Environ Monit Assess (2023) 195:1346
https://doi.org/10.1007/s10661-023-11954-8

RESEARCH

Regression-based analytical models for dissolved oxygen in wastewater

Hisham A. Maddah

Received: 21 November 2022 / Accepted: 5 October 2023

Abstract The limited freshwater resources and increasing demand for clean water require minimizing organic contamination in wastewater. High levels of biochemical oxygen demand (BOD) in water reduce available oxygen, harm ecosystem biodiversity, and degrade water quality. Here, regression-based analytical models are suggested to minimize organic contamination by estimating desired dissolved oxygen (DO) and dilution factors (df) correlated to the organic decomposition. Training datasets of defined independent inputs (i) ultimate biochemical oxygen demand (UBOD), (ii) minimum BOD_T (BODM), (iii) average BOD_T (BODA), (iv) COD, (v) O_2 consumption (X), and (vi) time (T) were collected and/or calculated based on literature. Results showed that there should be specified oxygen dosing amounts dependent upon BOD_5 levels, noting that BOD_5 and DO_5 are inversely proportional (proportionality might differ based on the microbial concentration). An increase in df is predominated by BOD_5, with df≈9.2 for storm (STM), df≈12 × 10^3 for industrial (IND), and df≈18.5–28.5 for domestic (DOM) wastewaters. Mixing/matching between the input features used in training regressors including medium trees (MT) and ensembles boosted trees (EBT) showed high accuracy > 94% for predictor combinations: (i) MT-[UBOD-X], MT-[UBOD-X-T-COD], and EBT-[UBOD-X-T-COD] for DO_5 predictions, and (ii) EBT-[BODM-BODA] and EBT-[BODM-BODA-UBOD-X-T-COD] for df predictions, knowing the general term XX-[a-b-c-d-e-f] has XX = regressor and a,b,c,d,e,f = predictors for the training parameters used as inputs. The models are capable of predicting changes in DO_5 against BOD with deviations 5–10%, whereas a suggested correction factor $\pm\left(\frac{UBOD_i}{BODM_i}\right)^{\alpha}$ further reduced this deviation to < 5%, where $i=0, 1, 2\ldots6$ refers to the BODM datapoint and its corresponding UBOD with the constant $\alpha=f(i)$. The optimized collective models (cubic equations derived for df and DO_5 from BODM that is

Highlights
- Development of regression-based models to minimize organic contamination.
- Machine learning training with input-based predictors for dissolved oxygen (DO).
- df≈9.2 for STM, df≈12 × 10^3 for IND, and df≈18.5–28.5 for DOM wastewater.
- Support vector machines (SVMs) trained models showed that $R^2 > 0.95$.
- Derived cubic equations for predicting DO_5 against BOD with deviations 5–10%.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10661-023-11954-8.

H. A. Maddah
Department of Chemical Engineering, Faculty of Engineering—Rabigh Branch, King Abdulaziz University, Jeddah 21589, Saudi Arabia

H. A. Maddah (✉)
Energy and Water Research Center (EWRC), Al-Maddah Group, Jeddah 23613, Saudi Arabia
e-mail: hmaddah@kau.edu.sa

Published online: 19 October 2023

an exponent function in UBOD) would enable effluent quality evaluation to manage organic contamination, bridging the gap between science and industry best practices.

Keywords Organic contamination · Wastewater · Dissolved oxygen · Quality control · Regression

Introduction

The quantity of oxygen needed by microorganisms to decompose organic matter is known as the biochemical oxygen demand (BOD). Using experimentally observed BOD levels from sample tests, which are directly related to microbial contamination, organic pollution in freshwater bodies can be determined (Maddah, 2016a, 2016b, 2016c; Maddah, 2018a; Vörösmarty et al., 2010). Due to the deterioration of the worldwide "sanitation problems," current population estimates have raised concerns about surface water quality due to the presence of organic wastes. The increase in organic waste is because of the growing demand for dairy products and meat consumption (Wen et al., 2017). Untreated wastewater and sanitary problems arising in surface water bodies (e.g., rivers) are mainly due to organic wastes which result in accelerating climate change. This is because untreated discharge into rivers can directly increase the rate of greenhouse gas (GHG) emissions from downstream rivers (Kim et al., 2019). Warming (2020) has stated that "untreated wastewater running directly out into the environment generates a GHG footprint roughly 3 times higher than that of the GHG footprint when the same wastewater is treated in a traditional wastewater facility" (Warming, 2020). Such consequences necessitate a reduction in discharge volumes of wastewaters or solving the discharge issues by improving our diluting capabilities. Chemical wastes and organic discharges will have a significant impact on the global economy requiring unprecedented advancement in treatment and monitoring technology to deal with the projected BOD rates on a worldwide scale (Maddah, 2016a, 2016b, 2016c; Maddah, 2020; Maddah et al., 2017; Maddah & Chogle, 2015; Maddah & Shihon, 2018; Wen et al., 2017).

Previous research in Europe predicted a rise in organic pollution, particularly in the southern nations where the majority of rivers no longer function as diluents due to wastewater discharge in water-scarce areas with no diluting capacity (Voß et al., 2012). However, this is dependent upon the organic's concentration of the receiving water which must be the same or higher than the organic's concentration of the river water where the discharge can be diluted. Reduced river dilution capacity and probable water quality deterioration are expected to have a greater impact on Eastern Europe and the Black Sea, according to the study (Maddah, 2021a; Maddah et al., 2018; Voß et al., 2012). For BOD testing, the standard 5-day interval for measuring the BOD_5 parameter is employed by Europe. This is because of the length of time it takes for river water to flow from its source to its delta in the UK (outlet end meeting a bay or an ocean). In 1936, the committee "American Public Health Association," known as APHA, suggested using the BOD parameter as a reference indication to evaluate the biodegradation of chemicals and hazardous substances (Nagel et al., 1992). BOD_5 to chemical oxygen demand (COD) represents the biodegradable proportion of effluent; BOD_5/COD can also be employed for sizing a wastewater treatment facility (Adedeji & Olayinka, 2013; Langeveld et al., 2012; Łapiński & Wiater, 2018; Maddah, 2016a, 2016b, 2016c; Nagel et al., 1992).

BOD_5 is the quantity of oxygen utilized by bacteria and other microorganisms in a water medium for 5 days at a standard temperature of 20°C for aerobic decomposition. As a result, the BOD_5 is an indirect measure for assessing existing organic or chemical wastes that are biodegradable in water in the presence of oxygen, expressed in mg O_2/L (Abdulla et al., 2016; Alagha et al., 2020; Al-Sulaiman & Khudair, 2018; Lewis, 2006). The amount of 5-day dissolved oxygen (DO_5) in mg/L necessary for the biodegradation of organics is indicated by BOD_5. The progression of the decomposition of organic waste determines the significance of the DO_5 parameter where the oxygen is totally or partially utilized by bacteria to break down the organic matter (Adedeji & Olayinka, 2013; Langeveld et al., 2012; Łapiński & Wiater, 2018; Maddah, 2018b).

The most common and recent BOD measurement method is the dilution method, which is based on the APHA standards. These standards have been certified by the US Environmental Protection Agency (USEPA) including the manometric system which has been widely used in many sewage plants for over 75 years (Hach et al., 1997). The USEPA, on the other

hand, has not certified the latter approach for wastewater analysis, even though it has authorized the manometric method in some circumstances owing to a lack of data consistency and improvement in related laboratory procedures (Attiogbe et al., 2009).

The same elements that impact DO also affect BOD. By 2050, high BOD-containing wastewater will affect the health condition of at least 2.5 billion people (if organics are not treated and BOD levels are not kept under control) (Voß et al., 2012). High BOD concentrations in water (i.e., attributed to DO = 5 mg O_2/L) reduce oxygen availability, pollute aquatic habitats (limiting aquatic life), impact ecosystem biodiversity, deteriorate water quality, and contaminate freshwater (Al-Sulaiman & Khudair, 2018; Voß et al., 2012). Topsoil leaves, woody debris, animal dung, and effluents from pulp and paper mills are all sources of BOD (Abdulla et al., 2016; Alagha et al., 2020; Al-Sulaiman & Khudair, 2018; Lewis, 2006). Anthropogenic causes of high BOD loadings to freshwater and/or watershed systems include home and livestock (animal) waste, industrial discharge, agricultural pollutants, and combined or mixed sewage overflows. BOD concentrations decrease as they travel through the stream network due to continual microbial breakdown. This results in river self-purification and self-revitalization, as well as the diluting of BOD-containing wastewater before it reaches the oceans. However, this mechanism is constrained by the fact there will be no additional BOD further downstream as per the Streeter-Phelps model (Voß et al., 2012).

The higher the BOD concentration, the faster the oxygen in the stream is lost depending on the water temperature (Lewis, 2006). This implies that larger marine animals will have less oxygen available for aerobic use in the existence of high BOD. Consequently, dissolved oxygen in water will be lower causing marine life to suffocate. In other words, the discharge of wastewater with high BOD would harm marine life in two ways: (i) organics are mostly toxic when consumed, and (ii) high BOD depletes dissolved oxygen that is required for the survival of aquatic life. Organic wastes in water can induce a surge in BOD at wastewater treatment facilities, feedlots, food-processing companies, and urban stormwater runoff (Abdulla et al., 2016; Alagha et al., 2020; Al-Sulaiman & Khudair, 2018; Lewis, 2006).

There have been several predictive models that were built for the estimation of BOD_5 and/or dissolved oxygen. Multi-linear regression (MLR) was applied by Qasaimeh and Al-Ghazawi (2020) to 10-year data of Irbid wastewater treatment plant to correlate effluent parameters including TSS and COD to BOD_5 (or DO_5) (Qasaimeh & Al-Ghazawi, 2020). The MLR approach allows correlating the variables of interest to build accurate models using indicators (inputs) or quality parameters data. The prediction of BOD_5 is possible without the need to conduct experiments of the standard 5-day test. Such prediction models would enhance the control and automation of biological treatment. The high model accuracy with a coefficient of determination (R^2) > 0.94 showed that the model $BOD_5 = 0.5TSS + 0.052COD + 10.1$ can be incorporated into wastewater treatment plants (WWTPs) for monitoring BOD_5 changes in the effluents (Qasaimeh & Al-Ghazawi, 2020). Another work applied artificial neural networks (ANNs) in the prediction of the influent BOD_5 with top-performing models achieved $R^2 > 0.75$. The ANNs' models enabled controlling the effluent quality in WWTPs and were viable to be used as soft sensors for on-time controlling. Data-driven simulation showed that BOD models outperformed COD and TSS models in optimizing the plant performance. The monitoring of associated quality parameters included influent COD, temperature, and conductivity as input parameters. The order of significance of the input parameters was found as follows: COD > temperature > conductivity > TSS > pH, indicating the importance of COD in predicting BOD_5 (Alsulaili & Refaie, 2021).

Qambar et al. (2022) proposed machine learning (ML)-based approaches using Askar and Al Dur WWTPs dataset to predict municipal wastewater influent BOD_5. Their models are built using decision trees, random forest, adaptive boosting, gradient boost, and extreme gradient boosting algorithms. The built models were robust and achieved $R^2 = 1$ (overfitting might be a possibility). Therefore, high-accuracy models were used for real-time prediction of BOD_5 to mitigate environmental risks and ensure an effective treatment process (Qambar et al., 2022). Rustum et al. (2007) developed a prediction of the effluent BOD_5 of a primary clarifier based on the Kohonen self-organizing map (KSOM) and multi-layered perceptron artificial neural networks (MLP-ANN) in MATLAB. A six-input model involved COD and TSS which were correlated to the output BOD_5 in the Seafield WWTP in Edinburgh, UK (Rustum et al., 2007). Similarly,

Obaid et al. (2015) suggested regression equations using MLR analysis methods to show the relationship between population and rainfall with maximum and average BOD_5 in the sewer networks of Karbala city center. The results showed that BOD_5 concentration rises by 9–19 mg/L for an increase in rainfall by 1 mm during festival periods and 4–17 mg/L for each increase of 10,000 population (Obaid et al., 2015). Khusravi (2013) studied kinetic models to combine Monod's kinetic equation and plug flow pattern in wastewater stabilization ponds (WAPs). The kinetic models estimated BOD_5 removal coefficients and the decrease rate of organic materials where BOD/COD > 0.5 indicates the biodegradability of organic materials (Khusravi, 2013).

