysis (SDA). But these studies can barely show 71% accuracy. In another research, Yilma et al. (2018), applied an artificial neural network for predicting WQI in Ethiopia's Akaki River. In this analysis, an artificial neural network with eight hidden layers and 15 hidden neurons predict WQI with more than 90% accuracy. Also, Imani et al. (2021), applied an artificial neural network with a single hidden layer for predicting water quality resilience in São Paulo, Brazil. Applying neural networks to predict WQI required lots of water quality data, which is expensive and time- consuming. Ho et al. (2019), applied a decision tree for classifying water quality status in Klang River, Malaysia. They considered three scenarios where they used six water quality variables in the first scenario. After that, in each procedure, they removed water quality parameters such as NH3-N, pH, and SS to evaluate the decision tree algorithm's ability in different situations. They achieved 84.09%, 81.82%, and 77.27% classification accuracy in each scenario, which is higher than the 75% classification accuracy benchmark. Besides, to predict the WQI, Ahmed et al. (2019), used several supervised machine learning methods. They conducted their model on four water quality parameters. They found that by using gradient boosting and polynomial regression, the WQI is more successfully predicted where a multilayer perceptron classifies the water quality category more effectively. However, this study worked with fewer water quality parameters, but both proposed prediction and classification models did not show more than 75% accuracy. On the other hand, Wang et al. (2017), applied support vector regression to predict WQI. More than 90% of accuracy was achieved in this analysis. In this study, 22 specimens of water quality were used, which makes the model computationally costly. Li et al. (2019), proposed an amalgam model for the study of time-series water quality data by integrating a recurrent neural network with the Dempster- Shafer Theory (DST), where the RNN is capable of analyzing time-series data effectively to predict WQI and DST, which is a probability method used to amalgamate the outcome of RNNs. It can be challenging to predict WQI using RNN and DST since specialized handling of the data is required when fitting and testing the model. Besides, Ahmed et al. (2019), proposed a neuro-fuzzy inference method based on a wavelet-de-noise technique to predict water quality parameters. Results indicated that

this model outperformed the other neural network model, such as RBF and MLP. But the neuro- fuzzy inference method causes a curse of dimensionality problem, which occurs when high dimensional data is analyzed and classified.

In summary, from the above studies, most of the current approaches were based on a predictive model but did not provide any classification model comparisons. Most of the models showed less accuracy and used many water quality specimens. The proposed method is applied to address the limitations described in the current approaches above. Also, the proposed model gives a dynamic approach to use any number of water quality specimens.

To protect the environment and human health, treated wastewater discharge must be sampled and monitored in most developed countries to assure discharge permits are met [1]. In the past, scientists had to collect and analyze a large number of wastewater samples to understand how wastewater discharges components impacted the environment. The collection and analysis of treated wastewater effluents is time- consuming and costly. Machine learning methods are proposed to address the problem. The usage of machine learning methods would result in a reduction in sampling frequency and minimization of costs associated with analysis. At first, deterministic models and multivariate linear regression (MLR) analysis were used to speed up the process of evaluating the quality of wastewater effluent discharges [2] [3]. As a water quality dataset can be considered as a time series dataset, which is likely to have a complicated nonlinear relationship, the performance of deterministic and MLR models is expected to be poor. In the past decade, many machine learning techniques have been proposed to address the problem. Artificial neural networks (ANN) are adopted to explore the non- linear relationships residing in water quality datasets [3][4]. Various ANN models have been designed to predict water and waste water discharge quality based on previous existing datasets. A comprehensive comparison between ANN and MLR models for oxygen demand prediction has been performed [3]. The experimental results show that a three-layer neural network model outperforms an MLR model. In [4], neural network models are used to predict four parameters in the Qiantang River and the proposed model has higher accuracy and better stability in the experiment.

