2023 2nd International Conference on Paradigm Shifts in Communications Embedded Systems, Machine Learning and Signal Processing (PCEMS) | 979-8-3503-1071-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/PCEMS58491.2023.10136050

A TinyML Approach for Quantification of BOD and COD in Water

Sanket Soni
Department of Electronics,
Shri Ramdeobaba College of Engineering and Management
Nagpur, India
sonisv@rknec.edu

Aleefia Khurshid
Department of Electronics,
Shri Ramdeobaba College of Engineering and Management
Nagpur, India
khurshidaa@rknec.edu

Anushree Mrugank Minase
Department of Electronics,
Shri Ramdeobaba College of Engineering and Management
Nagpur, India
minaseam@rknec.edu

Ashlesha Bonkinpelliwar
Department of Electronics,
Shri Ramdeobaba College of Engineering and Management
Nagpur, India
bonkinpelliwaraa_1@rknec.edu

***Abstract*—Water quality prediction is a crucial process before any consumption of water. Prediction and modeling methods are used for pollutants in water to deal with water pollution control. This work involves the use of a random forest learning algorithm to quantitate BOD and COD using parameter tuning to establish the importance of input variables. It uses minimal sensed quantitative parameters such as Temperature, pH, DO, and Conductivity along with categorical parameters. The trained model shows excellent efficiency compared to other models and is validated using the laboratory test results with a maximum error of 10%. It is computationally low-cost, requires minimal parameters, and is pruned to integrate and implement in an IoT hardware system, reducing the cost of expensive sensors.**

Keywords— Decision Tree Algorithm, Random Forest, K-Nearest Neighbor, Support Vector Machine

I. INTRODUCTION

Water is considered to be one of the most essential resources and the ultimate survival necessity. Numerous individuals pass away each year and with agricultural land sinks, there is a lot of water contamination. Therefore, an efficient mechanism must be introduced to examine the quality of water. A lot of industrial trash is dumped in rivers and lakes, making it highly hazardous. This led to a severe problem of polluting the water. Some local tests and field kits are used to evaluate the water quality, giving expected results but can take up to 3-5 business days.

This simple procedure is time-consuming and cannot be utilized on a commercial foundation. To determine the water grade, a real-time water quality system is required. Another applicable method is the usage of sensors. Two of the most important indicators of the total amount of pollutants in water are chemical oxygen demand (COD), which is the oxygen equivalent measurement of the organic matter in a water sample that is susceptible to oxidation by potent chemical oxidants, and biological oxygen demand (BOD), which is a calculation of the dissolved oxygen needed by aerobic biological organisms in a water body to break down the organic material present in it. The commercially available online sensors for the prediction of BOD and COD are very costly.

To determine the aforementioned parameters, the system uses different machine learning methods. The created model can be used as a real-time monitoring tool for estimating water quality for effective river pollution management in metropolitan settings. Therefore, the objectives of this work were to examine variations in water parameters over six months and estimate the degree of organic pollution based on the correlations between the acquired components using minimal water parameters. A minimal Random Forest

979-8-3503-1071-9/23/$31.00 ©2023 IEEE

based regression model is established for the purpose based on the samples collected from the river Ganga, where the water is infected by the effluents and various other urban activities.

II. RELATED WORK

The related literature reviewed is abridged below. A data-driven Machine Learning technique is implemented to quantify BOD in real-time using a soft sensor BOD model. With the use of a case study from a real-world application, this research proves that BOD soft sensors are effective, and the IBK ML method proves to be the most effective for predicting BOD. To examine the effectiveness of the IBK technique, 100 test readings of STP water samples were used in the experiment, and the statistical metric, RMSE is equal to 0.1994, with an edge response time of 0.15 s only. [1].

Research has provided new models as a result of which the requirement for greater accuracy in water quality modeling is expected. The latest research work on optimization algorithms [2, 3, 4] indicates that researchers have developed hybrid methods based on wavelet transforms.

After a lot of research, some findings have suggested a novel hybrid strategy that uses a modular neural network (MNN) can provide high accuracy. The samples can be grouped and combined with the weather state parameters for BOD prediction. The training RMSE was less than 0.01. The training time is reduced and the network is compressed as presented by Wenjing Li et al. [5].

