International Journal of Innovative Computing, Information and Control
Volume **19**, Number **5**, October **2023**

ICIC International ©2023 ISSN 1349-4198
pp. **1407–1424**

BIOLOGICAL OXYGEN DEMAND (BOD) AND CHEMICAL OXYGEN DEMAND (COD) MEASUREMENT OF WASTEWATER USING MACHINE LEARNING REGRESSION TECHNIQUES IMPLEMENTED ON THE EMBEDDED SYSTEM

Aryuanto Soetedjo[1], Evy Hendriarianti[2]
and Renaldi Primaswara Prasetya[3]

[1]Department of Electrical Engineering
[3]Department of Informatics Engineering
National Institute of Technology (ITN) Malang
Jalan Raya Karanglo KM 2, Malang 65143, Indonesia
{ aryuanto; renaldipp }@lecturer.itn.ac.id

[2]Department of Environmental Engineering
National Institute of Technology (ITN) Malang
Jalan Bendungan Sigura-gura No. 2, Malang 65145, Indonesia
evyhendriarianti@lecturer.itn.ac.id

Received January 2023; revised May 2023

Abstract. *Biological Oxygen Demand (BOD) and Chemical Oxygen Demand (COD) are the common parameters of the wastewater that should be monitored regularly to maintain the quality standard. This paper presents a real-time BOD and COD measurement approach based on machine learning. The system employs low-cost multi-sensor consisting of water and gas sensors and an embedded device for the machine learning implementation. Several machine learning regression techniques, namely AdaBoost, Random Forest, XGBoost, Support Vector Regression, and K-Nearest Neighbors (KNN) algorithms, are evaluated under the different combinations of features generated by the multi-sensor. The evaluation results using the samples from six different wastewater resources show that the KNN algorithm is superior to others, achieving a perfect BOD and COD prediction (Mean Absolute Percentage Error (MAPE) is zero and R-squared is one). The combining features from the sensory systems are suitable for the machine learning regression to predict the BOD and COD of the wastewater. Furthermore, the algorithm's execution time implemented on the embedded device is less than a hundredth millisecond, allowing the real-time measuring system.*
Keywords: BOD, COD, Wastewater, Machine learning, Embedded system

1. **Introduction.** Measuring the wastewater quality is essential to provide helpful information about the water quality standard, where the quality should be within the allowable range. Since wastewater quality measuring systems using the conventional methods are usually time-consuming and complex, measurements based on soft sensor and machine learning techniques are developed.

The soft sensor in [1] used Artificial Neural Network (ANN) to predict Biological Oxygen Demand (BOD), Chemical Oxygen Demand (COD), and Total Suspended Solids (TSS) of the effluent in the Wastewater Treatment Plant (WWTP). The inputs of the ANN were temperature, pH, conductivity, influent TSS, influent COD, and influent BOD. In [2], the ANN was used to predict the BOD, COD, and TSS of effluent water based on the BOD, COD, and TSS of influent water on the WWTP. In [3], a hybrid Convolutional

DOI: 10.24507/ijicic.19.05.1407

Neural Network (CNN) – Long Short-Term Memory (LSTM) was used to predict the effluent COD of the wastewater based on the temperature, pH, NH_3-N, sewage inflow, and influent COD.

The ANN in [4] was used to predict the next-day COD in the WWTP, where the inputs were pH, flow, COD, BOD, TSS, turbidity, Dissolved Oxygen (DO), and Oxidation Reduction Potential (ORP). In [5], the ANN was used to predict the one-day interval of total nitrogen of the effluent wastewater on the WWTP using the monthly data of influent flow rate, pH, temperature, suspended solids, total nitrogen of influent, and pretreated supernatant. In [6], the Recurrent Neural Network (RNN) was used to predict the influent flow, influent temperature, influent BOD, effluent chloride, effluent BOD, and power consumption of the WWTP. In [7], the ANN was used to estimate the COD in industrial sewage, where the inputs of the ANN were pH, temperature, DO, conductivity, and turbidity. The ANN in [8] was used to predict the BOD of river water using the inputs of temperature, hardness, conductivity, and DO. In [9], the ANNs were developed to predict river water quality in two scenarios, where the first scenario predicts the water quality in each station, and the second scenario predicts the water quality on the downstream stations based on the data on the upstream stations.

In [10], the Support Vector Regression (SVR) was used to predict the BOD and COD of the wastewater using the National Stormwater Quality Database (NSQD). The database consisted of the data of drainage area, precipitation depth, runoff, percentages of the residential area, institutional area, commercial area, industrial area, open space area, freeway, and impervious area. The Least Square – Support Vector Machine (LS-SVM) was used in [11] to predict the BOD and COD of the river water. The inputs of the algorithm were pH, turbidity, conductivity, sodium, calcium, magnesium, orthophosphate, nitrite, and nitrate nitrogen. The LS-SVM in [12] was used to predict the COD in the WWTP, where the inputs were the flow, concentration of suspended solids, and concentration of ammoniacal nitrogen in the influent, and Oxidation-Reduction Potential (ORP) and DO in the WWTP's reactors.

In [13], multiple linear regression was used to predict the BOD and COD of the treated wastewater in the fruit and vegetable processing industry. The algorithm's inputs were the level of process and treatment, the BOD, COD, TDS, and Total Nitrogen (TN) of the raw wastewater. The Random Forest Regression in [14] used the inputs of BODs and CODs from the previous days to predict the BOD and COD in the WWTP. In [15], Extreme Gradient Boosting was used to predict the BOD in the influent and effluent on the WWTP using simple sensors, namely pH, temperature, flow rate, and minimal complex sensors: COD and nutrients.

Based on previous studies, machine learning applications for measuring water quality can be divided into three categories. First, it is used to predict the water quality of the effluent wastewater using the data collected on the influent wastewater. The machine learning applications to predict the water quality of the downstream stations on the river water based on the data on the upstream stations fall into this category. Second, it is used to predict the next day's water quality based on the data from previous days or months. The data are collected from the same place/location of the predicted parameters, but data are collected at the previous time to predict the parameters in the future. Third, it is used to predict the specific parameters of the water quality, such as BOD and COD, based on the other parameters. It is also called indirect measurement. This method is usually employed to overcome the drawbacks of using the existing sensors, namely the high cost or other complex procedures.

Most works described previously perform the measurement or prediction of water quality on the personal computer, where the datasets are not collected in real time. Only a few implemented the algorithm on the embedded systems, as developed in [7,16-20].

The embedded platform powered by the Exynos4412 CPU chip was employed in [7] to measure the COD of sewage water using the ANN. It used five sensors: pH, temperature, DO, conductivity, and turbidity. The sensor devices were connected to the Exynos CPU through RS-485 communication using the Modbus-RTU protocol. The ANN algorithm was implemented in C language and Java Native Interface (JNI) in the Android operating system.

In [16], the machine learning algorithms to estimate the BOD were implemented on a Raspberry Pi 3, which is interfaced with the sensor devices for measuring the DO, temperature, pH, Conductivity, ORP, and turbidity. The algorithms were written using Python and run in the Raspbian operating system.

The low-cost optical sensor was developed in [17] to measure water quality. The sensor consisted of a multi-wavelength light source: green, red, amber, blue, and infrared. The optical sensor was connected to the Raspberry Pi unit for the data recording. The low-cost sensor was tested and showed the linear correlation between the sensor response and the turbidity concentration.

In [18,19], an electronic Nose (e-Nose) was developed to measure water quality (BOD) based on gas sensors. The e-Nose consisted of metal oxide sensors that detect hydrogen gas, organic solvent vapors, hydrogen sulfide, ammonia, air contaminants, liquefied petroleum gas, and organic vapors. The experiments showed that the relationship between the e-Nose sensing response and the BOD tended to be linear. According to [20], the e-Nose can be used for environmental parameter measuring when designed specifically and fine-tuned.