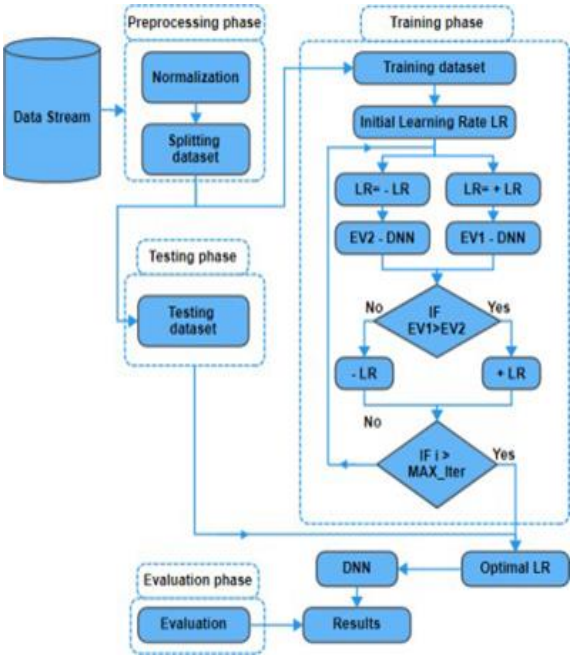
Although ANN models can effectively improve the prediction accuracy of water quality parameters, shortcomings still exist. Especially in some scenarios where the input parameters are ambiguous, neural networks struggle to formulate a non-linear relationship. Many studies have proven that an adaptive neuro-fuzzy inference system (ANFIS), which can integrate linear and non-linear relationships hidden in the dataset, is a better option in this scenario [5]. The experimental results in [6] show that an ANFIS model worked much better than an ANN model in predicting dissolved oxygen, even though there were only 45 data samples available. The experimental results confirm that the proposed method works. The ANFIS model has also been applied in effluent quality prediction, and an experiment with a dataset of around 150 data samples has proven that the ANFIS model is better than the ANN model [7]. Classifying the quality of drinking water accurately based on their physicochemical and microbiological parameters is a challenging problem in machine learning. It is in this lineage that the present contribution fits. In this paper, we used a real dataset to train different machine learning algorithms. From this research, we compared the accuracy and the precision of different algorithms.

PROBLEM STATEMENT

The water quality classification and predicting system using machine learning is not implemented with the efficient method to predict the water quality index (WQI).

The water quality parameters are not reduced in the existing approach, that allows for any water quality specimens.

For classifying the water quality status, a efficient classification model is not presented.



The efficient analysis is also not carried out on the dataset to determine the most dominant WQI parameter.

SCOPE

In recent decades most of the models, including artificial neural network, wavelet neural network, recurrent neural network, and decision tree, required lots of input parameters and computational power, which are considered expensive to construct such models. With this motivation, this project used water quality index (WQI), a combination of different water quality metrics that demonstrates the water quality condition of a particular region, and is applied both prediction and classification models to predict WQI and classify the water quality status. Principal component regression (PCR) is used to predict WQI in this analysis, combining both supervised and unsupervised techniques.

PCR's basic concept is that the principal component analysis is applied to the dataset to minimize the dimension. At the same time, a different algorithm is used for the PCA output. Since it can solve dataset multi-collinearity issues and allows fewer water quality specimens, PCR can predict the WQI more efficiently than the other techniques. The Gradient Boosting Classifier method is used in this study for the classification task. It is an ensemble technique that can operate with a small amount of data.

OBJECTIVES OF THE PAPER

To do extensive literature survey on water quality monitoring system.

The data set collection from the www.kaggle.com and also internet Sources.

Designing the model based on the appropriate algorithm.

Implementation for the Deigned model.

Training and testing the model with the data set.

To do the performance analysis for the model developed with best accuracy.

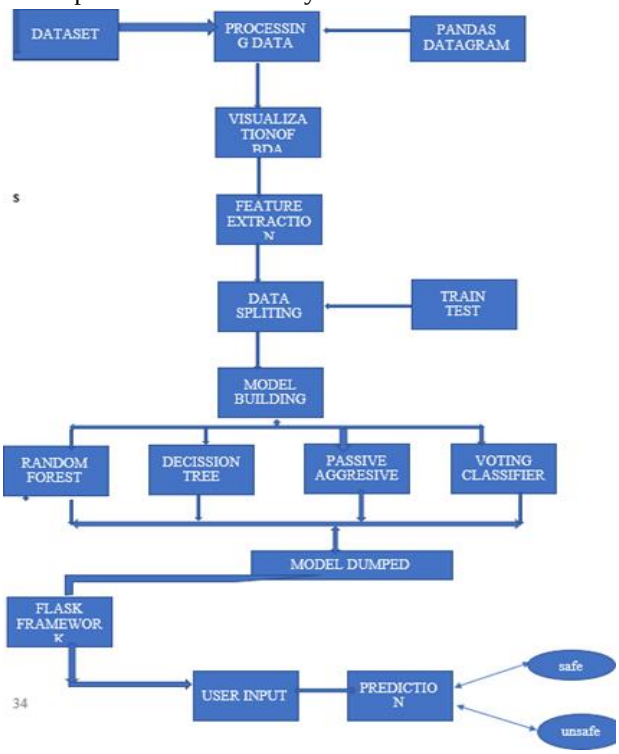


Fig: Architecture of Proposed Model

The Proposed model accepts the data set followed by the preprocessing, testing and training phase. The learning rate and evaluated rate is compared with different model to classify and predict the water quality