Arunima Pattanayak et. al. [6] demonstrated that machine learning algorithms output indicates that the KNN technique is currently best suited for the COD quantification regarding reaction time and error parameters for the water samples taken from the Ganga River using ten water variables as input to the model.

Gradient-boosted decision trees are suggested by Zifei Wang et al. [7] as a method of predicting COD load in wastewater. This algorithm is memory and time-consuming. To estimate COD, Sani Isa Abba et al. [8] created an ANN model with notable accuracy utilizing eight different parameters. The review of the literature led to the conclusion that one of the key elements in determining content is COD.

Davut Hanbay et al. [9] suggest using wavelet and neural networks for COD prediction. This method requires a bulky computation and is a shift variant. This research [10] uses regression algorithms to determine results and six minimal remotely sensed parameters to predict COD. Weighted multiple linear regression is used to create the model.

Greywater can create some critical errors and henceforth can't be used straightforwardly making it time-consuming and giving inaccurate results. The following study [11] suggests the multivariate Model for COD and BOD load calculation with R-squared equal to 0.9973 using four parameters. Here the model fitting on a straight line becomes more challenging as the number of independent variables increases.

The work presented in [12] analyzed the efficiency of multivariate linear regression (MLR) and artificial neural network (ANN) models. The experimentation indicates that pH has more effect on BOD and COD load calculations in comparison to other parameters. Also, the demonstrated models predict BOD well than COD.

Regression equations, according to the researchers [13], would make it easier to evaluate the effluent quality and enable the choice of the best procedures for reducing microbiological contamination or organic elements in wastewater. The authors propose the link between average BOD and COD that will give a much more accurate and quick analysis.

In order to estimate Total Phosphorus (TP) and COD, Abhilash Nair et al. [14] suggested identifying significant relationships, which were then used to create a mathematical model.

From the literature survey, it can be concluded that several problems in record keeping and quantification of water quality parameters like BOD and COD exist, and the developed models are either computationally intensive or require a large number of parameters to predict the load. The efforts are therefore directed in this study towards finding an optimized topology to predict complex water parameters (BOD & COD) using minimal input so that the treatment plants utilize minimal energy to accurately control the effluents. To accomplish this goal, flexible mathematical structures are investigated. By comparing the results of these models on the dataset from the Ganga river, the best successful prediction model is then selected [15]. The developed model with minimal input parameters based on Random Forest Regression is implemented on Raspberry Pi to validate the objective of this experimentation which can be used at the edge in an IoT environment.

III. PROPOSED METHODOLOGY

A. Dataset

The "Namami Ganga" project's live data streaming was used to compile the database (30,000 samples), which includes crucial variables including pH, DO, BOD, COD, Temperature, and Conductivity (EC) for six months to include variations due to weather changes. Table 1 shows the scrapped sample data.

TABLE.1. Sample Data

BOD	DO	EC	pH	Temp.	COD
0.70	6.93	298.0	8.11	30.5	11.30
6.08	0.44	1961.0	7.26	29.7	43.42
2.37	7.43	269.0	7.76	30.7	16.87
2.91	5.91	309.0	7.09	30.6	13.56
19.40	0.88	802.0	8.31	29.2	60.40
27.11	1.09	725.0	8.02	29.4	105.59
3.55	5.85	204.0	8.30	30.3	15.22
4.06	2.66	237.0	7.64	30.0	18.28
2.05	6.48	264.0	8.36	30.7	12.17
16.87	0.39	587.0	7.71	28.2	49.06

B. Data Pre-processing

For improved quality of data, pre-processing is undertaken. Data cleaning and transformation increase the data quality. Using the linear scaling approach, data was normalized and about 23,000 samples of data were collected after the cleaning process.

C. Correlation Analysis and Parameter Tuning

In order to understand the relationship between the parameters used in this study, a correlation heat map is used to derive the importance of each variable. The variables with a correlation coefficient greater than 0.25 are used to simulate the different learning models. The correlation heat map for the derived parameters is shown in figure 1.

