

Modeling the organic matter of water using the decision tree coupled with bootstrap aggregated and least-squares boosting

H. Tahraoui, Abdeltif Amrane, A.-E. Belhadj, J. Zhang

▶ **To cite this version:**

H. Tahraoui, Abdeltif Amrane, A.-E. Belhadj, J. Zhang. Modeling the organic matter of water using the decision tree coupled with bootstrap aggregated and least-squares boosting. Environmental Technology and Innovation, 2022, 27, pp.102419. 10.1016/j.eti.2022.102419 . hal-03632799

HAL Id: hal-03632799
https://hal.science/hal-03632799

Submitted on 31 May 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Environmental Technology & Innovation 27 (2022) 102419

Contents lists available at ScienceDirect

Environmental Technology & Innovation

journal homepage: www.elsevier.com/locate/eti

Modeling the organic matter of water using the decision tree coupled with bootstrap aggregated and least-squares boosting

Hichem Tahraoui [a,*], Abdeltif Amrane [b], Abd-Elmouneïm Belhadj [a], Jie Zhang [c]

[a] Laboratory of Biomaterials and Transport Phenomena (LBMPT), University of MÉDÉA, Nouveau Pôle Urbain, Médéa University, 26000 Médéa, Algeria
[b] Univ Rennes, Ecole Nationale Supérieure de Chimie de Rennes, CNRS, ISCR – UMR6226, F-35000 Rennes, France
[c] School of Engineering, Merz Court, Newcastle University, Newcastle Upon Tyne NE1 7RU, UK

ARTICLE INFO

Article history:
Received 10 September 2021
Received in revised form 23 December 2021
Accepted 8 February 2022
Available online 16 February 2022

Keywords:
Water
Physic-chemical parameters
Organic matter
Modeling
Decision tree
Bootstrap aggregates
Least-squares boosting

ABSTRACT

The purpose of the work is to investigate the use of decision tree (DT) enhanced by bootstrap aggregates (Bag) and least-squares boosting (Lsboost) in modeling the organic matter of water according to its physicochemical parameters. An entire database of 500 samples of 21 physicochemical parameters, including organic matter, was used to build the DT, DT_Bag, and DT_Lsboost models. Training data (364 data points) is resampled using a bootstrap technique to form different training datasets to train different models. The models built were validated by a dataset of 91 samples. The predicted outputs obtained from the developed DT models are then combined by simple averaging. On the other hand, the data was also boosted with the Lsboost technical aid to increase the strength of a weak learning algorithm. The model trains the first weak learner with equal weight across all data points in the training set, then trains all other weak learners based on the updated weight aimed at the validation result to minimize the squared error medium. Good agreement between the predicted and experimental organic matter concentrations for the DT_Lsboost model was obtained (the correlation coefficient for the validation dataset was 0.9992), followed by the DT-Bag model with a correlation coefficient of 0.9949. The comparison between DT, DT_Bag, and DT_Lsboost revealed the superiority of the DT_Lsboost model (the mean root of the squared errors for the dataset were 0.1295 for the DT_Lsboost, 0.1664 for the DT_Bag, and 0.5444 for the DT). These results show that Lsboost technology dramatically improved the DT model. This result is also confirmed by the results of tests on models (interpolation data of 45 points). It should also be noted that the Bag technique was also very effective in optimizing the DT model, as the results obtained with this technique were very close to the DT_Lsboost model.
© 2022 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Water treatment is essential due to the unavailability of potable water and its scarcity in developing countries in Africa (Tahraoui et al., 2021b). Algeria, like these countries, faces the problem of drinking water due to the overloading of the

* Corresponding author at: Laboratory of Biomaterials and Transport Phenomena (LBMPT), University of MÉDÉA, Nouveau Pôle Urbain, Médéa University, 26000 Médéa, Algeria.
E-mail addresses: tahraoui.hichem@univ-Medea.dz (H. Tahraoui), abdeltif.amrane@univ-rennes1.fr (A. Amrane), belhadj_1@yahoo.fr (A.-E. Belhadj), jie.zhang@newcastle.ac.uk (J. Zhang).
URLs: http://www.univ-medea.dz (H. Tahraoui), http://www.univ-Rennes1.dz (A. Amrane), http://www.ncl.ac.uk (J. Zhang).