From the above discussions, a real-time BOD and COD monitoring system is an interesting research area. It offers several challenging issues, such as low-cost hardware implementation, an effective algorithm, and sensor configuration. In this work, we develop a system for the real-time BOD and COD measurement of the wastewater from the multi-sensor. Like [16], we employ the machine learning implemented on Raspberry Pi embedded device. However, we combine water sensors used in [16] and gas sensors [18,19] for the sensory system. The motivation for combining the sensors is to improve the accuracy as proposed by [21], where the fusion of camera and radar sensors is employed to improve the detection accuracy in the collision warning system. Furthermore, unlike the existing approaches, our system exploits the sensor's raw data and the combination of features for better measurement. The main contributions of our work are as follows: a) It implements the machine learning techniques, namely AdaBoost, Random Forest, Extreme Gradient Boosting, SVR, and K-Nearest Neighbors (KNN) on the embedded platform for real-time BOD and COD measurement; b) It exploits the combination of various sensor devices consisting of water and gas sensors for accurate prediction.

2. **Proposed System.**

2.1. **Hardware architecture.** The proposed system hardware is depicted in Figure 1. The developed multi-sensor consists of the water sensor module, gas sensor-1 module, gas sensor-2 module, IoT cloud, and embedded machine learning module. Each sensor module is equipped with the sensors interfaced to an Arduino Nano 33 IoT microcontroller, which reads the sensors and sends the data wirelessly to the IoT cloud or embedded machine learning module. It is noted that data are sent from the microcontroller to the IoT cloud during data collection. It is indicated with the dashed arrow in the figure. Meanwhile, the data are sent to the embedded machine learning module during the online measurement,

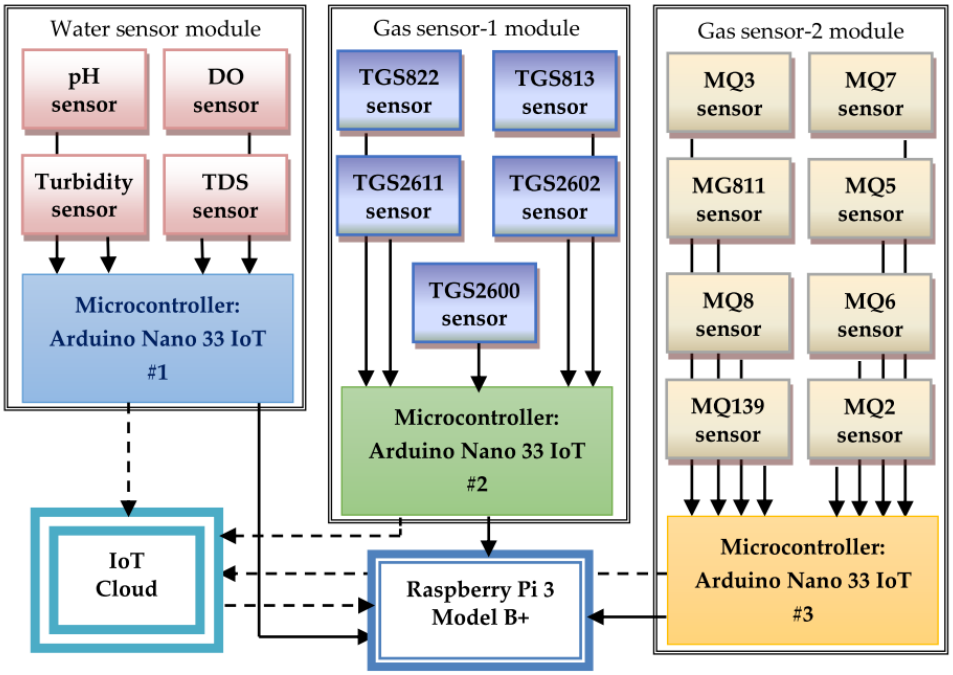


FIGURE 1. Hardware architecture

which is indicated with the solid arrow. Once the data are available in the cloud, they can be downloaded anywhere for other processes. Embedded machine learning can retrieve the data for the training process.

The water sensor module consists of pH, turbidity, TDS, and DO sensors. The gas sensor-1 module consists of the five TGS sensors (Figaro gas sensors): TGS2600, TGS-2602, TGS2611, TGS813, and TGS822. The gas sensor-2 module consists of eight MQ sensors (Hanwei Electronics Co., Ltd.): MQ139, MQ2, MQ3, MQ5, MQ6, MQ7, MQ8, and MG811. Each sensor comprises a sensor probe and an electronic signal conditioning unit. The specifications of the water sensor, gas sensor-1, and gas sensor-2 modules are given in Tables 1, 2 and 3, respectively.

TABLE 1. Specification of water sensor module

Sensor name	Measured parameter		Output voltage range
	Parameter	Measurement range	
DFRobot: Analog pH sensor	pH	0-14	0-3.0 V
DFRobot: Analog Turbidity sensor	Turbidity	0-3000 NTU	0-4.5 V
DFRobot: Analog TDS sensor	TDS	0-1000 ppm	0-2.3 V
DFRobot: Analog DO sensor	DO	0-20 mg/L	0-3.0 V

The sensor output is the voltage representing the measured parameter. In a typical application, a microcontroller is programmed to convert the voltage value to the respective parameter using a particular formula. However, most sensors have a non-linear relationship between the voltages and measured parameters. Therefore, to diminish the non-linearity, instead of converting to the parameter value, we directly use the raw data from the voltage output as the input to the machine learning algorithm.

The embedded machine learning is implemented on a Raspberry Pi 3 Model B+, powered by Broadcom BCM2837B0, Cortex-A53 64-bit SoC @ 1.4GHz, 1GB LPDDR2 SDRAM. The ThingSpeak platform [22] is employed as the IoT cloud server.

Table 2. Specification of gas sensor-1 module

Sensor name	Measured parameter		Output voltage range
	Parameter	Measurement range	
TGS2600	Hydrogen, carbon monoxide	0-10 ppm	0-5.0 V
TGS2602	Ammonia, hydrogen sulfide	0-10 ppm	0-5.0 V
TGS2611	Methane	500-10000 ppm	0-5.0 V
TGS813	Liquefied gas, natural gas, city gas and smog	500-10000 ppm	0-5.0 V
TGS822	Organic solvent vapors	50-5000 ppm	0-5.0 V

Table 3. Specification of gas sensor-2 module

Sensor name	Measured parameter		Output voltage range
	Parameter	Measurement range	
MQ139	Freon	10-1000 ppm	0-5.0 V
MQ2	Methane, Butane, LPG	5000-20000 ppm (Methane); 300-5000 ppm (Butane); 200-5000 ppm (LPG)	0-5.0 V
MQ3	Alcohol	0.05-10 mg/L	0-5.0 V
MQ5	LPG, Natural gas	200-10000 ppm	0-5.0 V
MQ6	Iso-butane, Propane	200-10000 ppm	0-5.0 V
MQ7	Carbon monoxide	20-2000 ppm	0-5.0 V
MQ8	Hydrogen	100-10000 ppm	0-5.0 V
MG811	Carbon dioxide	350-10000 ppm	0-5.0 V

The multi-sensor hardware is depicted in Figure 2, where Figures 2(a), 2(b), and 2(c) show the gas sensors, water sensors (electronic parts), and sensor measurement arrangement, respectively. The data collection process is as follows:

1) The wastewater samples are taken from the sites to the campus laboratory;

2) The wastewater samples are measured using the multi-sensor, as illustrated in Figure 2;

3) The gas and water sensors read the samples and send the data to the ThingSpeak IoT cloud every 5 seconds;

4) Each sample is measured by the multi-sensor for 5 minutes;

5) After measuring using the multi-sensor, the wastewater samples are sent to the Water Quality Laboratory for the BOD and COD measurement using the standard instrument and procedure;

6) Both multi-sensor data from the ThingSpeak and BOD and COD data from the Water Quality Laboratory are combined in the spreadsheet to compose the dataset for the machine learning training and testing.