Al-Ghazawi and Alawneh (2021) utilized ANNs to develop predictive models for the quality of treated effluent for irrigation from Wadi Arab WWTP. The feed flow rate, temperature, pH, BOD_5, COD, TSS, and NH_4-N were the input parameters for each ANN model, whereas a sensitivity analysis showed that the accuracy is dependent on the selected inputs and the mix/match. The models were highly sensitive to pH and slightly sensitive to influent TSS (Al-Ghazawi & Alawneh, 2021). Szelag et al. (2017) used a data mining approach to propose models that would be able to predict influent quality indicators including BOD, COD, TSS, total nitrogen (TN), and total phosphorus (TP). Three-year daily data were used from a WWTP located in Rzeszów to train the models according to the collected data of the quality indicators. Such models built *via* ANN can simulate TSS, TN, and TP by applying support vector machines (SVM), random forests (RF), and multivariate adaptive regression splines (MARS). The models were found to be viable for quality monitoring and controlling for reliable operation from the simulated results. The models' results were compared to the measured influent quality to adopt hybrid models that recorded minimum prediction errors (Szelag et al., 2017). Furthermore, Asami et al. (2021) adopted data mining and ANN to build BOD_5 models for the WWTP of Ramin thermal power covering 3 years (2013–2015) daily dataset. The model R^2 of 0.95 was observed for the prediction of either BOD_5 or COD to reduce the cost of monitoring based on the model's reliability and high generalization capability. In the BOD_5 modeling, it was determined that DO, COD, TSS, temperature, and turbidity at the influent in a WWTP were the most important parameters affecting BOD_5 (Asami et al., 2021).

In the present study, the aim was to establish analytical predictive models that would be able to predict DO and/or dilution factors (df) for wastewaters to minimize organic contamination. The goal is to statistically investigate the possibility of constructing highly accurate models that can estimate DO_5 and df in biologically active wastewaters. Levels of BOD or COD are defined by the concentrations of organics correlated to the DO or df which would allow controlling existing organics. Determination of the average laboratory df was carried out from ΔDO based on the O_2 consumption (X) = 60–70%. This would allow wastewater engineers to estimate the corresponding outputs (DO_5 and df) from knowing BOD_5, COD, DO_0, and O_2 consumption (X) information for training in MATLAB (mix/match of selected independent parameters). Accurate machine learning models were built from the independent variables (inputs) including (i) ultimate biochemical oxygen demand (UBOD), (ii) minimum BOD_T (BODM), (iii) average BOD_T (BODA), (iv) COD, (v) O_2 consumption (X), and (vi) time (T) to estimate DO_5 and df. Residual analysis and inter-quartile range (IQR) based on the 1.5IQR range-median decision rule would guide researchers towards models with minimum statistical errors to establish useful correlations.

Methodology and equations

Data collection and curation

The data collection of domestic wastewater was carried out by gathering information from Al-Diwaniyah Wastewater Treatment Plant, Wastewater Treatment Plants of Jordan and North Sewage Treatment Plant in Dhahran, Eastern Province, Saudi Arabia (Abdulla et al., 2016; Alagha et al., 2020; Al-Sulaiman & Khudair, 2018). Such organic contamination can be investigated or linked to the BOD_5 level in collected samples by changing organic loading rates and designed hydraulic, or operating conditions. The BOD_5 investigation should be done after measuring quality parameters such as pH, turbidity, total suspended solids (TSS), and COD for the successful prediction of potential risks of effluent's organic pollution to protect the environment. The curated raw datasets

can be found in Table S1 through Table S5 in the Supplementary (Abdulla et al., 2016; Alagha et al., 2020; Al-Sulaiman & Khudair, 2018; Lewis, 2006; Maddah, 2021b; Maddah et al., 2020).

The Kumasi Abattoir, Coca-Cola, and GGL plants in Ghana (Kumasi Metropolis, the capital of Ashanti) served as a database for the curation of industrial wastewater datasets (Al-Sulaiman & Khudair, 2018; Attiogbe et al., 2009). The DO_5 level of the water was measured in earlier works by the Azide modification of Winkler's method before and after incubation for 5 days at 20°C using sampling techniques. After the selection of a reasonable dilution factor (df), the estimated DO differences would reveal the sample's BOD_5 levels. To ensure consistent analysis, the samples' pH levels were maintained in the range of 6.5–8 for ideal biochemical oxidation (APHA, AWWA, WEF, 2012). The BOD_5 and COD data of industrial wastewater were then collected from the same three facilities. The dilution factors were calculated using the average BOD_5/COD of each wastewater facility. Datasets have been gathered from various stormwater sources located in several locations. The author gathered the stormwater quality from completed stormwater runoff sampling (from previous studies) in the streets, highways, surfaces, and parking lots. The samples were collected from various catchments in three different countries (Bialystok, Poland, Abeokuta, Nigeria, and Luxembourg). The parameter (BOD_5) was studied and correlated to the initial dissolved oxygen (DO_0) and the utilized dilution factors. The curated raw datasets can be found in Table S1 through Table S5 in the Supplementary (Adedeji & Olayinka, 2013; Langeveld et al., 2012; Łapiński & Wiater, 2018).

BOD estimation

Assessing BOD necessitates taking a minimum of two measurements: (i) for current (immediate) or DO_0 and (ii) for the residual quantity of final 5-day dissolved oxygen (DO_5) following incubation of water samples in the lab for 5 days. Such tests would allow estimation of the quantity of oxygen utilized by microbes throughout the incubation time to break down the organic matter contained in the sample (Abdulla et al., 2016; Alagha et al., 2020; Al-Sulaiman & Khudair, 2018; Lewis, 2006).

Earlier BOD sample investigations of diverse wastewater types, including domestic (household), industrial, and storm (surface) wastewaters, served as a database based on the literature data. No combination or separation of wastewater was considered or required because each study dealt with a one-type plant of wastewater. The calculated minimum and average constant reaction rates (k_1) guided the author towards the average values of k_1 = 0.14, 0.731, and 0.16 day^{-1} for domestic (DOM), industrial (IND), and storm (STM) wastewaters, respectively. The minimum reaction rates ($k_{1,\,min}$) and average reaction rates ($k_{1,\,avg}$) would quantify organic decomposition rates to calculate the corresponding minimum and average BOD_T values at their UBOD. The minimum BOD_T (BODM) was calculated from $k_{1,\,min}$ and the average BOD_T (BODA) was calculated from $k_{1,\,avg}$ using Eq. (1), (see Table S3 through Table S5 in the Supplementary). The reaction rates were taken to be different constants (average) for the various wastewater types considering three wastewater treatment plants for each wastewater (i.e., each wastewater type would have similar wastewater characteristics). UBOD and COD values (from the three different plants for each wastewater type) were kept constant over the studied time (60 days) since they were found not to be impacted by pH, organic-to-inorganic content, nutrients, and existing oxygen. However, BODM and BODA factors change over time and were determined based on the literature (Abdulla et al., 2016; Adedeji & Olayinka, 2013; Alagha et al., 2020; Al-Sulaiman & Khudair, 2018; Attiogbe et al., 2009; Langeveld et al., 2012; Łapiński & Wiater, 2018) from $k_{1,\,min}$ and $k_{1,\,avg}$, respectively, and the corresponding UBOD substituted in Eq. (1).

$$BOD_T = UBOD\left(1 - e^{-k_1 T}\right) \tag{1}$$

Dissolved oxygen

As per the generalized DO_0 averages from the literature (Anggraini & Herdiansyah, 2019; Lewis, 2006; Metcalf & Eddy, 2003), it was found that DO_0 at 20°C = 9.1 mg/L for BOD_5 (for domestic and industrial wastewaters) and DO_0 at 26.3°C = 8.1 mg/L for BOD_5 (for storm wastewater). The previously conducted analysis followed a selected DO from O_2 consumption (X) = 60–70% in Eq. (2) which was

determined from the initial dissolved oxygen (DO_0). Notations in Eq. (2) are defined as the following: P is the volumetric fraction of wastewater or the fraction of wastewater sample volume to total combined volume (i.e., waste sample volume divided by BOD incubation-bottle volume). The ΔDO corresponds to organic decomposition associated with oxygen consumption (X) in water that is the initial dissolved oxygen minus the final dissolved oxygen. DO_5 refers to the dissolved oxygen in water after the incubation period (5 days) compared to DO_0 which is the initial dissolved oxygen in water before the incubation period. The 5-day BOD is attributed to the measurement of dissolved oxygen that would be utilized by microorganisms in the biochemical oxidation of organic matter. BOD and DO are interrelated because a specific quantity of oxygen is required to biologically stabilize organic matter giving an approximation to the necessary size of a wastewater facility.

The organic decomposition attributed to the O_2 consumption (X) = 60–70% was selected since it has been successfully applied for the modeling of oxygen saturation levels in wastewater at a sludge treatment facility in Indonesia. It was found that the standard DO_0 has a maximum oxygen concentration of 9.1 mg/L and a minimum oxygen concentration of 7.5 mg/L at 20°C and 30°C, respectively, but the exact DO might differ depending upon the type of wastewater and the water temperature (Anggraini & Herdiansyah, 2019; Lewis, 2006; Metcalf & Eddy, 2003).

$$BOD \approx \frac{\Delta DO}{P} = \frac{DO_0 - DO_5}{P} \iff DO_5 = \mathrm{X}DO_0 \tag{2}$$

Dilution factor

The average BOD_5/COD ratios were then computed, and these values were subsequently utilized to find the typical laboratory dilution factors (df). The estimated df from Eq. (3) was computed from the known BOD_5, COD, DO_0, and the determined DO_5 at X = 65% (average) by using Eq. (2).

$$df = \frac{\frac{BOD_5}{COD}\Big|_{avg} \times COD_{avg}}{DO_0 - DO_5\ (mg/L)} \tag{3}$$

After acquiring BOD_5, COD, DO_0, and O_2 consumption (X) information (inputs for training) as shown in Table S1 through Table S5 (in the Supplementary) and Table 1, the desired outputs [DO_5 and df] were calculated. This enabled the curation of the original datasets which contain 108 data points in which the literature reported raw data of industrial and domestic wastewaters and stormwater. The collection of both BODM and BODA was carried out as per the definition of BOD_T. The BOD_T shows the impact of time on BOD where time $T = 0$, $T = 5$, and $T = 30$ days would give BODM, BOD_5 (or BODA), and UBOD, respectively, from the literature data (Abdulla et al., 2016; Adedeji & Olayinka, 2013; Alagha et al., 2020; Al-Sulaiman & Khudair, 2018;

Table 1 The reported and/or calculated values for BOD-related parameters and other important wastewater characteristics for the various studied wastewater types

Type metric*	Domestic wastewater			Industrial wastewater			Storm wastewater		
	Diwaniyah	Jordan	North	Kumasi	Coca-Cola	GGL	Poland	Luxembourg	Nigeria
BODM (mg/L)	107	4.25	48	10,000	75	10,00	5	5	94.50
BODA (mg/L)	192	147	72	120,000	909.5	20,500	27.5	30	39.40
UBOD (mg/L)	354	290	144	230,000	1744	40,000	50	90	66.35
COD (mg/L)	428	304.28	179	240,000	1851	56,678	279.5	138	192.84
TSS (mg/L)	–	103	887	–	–	–	235	–	158
BODA/COD*	0.45	0.48	0.40	0.50	0.49	0.36	0.1	0.22	0.34
pH*	6.5–8*	7.91	7.44	7.25	6.5–8*	6.5–8*	8.1	6.5–8*	9.1

*Ratio or number only, unitless. pH with (*) was assumed to be in the range 6.5–8 since no given values were reported in the literature and such values were taken per the recommendations from (APHA, AWWA, WEF, 2012) for consistent analysis

Abbreviations: *BODM* is minimum BOD_T, *BODA* is average BOD_T (or BOD_5)

Attiogbe et al., 2009; Langeveld et al., 2012; Łapiński & Wiater, 2018). Both BODM and BODA were correlated to the organic decomposition rate attributed to the O_2 consumption (X) = 60–70% based on the well-known inverse proportionality that exists between BOD and O_2 consumption (X) value. The developed models' accuracy in predicting DO_5 and df was then tested by randomly assigning 67% of the selected datasets to "training" and 33% to "testing".

Machine learning (training)

Various supervised ML regression learners in MATLAB (Mathworks., 2017) were selected in training/testing the labeled datasets. Ensembles and tree regression models as well as support vector machines (SVMs) were utilized to establish the trained models. A cross-validation (CV) of 5-fold was applied to identify the optimal model for training (Baştanlar & Ozuysal, 2014; Kotsiantis, 2007; Mathworks., 2017; Simeone, 2018). Defined independent variables (inputs) included (i) UBOD, (ii) BODM, (iii) BODA, (iv) COD, (v) O_2 consumption (X), and (vi) time (T). The two investigated outputs were DO_5 and df (Maddah et al., 2020). Training analysis would display the association between BOD-related parameters for the analyzed wastewater types based on BODM and df, BODM against both UBOD and BOD_5, and DO_5 against both BODM and BOD_5 at $X = 65\% \pm 5\%$. Such proposed models might allow wastewater engineers to select the optimal df for sampling analysis (after testing the viability of obtained trendlines). This would enable adjusting the operating conditions according to the desired DO corresponding to the feed BOD_5/COD ratio for maximum organic decomposition. The DO level that is investigated here is the dissolved oxygen that exists in the wastewater before being processed in a WWTP. Thus, it should not be the DO value after aeration but the DO before the aeration. With aeration advanced technology, one would be able to quantify how much oxygen is added to expedite the degradation process of organics. It is worth mentioning that COD is a parameter used to determine the ratio (BOD_5/COD) to quantify existing inorganics (chemicals) with respect to available organics. Figure 1 shows the step-by-step study framework starting with datasets collection, followed by utilized regressors, and statistical analysis methods for models creation.