https://doi.org/10.1016/j.eti.2022.102419
2352-1864/© 2022 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

coastline, disparities between rural and urban areas, periods of drought, and increased pollution (Teghidet, 2004). These factors are all equally destabilizing the already precarious balance of the environment. However, industrial, agricultural and urban development remains the major problem that alters water quality and makes it dangerous (Tahraoui et al., 2020). This is the case of the Médéa region, which is experiencing diversification and increasing the number of pollutants released into the aquatic environment without treatment (Tahraoui et al., 2020), especially organic matter. To protect and meet the needs of drinking water consumers for people and on an industrial scale, there is a great demand to harness water purification and make it potable. This is why treatment plants and water quality monitoring have been installed in many important water sources, intending to improve the ecosystem and human health and help to support drinking water production (Ding et al., 2014; Ho et al., 2019). However, the parameter analysis of organic matter still requires special attention (Baptista et al., 2015) because organic matter is currently a problem for drinking water treatment plants. Indeed, the main problems generated are globally the degradation of the organoleptic quality, the bacterial development in the pipes of the distribution network and consequently the aggravation of corrosion, and significant consumption of chlorine during disinfection (LeChevallier, 1990). However, monitoring the organic matter and sampling water satisfactorily over an uninterrupted period is expensive and time-consuming. This limits datasets to scattered sampling points throughout the year and restricts the habits of water resources management studies and the calibration and validation of water quality models (Libera and Sankarasubramanian, 2018). To overcome this problem in better water management, improving water quality is a vital step (Noori et al., 2020). Various modeling methods have been developed over the past decades to improve the accuracy of predictions of water quality parameters (Noori et al., 2020). The modeling of exceptional water variables in the surface water is today considered a water-saving technique that can lead to significant price savings thanks to the capacity of these indirect models, which succeed in predicting the values water variables (Rajaee et al., 2020). However, older models have the complexity of precise water quality resolution and the sophistication of water quality time series (Rajaee et al., 2020). For example, multiple linear regression (MLR) (Melesse et al., 2011; Rajaee et al., 2009), multiple nonlinear regression (NLMR) (Melesse et al., 2011), the Mann–Kendall trend test (MK) (Kisi and Ay, 2014) and the built-in auto-regression moving average models (ARIMA) (Faruk, 2010; Melesse et al., 2011).

Currently, several researchers have used machine learning to predict water quality due to their ability to solve environmental problems (Tahraoui et al., 2021a), which can solve highly nonlinear problems and time series sophistication problems of water quality and does not require any knowledge of process physics (Noori et al., 2020). Some of these studies include environmental assessment based surface water quality : A study has been initiated to tune hyper-parameters of Direct-Acting Neural Network (FFNN) and gene expression programming (GEP) with particle swarm optimization (PSO) to predict levels of the dissolved oxygen (DO) and total dissolved solids (TDS) (Shah et al., 2021b). The most influential input parameters for DO and TDS prediction were determined using principal component analysis (PCA). The results of the modeling indicated an excellent search efficiency of the PSO algorithm in the optimization of the structure and the hyper-parameters of the FFNN and GEP. PCA results revealed that magnesium, chloride, sulfate, bicarbonates, specific conductivity, and water temperature are suitable inputs for DO modeling, while; calcium, magnesium, sodium, chloride, bicarbonates and specific conductivity remained the influential parameters for TDS. The two hybrid models offered showed better accuracy in the prediction of DO and TDS; however, the hybrid PSO-GEP model achieved better accuracy than the PSO-FFNN with an R value greater than 0.85, the error root mean square (RMSE) less than 3 mg/ L and performance index value close to 1. The external validation criteria confirmed the overfitting problem solved and the generalized results of the models. Cross-validation of the model output achieves the best statistical metrics, i.e. ($R = 0.87$, RMSE $= 2.67$) and ($R = 0.895$, RMSE $= 2.21$) for PSO-FFNN and PSO-GEP models, respectively. The results demonstrated that implementing artificial intelligence models with an optimization routine can lead to models optimized for accurate prediction of water quality. Another study examined the ability of several types of Artificial Intelligence (AI) models to model four key water quality variables namely electrical conductivity (EC), sodium adsorption ratio (SAR), Total dissolved solids (TDS) and sulfate (SO_4^{2-}) using a dataset obtained from 90 wells in the plain of Tabriz, Iran; evaluated by the k-fold test (Shiri et al., 2021). Two different modeling scenarios were established to make simulations using other quality parameters and geographic information. The results obtained confirmed the capacities of AI models to model the quality variables of groundwater from wells. Among all the applied AI models, the developed machine-firefly hybrid support vector algorithm (SVM-FFA) model led to the best predictability performances for the two scenarios studied. The introduced computer aid methodology provided reliable technology for groundwater monitoring and assessment. A study was aimed to predict total dissolved solids (TDS), potential salinity (PS), sodium adsorption ratio (SAR), percent exchangeable sodium (ESP), magnesium adsorption ratio (MAR) and residual sodium carbonate (RSC) parameters by Conductivity (EC), Temperature (T) and pH as inputs (Bilali et al., 2021). To achieve this goal, adaptive amplification (Adaboost), random forest (RF), artificial neural network (ANN) and support vector regression (SVR) models using 520 samples of data related to fourteen groundwater quality parameters in the Berrechid aquifer (Morocco) were developed and evaluated. The results revealed that the overall prediction performance of the Adaboost and RF models is superior to that of SVR and ANN. However, the ability to generalize and the sensitivity to input variables show that the ANN and SVR models were more generalizable and less sensitive to input variables than Adaboost and RF. Overall, the models developed are valuable for predicting irrigation water quality parameters and could help farmers and policy makers to manage irrigation water strategies. The approaches developed in this study showed to be promising in low-cost, real-time prediction of groundwater quality through the use of physical parameters as input variables. Another study aimed to identify the relationship between water quality parameters and sources of contamination using basic radial function