HARDWARE REQUIREMENTS

Processor : i3/i5

RAM : 4GB/8GB RAM

Hard Disk : 40 GB Hard Disk Space.

SOFTWARE REQUIREMENTS

Operating System : Windows 10/11 or Ubuntu 20.04

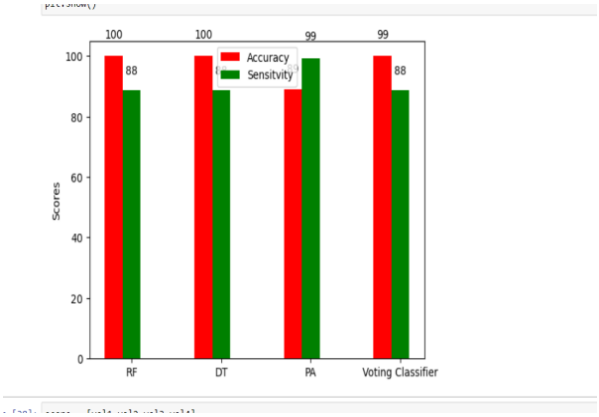
Coding Language : Python and HTML Front End : HTML, CSS, JS

Back End : Python and Sqlite

Tools :Jupyter Notebook and VS Code

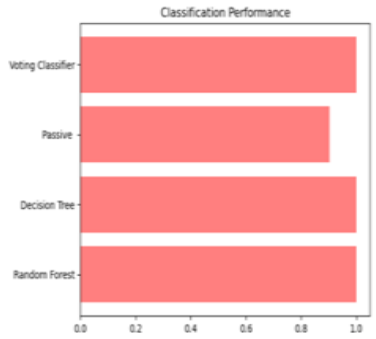
RESULTS AND DISCUSSION

This section explains the results attained by applying the proposed model (which consists of data set) to train the network by the model and hence, reducing the error rate as the training progress (models).



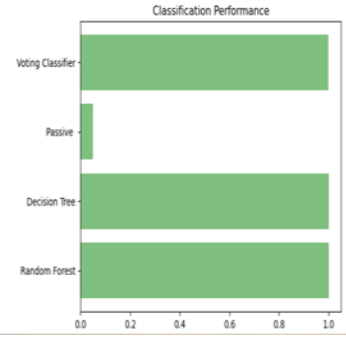
Precision

```
In [44]: plt2.barh(y_pos, score1, align='center', alpha=0.5,color='red')
         plt2.yticks(y_pos, classifier)
         plt2.xlabel('Precision Score')
         plt2.title('Classification Performance')
         plt2.show()
```



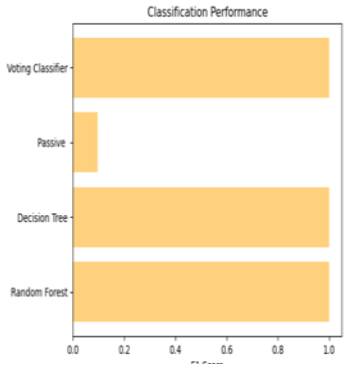
Recall

```
In [45]: plt2.barh(y_pos, score2, align='center', alpha=0.5,color='green')
         plt2.yticks(y_pos, classifier)
         plt2.xlabel('Recall Score')
         plt2.title('Classification Performance')
         plt2.show()
```



F1Score

```
In [46]: plt2.barh(y_pos, score3, align='center', alpha=0.5,color='orange')
         plt2.yticks(y_pos, classifier)
         plt2.xlabel('F1 Score')
         plt2.title('Classification Performance')
         plt2.show()
```



CONCLUSION

This paper proposed a performance analysis of the famous classification algorithms in the literature namely Decision Tree and SVMs using a real dataset retrieved from the water treatment station “Ghadir El Golla” of Tunis-Tunisia. The experimentation results gave us a good proof of the performances of the classification techniques. In addition, it's found that linear SVM seems adequate for our water quality monitoring system.