Training steps were carried out using four input-parameters models: (i) [UBOD-X], [T-COD], and [UBOD-X-T-COD] for DO_5 output; and (ii) [BODM-BODA], [UBOD-X], [T-COD], [UBOD-X-T-COD], and [BODM-BODA-UBOD-X-T-COD] for df output (i.e., combining or contrasting chosen independent factors, knowing the general term [a-b-c-d-e-f] has a,b,c,d,e,f = predictors or input features for the training. To develop controlling elements that would primarily lead to modifying DO_5 or df predictions based on wastewater characteristics analysis, it is important to use a variety of input parameters. The logic of this selection was made based on finding the most common factors that would play a key role in impacting both DO and df. According to the literature (Abdulla et al., 2016; Adedeji & Olayinka, 2013; Alagha et al., 2020; Al-Sulaiman & Khudair, 2018; Attiogbe et al., 2009; Langeveld et al., 2012; Łapiński & Wiater, 2018), BOD, O_2 consumption (X), COD, and time are the prominent factors that would contribute to noticeable changes in DO and df. Other factors like pH, salinity, and alkalinity might have an indirect impact and could be further investigated in future studies. Therefore, for choosing the best-identified regressors, training analysis steps were performed using existing supervised regression models in MATLAB. The mixing of inputs was arbitrary to choose the best-trained models capable of predicting experimental results at high accuracy. The model's accuracy was determined from prediction errors that are the deviation of the model results from the experiments. Trained models were checked and compared among each other from their determined statistical errors to keep only the accurate models that satisfied a coefficient of determination (R^2) > 0.90 for further analysis of response patterns and residuals. Equations (4) and (5), respectively, represent the detected statistical errors from the R^2 and residual.

$$\mathrm{R}^2 = \frac{\left[\sum_{i=1}^{n}\left(x_{o,i}-\bar{x}_o\right)\left(x_{p,i}-\bar{x}_p\right)\right]^2}{\sum_{i=1}^{n}\left(x_{o,i}-\bar{x}_o\right)^2 \times \sum_{i=1}^{n}\left(x_{p,i}-\bar{x}_p\right)^2} \quad (4)$$

$$\text{Residual} = x_o - x_p \quad (5)$$

$x_{o,\,i}$ and/or x_o: the observed values from experiments

$x_{p,\,i}$ and/or x_p: the predicted values from the model

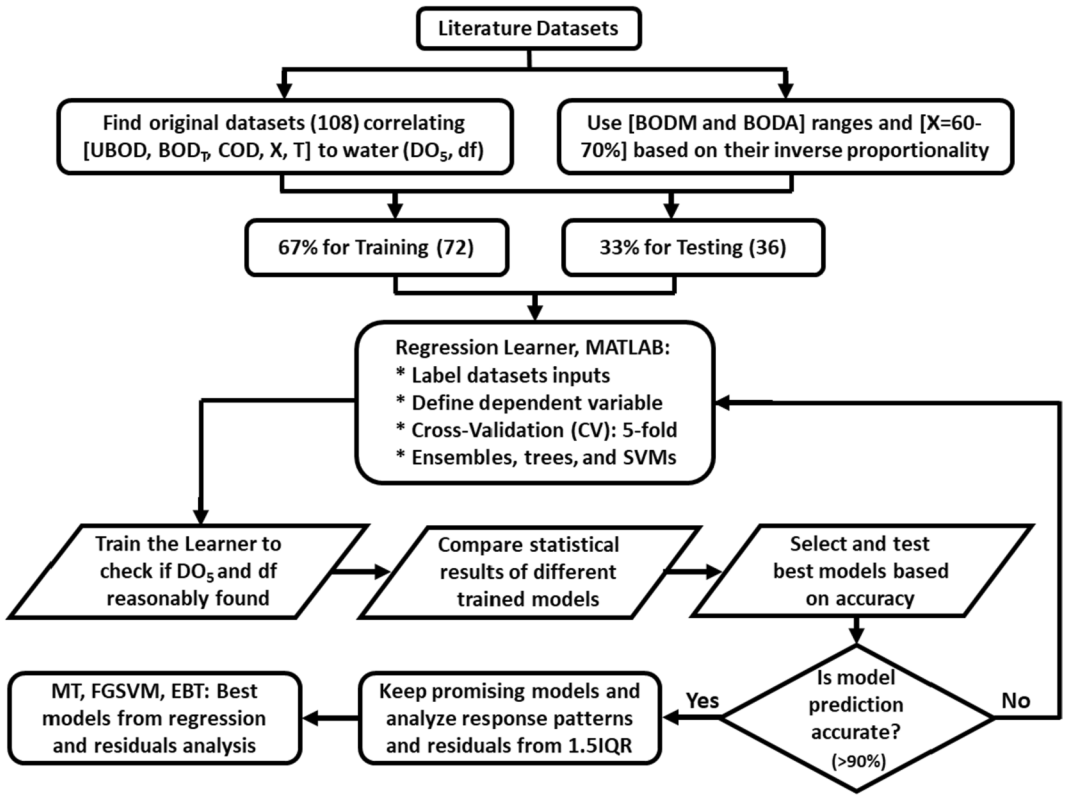


Fig. 1 Study framework starting with datasets collection, defining input and output parameters, and selection of *X* correlated with BDOM and BODA for regression analysis of training and testing datasets to select the best models based on different wastewater types. Acronyms: UBOD, ultimate biochemical oxygen demand; BOD, biochemical oxygen demand; COD, chemical oxygen demand; *X*, O_2 consumption; *T*, time; DO_5, 5-day dissolved oxygen; df, dilution factor; BODM, maximum biochemical oxygen demand; BODA, average biochemical oxygen demand; CV, cross-validation; SVMs, support vector machines; MT, medium trees; FGSVM, fine Gaussian support vector machines; EBT, ensembles boosted trees; IQR, inter-quartile range

$\bar{x}_o$: the observed value from experiments by averaging the returned estimations

$\bar{x}_p$: the theoretically estimated or predicted values from the model by averaging

n: dataset size or the number of experimental observations (dataset range)

Residual analysis (QR method)

To identify the most effective models and explore outliers from the trained models, the inter-quartile range (IQR) decision rule, residual analysis, quartile range (QR), and 1.5IQR range-median approaches were used. Based on the 1.5IQR range-median choice criteria, the analysis identifies models with the least statistical errors (least discrepancies) or least deviation of the model's predictions from those values found from the laboratory analysis (Chaudhary, 2019). The "QR method" is used to visualize a box plot where the median is the center point, the first quartile (Q_1) has 25% of the data between the minimum and Q_1, and the third quartile (Q_3) has 75% of the data between the minimum and Q_3. In light of the acquired ranges for residuals from each model type, the IQR may be calculated from the difference between Q_3 and Q_1 (IQR = Q_3-Q_1) to identify outliers. The lower limit and upper bound are computed with a scale of 1.5 as ($Q_1-1.5\times$IQR) and ($Q_3-1.5\times$IQR), respectively. Any data point that was discovered to be situated beyond the specified ranges is regarded as an outlier (Chaudhary, 2019). This is the same as just taking into account outliers for data that deviates more than 2.7 standard deviations (σ) from the mean (μ) on either side

of a normal distribution "bell curve" (Barbato et al., 2011; Chaudhary, 2019; Moska et al., 2020). Therefore, when the 1.5IQR range-median decision rule is used, the Gaussian distribution for outlier identification is relevant. In other words, the 1.5IQR analysis would enable us to check for data outliers and make judgments about the correctness of the developed trained models.

The five (or six with BODM) selected independent variables (inputs) and their corresponding values had a spike in BOD and/or COD for 12 out of 70 points in the normalized results, Fig. 2 (the figure also shows the defined factors affecting DO_5 and df). This is because of the presence of industrial wastewater in both training and testing datasets [with high BOD_T~230 g/L and high COD >240 g/L]. COD is normally higher than BOD because more organic compounds can be chemically oxidized than biologically oxidized. Knowing that the greater the pollution, the higher the COD and BOD. Industrial wastewater is in the high-to-ultra-high range of pollution of > 3 to 15 g/L of BOD, which is equivalent to approximately > 6 to 30 g/L of COD (NIHON KASETSU CO, 2023). According to Jain and Singh (2003), industrial wastewater may have COD up to 60 g/L, or can be only around 5 g/L, depending on the type of industry (Jain & Singh, 2003). In the author's earlier study (Maddah, 2022), COD for industrial wastewater was found to be very high (56 g/L to 240 g/L) because of the type of industry. Storm, industrial, and domestic wastewaters had an average range of BOD_5/COD ratios of 0.1~0.35, 0.36~0.5, and 0.4~0.48, respectively. Obtaining the average BOD_5/COD range is very useful since it can be used as an indicator tool that would help experimentalists accomplish accurate sampling analysis.

The selection of various input parameters is important to define controlling factors that would chiefly result in changing DO_5 and df based on the conducted attributes analysis. The goal is to statistically investigate the impact of each independent variable to establish useful correlations: (i) BODM against UBOD, and (ii) BODM (or BOD_5) against DO_5 and df for wastewater engineers.

Model formulation

A formulation of various models is suggested based on $UBOD = f(BODM)$ and that would give $BOD_T = f(BODM)\left(1 - e^{-k_1 T}\right)$; each model is for one of the studied wastewater types. Again from Eq. (1), the following correlations between BODM and UBOD were obtained from the fitting analysis and resulting in the following:

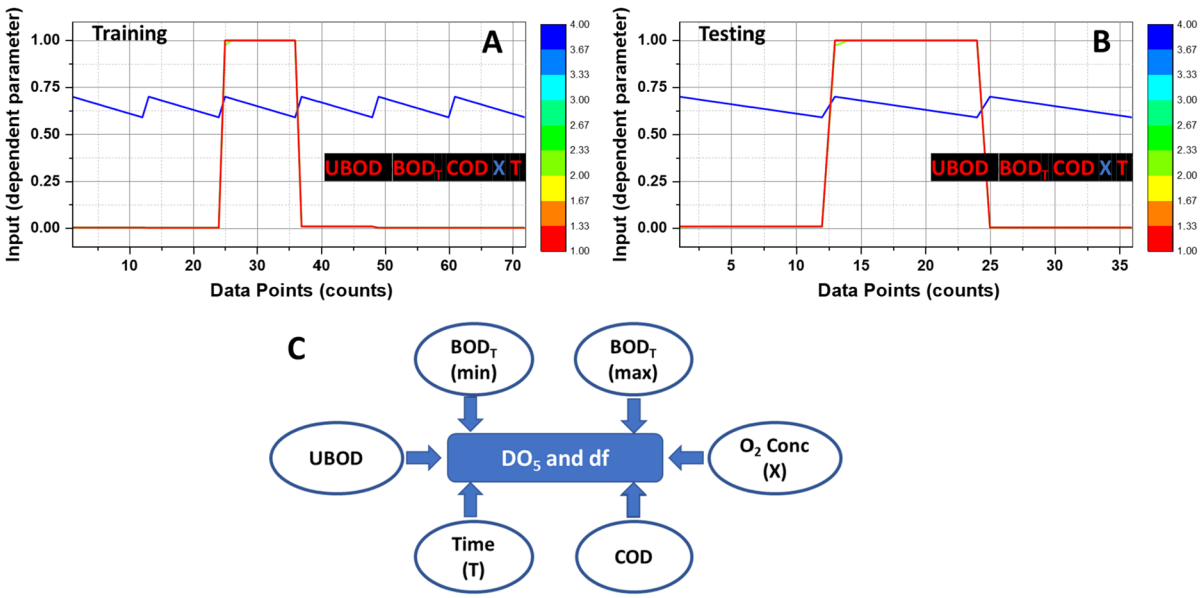


Fig. 2 The identified similar patterns of selected data points used in machine learning analysis. **A** Training and **B** testing; **C** the defined input parameters correlated with the two outputs DO_5 and df based on data from different studied wastewater types

Domestic wastewater,

$$BODM = e^{18-0.13UBOD+2.7\times10^{-4}UBOD^2} \quad (6)$$

$$2.7\times10^{-4}UBOD^2 - 0.13UBOD + 18 - \ln BODM = 0 \quad (7)$$

Constraints,

$$BODM < UBOD$$

$$150\ mg/L < BODM < 350\ mg/L$$

Industrial wastewater,

$$BODM = 3.44\times10^{-8}e^{\frac{-34994}{UBOD-1802.7}} \quad (8)$$

$$\ln\left(\frac{BODM}{3.44\times10^{-8}}\right)UBOD - 1802.7\ln\left(\frac{BODM}{3.44\times10^{-8}}\right) + 34994 = 0 \quad (9)$$

$$UBOD = \left[1802.7\ln\left(\frac{BODM}{3.44\times10^{-8}}\right) - 34994\right] / \ln\left(\frac{BODM}{3.44\times10^{-8}}\right) \quad (10)$$

Constraints,

$$BODM < UBOD$$

$$0\ mg/L < BODM < 160{,}000\ mg/L$$

Storm wastewater,

$$BODM = e^{29-0.8UBOD+0.0058UBOD^2} \quad (11)$$

$$0.0058UBOD^2 - 0.8UBOD + 29 - \ln BODM = 0 \quad (12)$$

Constraints,

$$BODM < UBOD$$

$$20\ mg/L < BODM < 90\ mg/L$$

BODM numbers do not cover the whole expected range for domestic, industrial, and storm wastewaters. The introduced constraints for the BODM would ensure a good approximation of UBOD based on the fitting analysis. The model equations can be used to calculate UBOD from BODM or vice versa as long as the given constraints are met. This has been backed up by the fitting analysis and algebraic rearrangement regardless of having BODM in a large range (i.e., industrial wastewater) or in a small range (i.e., stormwater).