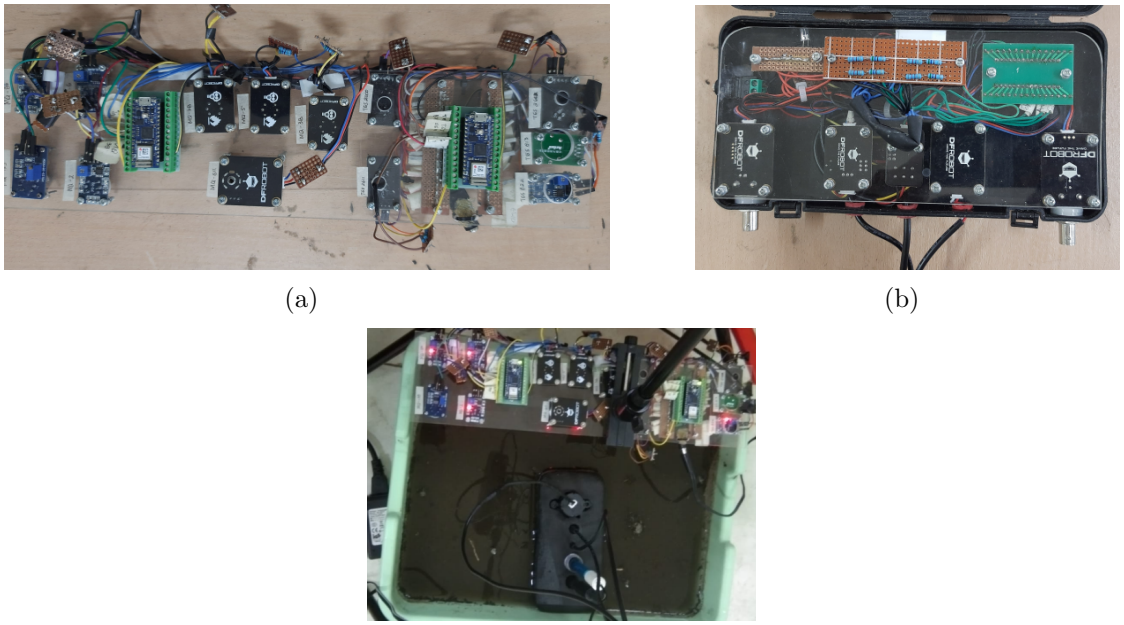
2.2. **Dataset preparation.** The datasets required to train and test the machine learning are prepared as described in the following. At first, the wastewater samples are collected from six wastewater resources:

1) River water;

2) Car wash wastewater;

3) Laundry wastewater;

4) Wastewater from the Sedimentation unit of the WWTP;



(a) (b)

(c)

FIGURE 2. Multi-sensor hardware: (a) Gas sensors; (b) water sensors (electronic parts); (c) sensor measurement arrangement

5) Wastewater from the Anaerobic Baffled Reactor (ABR) unit of the WWTP;
6) Wastewater from the Anaerobic Filter (AF) unit of the WWTP.

The datasets are composed of the data from multi-sensor (the ADC values) and the BOD and COD data from the Water Quality Laboratory, which are collected as discussed previously. In this work, to evaluate the effectiveness of the features used by machine learning, we consider the different combinations of the sensors for creating the datasets. Therefore, the datasets consist of the water sensors, gas sensor-1 (TGS gas sensors), gas sensor-2 (MQ gas sensors), a combination of gas sensor-1 and gas sensor-2, a combination of water sensors and gas sensor-1, a combination of water sensors and gas sensor-2, and a combination of water sensors, gas sensor-1, and gas sensor-2. The combinations of datasets are applied for both BOD and COD measurement. Thus, there are fourteen datasets (seven for BOD and seven for COD), as described in Table 4.

2.3. **Machine learning regression techniques.** As described previously, our machine learning technique is implemented on an embedded module, the Raspberry Pi 3 Model B+, for easy installation in the natural environment. Due to the limitation of the processor speed time and the memory of the embedded module, we only consider some machine learning techniques that fulfill this limitation. Therefore, in this work, we examine five machine learning techniques with low computation and memory requirements, i.e., AdaBoost, Random Forest, Extreme Gradient Boosting, SVR, and K-Nearest Neighbors (KNN).

AdaBoost (Adaptive Boosting) [23,24] is categorized as the ensemble learning algorithm, a machine learning technique that combines several models to find the optimal prediction. The AdaBoost is the boosting algorithm that combines multiple weak learners to build a strong learner. The AdaBoost is originally used to solve binary classification problems. Several versions of the AdaBoost algorithms are called AdaBoost.M1 and AdaBoost.M2 for multi-class classification, and AdaBoost.R for solving regression problems

Table 4. Description of the datasets

No.	Dataset name	Description
1	BOD_water	BOD data and water sensor data
2	BOD_gas-1	BOD data and gas sensor-1 data
3	BOD_gas-2	BOD data and gas sensor-2 data
4	BOD_gas	BOD data, gas sensor-1 data and gas sensor-2 data
5	BOD_water_gas-1	BOD data, water sensor data and gas sensor-1 data
6	BOD_water_gas-2	BOD data, water sensor data and gas sensor-2 data
7	BOD_water_gas	BOD data, water sensor data, gas sensor-1 data and gas sensor-2 data
8	COD_water	COD data and water sensor data
9	COD_gas-1	COD data and gas sensor-1 data
10	COD_gas-2	COD data and gas sensor-2 data
11	COD_gas	COD data, gas sensor-1 data and gas sensor-2 data
12	COD_water_gas-1	COD data, water sensor data and gas sensor-1 data
13	COD_water_gas-2	COD data, water sensor data and gas sensor-2 data
14	COD_water_gas	COD data, water sensor data, gas sensor-1 data and gas sensor-2 data

[24]. In this paper, we employ the AdaBoost for regression, namely AdaBoost.R2 proposed by [25].

Extreme Gradient Boosting (XGBoost) [26] is the most popular implementation of the Gradient Boosting Machine (GBM) [27,28]. The algorithm is a scalable tree boosting system designed for high speed and performance. The main features are a) regularized learning objective, which helps to smooth the final weights to avoid over-fitting; b) using an additive manner to train the model; c) in addition to the regularization, the shrinkage and column subsampling is used to avoid over-fitting [26].

Random Forest is a tree-based ensemble machine learning technique developed by [29,30]. The algorithm combines the ensemble learning with the decision tree and averages the results to improve the accuracy and avoid over-fitting.

Support Vector Regression (SVR) [31] is the regression technique that uses the Support Vector Machine (SVM). The algorithm finds the best fit line, i.e., the hyperplane with a maximum number of points. The SVR has a good generalization, robust to outliers, and easy to be implemented.

K-Nearest Neighbors (KNN) is simple supervised learning for classification and regression. The KNN regression algorithm calculates the average of the target values of the K-nearest neighbors. The neighbors are calculated using the distance function, such as the Euclidean distance, Manhattan distance, and Minkowski distance.

