Predictive Sensor for Biological Oxygen Demand in water using Active Learning based Random Forest Algorithm

***Rishi Khare, *# A.A.Khurshid, *Aditya Jain, *Shivam Shukla**

*Shri Ramdeobaba College of Engineering and Management, Nagpur, #khurshidaa@rknec.edu

Abstract

The environmental pollution caused by human beings, and industries is interfering with water sources that bound the living zones in marine environments. Greywater reuse after treatment can be carried out for non-consumable water use and hence Biochemical oxygen demand (BOD) is a vital parameter in deciding the nature of waste water created. The measurement of BOD involves difficulties and performing the BOD test takes too long and the result does not remain relevant for the current wastewater. This work proposes an ensemble learning-based random forest model with active learning to iteratively select samples with large ambiguity for reducing errors in predicting BOD. The model uses minimal basic physical & chemical parameters of water and hence does not require specialized ion-selective and costly sensors for measurement. Feature importance is exploited to enhance the efficiency of the machine learning model leading to a prediction accuracy of 81.21%. The developed model could be used as a predictive sensor at the edge of the water quality monitoring system.

Keywords : Water Quality, BOD, Machine Learning, Active Learning

DOI Number: 10.14704/nq.2022.20.9.NQ44230 **Neuro Quantology 2022; 20(9):1983-1988**

1.Introduction

The environmental pollution caused by human beings, and industries have a detrimental effect on living creatures. Different water management approaches are been set up in dry regions . Though the biosphere's water supply is from seas, still a limited percentage of it can be used directly [1]. Therefore, the necessity of monitoring water quality is essential. Greywater reuse after treatment can be carried out for non-consumable water use and hence Biochemical oxygen demand (BOD) is an important parameter in deciding the nature of wastewater created. BOD can be used as a measure of the effectiveness of wastewater treatment plants also. BOD is an estimated measure of biochemically degradable organic matter in water. The measurement of BOD is difficult as performing the BOD test takes too long and the result does not remain applicable for the present water. For safe usage of water , it is essential to develop a model for forecasting the BOD value on the basis of previous observations of water quality parameters. As the parameters exhibit nonlinear behaviors, it is hard to define them by mathematical models[2].

2.Literature Review

Prediction of BOD values using traditional statistical methods, may lead to low performance in comparison to machine learning-based methodology. Several researchers have directed their work towards evolving a model which could rapidly predict BOD based on statistical and deterministic methods [3–5]. Artificial neural networks also have been successfully used in this connection and some current experiments prove its validation [6-7]. Also, different intelligence networks based on convolutional neural network and long-short-term memory [8], deep belief networks [9] ,modular neural network [10] are used in the estimation of parameters in wastewater . Latest research work on optimization algorithms [11,12, 13] indicate that researchers have developed hybrid methods based on wavelet transforms. It was found that hybrid models yield better models . Reference [14] also investigates the use of ANN model to increase the efficiency of assessment process for BOD and the limit of ANN technique for river water was investigated. The researchers in reference[15] four layer learning model with machine learning methods such as random forest

(RF), support vector regression (SVR) and multilayer perceptron in the second layer , third layer to tune the hyperparameters for optimisation and fourth layer to predict the BOD values with the model having highest correlation coefficient in water samples .The number of features used are 48. The authors also presented the performance of models using selected features and concluded that the significant features with reference to the correlation of the BOD values include Ammonium, COD and nitrite. Researchers in reference[16] have developed an adaptive neuro fuzzy system to predict BOD using phosphate, total dissolved solid , chemical oxygen demand(COD), dissolved oxygen and total inorganic nitrogen as inputs to the proposed system. The SVM model proposed in [17] accounts for use of pH, Temperature, Free Ammonia, COD, Total Solids and Ammonia Nitrogen as predictors.