(RBF) networks (Panneerselvam et al., 2021). The main objective of the study was to understand the prioritization of the ionic concentration in terms of groundwater quality. In addition, multivariate statistical analysis of groundwater for 30 sampling locations was performed. Piper and Gibbs revealed that the rock–water interaction, the ion exchange process, the weathering of host rocks were the main processes that influence the quality of groundwater in the studied area. Principal component analysis (PCA) of groundwater confirmed that two phenomena, namely geogenic and anthropogenic activities govern the chemical composition of groundwater. These major phenomena were validated by hierarchical cluster analysis (HCA) and K-mean cluster analysis. The RBF network revealed that the anion ($R^2 = 0.95$) plays a more important role in the quality of groundwater than the cations ($R^2 = 0.901$). The major sources of contamination observed in the studied area are weathering of host rocks, ion exchange process, and excess utilization of synthetic fertilizers. The leachate from landfills and agricultural fields could be the source of a higher concentration of anions in groundwater chemistry, and this leads to an increase in the pH value in groundwater. The authors suggested an integrated approach involving a corrective measure to be followed to avoid the contamination of groundwater in the study area. The results can provide useful information to the decision maker regarding the risk management of groundwater pollution mainly in arid and semi-arid regions. A methodology was also previously presented for the pre-processing of datasets and the optimization of inputs to reduce the complexity of modeling (Shah et al., 2021a). The objective of this study was achieved by employing a two-tailed detection approach for the elimination of outliers and an exhaustive search method to select essential modeling inputs. Subsequently, the Adaptive Neuro-Fuzzy Inference System (ANFIS) was applied to model electrical conductivity (EC) and total dissolved solids (TDS) in the upper Indus River. A larger dataset from a 30-year historical period, measured monthly, was used in the modeling process. The predictive capacity of the models developed was estimated by statistical evaluation indicators. In addition, the 10-fold cross-validation method was used to solve the modeling overfitting problem. The results of the input optimization indicate that Ca^{2+}, Na^{+} and Cl^{-} are the most relevant inputs to use for CE. Meanwhile, Mg^{2+}, HCO^{3-} and SO_4^{2-} were selected to model TDS levels. The optimal ANFIS models for EC and TDS data showed R values of 0.91 and 0.92 and root mean square error (RMSE) results of 30.6 μS/ cm and 16.7 ppm, respectively. The optimal ANFIS structure included a hybrid learning algorithm with 27 fuzzy rules of triangular fuzzy membership functions for EC and a Gaussian curve for TDS modeling, respectively. Obviously, the results of the study revealed that ANFIS modeling, aided by the technique of data pre-processing and input optimization, is an appropriate technique to simulate the quality of surface water. This could be an effective approach to minimize the complexity of modeling and develop appropriate management and mitigation measures.