3. **Results and Discussion.** The proposed approach is evaluated using the datasets described in the previous section. The numbers of features and data points of the datasets are given in Table 5. Since each feature represents a value of each sensor, the number of features corresponds to the number of sensors used in the dataset. For instance, there are four features in BOD_water or COD_water dataset because the dataset uses four water sensors: pH, turbidity, TDS, and DO. Meanwhile, the number of data points depends on the data successfully sent to the ThingSpeak cloud for 5 minutes during data collection, as described in Section 2.1. It is noted that the BOD and COD are treated separately but use the same wastewater samples. Thus, the BOD and COD datasets have the same

Table 5. Datasets used in the evaluation

No.	Dataset name	Number of features	Number of data points
1	BOD_water, COD_water	4 (pH, turbidity, TDS, DO)	445
2	BOD_gas-1, COD_gas-1	5 (TGS2600, TGS2602, TGS2611, TGS813, TGS822)	503
3	BOD_gas-2, COD_gas-2	8 (MQ139, MQ2, MQ3, MQ5, MQ6, MQ7, MQ8, MG811)	503
4	BOD_gas, COD_gas	13 (TGS2600, TGS2602, TGS2611, TGS813, TGS822, MQ139, MQ2, MQ3, MQ5, MQ6, MQ7, MQ8, MG811)	503
5	BOD_water_gas-1, COD_water_gas-1	9 (pH, turbidity, TDS, DO, TGS2600, TGS2602, TGS2611, TGS813, TGS822)	396
6	BOD_water_gas-2, COD_water_gas-2	12 (pH, turbidity, TDS, DO, MQ139, MQ2, MQ3, MQ5, MQ6, MQ7, MQ8, MG811)	396
7	BOD_water_gas, COD_water_gas	17 (pH, turbidity, TDS, DO, TGS2600, TGS2602, TGS2611, TGS813, TGS822, MQ139, MQ2, MQ3, MQ5, MQ6, MQ7, MQ8, MG811)	396

data feature (ADC values of the multi-sensor) but different target data, where the target data of the BOD and COD datasets are the BOD and COD values, respectively.

Five machine learning algorithms, AdaBoost, Random Forest, XGBoost, Support Vector Regression (SVR), and KNN, are evaluated. The algorithms are written in Python and use the OpenCV [32] and Scikit-learn [33] libraries, which are implemented on a Raspberry Pi 3+ running on the Raspberry Pi OS operating system. Since all five machine learning algorithms have low computation times, both the training and testing phase can be performed on the embedded platform. In the experiments, data in each dataset is split randomly into training and testing data, where 67% are used for training, and the rest, 33%, are for testing.

The performance metrics for the evaluation are the Mean Absolute Percentage Error (MAPE) which measures the prediction accuracy, and R-squared (R^2), which measures the strength of the relationship between the dependent and independent variables. The high performance is indicated by the high value of R-squared and the low value of MAPE.

Since the performance of a machine learning algorithm is affected by hyperparameter, we employ the Random Search Cross Validation method (RandomizedSearchCV) provided by the Scikit-learn [33] library to find the best hyperparameter. The best values are then used in the evaluation.

3.1. **Evaluation results of BOD measurement.** The evaluation results of the BOD measurement for seven datasets are given in Tables 6 to 12. The MAPE, R-squared, training time, and testing time are shown in the table. Each dataset's best performance (highest R-squared and lowest MAPE) is indicated with the background of light gray in each table.

The AdaBoost achieves high performance in the BOD_water, BOD_gas-1, BOD_gas, BOD_water_gas-1, and BOD_water_gas datasets, with the MAPE lower than 0.25 and R-squared greater than 0.90. It achieves low performance in the BOD_gas-2, and BOD_water_gas-2 datasets, with the MAPE greater than 0.40. The tables show that combining the

Table 6. Evaluation of BOD_water dataset

	AdaBoost	Random Forest	XGBoost	SVR	KNN
MAPE	0.12	0.04	0.03	1.17	0.04
R-squared	0.98	0.92	0.98	0.06	0.93
Training time (s)	0.816	0.274	0.453	0.091	0.004
Testing time (s)	0.062	0.029	0.019	0.052	0.008

Table 7. Evaluation of BOD_gas-1 dataset

	AdaBoost	Random Forest	XGBoost	SVR	KNN
MAPE	0.16	0.23	0.19	0.92	0.07
R-squared	0.98	0.91	0.96	0.72	0.98
Training time (s)	0.829	0.292	0.317	0.327	0.005
Testing time (s)	0.070	0.036	0.017	0.026	0.010

Table 8. Evaluation of BOD_gas-2 dataset

	AdaBoost	Random Forest	XGBoost	SVR	KNN
MAPE	0.41	0.03	0.12	0.47	0.06
R-squared	0.82	0.91	0.91	0.19	0.83
Training time (s)	1.754	0.374	0.474	0.143	0.006
Testing time (s)	0.146	0.040	0.020	0.077	0.019

Table 9. Evaluation of BOD_gas dataset

	AdaBoost	Random Forest	XGBoost	SVR	KNN
MAPE	0.07	0.11	0.25	0.37	0.00
R-squared	0.96	0.97	0.94	0.94	1.00
Training time (s)	1.616	0.402	0.523	85.760	0.006
Testing time (s)	0.129	0.058	0.018	0.034	0.014

Table 10. Evaluation of BOD_water_gas-1 dataset

	AdaBoost	Random Forest	XGBoost	SVR	KNN
MAPE	0.05	0.13	0.11	0.49	0.00
R-squared	1.00	0.92	0.90	0.63	1.00
Training time (s)	1.311	0.738	0.586	0.100	0.006
Testing time (s)	0.122	0.084	0.021	0.021	0.011

Table 11. Evaluation of BOD_water_gas-2 dataset

	AdaBoost	Random Forest	XGBoost	SVR	KNN
MAPE	0.53	0.38	0.15	0.45	0.00
R-squared	0.88	0.82	0.92	0.85	1.00
Training time (s)	0.961	0.391	0.712	64.895	0.006
Testing time (s)	0.077	0.036	0.028	0.023	0.011

BOD_water and BOD_gas_1 datasets increases the R-squared and decreases the MAPE. We may observe that combining the good datasets will increase performance. Meanwhile, the performance of combination of the good and bad datasets cannot be defined precisely.

Table 12. Evaluation of BOD_water_gas dataset

	AdaBoost	Random Forest	XGBoost	SVR	KNN
MAPE	0.05	0.09	0.09	0.26	0.00
R-squared	1.00	0.91	0.94	0.93	1.00
Training time (s)	1.515	1.009	0.751	68.517	0.003
Testing time (s)	0.121	0.080	0.019	0.025	0.028

The Random Forest achieves high performance in the BOD_water, BOD_gas-1, BOD_gas-2, BOD_gas, BOD_water_gas-1, and BOD_water_gas datasets, with the MAPE lower than 0.25 and R-squared greater than 0.90. The results show that the individual datasets (BOD_water, BOD_gas-1, BOD_gas-2) are suitable for the Random Forest. Combining them will slightly increase the performance in most cases, except the BOD_water_gas-2 (combination of BOD_water and BOD_gas-2), which has a lower performance.

The XGBoost achieves high performance in all seven datasets, with the MAPE lower than 0.25 and R-squared greater than 0.90. It is interesting to note that combining the individual datasets does not increase the performance. However, it still achieves high performance.

In contrast with the above results, the SVR achieves high performance only in the BOD_water_gas dataset. As shown in Tables 6, 7, and 8, the MAPE is high, about 0.9. Combining two individual datasets decreases the MAPE to about 0.4 (Tables 9, 10, 11). Finally, combining three individual datasets decreases the MAPE to 0.26 (Table 12).

The KNN is superior to the others. The performance is very high in all seven datasets. Moreover, the perfect prediction, i.e., MAPE is 0.0 and R-squared is 1.0, can be achieved in the BOD_gas, BOD_water_gas-1, BOD_water_gas-2, and BOD_water_gas datasets. It is worth noting that combining datasets (two or three combinations) achieves the perfect prediction.