From the literature , it can be concluded that most of the researchers are making use of parameters such as COD, ammonia, and nitrogen which require specialized ion-selective and costly sensors for measurement. Due to various issues in the assessment of BOD , this work is an attempt to find the best model for predicting BOD using basic physical and chemical parameters of water. The dataset used for experimentation is collected from different locations in India. It contained 1767 samples from various parts of India during the period from 2005 to 2014. The dataset consists of parameters, namely, fecal coliform, dissolved oxygen (DO), pH, biological oxygen demand (BOD), conductivity, nitrate, total coliform, and fecal streptococci etc. Data were collected from rivers by the Indian government to safeguard the supply of drinking water quality. Dataset is collected from http://www.cpcbenvis.nic.in/water_quality_data.html# which helped to train the model. A small snippet of the data set used is shown below in table 1

Table 1: Water Parameters from Indian Rivers

Station Code	Name of Monitoring Location	State Name	Temperature °C		Dissolved Oxygen (mg/L)		pH		Conductivity (µmhos/cm)		BOD (mg/L)	
			Min	Max	Min	Max	Min	Max	Min	Max	Min	Max
2354	SAMARLA KOTA CANAL, KAKINADA, EAST GODAVARI	ANDHRA PRADESH	22	30	2.5	6	7.1	8.4	208	536	1.2	4.2
2355	TULJE BAGH CANAL, TEKRI DRAIN, KAKINADA, EAST GODAVARI	ANDHRA PRADESH	26	29	1	7.3	6.5	7.8	758	39096	2.5	12.4
3051	BUDAMERU CANAL NR BDG AT NH-5, KEESARAPALLY, KRISHNA	ANDHRA PRADESH	25	26	5.1	5.9	6.8	7.6	899	1860	2.8	9.2
4356	ELERU CANAL NEAR PHARMA CITY	ANDHRA PRADESH	23	28	4.9	6.6	7.4	8.1	248	398	1.5	2.5
4370	KRISHNA CANAL AT HANUMAN NAGAR, NEAR SAIBABA TEMPLE, ELIRU	ANDHRA PRADESH	21	28	1	6	7.1	8.3	196	1408	1.8	4.8
4374	GOSTTA NADI- VELPURU CANAL AT HANUMAN TEMPLE, DOWNSTREAM OF TANUKU TOWN, ATTILI (M)	ANDHRA PRADESH	21	28	0.8	7	6.7	8.2	248	1980	1.7	6

3.Methodology

3.1 Preprocessing:

For designing a machine learning model with minimal computational complexity , it is essential to provide relevant and effective features at its input to generate the required single output from large amounts of data. These factors may affect the correctness of the output at different coefficients and degrees and hence can be sifted out based on their role in determining the output and redundancy. To understand the significance of each feature/parameter , statistical measure which can indicate the degree to which these features vary together is used. Pearson correlation statistics is used to quantify the degree of relationship amongst the various water parameters. It can be concluded from the below correlation matrix, figure 1 that dissolved oxygen(DO), temperature(T) ,and pH have the utmost dependencies on values in the range 0.01

- 0.3. Considering these threshold values, the above parameters are chosen which exhibit high dependencies with BOD.

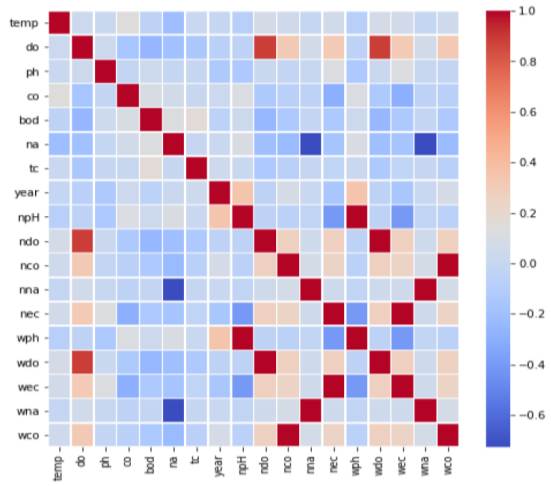

Figure 1: Correlation Matrix for various water quality parameters

The processing phase is crucial in data analysis to enhance the quality of data . For gaining greater accuracy, the data normalization procedure employed is z -score . Data cleaning and binning is performed to remove noisy data.

3.2 Model selection and Experimentation:

3.2.1 Weighted Random Forest with Active Learning:

Numerous models have been proposed to compute dissolved oxygen concentration in water using machine learning as presented in section 2. In this study, Random Forest(RF) regressor algorithm which makes use of a huge quantity of distinct decision trees functioning collectively with a specific tree predicting a class is employed. The majority class vote is then predicted by the model. The implementation of weighted RF is created on the basis of the traditional random forest algorithm proposed by Breiman [18] by adding weight to features based on their importance. As some features are significant to a learning problem, a measure of importance can be provided to them. Active learning is incorporated and the algorithm selects samples with large ambiguity and this criterion is used for assessing the "fitness" of the solution. The expected integrated squared difference is used as a cost function [19] to be minimized thus enabling the algorithm to find new training data with potential utility from specific sections of the input data space, leading to faster training and reduction in approximation error.