In this work, the decision tree (DT) method is coupled by two techniques: bagging bootstrap aggregation (DT_Bag) and gradient boosted using least-squares boosting (DT_Lsboost) for the prediction of organic matter based on physical and chemical parameters and to compare them to select the most effective modeling method. Indeed, to our knowledge, a complexation between the decision tree with least-squares boosting (DT_Lsboost) and bagging bootstrap aggregation has never been used to predict organic matter content in the Médéa region in Algeria. Moreover, such a comparison has never been made before. The paper is organized as following. Section 2 presents the analysis of raw water and treated water in the Medea region to create the database; modeling of organic matter using the decision tree (DT) method coupled by two techniques: bagging bootstrap aggregation (DT_Bag) and gradient boosted using least-squares boosting (DT_Lsboost). Section 3 presents a comparison between different batteries models: a comparison between DT, DT_Bag, and DT_Lsboost. The last section concludes the paper.

2. Materials and methods

2.1. Database

The data used in this study were those related to the analysis of raw water and treated water in the Médéa region. The analyses were carried out according to the water analysis book by Jean Rodier 9th edition (Rodier et al., 2009).

The output variable is the Organic matter. On the other hand, the inputs variables (the physicochemical parameters) are listed in Table 1.

2.2. Prediction methods

Several methods have been applied to solve problems associated with the prediction and modeling of complex nonlinear systems. These strategies are particularly beneficial when such systems are challenging to model by using traditional methods (Jamin, 2010). This study investigates and contrast the use of 3 techniques for predicting organic matter doses from physicochemical water parameters. These methods are DT, DT_Bag, and DT_Lsboost.

The database was normalized once in the interval [-1, + 1] using the mapminmax function of MATLAB [31], divided into two sections for three modeling methods (DT, DT_Bag, and DT_Lsboost): 80% of the dataset for training and 20% of the remaining samples, which were not currently involved in model training, were used for model validation (Badaoui et al., 2012). The data normalization is carried out as:

$$x_N = 2\left(\frac{x - x_{\min}}{x_{\max} - x_{\min}}\right) - 1 \quad (1)$$

Table 1
The variable physicochemical parameter inputs.

Variables	Symbol	Unit
Conductivity	X_1	µS cm^{-1}
Turbidity	X^2	NTU
Potential hydrogen	X_3	–
Hardness	X_4	mg L^{-1}
Calcium	X_5	mg L^{-1}
Magnesium	X_6	mg L^{-1}
Total alkalimetric titer	X_7	°F
Bicarbonate	X_8	mg L^{-1}
Chlorides	X_9	mg L^{-1}
Nitrogen dioxide	X_{10}	mg L^{-1}
Ammonium	X_{11}	mg L^{-1}
Nitrates	X_{12}	mg L^{-1}
Phosphate	X_{13}	mg L^{-1}
Sulfate	X_{14}	mg L^{-1}
Sodium	X_{15}	mg L^{-1}
Potassium	X_{16}	mg L^{-1}
Manganese	X_{17}	mg L^{-1}
Iron	X_{18}	mg L^{-1}
Aluminum	X_{19}	mg L^{-1}
Dry residues	X_{20}	mg L^{-1}

Where: X_N is the normalized value, x_{max} and x_{min} are the maximum and minimum values respectively; and x is the actual value. The Correlation Coefficient (R), Coefficient of Determination (R^2), Adjusted Coefficient (R^2_{adj}), Root Mean Square Error (RMSE), Mean Square Error (MSE) and Mean Absolute Error (MAE) were used to estimate the performance of each model. These are calculated using the following equations (Belsley et al., 1980; Bousselma et al., 2021; Hong et al., 2007).

$$R = \frac{\sum_{i=1}^{N}\left(y_{\exp}-\bar{y}_{\exp}\right)\left(y_{pred}-\bar{y}_{pred}\right)}{\sqrt{\sum_{i=1}^{N}\left(y_{\exp}-\bar{y}_{\exp}\right)^2\sum_{i=1}^{N}\left(y_{pred}-\bar{y}_{pred}\right)^2}} \tag{2}$$

$$R^2_{adj} = 1-\frac{\left(1-R^2\right)(N-1)}{N-K-1} \tag{3}$$

$$RMSE = \sqrt{\left(\frac{1}{N}\right)\left(\sum_{i=1}^{N}\left[\left(y_{\exp}-y_{pred}\right)\right]^2\right)} \tag{4}$$

$$MSE = \left(\frac{1}{N}\right)\left(\sum_{i=1}^{N}\left(y_{\exp}-y_{pred}\right)^2\right) \tag{5}$$

$$MAE = \left(\frac{1}{N}\right)\sum_{i=1}^{N}\left|y_{\exp}-y_{pred}\right| \tag{6}$$

Where N is the number of data samples; K is the number of variables (inputs); $y_{\exp}$ and y_{pred} are the experimental and the predicted values respectively; $\bar{y}_{\exp}$ and $\bar{y}_{pred}$ are respectively the average values of the experimental and the predicted values (Dolling and Varas, 2002; Manssouri et al., 2014, 2011).