The best BOD algorithm for each dataset is given in Table 13, which summarizes the best MAPE and R-squared values indicated with the light gray color in Tables 6 to 12. Table 13 shows that the KNN is the best algorithm for most of the datasets. The table also shows that the proposed multi-sensor system is suitably used by the machine learning algorithm for BOD prediction, as indicated by the low MAPE and high R-squared for all datasets.

Table 13. Best BOD prediction algorithm for each dataset

Dataset name	MAPE	R-squared	Algorithm
BOD_water	0.03	0.98	XGBoost
BOD_gas-1	0.07	0.98	KNN
BOD_gas-2	0.03	0.91	Random Forest
BOD_gas	0.00	1.00	KNN
BOD_water_gas-1	0.00	1.00	KNN
BOD_water_gas-2	0.00	1.00	KNN
BOD_water_gas	0.00	1.00	KNN

Figures 3 to 6 depict the scatter plots of regression models for all seven BOD datasets. The horizontal axis represents the real BOD value, and the vertical axis represents the predicted BOD value. The blue line indicates the perfect prediction, where the predicted value is the same as the real one. The figures clearly show that the KNN points are almost aligned with the blue lines in most of the figures. It complies with the results shown in

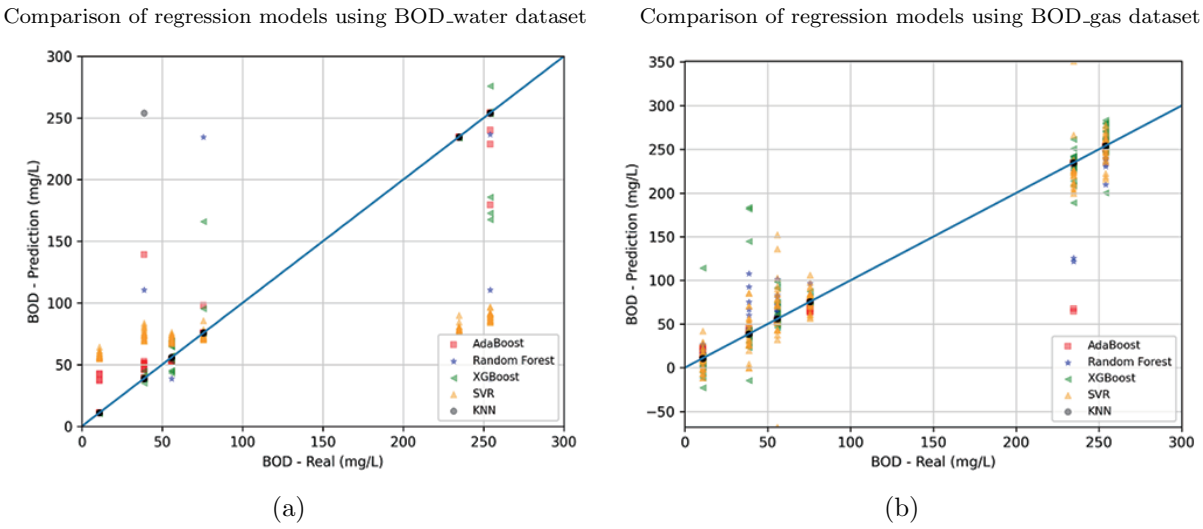


FIGURE 3. Comparison of regression models: (a) BOD_water dataset; (b) BOD_gas dataset

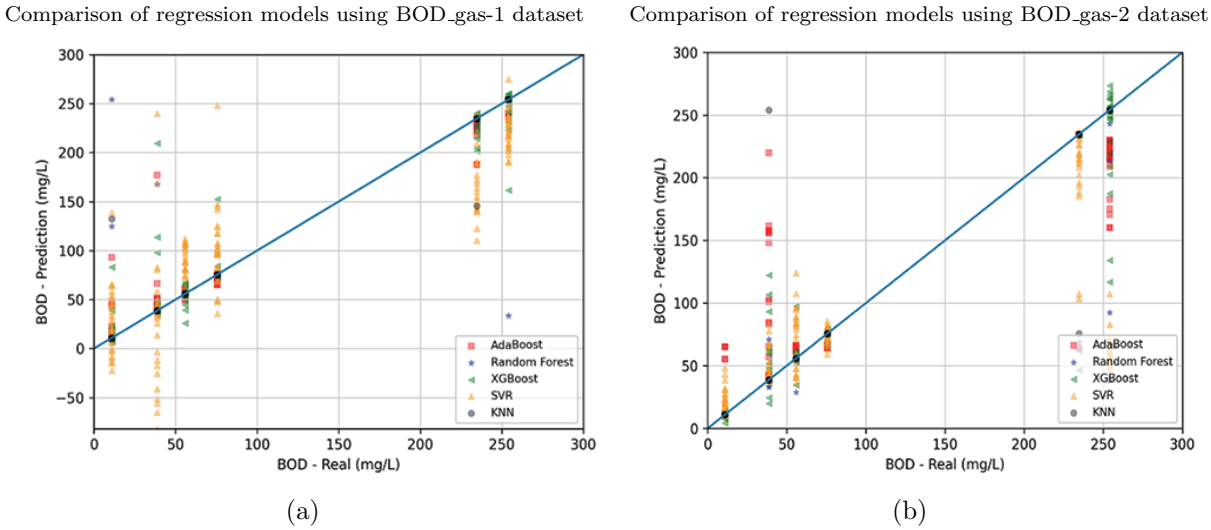


FIGURE 4. Comparison of regression models: (a) BOD_gas-1 dataset; (b) BOD_gas-2 dataset

Table 13. Meanwhile, the SVR points are mainly scattered from the blue lines in the figures, which denotes a lousy predictor. The other important finding is that the points in Figure 6 are less scattered from the blue lines than in the other figures. It suggests that the BOD_water_gas dataset is preferable for all seven machine algorithms in the prediction.

3.2. **Evaluation results of COD measurement.** The evaluation results of the COD measurement for seven datasets are given in Tables 14 to 20. To make a straightforward interpretation, the evaluation of the COD measurement follows the BOD ones.

The AdaBoost achieves high performance in the COD_water, COD_gas-1, COD_gas, and COD_water_gas-1 datasets, with the MAPE lower than 0.25 and R-squared greater than 0.90. It achieves low performance in the COD_gas-2, COD_water_gas-2, and COD_water_gas datasets with the MAPE greater than 0.30. Unlike the BOD datasets, combining

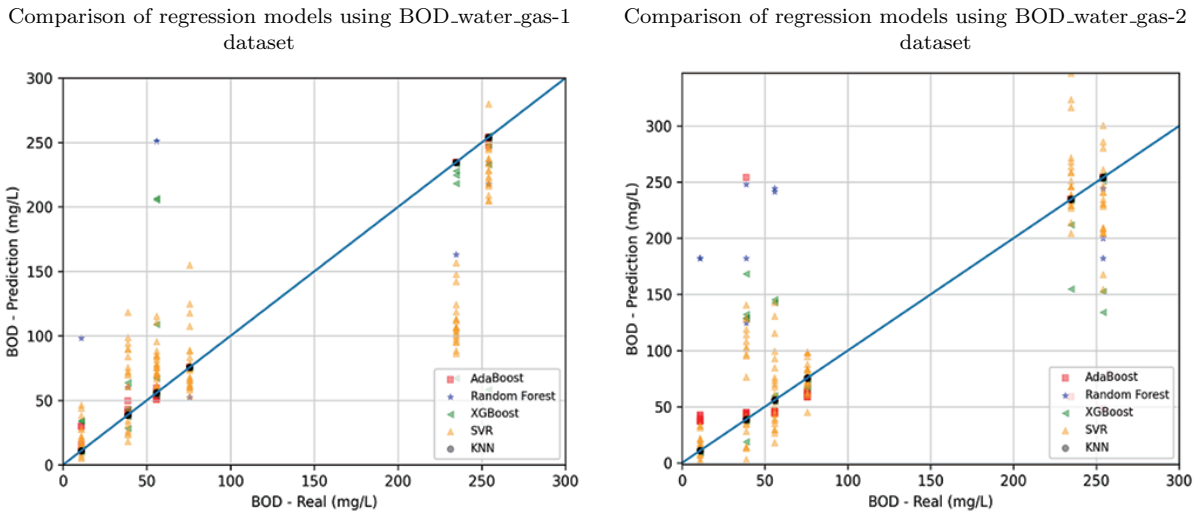


FIGURE 5. Comparison of regression models: (a) BOD_water_gas-1 dataset; (b) BOD_water_gas-2 dataset