Our work has some limits: in fact, the water quality monitoring system is reactive. It cannot anticipate the intrusion which can act on the water quality.

For further study, we are trying to integrate a new data aggregation algorithm to minimize the amount of the collected data to run the SVM classification algorithm.

REFERENCES

[1]. Narsimha Adimalla Groundwater quality for drinking and irrigation purposes and potential health risks assessment: a case study from semi-arid region of South India Exposure and Health, 11 (2) (2019), pp. 109-123

[2]. Pariatamby Agamuthu, Dennis Victor Policy trends of extended producer responsibility in Malaysia Waste Management & Research, 29 (9) (2011), pp. 945-953

[3]. B. Aghel, A. Rezaei, M. Mohadesi Modeling and prediction of water quality parameters using a hybrid particle swarm optimization–neural fuzzy approach International Journal of Environmental Science and Technology, 16 (8) (2019), pp. 4823-4832

[4]. Ali Najah Ahmed, et al. Machine learning methods for better water quality prediction Journal of Hydrology, 578 (2019), Article 124084
[5]. Umair Ahmed, et al. Efficient water quality prediction using supervised Machine Learning Water, 11 (11) (2019), p. 2210
[6]. Seyed Babak Asadollah, Haji Seyed, et al. River water quality index prediction and uncertainty analysis: A comparative study of machine learning models Journal of Environmental Chemical Engineering, 9 (1) (2021), Article 104599
[7]. Rahim Barzegar, Asghar Asghari Moghaddam Combining the advantages of neural networks using the concept of committee machine in the groundwater salinity prediction Modeling Earth Systems and Environment, 2 (1) (2016), p. 26
[8]. Barzegar, Rahim, Mohammad Taghi, Aalami, Jan, Adamowski, 2020. Short- term water quality variable prediction using a hybrid CNN–LSTM deep learning model. Stochastic Environmental Research and Risk Assessment, pp. 1–19.
[9]. Duie Tien Bui, et al. Improving prediction of water quality indices using novel hybrid machine-learning algorithms Science of The Total Environment, 721 (2020), Article 137612
[10]. Bahram Choubin, et al.Multiple linear regression, multi-layer perceptron network and adaptive neuro-fuzzy inference system for forecasting precipitation based on largescale climate signals
[11]. Gulshan Lake, 2016. Published on May 20, 2018.URL:http://www.doe.gov.bd/site/ publicatio ns/5132a8d7-68e9-469d-a9af-898130 6b3b9f/ Surface-and-Ground-Water-Quality-Report-2016.
[12]. Bloodless Dzwairo, et al. Assessment of the impacts of pit latrines on groundwater quality in rural areas: a case study from Marondera district, Zimbabwe
[13]. Salam Hussein Ewaid, Salwan Ali Abed, Safaa A. Kadhum Predicting the Tigris River water quality within Baghdad, Iraq by using water quality index and regression analysis Environmental Technology & Innovation, 11 (2018), pp. 390-398
[14]. Satyajit Gaikwad, et al.Geochemical mobility of ions in groundwater from the tropical western coast of Maharashtra, India: implication to groundwater quality Environment, Development and Sustainability, 22 (3) (2020), pp. 2591- 2624
[15]. Mohammed Hameed, et al. Application of artificial intelligence (AI) techniques in water quality index prediction: a case study in tropical region, Malaysia Neural Computing and Applications, 28 (1) (2017), pp. 893-905
[16]. Jun Yung Ho, et al. Towards a time and cost effective approach to water quality index class prediction Journal of Hydrology, 575 (2019), pp. 148-165
[17]. Robert K. Horton An index number system for rating water quality Journal of Water Pollution Control Federation, 37 (3) (1965), pp. 300-306
[18]. Imani, Maryam, et al., 2021. A novel machine learning application: Water quality resilience prediction Model. Science of the Total Environment 768, 144459.
[19]. A.K. Kadam, et al. Prediction of water quality index using artificial neural network and multiple linear regression modelling approach in Shivganga River basin, India Modeling Earth Systems and Environment, 5 (3) (2019), pp. 951-962
[20]. Devashish Kar Wetlands and Lakes of the World Springer New Delhi, India (2013) Kar, 2019. Wetlands and their Fish Diversity in Assam (India). Transylvanian Review of Systematical and Ecological Research 21 (3), 47–94.
[21]. Khadr, Mosaad, 2017. Modeling of water quality parameters in Manzala lake using adaptive neuro-fuzzy inference system and stochastic models. In: Egyptian Coastal Lakes and Wetlands: Part II. Springer, pp. 47–69.
[22]. Ozgur Kisi, et al. Modeling groundwater quality parameters using hybrid neuro- fuzzy methods Water Resources Management, 33 (2) (2019), pp. 847-861
[23]. Wei Cong Leong, et al. Prediction of water quality index (WQI) using support vector machine (SVM) and least square- support vector machine (LS-SVM) International Journal of River Basin Management (2019), pp. 1-8
[24]. Zhenbo Li, et al. Water quality prediction model combining sparse auto-encoder and LSTM network IFAC- PapersOnLine, 51 (17) (2018), pp. 831-836
[25]. Lei Li, et al. Water quality prediction based on recurrent neural network and improved evidence theory: a case study of Qiantang River, China Environmental Science and Pollution Research, 26 (19) (2019), pp. 19879-19896