2.2.1. Decision tree

The decision tree method is a universal and well-known classification and regression algorithm, where testers can verify internal constructs or application performance. The decision tree includes an algorithm that is tree-like in nature with branches from the base of the root node on the Greed Recursive algorithm, takes advantage of the attribute with the best weight to obtain facts, verifies the report facts ratio and the Gini coefficient at each division step until the section node is taken into account (the target attribute) (Ahmadi et al., 2014; Al-Anazi and Gates, 2010; Lukoševičius and Jaeger, 2009; Yu et al., 2017a,b). Information gain (Lee and Lee, 2006), information gain ratio (Dai and Xu, 2013) and Gini coefficient (Yitzhaki, 1979) were used to calculate the weighting of chemical variables simultaneously. Facts gain measures the importance of traits in the evaluated characteristic, such as the difference between the actual requirement and the new requirement as given by Eq. (7).

$$Gain\,(A) = Info\,(D) - Info_A\,(D) \tag{7}$$

Where *Info* is the original expectation information, and $Info_A$ is the new expectation information. The information gain ratio corrected the shortcoming of information gain, which was always prone to choosing multi-valued attributes by using split information value and reduced the bias caused by the information gain. The split information value represents the potential information generated by splitting the training dataset D into v partitions, corresponding to v outcomes on attribute A

$$SplitInfo_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} - \log_2\left(\frac{|D_j|}{D}\right) \tag{8}$$

And the gain ratio is defined as:

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo(A)} \tag{9}$$

$$SplitInfo_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} - \log_2\left(\frac{|D_j|}{D}\right) \tag{10}$$

And the gain ratio is defined as:

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo(A)} \tag{11}$$

Whereas the Gini coefficient measured the impurity of the samples:

$$Gini(D) = 1 - \sum_{j=1}^{v} p_i^2 \tag{12}$$

The average Gini index is defined as:

$$Gini(D) = \sum_{j=1}^{v} \frac{|D_i|}{|D|} \times Gini(D) \tag{13}$$

Based on the results provided the attribute with the highest weight among them at each calculation step is chosen as the division attribute.

The development of the decision tree model involves the following steps:

(1) Collection of experimental data.

(2) Define the input variables and the corresponding output variables.

(3) Data pre-processing and analysis (normalization of database).

(4) Scaling and splitting of data for the learning, validation and test phases.

(5) Optimization of the decision tree parameters (Breiman et al., 1984; Loh, 2002; Loh and Shih, 1997): The selection of decision tree parameters is affected by three main factors:

- Max Number Splits: Maximal number of decision splits.
- Min Leaf Size: Minimum number of leaf node observations.
- Min Parent Size: Minimum number of branch node observations.

Where, optimization of the parameters (Min Leaf Size, Min Parent Size and Max Number Splits) were carried out in order to obtain a good result (Fig. 1).

2.2.2. Bootstrap aggregation and least-squares boosting tree

Regression tree ensembles are predictive models constructed as weighted combinations of multiple individual regression trees (Zheng, 2006).

The main objective of the overall methodology is to improve the performance of a single model by aggregating several weak/primary learners (Bauer and Kohavi, 1999; Mendes-Moreira et al., 2012). It has been proven that aggregate learning has good predictive performance on unbalanced datasets (Galar et al., 2013). The performance of ensemble methods is highly dependent on aggregation methods and the diversity and precision of essential learners. However, there is no consensus on how to build an optimal model due to flexibility.

It is often possible to increase the prediction accuracy by averaging the decisions of an ensemble of classifiers. Boosting and bagging are two techniques for such a purpose, and they work better for unstable learning algorithms such as neural networks, logistics regression, and decision trees (Zheng, 2006).