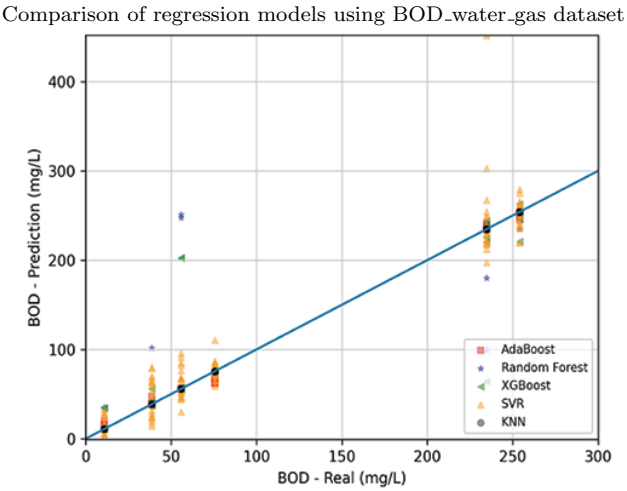


FIGURE 6. Comparison of regression models using BOD_water_gas dataset

TABLE 14. Evaluation of COD_water dataset

	AdaBoost	**Random Forest**	**XGBoost**	**SVR**	**KNN**
MAPE	0.12	0.04	0.03	1.31	0.03
R-squared	0.96	0.93	0.95	0.65	0.96
Training time (s)	1.119	0.282	0.581	0.129	0.005
Testing time (s)	0.090	0.038	0.019	0.022	0.008

COD datasets does not show significant improvement. Significantly, the three combining datasets (COD_water_gas) achieve a high MAPE of 0.41.

Like the BOD datasets, the Random Forest achieves high performance in the COD_water, COD_gas-1, COD_gas-2, COD_gas, COD_water_gas-1, and COD_water_gas datasets, with the MAPE lower than 0.25 and R-squared greater than 0.90. The property of combining the datasets follows the ones in the BOD datasets, i.e., increases the performance.

TABLE 15. Evaluation of COD_gas-1 dataset

	AdaBoost	Random Forest	XGBoost	SVR	KNN
MAPE	0.2	0.16	0.22	0.94	0.06
R-squared	0.94	0.97	0.94	0.59	0.98
Training time (s)	0.952	0.251	0.211	0.209	0.005
Testing time (s)	0.070	0.030	0.017	0.027	0.010

TABLE 16. Evaluation of COD_gas-2 dataset

	AdaBoost	Random Forest	XGBoost	SVR	KNN
MAPE	0.7	0.02	0.08	0.91	0.05
R-squared	0.86	0.92	0.92	0.67	0.89
Training time (s)	1.312	0.387	0.742	3.769	0.005
Testing time (s)	0.102	0.043	0.027	0.031	0.013

TABLE 17. Evaluation of COD_gas dataset

	AdaBoost	Random Forest	XGBoost	SVR	KNN
MAPE	0.05	0.12	0.09	0.37	0.00
R-squared	0.94	0.98	0.92	0.92	1.00
Training time (s)	1.214	0.552	0.822	19.195	0.006
Testing time (s)	0.090	0.070	0.018	0.035	0.014

TABLE 18. Evaluation of COD_water_gas-1 dataset

	AdaBoost	Random Forest	XGBoost	SVR	KNN
MAPE	0.09	0.15	0.1	0.73	0.00
R-squared	1.00	0.98	0.93	0.77	1.00
Training time (s)	1.500	0.972	0.832	0.303	0.005
Testing time (s)	0.125	0.126	0.031	0.019	0.011

TABLE 19. Evaluation of COD_water_gas-2 dataset

	AdaBoost	Random Forest	XGBoost	SVR	KNN
MAPE	0.39	0.27	0.07	0.58	0.00
R-squared	0.93	0.89	0.95	0.87	1.00
Training time (s)	1.177	0.414	0.694	10.779	0.005
Testing time (s)	0.106	0.042	0.019	0.022	0.011

TABLE 20. Evaluation of COD_water_gas dataset

	AdaBoost	Random Forest	XGBoost	SVR	KNN
MAPE	0.41	0.06	0.15	0.31	0.00
R-squared	0.98	0.99	0.97	0.94	1.00
Training time (s)	1.731	0.300	0.490	13.684	0.003
Testing time (s)	0.133	0.039	0.029	0.024	0.029

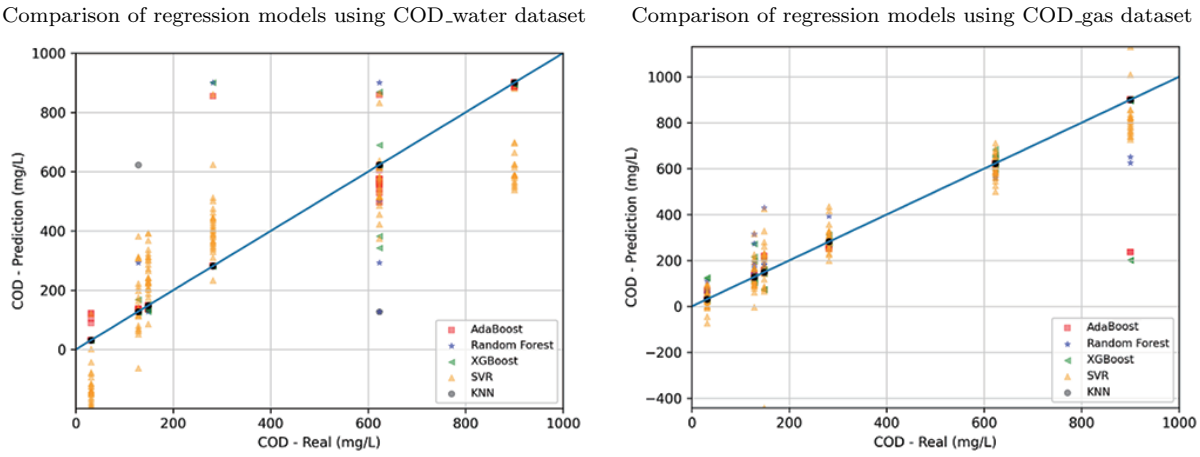
The properties of XGBoost and SVR are also similar to the BOD datasets. Again, the KNN shows superiority, similarly to the BOD datasets.

The best COD algorithm for each dataset is given in Table 21. Similar to the BOD datasets, the table shows that the KNN is the best algorithm for most of the datasets. The machine learning algorithm suitably uses the proposed multi-sensor system for COD prediction.

TABLE 21. Best COD prediction algorithm for each dataset

Dataset name	MAPE	R-squared	Algorithm
COD_water	0.03	0.96	KNN
COD_gas-1	0.06	0.98	KNN
COD_gas-2	0.02	0.92	Random Forest
COD_gas	0.00	1.00	KNN
COD_water_gas-1	0.00	1.00	KNN
COD_water_gas-2	0.00	1.00	KNN
COD_water_gas	0.00	1.00	KNN

Figures 7 to 10 depict the scatter plots of regression models for all seven COD datasets. The figures show the exact property of the BOD datasets discussed previously.