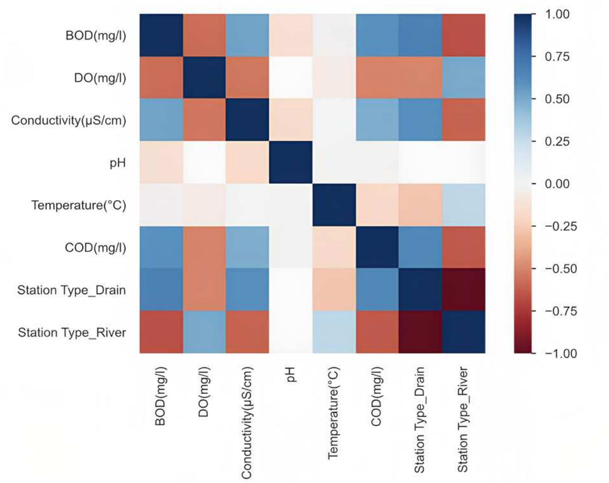


Fig. 1. Correlation Heat Map

The outcome is driven by variables of great importance, and the values of these variables have a big impact on the output. The weighting assigned to each variable depends on how much accuracy is lost when the variable is removed. One of the most crucial aspects of modeling is choosing the input parameters and after parameter tuning using scoring error, the relative importance of each variable is thus derived as shown in figure 2.

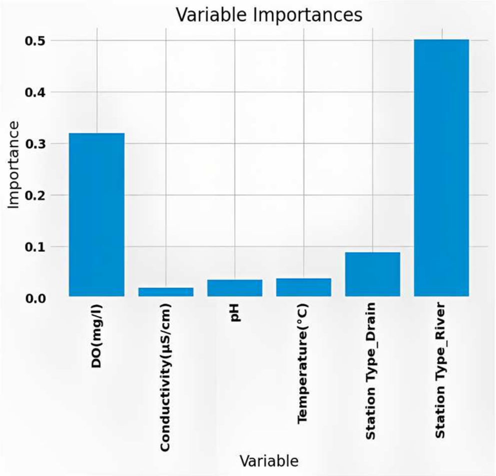


Fig.2. Importance of Derived Input Parameters

D. Data Partition

The train-test split method is applied to this data. 20% of the data is used in testing and 80% data is in training.

E. Machine Learning Algorithms

Water quality prediction can be done using various learning algorithms. The characteristics of data in training and testing can affect the efficiency of learning algorithms. For a low-cost model, intelligent algorithms such as Regression, SVM, KNN, Decision Tree, and Random Forest are tested for the prediction of the BOD and COD values. Based on the correlation map and the interaction between different parameters and their importance, four water parameters i.e., pH, DO, EC, and temperature form the input to the model.

In order to streamline the data analysis, the categorical data also forms the input to the trained model. The random forest learning method (figure 3) required fewer resources than other models and even produced better prediction results where the final output is determined based on the averaging method for both the output parameter.

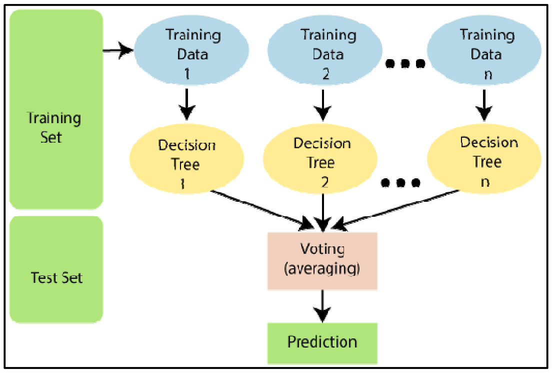


Fig. 3. Steps involved in Random Forest Algorithm

IV. RESULTS

A. Performance of Machine learning algorithms

Checking the accuracy of learning algorithms is an important step to determine their efficiency. The evaluation metrics used for this purpose are Mean squared error (MSE), Mean absolute error (MAE), and Root mean square error (RMSE). These metrics depict that Random Forest is the best-performing model among the applied algorithms. Table 2 lists the performance metric values of all these models.