Bagging (Bootstrap Aggregation—Bag) involves fitting models on the bootstrap re-sampling replications of the original training set and then combining the models. Bootstrap re-sampling with replacement of the original training set is used to generate replications of the original training dataset. Some of the original data points can appear more than once, while others do not appear at all in a particular replication. By averaging across the developed models, bagging effectively

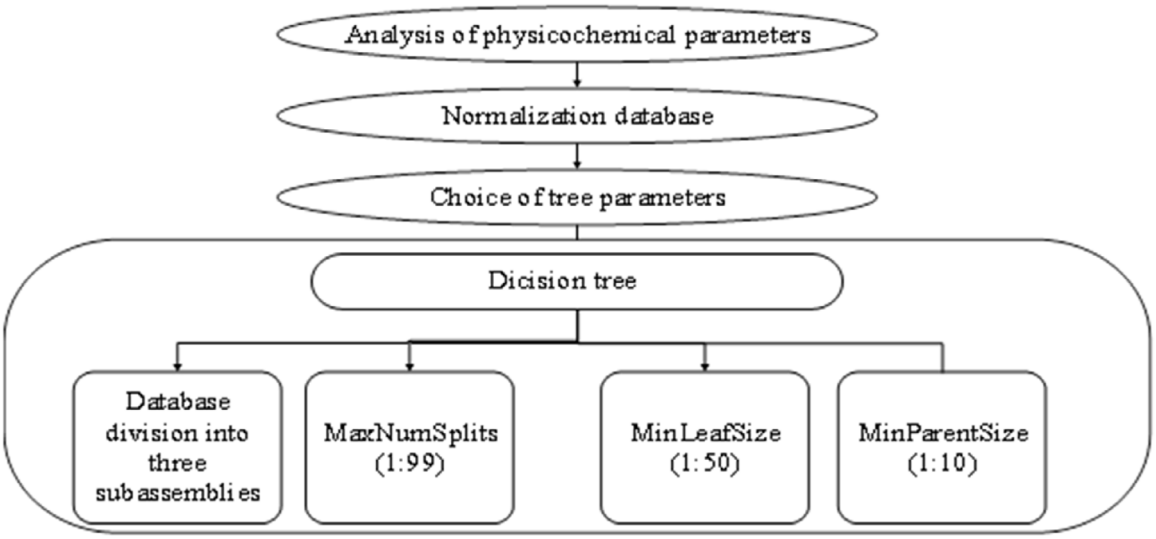


Fig. 1. Organization chart for the development and optimization of the decision tree model.

removes the instability of the decision rule. Thus, the variance of the bagged prediction model is smaller than that of fitting only one classifier to the original training set (Kilian and Inoue, 2005). Bagging also helps to avoid overfitting (Zheng, 2006).

That is to say, to create bootstrap aggregated decision tree models, the training data was resampled using bootstrap resampling with replacement (Zhang, 1999) to form n training sets. For each training set, a decision tree model is built and validated with the validation dataset. The developed decision tree models are then combined by averaging using the following equation:

$$y = \frac{\sum_{i=1}^{n} y_i}{n} \tag{14}$$

Where y_i is the calculated output of the ith decision tree model, y is the expected output of the final model, and n is the number of decision tree models.

The idea of boosting is to increase the strength of a weak learning algorithm. According to a rule of thumb, a weak learning algorithm is considered to be better than random guessing. For a binary classifier, the weak learning hypothesis is getting 50% right, boosting trains a weak learner many times, using a reweighted version of the original training set. Boosting trains the first weak learner with equal weight on all the data points in the training set then trains all other weak learners based on the updated weight. The data points wrongly classified by the preview weak learner get heavier weight, and the correctly classified data points get lighter weight (Zheng, 2006). This way, the next classifier will attempt to fix the errors made by the previous learner. There are several boosting algorithms, including Least-Squares Boosting (Lsboot) (Zheng, 2006). Lsboost is one of the regression ensembles that aims at minimizing the mean-squared error. Each step in the process fits a new learner to the difference between the target value observed and all previous-grown learners' aggregated prediction (Zhang and Xu, 2021).

Lsboost is a tree-based algorithm that uses a gradient boosting strategy to create a robust regression model. It was developed by Friedman (Friedman, 2001) to create robust gradient boosting, using square loss $L(y, F) = (y-F)^2/2$, where F is the actual training output and y is the current cumulative output $y_i = \beta_0 + \sum_{j=1}^{i=1} \beta_i h_j + \beta_i h_i = y_{i-1} + \beta_i h_i$(Ashqar et al., 2021). Another parameter was then proposed namely, $\hat{F}$, to minimize the loss using the following equation (Barutçuoğlu and Alpaydın, 2003) :

$$E = \sum_{t=1}^{N} \left[\beta_i h_i^t - \hat{F}^t \right] \tag{15}$$

where $\hat{F}$ is defined as the current residual error and the combination coefficients β_i are determined by solving $\partial E/\partial \beta_i = 0$ (Ashqar et al., 2021). The Lsboost pseudocode is presented in Algorithm 1, as presented by Friedman (Alajmi and Almeshal, 2021).