To demonstrate the effectiveness, the active learning-based RF is compared with other tree-based learning algorithms which provide faster training and higher efficiency ie. light GBM and Support vector machine (SVM). The prediction accuracy is shown in Table 2. Table 2 also indicates the performance of the RF algorithm with a wide range of eighteen (physical, biological and chemical, variables of water) and three (physical and chemical variables of water) input features to the model. The model is trained on the dataset as mentioned in section 2 of Indian rivers and cross-validation is used for testing. The process is repeated with different test data each time and the average performance is noted. The geographical diversity of the data is explored to develop a robust and accurate model and the prediction error is depicted in Figure 3 and Figure 4.

Table 2: Prediction Accuracy for different models

Sr. No.	Model Name	Accuracy (random sample)
1.	Random Forest Regressor (18 input parameters)	65.05 %
2.	Light GBM (3 input parameters)	34.90 %
3.	Support Vector Regressor (3 input parameters)	64.11 %
4.	Active learning-based Random forest (3 input parameters)	72.45%
5.	Weighted active learning-based Random Forest Regressor	81.21 %

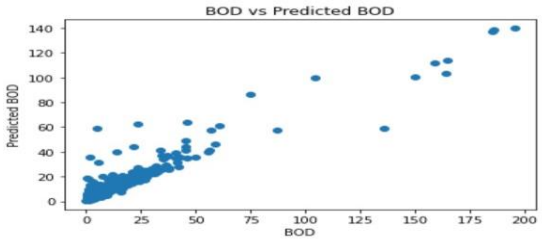

Figure 2 : Error plot for random forest regressor

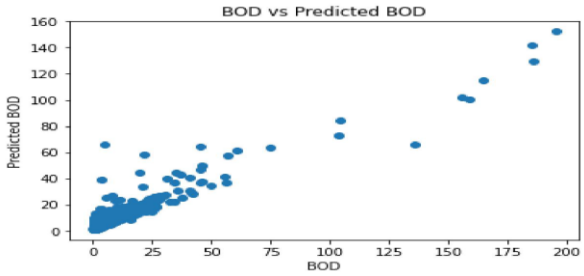

Figure 3 : Error plot for active learning-based random forest regressor with weighted features

Table 3 compares the features of different proposed models in the literature surveyed with the weighted Random Forest active learning model with three water quality variables as input. The proposed active learning framework can be a potential candidate for BOD forecasting and meets the desired objective by utilizing a minimum number of input water quality variables.

Table 3: Comparison of weighted RF with active learning model with different proposed models

Parameters	[6]	[10]	[15]	[20]	Proposed
Algorithm	Feed Forward Artificial Neural network	Constructive Radial Basis Function neural network	multiple machine learning methods RF, SVR and multilayer perceptron with the optimization of hyper - parameters carried out using a genetic algorithm	Backpropagation neural network	Weighted RF with active learning
No. of input parameters	8	10	6	10	3
Type of water quality parameters	Physical + Biological+ Chemical	Chemical +Physical + Biological	Biological+ Physical +Chemical	Physical + Chemical + Biological	Chemical +Physical

4. Conclusion

Active learning and feature importance-based BOD concentration prediction model is implemented using water quality data set of rivers with minimal input parameters and computational cost. After testing several algorithms like Linear Regression, Decision Tree, and Neural Network, Random forest has been the best one to move forward with a predictive model with improved performance. The advantage of random forest is its random selection of features in the training process. Hence, it is independent of a definite feature set. The proposed model is able to capture long-term trends detected in the presented data of water quality parameters.

The correlation matrix helped to understand what factors are the most important for the prediction of BOD. With the help of limited parameters in the dataset, a model for the prediction of water BOD is built which can also be used in the sewage treatment plant and other industries. The main motive behind this work to propose a virtual sensor with minimal input parameters is achieved. This approach can be valuable for areas that are appropriate for instance-based learning with features of unequal relevance.

Future work can be planned to explore different ways to find feature importance from the network itself. In the future study, the effect of feature importance factors at different layers/levels can be explored with different types of water bodies too.