(a) (b)

FIGURE 7. Comparison of regression models: (a) COD_water dataset; (b) COD_gas dataset

4. **Conclusions.** A real-time system for measuring the BOD and COD of the wastewater was proposed. The approach adopted the low complexity machine learning techniques, which can be implemented on the embedded device. A developed multi-sensor system consisting of the water and gas sensors provides the flexibility for feature generation required by the machine learning algorithm. The AdaBoost, Random Forest, XGBoost, and KNN algorithms show high performance in the BOD and COD prediction, with the MAPE less than 0.25 and the R-squared greater than 0.90. Among them, the KNN is the superior one, which achieves the perfect prediction.

The developed system will be extended in the future to cope with the large data samples. Furthermore, the actual implementation on the site will be investigated.

Acknowledgment. This work was supported by the Indonesia Endowment Funds for Education (LPDP), Republic of Indonesia, No.: 192/E4.1/AK.04.RA/2021.

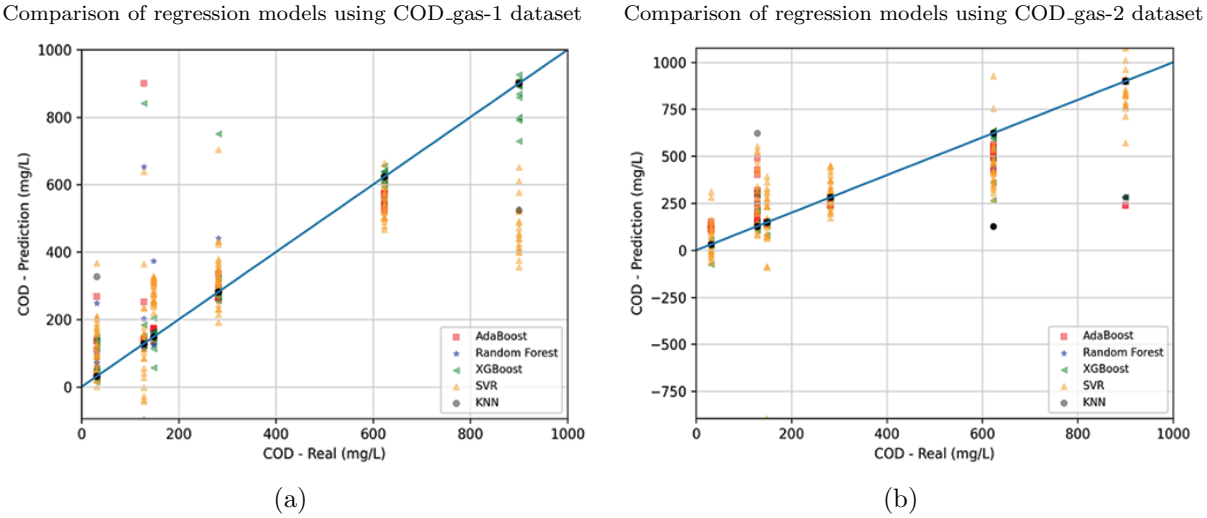


FIGURE 8. Comparison of regression models: (a) COD_gas-1 dataset; (b) COD_gas-2 dataset

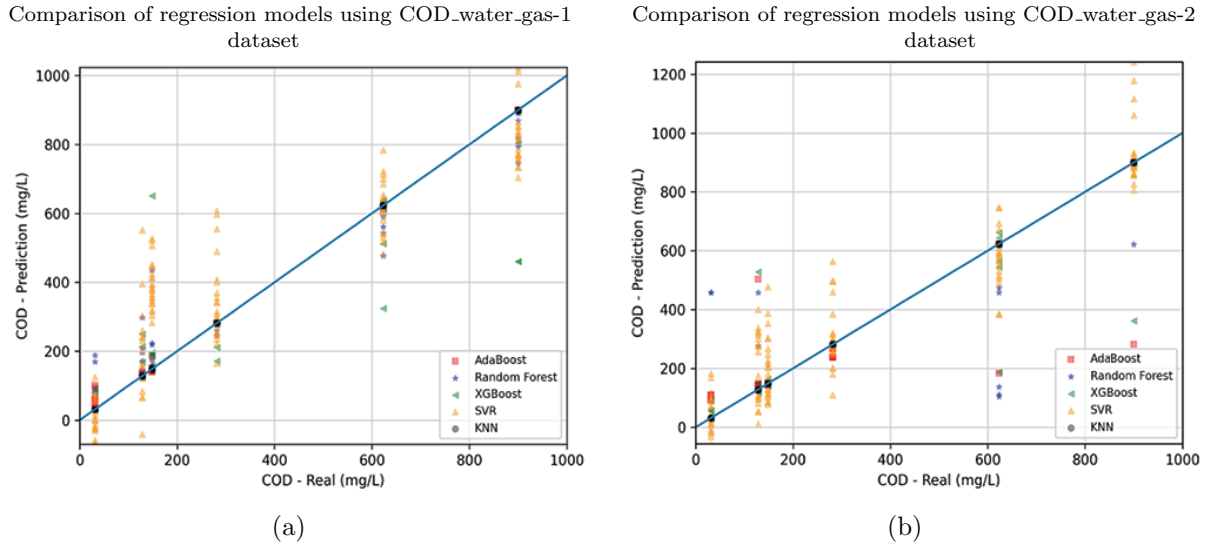


FIGURE 9. Comparison of regression models: (a) COD_water_gas-1 dataset; (b) COD_water_gas-2 dataset

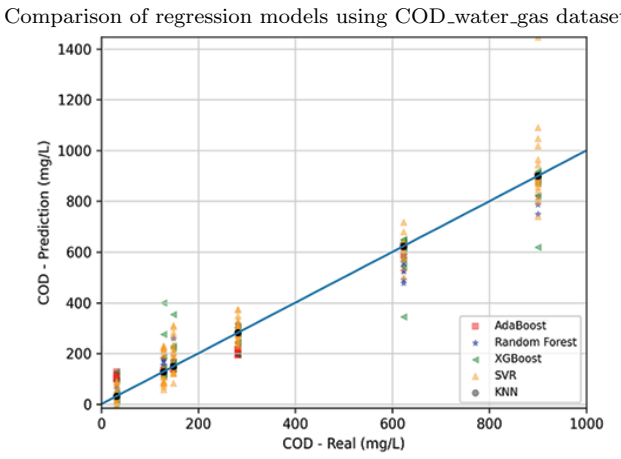


FIGURE 10. Comparison of regression models using COD_water_gas dataset

REFERENCES

[1] A. Alsulaili and A. Refaie, Artificial neural network modeling approach for the prediction of five-day biological oxygen demand and wastewater treatment plant performance, *Water Supply*, vol.21, no.5, pp.1861-1877, 2021.

[2] Q. Liu, A. Ibeas and R. Vilanova, Neural network identification of wastewater treatment plants, *Proc. of the 23rd Mediterranean Conference on Control and Automation (MED2015)*, Torremolinos, Spain, pp.840-846, 2015.

[3] Z. Wang, Y. Man, Y. Hu, J. Li, M. Hong and P. Cui, A deep learning based dynamic COD prediction model for urban sewage, *Environ. Sci. Water Res. Technol.*, vol.5, no.12, pp.2210-2218, 2019.

[4] L. Arismendy, C. Cárdenas, D. Gómez, A. Maturana, R. Mejía and C. G. Quintero M., Intelligent system for the predictive analysis of an industrial wastewater treatment process, *Sustainability*, vol.12, no.16, DOI: 10.3390/su12166348, 2020.