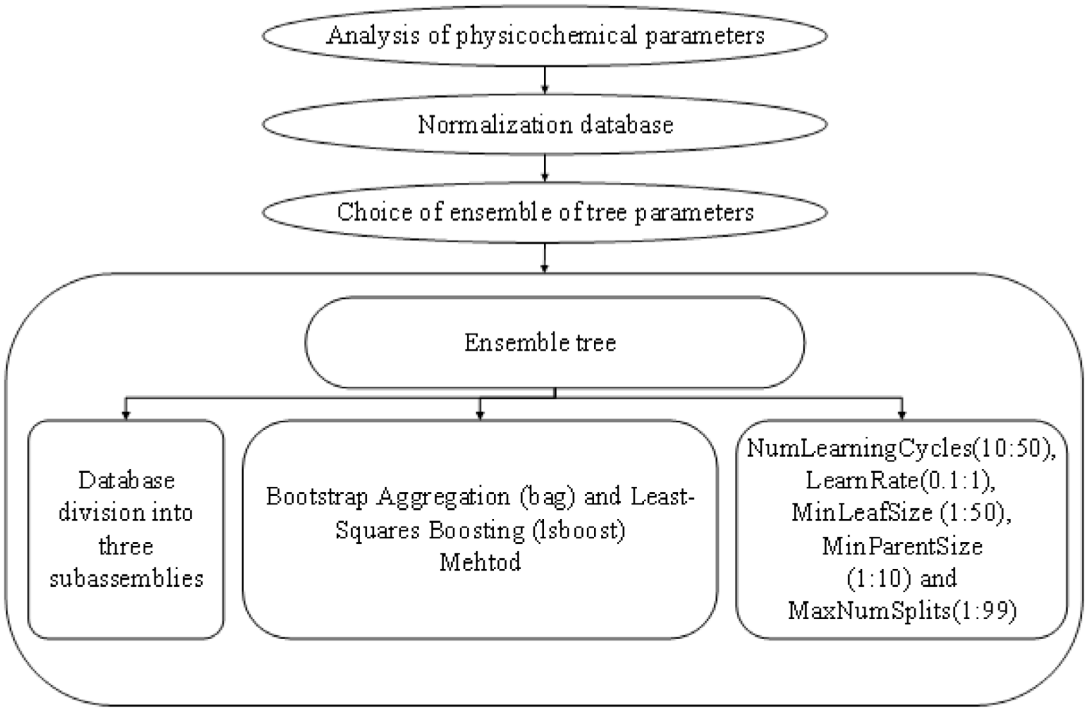


Fig. 2. Organization chart for the development and optimization of the ensemble tree model.

Algorithm 1 Lsboost Algorithm

Define x_i and y_i as explainable variables and M as the number of iterations

Define the training set $\{(x_i, y_i)\}_{i=1}^{n}$, a loss function as $L(y, F) = \frac{(y-F)^2}{2}$ and $F_m(x)$ as the regression function.

Initialization: $F_0(x) = \bar{y}$

For $m = 1$ to M do:

$\tilde{y}_i = y_i - F_{m-1}(x_i)$ *for* $i = 1, 2, \ldots, N$

$$(\rho_m, \alpha_m) = \underset{\rho,\alpha}{\operatorname{argmin}} \sum_{i=1}^{N} \left[\tilde{y}_i - \rho h(x_i; \alpha)\right]^2$$

$$F_m(x) = F_{m-1}(x) + \rho_m h(x; \alpha_m)$$

End

Where :

h corresponds to an activation function.

α is drawn from a sequence of pseudo-random U ([0; 1]) numbers.

ρ is learning rate (0 < ρ < 1).

In this study, the grid search strategy was conducted on the Bagging method based on Bootstrap Aggregation (Bag) and boosting method based on Least-Squares Boosting (Lsboot) to improve the result the previous decision tree and compare the results between them. For this, several parameters are generally considered in building an Lsboost and Bag models; for this, the parameters (Number of Learning Cycles, Learn Rate, Min Leaf Size, Min Parent Size and Max Number of Splits) have been optimized for both methods (Lsboost and Bag) to obtain a good result (Breiman, 2001, 1996; Freund and Schapire, 1997; Friedman, 2001; Hastie et al., 2009). These parameters are shown in Fig. 2:

- Number of Learning Cycles: Number of ensemble learning cycles.
- Learn Rate: Learning rate for shrinkage.
- Max Number Splits: Maximal number of decision splits.
- Min Leaf Size: Minimum number of leaf node observations.
- Min Parent Size: Minimum number of branch node observations.