Results and discussion

Benchmarking of DO_5 against the O_2 consumption (X) was established from the fitted linear trendlines for both training and testing datasets as shown in Fig. 3A and B, respectively. There is a possible proportional correlation between O_2 consumption and organic decomposition, but the two factors do not necessarily change with the same magnitude. The average accuracy (%) of the built and trained models shown in Fig. 3C based on the applied algorithm and the involved number of features confirmed the optimal models: the fine Gaussian support vector machines (FGSVM), medium trees (MT), and ensemble boosted trees (EBT)-based models had the highest accuracies (with $R^2 > 0.82$) among the other tested supervised models used in the prediction of DO_5. Less than 20% of the FGSVM plotted data points of the observed (predicted) responses against true responses of DO_5 had shown strong deviations. This indicated the highly observed accuracies for DO_5 prediction determined from the three models (FGSVM, MT, and EBT) based on the different combinations and numbers of predictors as shown in Fig. 3D–F. The maximum accuracy of 95% ($R^2 = 0.95$) was achieved from FGSVM-[UBOD-X] to FGSVM-[UBOD-X-T-COD] whereas the minimum accuracy of 80% ($R^2 = 0.8$) was found from the use of EBT-[UBOD-X] and EBT-[T-COD], knowing the general term XX-[a-b-c-d-e-f] has XX = regressor and a,b,c,d,e,f = predictors (inputs). Thus, it was concluded that the acquired datasets that were curated from earlier studies (Abdulla et al., 2016; Alagha et al., 2020; Al-Sulaiman & Khudair, 2018) would be better modeled *via* SVMs rather than ensembles/decision tree algorithms.

The identified raw datasets from Abdulla et al., 2016, Al-Sulaiman & Khudair, 2018, Alagha et al., 2020 conveyed the relationship between BODM and df which were then utilized to estimate the trends of the changing patterns for each of the studied wastewater (domestic,

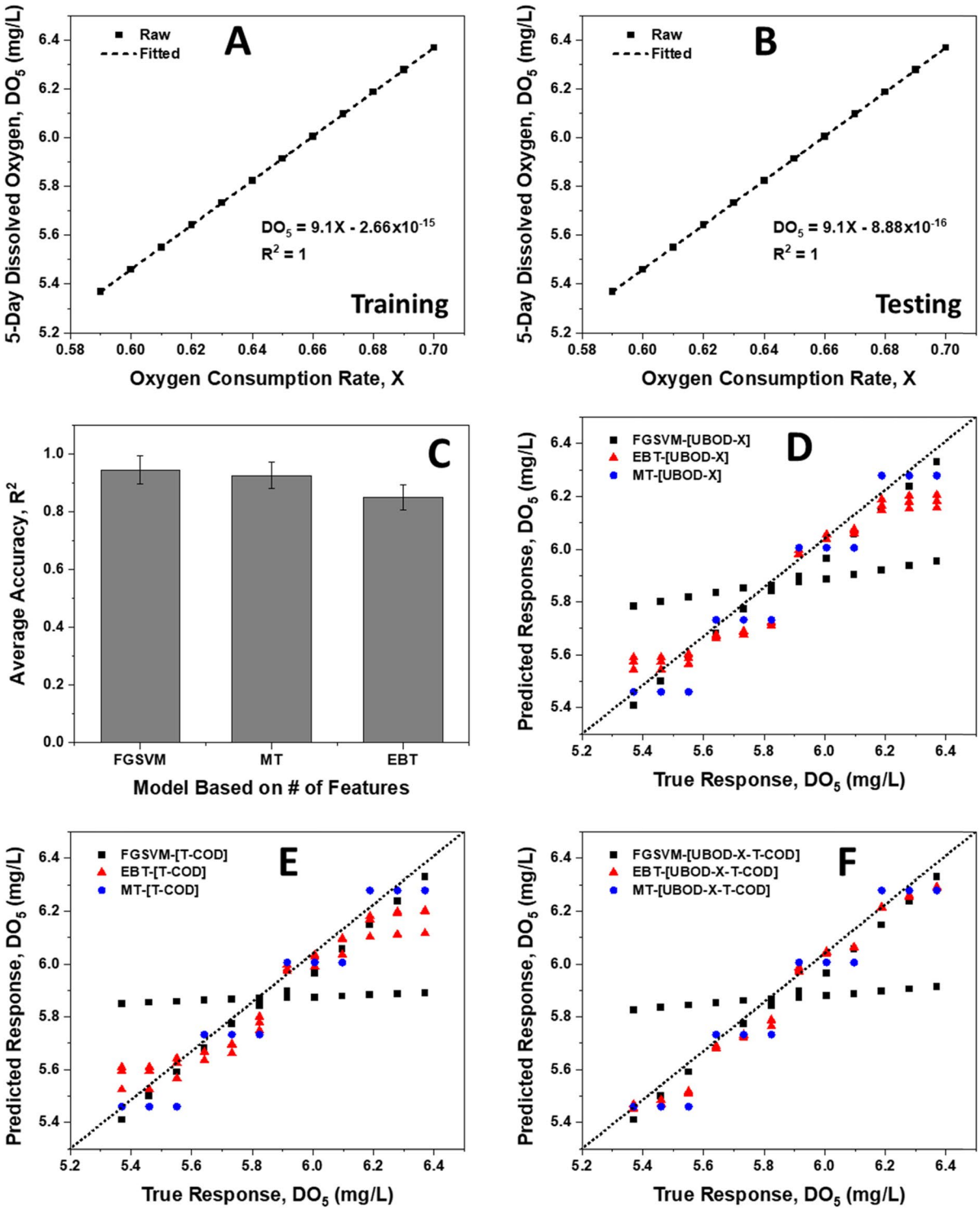


Fig. 3 The determined fitted linear relationships between DO_5 and X = 60–70% according to raw data used in the machine learning analysis. **A** Training and **B** testing; **C** average accuracy of various built and trained models used in the prediction of DO_5; the observed (predicted) responses against true responses of DO_5 using various models based on different combinations and numbers of predictors: **D** UBOD-X as the only two features, **E** T-COD as the only two features, and **F** UBOD-X-T-COD as four features

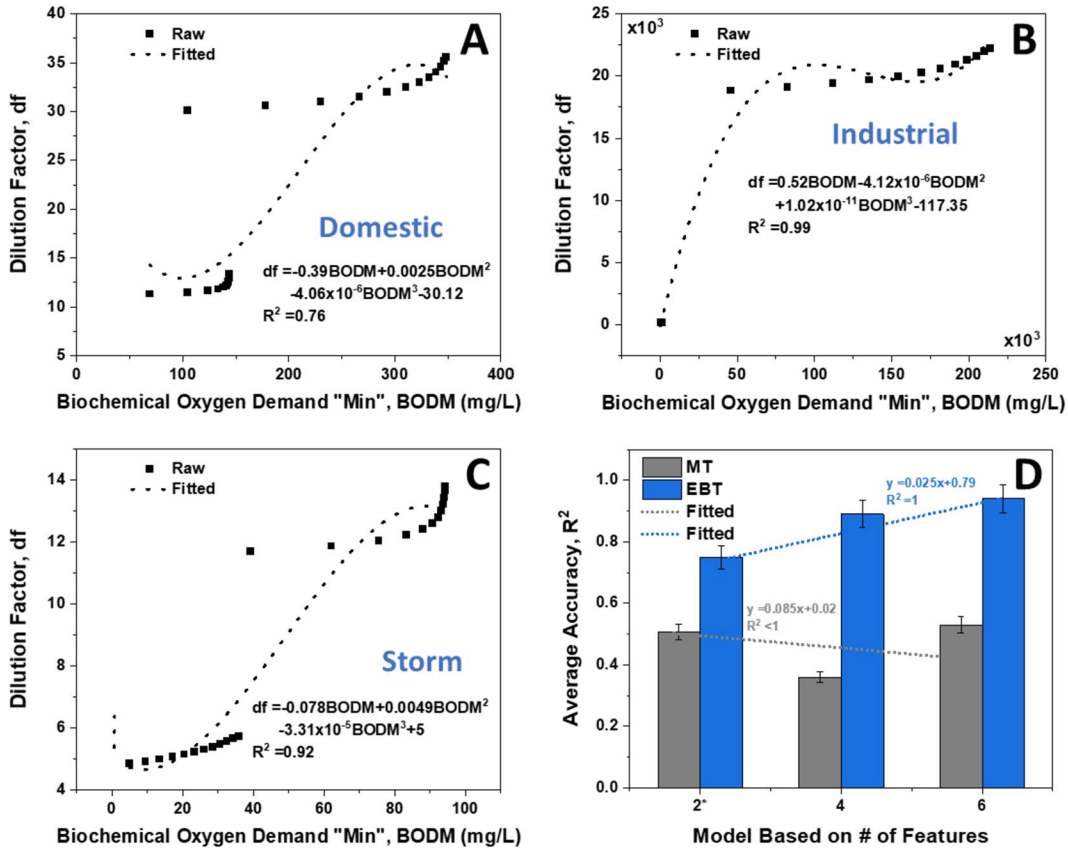


Fig. 4 The determined fitted third-order polynomial relationships between df and BODM (mg/L) according to raw data collected from influents in various wastewater treatment plants. **A** Domestic wastewater (DOM), **B** industrial wastewater (IND), **C** storm wastewater (STM), and **D** comparison of average accuracy of MT- and EBT-trained models in predicting df using different number of features for regression training, which show an increase in EBT accuracy and a slight decrease in MT accuracy with more included predictors

industrial, and storm), as shown in Fig. 4A–C. As expected, stormwater showed the lowest required df among the other studied wastewater types with the minimum observed BOD (or BODM < 100 mg/L) as illustrated in Fig. 4C. Conversely, the industrial wastewater with its specific characteristics reported in the "Methodology and Equations" section, and from Al-Sulaiman & Khudair, 2018; Attiogbe et al., 2009, had the maximum required df for BOD analysis as shown in Fig. 4B. Industrial wastewater requires three orders of magnitudes higher df values than that for domestic and storm wastewaters. The fitted third-order polynomial relationships enabled the initiation of model formulation analysis. The established collective model describes changes in df and DO_5 according to BOD. The EBT-trained models were found to have much better average prediction accuracies than MT. The models' reliability and prediction capability were generally enhanced when more predictors were included in the training (Fig. 4D). The given asterisk for #2* features (in the x-axis) refers to the unreported results from MT or EBT with an accuracy of less than 60% (excluded). The use of different numbers of predictors is also referred to as "sensitivity analysis" in which it was performed to understand the impact of variations in the input on the results from the trained models with a CV of 5-fold. This approach enables the selection of appropriate features by understanding their importance by trying each input feature one at a time and checking/examining the response of the ML models to determine the feature's rank.