[26]. Reza Mohammadpour, et al. Prediction of water quality index in constructed wetlands using support vector machine Environmental Science and Pollution Research, 22 (8) (2015), pp. 6208-6219

[27]. Sang-Ki Moon, Nam C. Woo, Kwang S. Lee Statistical analysis of hydrographs and water-table fluctuation to estimate groundwater recharge Journal of Hydrology, 292 (1-4) (2004), pp. 198-209

[28]. Oelen, Allard, van Aart, Chris J., De Boer, Victor, 2018. Measuring surface water quality using a low-cost sensor kit within the context of Rural Africa. In: P- ICT4D@ WebSci.

[29]. Johnson O. Oladipo, et al. Comparison between fuzzy logic and water quality index methods: A case of water quality assessment in Ikare community, Southwestern Nigeria Environmental Challenges, 3 (2021), Article 100038

[30]. Shafkat Shamim Rahman, Md Mahboob Hossain Gulshan Lake, Dhaka City, Bangladesh, an onset of continuous pollution and its environmental impact: a literature review Sustainable Water Resources Management, 5 (2) (2019), pp. 767-777

[31]. Taher Rajaee, Salar Khani, Masoud Rav ansalar Artificial intelligence-based single and hybrid models for prediction of water quality in rivers: A review Chemometrics and Intelligen Laboratory Systems, 200 (2020), Article 103978

[32]. S. Mehdi Saghebian, et al. Ground water quality classification by decision tree method in Ardebil region, Iran

[33]. Marjan Salari, et al. Quality assessment and artificial neural networks modeling for characterization of chemical and physical parameters of potable water Food and Chemical Toxicology, 118 (2018), pp. 212-219

[34]. Mohd Saiful Samsudin, et al. Comparison of prediction model using spatial discriminant analysis for marine water quality index in mangrove estuarine zones

[35]. Tadesse A. Sinshaw, et al. Artificial neural network for prediction of total nitrogen and phosphorus in US Lakes

[36]. Shweta Tyagi, et al. Water quality assessment in terms of water quality index

[37]. Xiaoping Wang, Fei Zhang, Jianli Ding Evaluation of water quality based on a machine learning algorithm and water quality index for the Ebinur Lake Watershed, China World Bank Report, 2018 World Health Organization et al., 2004

[38]. Yiping Wu, Shuguang Liu Modeling of land use and reservoir effects on nonpoint source pollution in a highly agricultural basin

[39]. Longqin Xu, Shuangyin Liu Study of short-term water quality prediction model based on wavelet neural network

[40]. Hiroshi Yajima, Jonathan Derot Application of the Random Forest model for chlorophyll-a forecasts in fresh and brackish water bodies in Japan, using multivariate long-term databases

[41]. Mulugeta Yilma, et al. Application of artificial neural network in water quality index prediction: a case study in Little Akaki River, Addis Ababa, Ethiopi

[42]. Yanyang Zhang, et al. Integrating water quality and operation into prediction of water production in drinking water treatment plants by genetic algorithm enhanced artificial neural network

[43]. Senlin Zhu, et al Two hybrid data- driven models for modeling water-air temperature relationship in rivers