[5] H. Guo, K. Jeong, J. Lim, J. Jo, Y. M. Kim, J. P. Park, J. H. Kim and K. H. Cho, Prediction of effluent concentration in a wastewater treatment plant using machine learning models, *J. Environ. Sci. (China)*, vol.32, pp.90-101, 2015.

[6] T. Cheng, F. Harrou, F. Kadri, Y. Sun and T. Leiknes, Forecasting of wastewater treatment plant key features using deep learning-based models: A case study, *IEEE Access*, vol.8, pp.184475-184485, 2020.

[7] Y. Li, Y. Shi, K. Wang, D. Sun and D. Yang, Design of online monitoring device for COD parameter in industrial sewage based on soft measurement method, *Proc. of the 32nd Youth Acad. Annu. Conf. Chinese Assoc. Autom.*, Hefei, China, pp.959-964, 2017.

[8] R. Roshni and E. C. Kuruvila, BOD modelling using artificial neural network, *Int. J. of Adv. Res. and Innov. Ideas in Ed.*, no.4, pp.32-41, 2017.

[9] A. N. Ahmed, F. B. Othman, H. A. Afan, R. K. Ibrahim, C. M. Fai, M. S. Hossain, M. Ehteram and A. Elshafi, Machine learning methods for better water quality prediction, *J. Hydrol.*, vol.578, pp.1-18, 2019.

[10] F. Granata, S. Papirio, G. Esposito, R. Gargano and G. de Marinis, Machine learning algorithms for the forecasting of wastewater quality indicators, *Water (Switzerland)*, vol.9, no.2, pp.1-12, 2017.

[11] M. Najafzadeh and A. Ghaemi, Prediction of the five-day biochemical oxygen demand and chemical oxygen demand in natural streams using machine learning methods, *Environ. Monit. Assess.*, vol.191, no.6, 2019.

[12] Q. Cong and W. Yu, Integrated soft sensor with wavelet neural network and adaptive weighted fusion for water quality estimation in wastewater treatment process, *Meas. J. Int. Meas. Confed.*, vol.124, pp.436-446, 2018.

[13] G. Mundi, R. G. Zytner, K. Warriner, H. Bonakdari and B. Gharabaghi, Machine learning models for predicting water quality of treated fruit and vegetable wastewater, *Water (Switzerland)*, vol.13, no.18, pp.1-17, 2021.

[14] A. Sharafati, S. B. H. S. Asadollah and M. Hosseinzadeh, The potential of new ensemble machine learning models for effluent quality parameters prediction and related uncertainty, *Process Safety and Environmental Protection*, vol.140, pp.68-78, 2020.

[15] P. M. I. Ching, X. Zou, D. Wu, R. H. Y. So and G. H. Chen, Development of a wide-range soft sensor for predicting wastewater BOD5 using an extreme gradient boosting (XGBoost) machine, *Environ. Res.*, vol.210, 2022.

[16] B. S. Pattnaik, A. S. Pattanayak, S. K. Udgata and A. K. Panda, Machine learning based soft sensor model for BOD estimation using intelligence at edge, *Complex Intell. Syst.*, vol.7, no.2, pp.961-976, 2021.

[17] K. Murphy, B. Heery, T. Sullivan, D. Zhang, L. Paludetti, K. T. Lau, D. Diamond, E. Costa, N. O'Connor and F. A. Regan, Low-cost autonomous optical sensor for water quality monitoring, *Talanta*, vol.132, pp.520-527, 2015.

[18] S. Siyang, R. Palasuek and T. Kerdcharoen, Development of IoT indirect BOD monitoring system based on electronic nose technology, *Proc. of 2018 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW 2018)*, Taichung, Taiwan, pp.2-3, 2018.

[19] R. Palasuek, T. Seesa-Ard, C. Kunarak and T. Kerdcharoen, Electronic nose for water monitoring: The relationship between wastewater quality indicators and odor, *Proc. of ECTI-CON 2015 – 2015 12th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology*, Hua Hin, Thailand, 2015.

[20] L. Capelli, S. Sironi and R. Del Rosso, Electronic noses for environmental monitoring applications, *Sensors (Basel, Switzerland)*, vol.14, no.11, pp.19979-20007, 2014.
[21] Y.-H. Lai, Y.-W. Chen and J.-W. Perng, Sensor fusion of camera and MMW radar based on machine learning for vehicles, *International Journal of Innovative Computing, Information and Control*, vol.18, no.1, pp.271-287, 2022.
[22] *ThingSpeak*, https://thingspeak.com, Accessed on 01-04-2022.
[23] Y. Freund and R. E. Schapire, Experiments with a new boosting algorithm, *Proc. of the 13th Int. Conf. Mach. Learn*, Bari, Italy, pp.148-156, 1996.
[24] Y. Freund and R. E. Schapire, A desicion-theoretic generalization of on-line learning and an application to boosting, in *Computational Learning Theory. EuroCOLT 1995. Lecture Notes in Computer Science*, P. Vitányi (ed.), Berlin, Heidelberg, Springer, DOI: 10.1007/3-540-59119-2_166, 1995.
[25] H. Drucker, Improving regressors using boosting techniques, *Proc. of the 14th Int. Conf. Mach. Learn.*, San Francisco, CA, United States, pp.107-115, 1997.
[26] T. Chen and C. Guestrin, XGBoost: A scalable tree boosting system, *Proc. of ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, San Francisco, CA, USA, pp.785-794, 2016.
[27] A. I. A. Osman, A. N. Ahmed, M. F. Chow, Y. F. Huang and A. El-Shafie, Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia, *JournalAin Shams Engineering Journal*, vol.12, no.2, pp.1545-1556, 2021.
[28] W. Dong, Y. Huang, B. Lehane and G. Ma, XGBoost Algorithm-based prediction of concrete electrical resistivity for structural health monitoring, *Autom. Constr.*, vol.114, 103155, 2020.
[29] L. Breiman, Random forests, *Mach. Learn.*, vol.45, no.1, pp.5-32, 2001.
[30] A. Cutler, D. R. Cutler and J. R. Stevens, Random forests, in *Ensemble Machine Learning*, C. Zhang and Y. Ma (eds.), Boston, MA, Springer, 2012.
[31] H. Drucker, C. J. C, Surges, L. Kaufman, A. Smola and V. Vapnik, Support vector regression machines, *Adv. Neural Inf. Process. Syst.*, vol.1, pp.155-161, 1997.
[32] *OpenCV*, https://opencv.org, Accessed on 15-04-2022.
[33] *scikit-learn: Machine Learning in Python*, https://scikit-learn.org/stable, Accessed on 15-04-2022.

Author Biography

Aryuanto Soetedjo received the B.Eng. and M.Eng. degrees in Electrical Engineering from Bandung Institute of Technology, Indonesia, in 1993 and 2002, respectively, and the Dr. Eng. degree in Information Science and Control Engineering from Nagaoka University of Technology, Japan, in 2006.

He is currently a full-time professor at the Department of Electrical Engineering, National Institute of Technology (ITN) Malang, Indonesia. His main research interests are image processing, artificial intelligence, machine learning, Internet of Things, control system, and robotics.

Evy Hendriarianti received the B.Eng. degree in Environmental Engineering, Master of Technology Management, and Doctor in Environmental Engineering from Sepuluh November Institute of Technology, Indonesia, in 1997, 2002, and 2016, respectively.

She is currently an associate professor at the Department of Environmental Engineering, National Institute of Technology (ITN) Malang, Indonesia. Her main research interests are environmental quality modeling and wastewater treatment technology.

Renaldi Primaswara Prasetya received the B.S. and M.S. degrees in Computer Science from Brawijaya University, Indonesia, in 2013 and 2018, respectively.

He is currently a lecturer at the Department of Informatics Engineering, National Institute of Technology (ITN) Malang, Indonesia. His main research interests are computer vision, speech recognition, and digital image processing.