EBT predictions were more accurate than MT predictions in nearly all cases as shown in Fig. 5. This observation is consistent with data shown earlier in Fig. 4D which shows the incapability of MT-trained

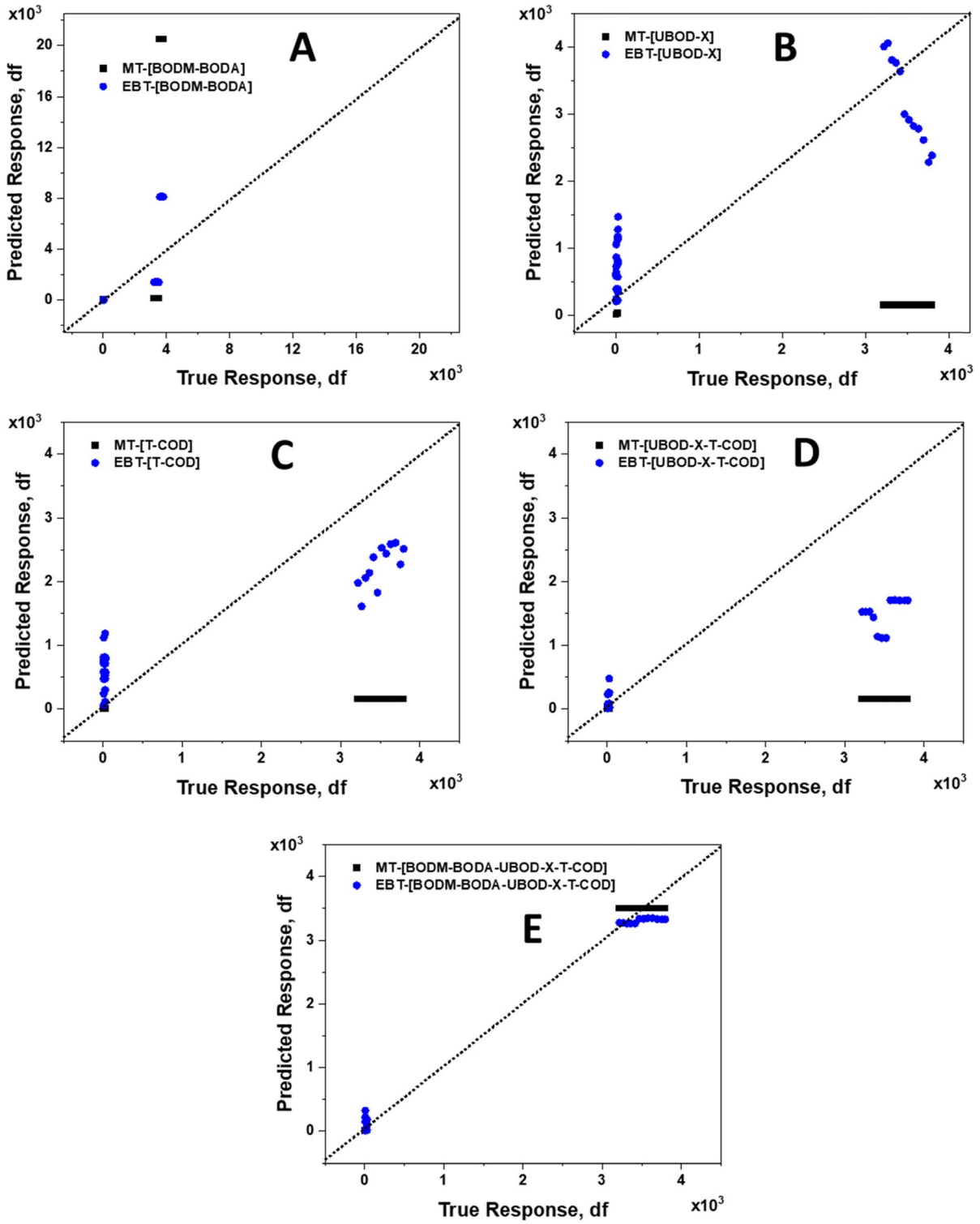


Fig. 5 The observed (predicted) responses against true responses of df from influents in various industrial wastewater treatment plants using various models based on different combinations and numbers of predictors. **A** BODM-BODA, **B** UBOD-X, **C** T-COD, **D** UBOD-X-T-COD, and **E** BODM-BODA-UBOD-X-T-COD

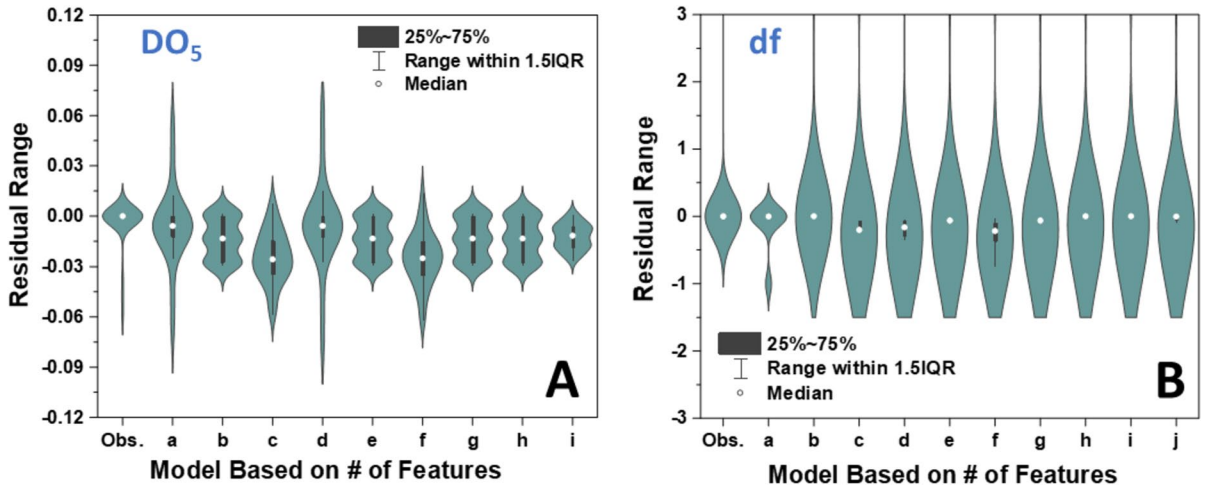


Fig. 6 The residual ranges within 1.5IQR associated with the expected errors from various trained models based on different combinations and numbers of predictors for the prediction of outputs: **A** DO_5 output, **B** df output. Note: the x-axis refers to { Obs. Observed, a. FGSVM-[UBOD-X], b. MT-[UBOD-X], c. EBT-[UBOD-X], d. FGSVM-[T-COD], e. MT-[T-COD], f. EBT-[T-COD], g. FGSVM-[UBOD-X-T-COD], h. MT-[UBOD-X-T-COD], i. EBT-[UBOD-X-T-COD] }

models to achieve accuracy higher than 60%. The low accuracy of MT models was reflected in the predicted responses deviating from the true responses for df values. Note that df is reported and found in high values due to the impact of industrial wastewater in the model building. Industrial wastewater mostly contains very high amounts of organic compounds that would cause a significant increase in df value. The closer the data points were found to the diagonal-dotted line in Fig. 5, the more accuracy we inferred from such combinations of predictors and this is in agreement with a high accuracy of $R^2 > 94\%$ reserved for the following predictors' combinations: (i) MT-[UBOD-X], MT-[UBOD-X-T-COD], and EBT-[UBOD-X-T-COD] for DO_5 predictions; and (ii) EBT-[BODM-BODA] and EBT-[BODM-BODA-UBOD-X-T-COD] for df predictions. It should be noted that the general term XX-[a-b-c-d-e-f] has XX = regressor and a,b,c,d,e,f = predictors (inputs). Over-fitting is a possibility, but it mostly occurs when adding many predictors (more than 4) for model training yielding unexpectedly high accuracies. However, it is not necessary that adding more predictors will always result in higher model prediction accuracies since a two-predictor model (e.g., EBT-[BODM-BODA]) was able to achieve the same high accuracy as a six-predictor model (e.g., EBT-[BODM-BODA-UBOD-X-T-COD]) for df estimations as shown in Fig. 5A and E. Similarly, these observations were found true for DO_5 MT-based models and were seen from a comparison between MT-[UBOD-X] against MT-[UBOD-X-T-COD]. These results suggest the ability to have a good fitting with fewer errors using fewer predictors. In other words, some 2-predictor models yielded the same or better accuracies than those (+3)-predictor models, opposing the expected impact of over-fitting.

The residual analysis, QR, and IQR methods were applied to confirm the selection of the best models. The 1.5IQR range-median criterion ensures finding models with the fewest statistical anomalies (outliers). Using these techniques, unusual responses from the trained models can be detected for both the training and testing datasets. Both MT and EBT outcomes from the +10 distinct input models (b. MT-[UBOD-X] i. EBT-[UBOD-X-T-COD]) had the least range of residuals for DO_5 as shown in Fig. 6A. This confirms the high accuracy owing to the minimum detected outliers (near the zero line). The same analysis was applied to df models and brought about an unexpected outcome, as shown in Fig. 6B, which suggested that only one model (a. MT-[BODM-BODA]) had the minimum outliers among the +10 models. However, this observation seems to be partially true since BODM-BODA was the optimum selection of predictors for a close-to-perfect estimation, but with EBT rather than the MT algorithm. This discrepancy

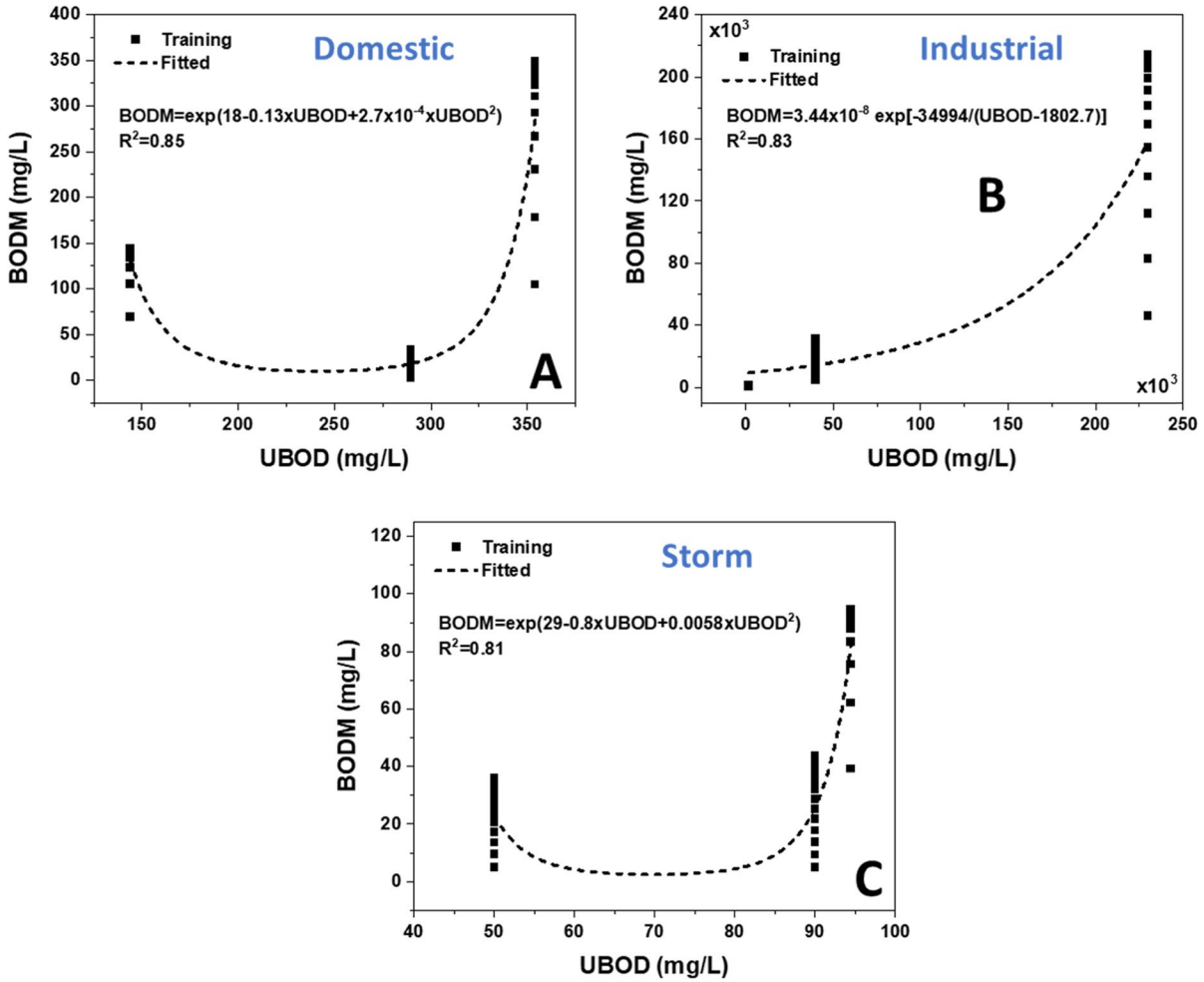
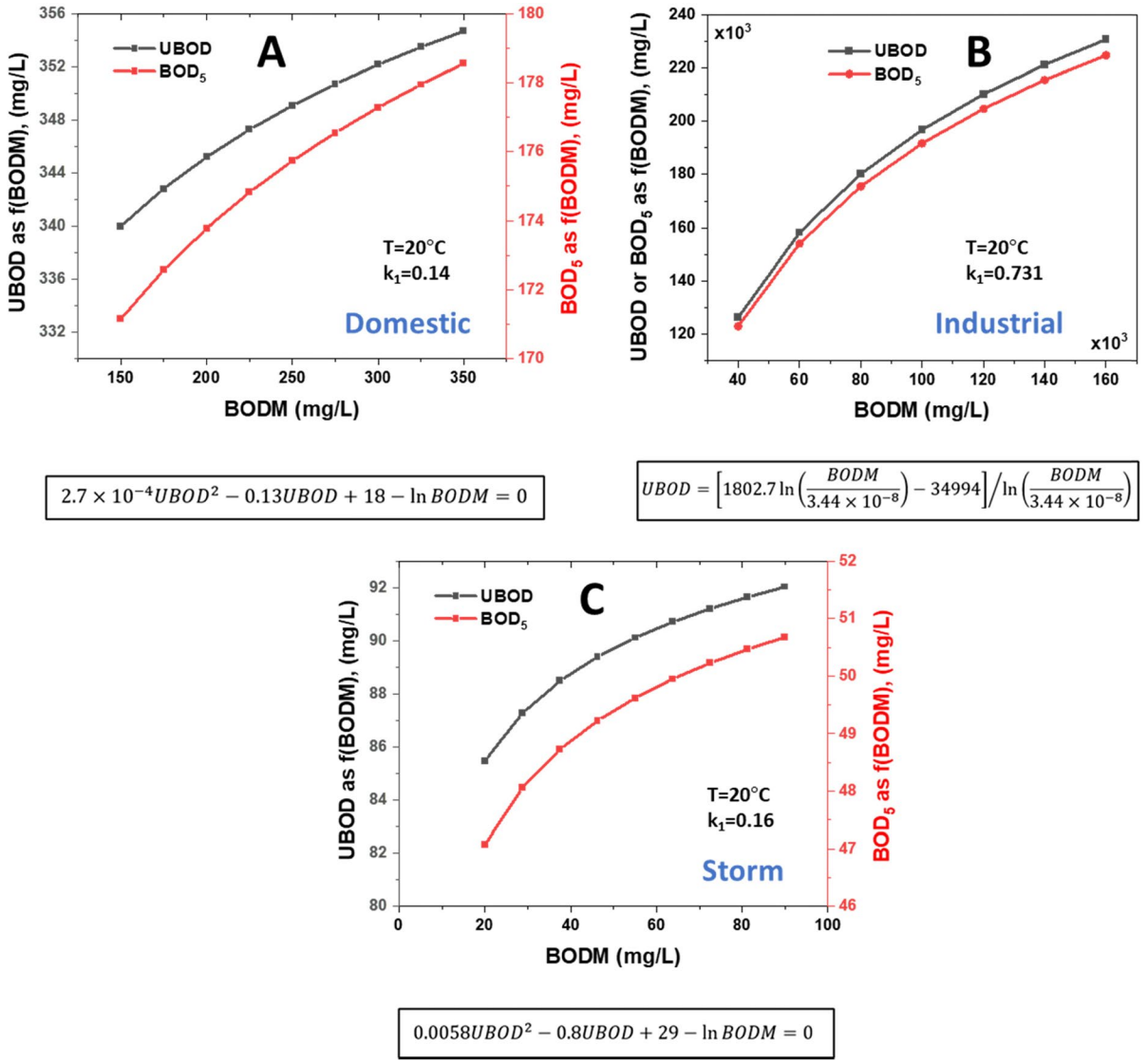