5. References

[1] Oteng-Peprah M, Acheampong M, DeVries N, (2018) "Greywater Characteristics, Treatment Systems, Reuse Strategies and User Perception—a Review". Water Air, & Soil Pollution; volume 229, Article number: 255

[2] I. Plazl, G. Pipus, M. Drolka, T. Koloini, Parametric sensitivity and evaluation of a dynamic model for single-stage wastewater treatment plant, Acta Chim. Slov., 46 (1999) 289–300

[3] K.P. Singh, A. Basant, A. Malik, G. Jain, Artificial neural network modeling of the river water quality a case study, Ecol. Model., 220 (2009) 888–895.

[4] X. Wen, J. Fang, M. Diao, C. Zhang, Artificial neural network modeling of dissolved oxygen in the Heihe River, Northwestern China, Environ. Monit. Assess., 185 (2013), 4361–4371.

[5] A.N.S. Tomić, D.Z. Antanasijević, M.Đ. Ristić, A.A. PerićGrujić, V.V. Pocajt, Modeling the bod of Danube River in Serbia using spatial, temporal, and input variables optimized artificial neural network models, Environ. Monit. Assess., 188 (2016).

[6] Dogan E., Sengorur B. & Koklu R. 2009 Modeling biological oxygen demand of the Melen River in Turkey using an artificial neural network technique. Journal of Environmental Management 90 (2), 1229–1235.

[7] O.T. Baki, E. Aras, Estimation of BOD in wastewater treatment plant by using different ANN algorithms, Membr. Water Treat., 9 (2018) 455-462

[8] R. Barzegar, M. T. Aalami, and J. Adamowski, "Short-term water quality variable prediction using a hybrid CNN–LSTM deep learning model," Stochastic Environmental Research and Risk Assessment, vol. 34, no. 8, pp. 1–19, 2020.

[9] Li Liang, "Water Pollution Prediction Based on Deep Belief Network in Big Data of Water Environment Monitoring", Scientific Programming, vol. 2021, Article ID 8271950, 11 pages, 2021

[10] Wenjing Li 1,2,* and Junkai Zhang," Prediction of BOD Concentration in Wastewater Treatment Process Using a Modular Neural Network in Combination with the Weather Condition", Faculty of Information Technology, Beijing University of Technology, Published: oct-2020

[11] Deng W., Xu J., Gao X.-Z. & Zhao H. 2020 An enhanced MSIQDE algorithm with novel multiple strategies for global optimization problems. IEEE Transactions on Systems, Man, and Cybernetics: Systems 1–10.

[12]Cai X., Zhao H., Shang S., Zhou Y., Deng W., Chen H. & Deng W. 2021 An improved quantum-inspired cooperative co-evolution algorithm with muli-strategy and its application. Expert Systems with Applications 171, 114629.

[13]Li P., Hua P., Gui D., Niu J., Pei P., Zhang J. & Krebs P. 2020 A comparative analysis of artificial neural networks and wavelet hybrid approaches to long-term toxic heavy metal prediction. Scientific Reports 10, 13439.

[14] Dogan E, Ates A, Ceren Yilmaz E, Eren B (2008), "Application of Artificial Neural Networks to Estimate Wastewater Treatment Plant Inlet Biochemical Oxygen Demand". Environ Prog, 27(4):439– 445

[15] Kai Sheng Ooi; ZhiYuan Chen; Phaik Eong Poh; Jian Cui, BOD5 prediction using machine learning methods , Water Supply (2022) 22 (1): 1168–1183

[16] Belouz Khaled; Aidaoui Abdellah; Dechemi Noureddine; Heddam Salim; Aguenini Sabeha, Modelling of biochemical oxygen demand from limited water quality variable by ANFIS using two partition methods, Water Quality Research Journal (2018) 53 (1): 24–40

[17] Manu, D.S., Thalla, A.K. Artificial intelligence models for predicting the performance of biological wastewater treatment plant in the

removal of Kjeldahl Nitrogen from wastewater. Appl Water Sci 7, 3783–3791 (2017).

[18] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[19] Kah Sung, Partha Niyogi, Active Learning for Function Approximation,vPart of Advances in Neural Information Processing Systems 7 (NIPS 1994)

[20] Hossein Banejad , Ehsan Olyaie, Application of an Artificial Neural Network Model to Rivers Water Quality Indexes Prediction – A Case Study, Journal of American Science, 2011;7(1)