TABLE 2. Evaluation metrics of all machine learning algorithms

Algorithm	**MAE**	**MSE**	**RMSE**
Multiple Linear Regression	5.82	342.26	18.50
Linear Regression	5.75	357.60	18.91
SVM	11.75	680.74	26.09
KNN	5.61	401.40	20.03
Decision Tree	1.46	8.50	2.91
Random Forest	0.55	3.45	1.86

B. Performance Analysis of Random Forest Algorithm

The Random Forest algorithm proves to be efficient compared to the other machine learning models. Figure 4 and Figure 5 show the predicted and actual values of the samples during the testing phase for BOD and COD indicating high accuracy for both parameters with a maximum error of 2%.

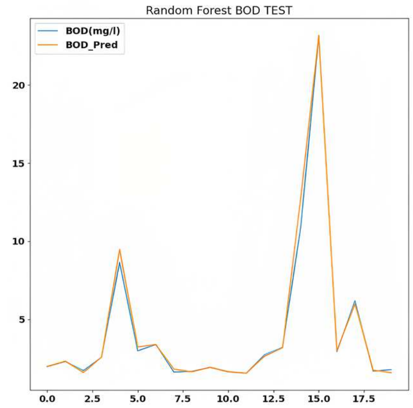


Fig.4. Predicted values of BOD for Random Forest model

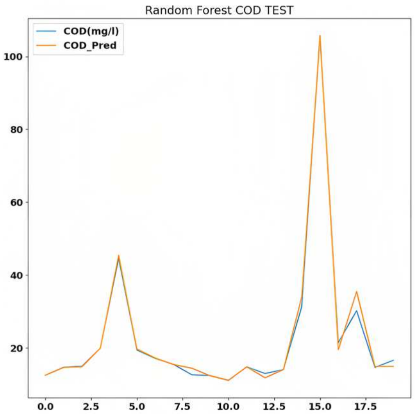


Fig.5. Predicted values of COD for Random Forest model

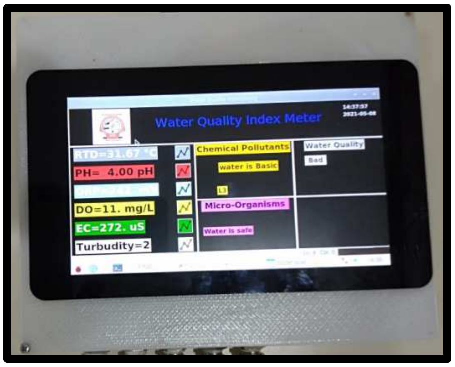
TABLE.3. Comparison table

Ref no.	Algorithm used	No. of inputs	Predicted Parameter	RMSE
[1]	Random forest	6	BOD	0.0359
[5]	MNN	9	BOD	0.0079
[6]	KNN	10	COD	2.93
[8]	ANN	8	COD	0.0108
[14]	SLSPQ	6	COD	3.48
-	Proposed method	4	BOD & COD	1.86

Table.3 above compares results of the proposed method with the most recent literature to the outcomes of the suggested approach. The proposed method delivers acceptable performance while using fewer input variables, as seen in the table 3. Moreover, it simultaneously forecasts BOD and COD, which amplifies its advantage.

Further, the computational complexity at test time for a Random Forest model depends on the number of decision trees (n_estimators) and the number of input features (m). The complexity was calculated to be as O (n_estimators * m log n). Where, n_estimators = 20, the number of input features (m) = 6, and the number of samples (instances) in the test data (n) = 4557.

Also, the additional samples collected from other urban water bodies, are tested in the NABL accredited laboratory and these un-seen samples are used to validate the trained model. The prediction error is in the range of 5 to 10% for these samples. The training time for the model run on an M1 processor operating at 3.2 GHz is 722 ns and the prediction time equals 48.7 ns. The pruned model is also implemented on Raspberry Pi 4 (figure 6) to integrate the developed soft sensor in an IoT environment. From the test results, it can be concluded that the model is best suitable for the quantification of COD and BOD values in water bodies specifically storage tanks, rivers, and lakes.