Fig. 7 The determined fitted exponential relationships between BODM and UBOD (mg/L) according to raw data collected from influents in various wastewater treatment plants: **A** domestic wastewater (DOM), **B** industrial wastewater (IND), **C** storm wastewater (STM)

can be explained by the fact that df for IND wastewater is three orders of magnitude higher than that of DOM and STM wastewaters which might impact outlier detection. In general, small ranges within 1.5IQR indicate minimal deviations from actual observations which would mean much less spread of residuals and fewer errors.

The relationships between BODM and UBOD (mg/L) have been determined for the three wastewaters according to raw data of influents. It has been found that the raw data followed exponential patterns which were utilized later in the model formulation analysis. The importance of the found exponential fittings from $UBOD = f(BODM)$ yielded $BOD_T = f(BODM)\left(1 - e^{-k_1 T}\right)$ for each wastewater type, which assisted in obtaining the correlations between df and DO_5 against BOD_T (or BODM) for the construction of collective models. For the correctness of the fitted relationships shown in Fig. 7, BODM must always be less than its corresponding UBOD (max BOD_T) as per the BODM constraints. BODM exponentially increases with UBOD as found for the three wastewater types implying that the BODM parameter becomes larger over time until reaching the UBOD.

Algebraic rearrangements of the exponential relationships of BODM against UBOD showed that both



$$2.7 \times 10^{-4} UBOD^2 - 0.13 UBOD + 18 - \ln BODM = 0$$

$$UBOD = \left[1802.7 \ln\left(\frac{BODM}{3.44 \times 10^{-8}}\right) - 34994\right] \Big/ \ln\left(\frac{BODM}{3.44 \times 10^{-8}}\right)$$

$$0.0058 UBOD^2 - 0.8 UBOD + 29 - \ln BODM = 0$$

Fig. 8 The change of both UBOD and BOD_5 (mg/L) as a function of BODM (mg/L) at a constant water temperature (20°C) with the corresponding UBOD's derived quadratic and/or logarithmic equations according to the determined fitted exponential relationships between BODM and UBOD: **A** domestic wastewater (DOM) with $k_{1,\,avg} = 0.14$ day^{-1}, **B** industrial wastewater (IND) with $k_{1,\,avg} = 0.731$ day^{-1}, **C** storm wastewater (STM) with $k_{1,\,avg} = 0.16$ day^{-1}

UBOD in domestic and storm wastewaters could be explained by quadratic equations knowing the logarithmic value of BODM as shown in Fig. 8. Conversely, industrial wastewater had a unique fractional relationship to correlate UBOD to the logarithmic value of BODM. Each of the identified equations was applied to visualize the change of both UBOD and BOD_5 as a function of BODM at a constant water temperature (20°C). The results confirmed that BOD_5 is always less than UBOD which is a function of BODM according to the derived quadratic and/or logarithmic equations. These findings confirmed the correctness of derived relationships since BODM is always less than the average BOD_5. The incrementing patterns of UBOD and BOD_5 (against BODM or time) are due to the near-complete organic decomposition attributed to 60–70% oxygen consumption from the available O_2 in wastewater.

For industrial wastewater, it was found that UBOD and BOD_5 were very close in values at different BODM which might be explained by high organic contents as illustrated in Fig. 8B. High amounts of organics would make organic decomposition occur within the first few days (5 days) leading to having close values of average and maximum BOD. High organic levels depend on the industry type in which food processing, pulping processes, and meat processing industries would have much higher levels of organics in wastewater than those from mineral processing. The progression of organic decomposition would be much faster over time (or based on BODM) for industrial wastewater because of the high organic contents, resulting in having BOD_5 approaching UBOD.

Collective model

Considering the three wastewater types together in the modeling analysis, it is possible to estimate a correlation between df and BODM. This ultimately enables finding the approximated relationship of DO_5 against both BOD_5 and BODM using the aid of Eq. (1). The analysis was carried out with $k_{1,avg}$ based on the selected wastewater for BOD_T @ $T = 5$ and $T = 60$ days. Note that BOD_5 and DO_5 are inversely proportional to each other as confirmed by Eq. (13) to (15) (i.e., a decline in DO_5 levels reflects a high level of BOD_5).

$$df = \frac{BOD_5}{DO_0 - DO_5} \tag{13}$$

The constructed collective model was used to plot how the 5-day dissolved oxygen (DO_5) changes concerning both BODM and BOD_5 for an organic decomposition extent of $X = 60$–70%. The proposed collective model equations enable the prediction of changes in DO_5 against the minimum and average BOD with deviations ranging from 5 to 10% and R^2 approaching unity as shown in Fig. 9. The red-dotted lines correspond to the linear fitting of model approximations for BODM vs. DO_5. The maximum observed errors in DO_5 predictions against BODM were ±0.30 (10%), +0.20 (5%), and +0.60 (10%) in mg/L for DOM, IND, and STM wastewaters, respectively, as shown in Fig. 9A–C. Furthermore, considering the existence of some outliers resulted in fitted model trends similar to the linear trends of those raw data points. When it comes to the average BOD_5 against DO_5 analysis (blue lines/markers), it was found that deviations from actual values were around 10% at maximum with −0.60, +0.40, and +0.70 in mg/L in DO_5 predictions for DOM, IND, and STM wastewaters, respectively, as shown in Fig. 9A–C. The regression-based analytical models from Eq. (15) enable prediction of DO_5 changes according to changes in BOD_5, while taking into consideration that IND-BOD numeric values should be multiplied by 10^3 for correct predictions.

The DO_5 for wastewater was initially estimated based on the linear fitting of observed changes. However, the collective model relationships Eq. (13) to (15) required imposing further constraints and

$$df = \begin{cases} DOM \rightarrow & -0.39BODM + 0.0025BODM^2 - 4.06\times10^{-6}BODM^3 - 30.12 \\ IND \rightarrow & 0.52BODM - 4.12\times10^{-6}BODM^2 + 1.02\times10^{-11}BODM^3 - 117.35 \\ STM \rightarrow & -0.078BODM + 0.0049BODM^2 - 3.31\times10^{-5}BODM^3 + 5 \end{cases} \tag{14}$$

$$DO_5 = \begin{cases} DOM \rightarrow & \frac{1}{(1/X-1)} \frac{BOD_5}{-0.39BODM+0.0025BODM^2-4.06\times10^{-6}BODM^3-30.12} \\ IND \rightarrow & \frac{1}{(1/X-1)} \frac{BOD_5}{0.52BODM-4.12\times10^{-6}BODM^2+1.02\times10^{-11}BODM^3-117.35} \\ STM \rightarrow & \frac{1}{(1/X-1)} \frac{BOD_5}{-0.078BODM+0.0049BODM^2-3.31\times10^{-5}BODM^3+5} \end{cases} \tag{15}$$

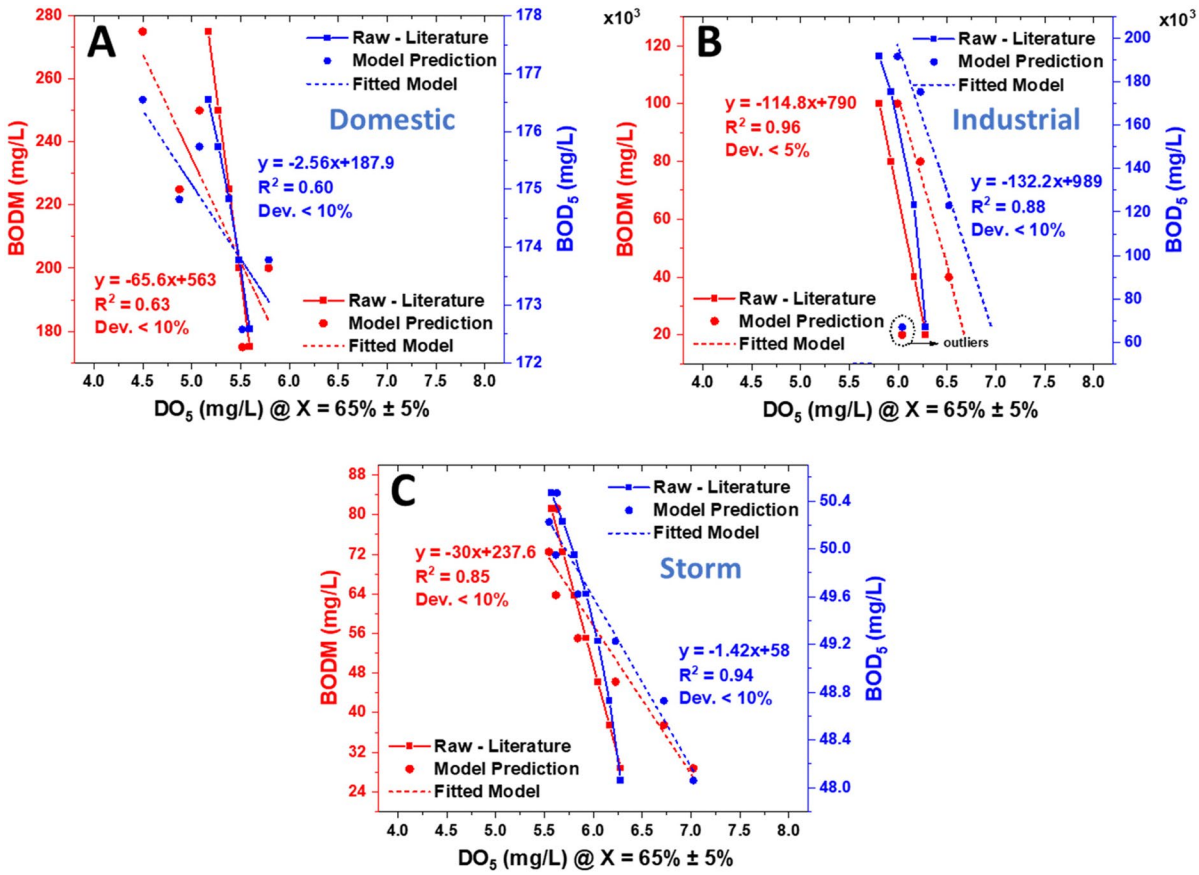


Fig. 9 The change of DO_5 (mg/L) as a function of both BODM (mg/L) and BOD_5 (mg/L) at a constant water temperature (20°C) according to the literature raw data and model prediction from the corresponding DO_5's derived equations determined from the fitted exponential relationships between BODM and UBOD: **A** domestic wastewater (DOM) with $k_{1,\,avg}$ = 0.14 day^{-1}, adopted from Abdulla et al., 2016; Al-Sulaiman & Khudair, 2018; Alagha et al., 2020, **B** industrial wastewater (IND) with $k_{1,\,avg}$ = 0.731 day^{-1}, adopted from Al-Sulaiman & Khudair, 2018; Attiogbe et al., 2009, **C** storm wastewater (STM) with $k_{1,\,avg}$ = 0.16 day^{-1}, adopted from Adedeji & Olayinka, 2013; Langeveld et al., 2012; Łapiński & Wiater, 2018. Note: dev. refers to deviation of model results from the experiments or the reported raw data in the literature

corrections to optimally predict the DO_5 changing patterns with the highest possible accuracy. The introduced correction factor $+\left(\frac{UBOD_i}{BODM_i}\right)^{\alpha}$ for the DOM wastewater relationship shown in Eq. (15) included two constraints, where i=0, 1, 2…6 refers to the selected BODM datapoint and its corresponding UBOD and [$=3.8+i$ for BODM ≤ 200 mg/L, $\alpha=6+2i$ for BODM > 200-275 mg/L]. Moreover, it was observed that adding a correction factor to the right-hand side (RHS) of the IND wastewater relationship shown in Eq. (15) as $+\left(\frac{UBOD_i}{BODM_i}\right)^{\alpha}$ would further eliminate existing prediction errors for DO_5 as shown in Fig. 9. It should be noted that for IND wastewater i=0, 1, 2…6 refers to the selected BODM datapoint and its corresponding UBOD and [$=1.6+0.3i$ for BODM ≤ 60×10^3 mg/L, $\alpha=0.8+0.7i$ for BODM > 60-120$\times10^3$ mg/L]. However, STM wastewater had a simple correction factor which resulted in a higher prediction accuracy. Such high predictions were achieved by the addition of the factor $-\left(\frac{UBOD_i}{BODM_i}\right)^{\alpha}$ to the STM wastewater relationship shown in Eq. (15), where i=0, 1, 2…6 refers to the selected BODM datapoint and its corresponding UBOD and [$\alpha=1.85+0.2i$] for the studied range of BODM (20–90 mg/L). The derived models are only applicable to wastewaters similar to the studied