Water Quality Index Meter

14:37:57
2021-05-08

RTD=31.67 °C
PH= 4.00 pH
DO=11. mg/L
EC=272. uS
Turbudity=2

Chemical Pollutants
water is Basic
L3

Micro-Organisms
Water is safe

Water Quality
Bad

Fig.6. GUI for the Random Forest Model implemented on Raspberry Pi

V. CONCLUSION

The capacity to discover and anticipate BOD and COD values is critical for ecological conservation. To predict data, various intelligent algorithms are implemented. All methods were tested on the scrapped data over different seasons and it can be concluded that the Random Forest model for COD and BOD prediction gives better accuracy and can be used as a soft sensor for predicting the load. The prediction accuracy is enhanced with the addition of categorical data. The suggested methodology provides an adequate level of accuracy while requiring the fewest possible parameters and performing computations more quickly. The trained model is implemented on Raspberry Pi, a resource-light microcontroller to validate its use at the edge in the IoT environment indicating the possibility for near real-time monitoring. Single BOD and COD prediction soft sensor is thus validated with the data set of rivers and drains with minimal input parameters and computational cost opening alternate possibilities while integrating the proposed approach to macro-systems used in large water bodies. Future efforts can be directed toward creating a low-cost hardware model for large water bodies that incorporates the presented soft sensor for real-time water quality monitoring.

VI. REFERENCES

[1] Bhawani Shankar Patnaik, Arunima Sambhuta, Pattanayak, Siba Kumar Udgata_Aiit Kumar Panda “Machine learning-based soft sensor model for BOD Estimation using Intelligence at edge”, Complex & Intelligent Systems (2021) 7:961–976, 2021.

[2] 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.

[3] 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.

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

[5] Wenjing Li 1,2, and Junkai Zhang." Prediction of BOD Concentration Wastewater Treatment Process Using a Modular Neural Network in Combination w the Weather Condition", *Appl. Sci.* 2020, *10*(21), 7477.

[6] A. S. Pattanayak, B. S. Pattnaik, S. K. Udgata, and A. K. Panda, "Development of Chemical Oxygen on Demand (COD) Soft Sensor Using Edge Intelligence," IEEE Sensors Journal, vol. 20, no. 24, pp. 14892- 14902, 2020.

[7] Zifei Wang, Yiman," Artificial intelligence algorithm application in wastewater treatment plants: A case study for COD load prediction, Chapter 7:2021, Pages 143-164.

[8] Sani Isa Abbaa, GozenElkiran," Effluent prediction of chemical oxygen demand from the wastewater treatment plant using artificial neural network application", Procedia Computer Science, Volume 120, 2017, Pages 156-16, 1073083.

[9] Davut Hanbay, Ibrahim Turkoglu, Yakup Demir," Prediction of Chemical Oxygen Demand (COD) Based on Wavelet Decomposition and Neural Networks", Clean - Soil Air Water, Volume 35, Issue 3, June 2007, Pages 250-254.

[10] Hemant Bhutada, Aleefia Khurshid, Manish Yadav, Nishant Yadav, Pankaj Baheti, "COD prediction in water using Edge Artificial Intelligence", 10th IEEE International Conference on Emerging Trends in Engineering & Technology Signal and Information Processing (ICETET SIP-22)

[11] Shaikh Samir, Shahapurkar Rekha, "Predicting COD and BOD Parameters of Greywater Using Multivariate Linear Regression", Recent Trends in Intensive Computing, IOS Press, 2021

[12] Zare Abyaneh, H. Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters. *J Environ Health Sci Engineer* 12, 40 (2014).

[13] Hisham A. Maddah, "Predicting Optimum Dilution Factors for BOD Sampling and Desired Dissolved Oxygen for Controlling Organic Contamination in Various Wastewaters", International Journal of Chemical Engineering, vol. 2022, Article ID 8637064, 14 pages,2022. https://doi.org/10.1155/2022/8637064

[14] Abhilash Nair, Aleksander Hykkerud and Harsha Ratnaweera, "Estimating Phosphorus and COD Concentrations Using a Hybrid Soft Sensor: A Case Study in a Norwegian Municipal Wastewater Treatment Plant", Water, vol. 14, pp. 332, 2022.

[15] "CPCB (2020) A report on the impact of lockdown on water quality of river Ganga", Central pollution control board Delhi. Ministry of Environment Forest and Climate Change Govt. of India.