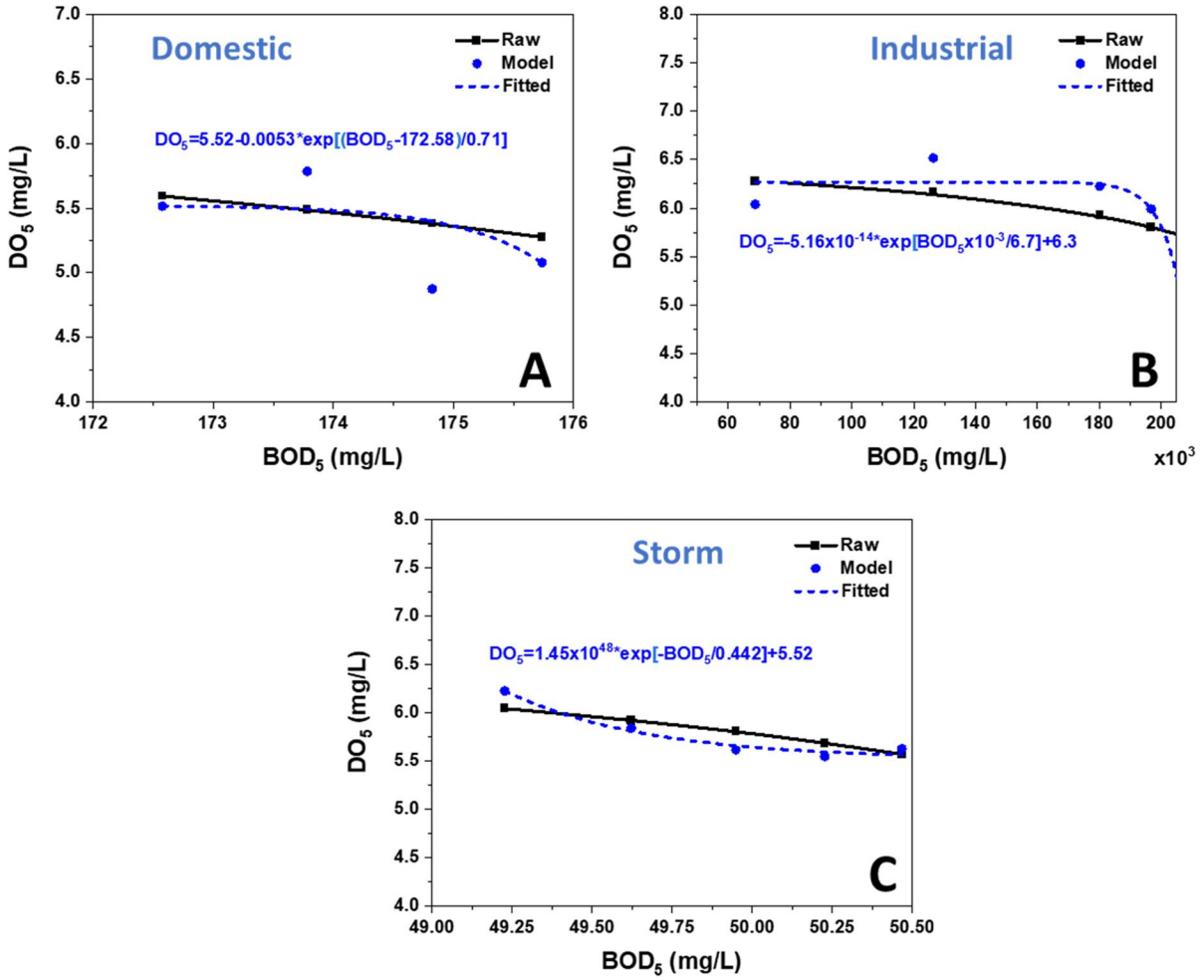


Fig. 10 The collective model predictions and the patterns of change of DO_5 (mg/L) as a function of BOD_5 (mg/L) at a constant water temperature (T=20 °C): **A** domestic wastewater (DOM) with $k_{1,\,avg}$ = 0.14 day^{-1}, **B** industrial wastewater (IND) with $k_{1,\,avg}$ = 0.731 day^{-1}, **C** storm wastewater (STM) with $k_{1,\,avg}$ = 0.16 day^{-1}

characteristics of the three different types considering the BODM constraints in model building.

According to the initial estimation for BOD_5 and BODM ranges as per the type of wastewater that WWTP engineers are dealing with, Eq. (14) and (15) enable wastewater operators and engineers to choose the optimal df and DO_5 to ensure the effectiveness of the experimental (sampling) measurement of BOD_T. By taking into account the developed charts for BODM vs. df (Fig. 4) and BOD_5 vs. DO_5 (Fig. 9), the average oxygen consumption rate would provide the best forecast for estimating df values for each of the analyzed wastewater types. An increase in the df is predominated by BOD_5 levels (i.e., more dilution is needed for higher BOD_5). Dilution techniques should be performed following the previously mentioned "standard methods for water and wastewater examination," which states that five samples should be created with a variety of dilutions. At least two samples should produce acceptable minimum DO depletion (> 2 mg/L uptake after a 5-day incubation period) and residual limits (> 1 mg/L) (APHA, AWWA, WEF, 2012). As a general rule, existing oxygen (also known as DO content) and consumption or organic degradation are always inversely correlated and are dependent upon the oxygen addition by using aerators/diffusers; the greater the DO_5 level in the water, the more organic breakdown takes place. Anaerobic digestion

Table 2 The estimated average and maximum BOD_T with the maximum observed errors in DO_5 predictions, proposed experimental dilution factor (df), average df range, and introduced corrections/constraints for the built collective models

Wastewater type (abbrev.)	BOD_T* (mg/L)	UBOD (mg/L)	DO_5 vs. BODM, max err	DO_5 vs. BOD_5, max err	df*	df range*	Correction factors and constraints
Storm (STM)	49.3	89.6	10% (+0.60)	10% (+0.70)	9.2	4.5–13.5	$-\left(\frac{UBOD_i}{BODM_i}\right)^{\alpha}$ [$\alpha = 1.85 + 0.2i$]
Industrial (IND)	150,000	174,000	5% (+0.20)	10% (+0.40)	12	1–22	$+\left(\frac{UBOD_i}{BODM_i}\right)^{\alpha}$ [$\alpha = 1.6 + 0.3i$ for BODM ≤ 60×10^3 mg/L, $\alpha = 0.8 + 0.7i$ for BODM > 60-120×10^3 mg/L]
Domestic (DOM)	175.4	348.4	10% (±0.30)	10% (−0.60)	18.5–28.5	12–35	$+\left(\frac{UBOD_i}{BODM_i}\right)^{\alpha}$ [$\alpha = 3.8 + 1i$ for BODM ≤ 200 mg/L, $\alpha = 6 + 2i$ for BODM > 200-275 mg/L]

*Average values from $k_{1,\,avg}$ at 20°C which correspond to the selected wastewater type experimental results with DO_5 and df obtained from near-complete organic oxidation and O_2 consumption at 60–70%. The df values for IND must be multiplied by 10^3 as per the model formulation analysis. Note that $i = 0, 1, 2 \ldots 6$ refers to the selected BODM datapoint and its corresponding UBOD

of organic matter can occur leading to organic degradation which yields methane production (Chynoweth, 1987). The average df values were indicated by Fig. 4, with df≈9.2 for STM water, df≈12×10^3 for IND wastewater, and df≈18.5–28.5 for DOM wastewater demonstrating the lack of organic elements in storm wastewater influents. The derived df values can be useful for wastewater engineers to start with when trying to find a cost-effective DO concentration that would enable maximal organic decomposition.

The implemented collective models (with introduced corrections) enabled improved prediction accuracy for DO_5 against BOD_5 as shown in Fig. 10. Both proposed collective model equations for DOM and STM wastewater were highly accurate when considering fitting the model observations against the raw literature results for DO_5 against BOD_5 (Abdulla et al., 2016; Adedeji & Olayinka, 2013; Alagha et al., 2020; Al-Sulaiman & Khudair, 2018; Attiogbe et al., 2009; Langeveld et al., 2012; Łapiński & Wiater, 2018). The fitted-model results showed similar patterns to the raw-data patterns suggesting the models' applicability for the studied BOD_5 ranges. However, some deviations from the raw data were evident as shown for IND wastewater and $BOD_5 = 140-180\times10^3$ mg/L, with a maximum of +5% errors. For better predictions, it is advised to use the identified exponential equations presented in Fig. 10 instead of those linear equations obtained previously as shown in Fig. 9. This is because the collective model equations were further optimized from the developed correction factors which resulted in the exponential fitted-model equations. Table 2 summarizes the estimated BOD_T as well as maximum errors in DO_5 predictions and df with the average df range for the various wastewaters. Moreover, the developed correction factors and constraints are reported in Table 2. The formulated models approximated the relationship of DO_5 against both BOD_5 and BODM, noting that BOD_5 and DO_5 are inversely proportional to each other (a decline in DO_5 levels reflects a high level of BOD_5). However, the proportionality will differ based on the microbial concentration at varying locations. Despite that the BOD_5 ranges in Fig. 10 are small, these are different for the various wastewaters and as per the constraints taken in building the collective model. The model's usefulness is in the possibility of estimating DO_5 knowing BOD_5 or the other way around.

Conclusion

This study investigated organic decomposition rates in BOD-containing wastewater from BOD_T literature datasets of domestic, industrial, and storm wastewaters. Determination of the average laboratory dilution factors (df) was carried out from ΔDO based on the O_2 consumption (X) in the range of 60–70%. Accurate machine learning models were built from the defined independent variables (inputs) including (i) UBOD, (ii) BODM, (iii) BODA, (iv) COD, (v) O_2 consumption (X) of 60–70%, and (vi) time (T). The proposed analysis enabled estimating the corresponding desired outputs [DO_5 and df] from knowing BOD_5, COD, DO_0, and O_2 consumption (X) information for training with mixing and matching independent parameters. Residual analysis and inter-quartile range (IQR) based on the 1.5IQR range-median decision rule would guide researchers towards useful correlations with minimum statistical errors. The formulated models approximated the relationship of DO_5 against both BOD_5 and BODM, noting that BOD_5 and DO_5 are inversely proportional to each other (a decline in DO_5 levels reflects a high level of BOD_5).

The maximum accuracy of 95% was achieved from FGSVM-[UBOD-X] and FGSVM-[UBOD-X-T-COD], indicating the potential of SVMs training. It should be noted that the general term XX-[a-b-c-d-e-f] has XX = regressor and a,b,c,d,e,f = predictors (input features). Such df and DO_5 numbers are useful as a starting point for sampling analysis to quantify organic decomposition rates attributed to the introduced oxygen. An increase in df is predominated by BOD_5 levels (more dilution needed for higher BOD_5), with df≈9.2 for storm (STM) wastewater, df≈12 × 10^3 for industrial (IND) wastewater, and df≈18.5–28.5 for domestic (DOM) wastewater. Stormwater had shown the lowest required df (BODM < 100 mg/L) whereas industrial wastewater had the maximum required df for BOD analysis. In models training, more predictors generally enhanced the model reliability with a high accuracy > 94% for predictors combinations: (i) MT-[UBOD-X], MT-[UBOD-X-T-COD], and EBT-[UBOD-X-T-COD] for DO_5 predictions; and (ii) EBT-[BODM-BODA] and EBT-[BODM-BODA-UBOD-X-T-COD] for df predictions. The least range of residual was found for MT-[UBOD-X] and EBT-[UBOD-X-T-COD]) for DO_5, but only MT-[BODM-BODA] had the minimum outliers among df models. It was found that BODM exponentially increases over time until reaching UBOD. BOD_5 was found to be a function of BODM from the derived quadratic and/or logarithmic equations according to the fitted exponential relationships. The proposed collective models were capable of predicting changes in DO_5 with deviations ranging from 5 to 10%. Moreover, imposing constraints and introducing correction factors as $\pm\left(\frac{UBOD_i}{BODM_i}\right)^{\alpha}$ resulted in achieving the highest accuracy for DO_5 estimations. The optimized collective models yield cubic equations derived for df and DO_5 from BODM which is an exponent function in UBOD. Proposed models would bridge the gap between science and industry best practices for optimal design and operation to minimize organic contamination and facilitate effluent quality assessment.

Acknowledgements The author would like to acknowledge the Deanship of Scientific Research (DSR) at King Abdulaziz University (KAU) for their technical and financial support to complete this work.

Author contribution HA M: conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, supervision, validation, visualization, roles/writing—original draft, and writing—review and editing.

Funding This research work was funded by General Research Fund Projects. Therefore, the author gratefully acknowledges technical and financial support from the Ministry of Education and King Abdulaziz University, Deanship of Scientific Research (DSR), Jeddah, Saudi Arabia.

Data availability The datasets generated during and/or collected during the current study are available from the corresponding author on reasonable request.

Declarations

Ethical approval The submitted work is original and has not been published elsewhere in any form or language.

Consent to participate Not applicable. The research does not involve human subjects.

Consent for publication Not Applicable. No case studies and the author does not object to having the manuscript data published.

Competing interests The author declares no competing interests.

References

Abdulla, F. A., Alfarra, A., Qdais, H. A., & Sonneveld, B. (2016). Evaluation of wastewater treatment plants in Jordan and suitability for reuse. *Academia Journal of Environmental Science, 4*(7). https://doi.org/10.15413/ajes.2016.0305

Adedeji, O. H., & Olayinka, O. O. (2013). Heavy metal concentrations in urban stormwater runoff and receiving stream. *Journal of Environment and Earth Science, 3*(7), 141–150.

Alagha, O., Allazem, A., Bukhari, A. A., Anil, I., & Mu'azu, N. D. (2020). Suitability of SBR for wastewater treatment and reuse: Pilot-scale reactor operated in different anoxic conditions. *International Journal of Environmental Research and Public Health, 17*(5), 1617.

Al-Ghazawi, Z., & Alawneh, R. (2021). Use of artificial neural network for predicting effluent quality parameters and enabling wastewater reuse for climate change resilience – A case from Jordan. *Journal of Water Process Engineering, 44*. https://doi.org/10.1016/j.jwpe.2021.102423

Alsulaili, A., & Refaie, A. (2021). Artificial neural network modeling approach for the prediction of five-day biological oxygen demand and wastewater treatment plant performance. *Water Supply, 21*(5). https://doi.org/10.2166/ws.2020.199

Al-Sulaiman, A. M., & Khudair, B. H. (2018). Correlation between Bod5 and Cod for Al-Diwaniyah wastewater treatment plants to obtain the biodigrability indices. *Pakistan Journal of Biotechnology, 15*, 423–427.

Anggraini, N., & Herdiansyah, H. (2019). COD values for determining BOD5 dilution factor in faecal sludge waste - Case study on the duri kosambi faecal sludge treatment plant in DKI Jakarta province. *AIP Conference Proceedings, 2120*. https://doi.org/10.1063/1.5115676

APHA, AWWA, & WEF. (2012). *Standard methods for examination of water and wastewater*. American Public Health Association.

Asami, H., Golabi, M., & Albaji, M. (2021). Simulation of the biochemical and chemical oxygen demand and total suspended solids in wastewater treatment plants: Data-mining approach. *Journal of Cleaner Production, 296*. https://doi.org/10.1016/j.jclepro.2021.126533

Attiogbe, F., Glover-Amengor, M., & Nyadziehe, K. (2009). Correlating biochemical and chemical oxygen demand of effluents – A case study of selected industries in Kumasi, Ghana. *West African Journal of Applied Ecology, 11*(1). https://doi.org/10.4314/wajae.v11i1.45722

Barbato, G., Barini, E. M., Genta, G., & Levi, R. (2011). Features and performance of some outlier detection methods. *Journal of Applied Statistics*. https://doi.org/10.1080/02664763.2010.545119

Baştanlar, Y., & Ozuysal, M. (2014). Introduction to machine learning second edition. *Methods in molecular biology (Clifton, N.J.)*. https://doi.org/10.1007/978-1-62703-748-8_7

Chaudhary, S. (2019). Why "1.5" in IQR method of outlier detection? Towards data science. https://towardsdatascience.com/why-1-5-in-iqr-method-of-outlier-detection-5d07fdc82097

Chynoweth, D. P. (1987). Anaerobic digestion of biomass.

Hach, C. C., Klein, R. L., Jr., & Gibbs, C. R. (1997). *Biochemical oxygen demand* (p. 7). Tech. Monogr..

Jain, S. K., & Singh, V. P. (2003). Chapter 13 Water quality modeling. In *Developments in Water Science* (Vol. 51, pp. 743–786). Elsevier. https://doi.org/10.1016/S0167-5648(03)80067-9

Khusravi, R. (2013). BOD5 removal kinetics and wastewater flow pattern of stabilization pond system in Birjand. *European Journal of Experimental Biology, 3*(2), 430–436.

Kim, D., Begum, M. S., Choi, J., Jin, H., Chea, E., & Park, J.-H. (2019). Comparing effects of untreated and treated wastewater on riverine greenhouse gas emissions. *APN Science Bulletin, 9*(1). https://doi.org/10.30852/sb.2019.872

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. *Informatica (Ljubljana)*. https://doi.org/10.31449/inf.v31i3.148

Langeveld, J. G., Liefting, H. J., & Boogaard, F. C. (2012). Uncertainties of stormwater characteristics and removal rates of stormwater treatment facilities: Implications for stormwater handling. *Water Research, 46*(20). https://doi.org/10.1016/j.watres.2012.06.001

Łapiński, D., & Wiater, J. (2018, 30). Contamination content introduced with rain water to the rivers after they have been cleaned in separators of petroleum compounds. *E3S Web of Conferences*. https://doi.org/10.1051/e3sconf/20183001019

Lewis, M. E. (2006). Dissolved oxygen: U.S. Geological Survey techniques of water-resources investigations. USGS, US. https://water.usgs.gov/owq/FieldManual/. Accessed 10 Jan 2023.

Maddah, H. A. (2016a). Optimal operating conditions in designing photocatalytic reactor for removal of phenol from wastewater. *ARPN Journal of Engineering and Applied Sciences, 11*(3), 1799–1802.

Maddah, H. A. (2016b). Application of finite fourier transform and similarity approach in a binary system of the diffusion of water in a polymer. *Journal of Materials Science and Chemical Engineering, 4*, 20–30.

Maddah, H. A. (2018a). Numerical analysis for the oxidation of phenol with TiO2 in wastewater photocatalytic reactors. *Engineering, Technology & Applied Science Research, 8*(5), 3463–3469.

Maddah, H. A. (2018b). Modeling the feasibility of employing solar energy for water distillation. In *Handbook of Environmental Materials Management*.

Maddah, H. A. (2021a). Predicting flux rates against pressure via solution-diffusion in reverse osmosis membranes. *Engineering, Technology & Applied Science Research, 11*(2), 6902–6906.

Maddah, H. A. (2021b). Simulating fouling impact on the permeate flux in high-pressure membranes. *International Journal of Advanced and Applied Sciences, 8*(8), 1–8.

Maddah, H. A., & Chogle, A. M. (2015). Applicability of low pressure membranes for wastewater treatment with cost study analyses. *Membrane Water Treatment, 6*(6). https://doi.org/10.12989/mwt.2015.6.6.477

Maddah, H. A. (2016c). Polypropylene as a promising plastic: A review. *American Journal of Polymer Science, 6*(1), 1–11. https://doi.org/10.5923/j.ajps.20160601.01

Maddah, H. A. (2020). Adsorption isotherm of NaCl from aqueous solutions onto activated carbon cloth to enhance membrane filtration. *Journal of Applied Science and Engineering*. https://doi.org/10.6180/jase.202003_23(1).0009

Maddah, H. A. (2022). Predicting optimum dilution factors for BOD sampling and desired dissolved oxygen for controlling organic contamination in various wastewaters. *International Journal of Chemical Engineering, 2022*. https://doi.org/10.1155/2022/8637064

Maddah, H. A., Alzhrani, A. S., Almalki, A. M., Bassyouni, M., Abdel-Aziz, M. H., Zoromba, M., & Shihon, M. A. (2017). Determination of the treatment efficiency of different ommercial membrane modules for the treatment of groundwater. *Journal of Materials and Environmental Science, 8*(6), 2006–2012.

Maddah, H. A., Alzhrani, A. S., Bassyouni, M., Abdel-Aziz, M. H., Zoromba, M., & Almalki, A. M. (2018). Evaluation of various membrane filtration modules for the treatment of seawater. *Applied Water Science*. https://doi.org/10.1007/s13201-018-0793-8

Maddah, H. A., Bassyouni, M., Abdel-Aziz, M. H., Zoromba, M. S., & Al-Hossainy, A. F. (2020). Performance estimation of a mini-passive solar still via machine learning. *Renewable Energy*. https://doi.org/10.1016/j.renene.2020.08.006

Maddah, H. A., & Shihon, M. A. (2018). Activated carbon cloth for desalination of brackish water using capacitive deionization. *Desalination and Water Treatment*. https://doi.org/10.5772/intechopen.76838

Mathworks. (2017). *Statistics and Machine Learning Toolbox ™ User's Guide R2017a*. MatLab.

Metcalf, W., & Eddy, C. (2003). *Metcalf and Eddy Wastewater Engineering: Treatment and reuse* (p. New York, NY). Wastewater Engineering: Treatment and Reuse McGraw Hill.

Moska, B., Kostrzewa, D., & Brzeski, R. (2020). Influence of the applied outlier detection methods on the quality of classification. *Advances in Intelligent Systems and Computing, 1061*. https://doi.org/10.1007/978-3-030-31964-9_8

Nagel, B., Dellweg, H., & Gierasch, L. M. (1992). Glossary for chemists of terms used in biotechnology (IUPAC recommendations 1992). *Pure and Applied Chemistry, 64*(1), 143–168.

NIHON KASETSU CO. (2023). BOD and COD to characterise wastewater. https://nihonkasetsu.com/bod-and-cod-to-characterise-wastewater/. Accessed 20 Jan 2023.

Obaid, H. A., Shahid, S., Basim, K. N., & Chelliapan, S. (2015). Modeling of wastewater quality in an urban area during festival and rainy days. *Water Science and Technology, 72*(6). https://doi.org/10.2166/wst.2015.297

Qambar, A. S., Khalidy, M. M., & Al. (2022). Prediction of municipal wastewater biochemical oxygen demand using machine learning techniques: A sustainable approach. *Process Safety and Environmental Protection, 168*. https://doi.org/10.1016/j.psep.2022.10.033

Qasaimeh, A., & Al-Ghazawi, Z. (2020). Regression modeling for rapid prediction of wastewater bod5. *Desalination and Water Treatment, 201*. https://doi.org/10.5004/dwt.2020.26043

Rustum, R., Adeloye, A., & Simala, A. (2007). *Kohonen self-organising map (KSOM) extracted features for enhancing MLP-ANN prediction models of BOD5*. IAHS-AISH Publication.

Simeone, O. (2018). A brief introduction to machine learning for engineers. *Foundations and Trends in Signal Processing*. https://doi.org/10.1561/2000000102

Szelag, B., Barbusiński, K., Studziński, J., & Bartkiewicz, L. (2017). Prediction of wastewater quality indicators at the inflow to the wastewater treatment plant using data mining methods. *E3S Web of Conferences, 22*. https://doi.org/10.1051/e3sconf/20172200174

Vörösmarty, C. J., McIntyre, P. B., Gessner, M. O., Dudgeon, D., Prusevich, A., Green, P., et al. (2010). Erratum: Global threats to human water security and river biodiversity (Nature (2010) 467 (555-561)). *Nature*. https://doi.org/10.1038/nature09549

Voß, A., Alcamo, J., Bärlund, I., Voß, F., Kynast, E., Williams, R., & Malve, O. (2012). Continental scale modelling of in-stream river water quality: A report on methodology, test runs, and scenario application. *Hydrological Processes, 26*(16). https://doi.org/10.1002/hyp.9445

Warming, M. (2020). *How can more water treatment cut CO2 emissions?* International Water Association. https://iwa-network.org/how-can-more-water-treatment-cut-co2-emissions. Accessed 15 Sept 2022.

Wen, Y., Schoups, G., & Van De Giesen, N. (2017). Organic pollution of rivers: Combined threats of urbanization, livestock farming and global climate change. *Scientific Reports, 7*. https://doi.org/10.1038/srep43289