Table 2
Performances of the decision tree model.

Min leaf size	Surrogate	Min parent size	Max number splits	Number of node	Coefficients of correlation			RMSE		
					APP	VAL	ALL	APP	VAL	ALL
1	ALL	2	91	183	0.9908	0.9907	0.9905	0.5381	0.5690	0.5444

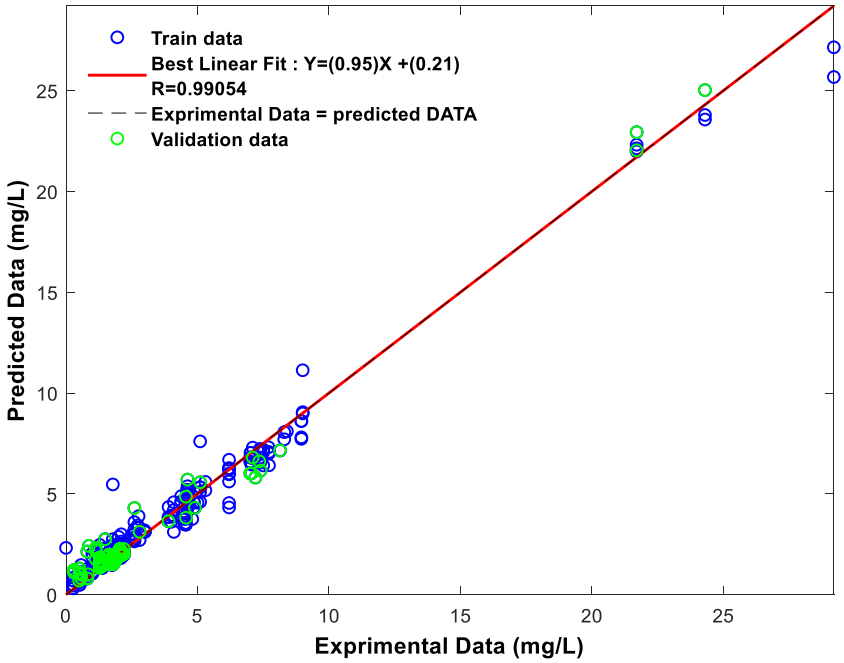


Fig. 3. Comparison between experimental and predicted values.

3. Results

3.1. Decision tree modeling

Table 2 shows the results of decision tree modeling. It shows correlation coefficients and RMSE for training and validation and parameters, MinParentsize, MinLeafSize, and MaxNumSplits. The results show efficiency in predicting the organic matter. Due to the high correlation coefficients and low RMSE obtained, $R = 0.9908$ and RMSE $= 0.5381$ in the learning phase, $R = 0.9907$ and RMSE $= 0.5690$ in the validation phase, and finally $R = 0.9905$ and RMSE $= 0.5444$ in all the data. The R values are close to 1 and the RMSE values are close to zero. It should be noted that this result was obtained under the following conditions: Min Leaf Size $= 1$, Min Parent Size $= 2$, and Max Num Splits $= 91$.

Fig. 3 represents the result of the decision tree model graphically (the predicted values vs the experimental values).

The advantage of this model is that it is possible to directly know the parameters, even their doses, and their positive or negative effects that had a significant influence on the Organic matter. It is possible to use the decision tree diagram to reduce the physicochemical analyses in the future to have the Organic matter doses by performing the analyses of parameters one by one, following the decision tree diagram and following the route according to the result of the analysis.

Fig. 4 shows the decision tree of organic matter composed of 183 nodes, each node gives a rule of division according to the physicochemical parameters (child) basing on the result of the information gain or the results of prediction of the organic matter (leafs). These rules of division results and the prediction results of Fig. 4 are described in Table 7 (Supplementary material). In addition, Table 7 (Supplementary material) includes the rules of each physico-chemical parameter or the prediction result of the organic matter in each node.

From table 7 (Supplementary material), it can be seen that Calcium (x5) (root in node 1) was identified as the most important parameter in the decision tree. In addition, all the parameters of the physicochemical analyses influenced the organic matter, with the exception of the following parameters: ammonium (x11), nitrates (x12), phosphate (x13) and aluminum (x19) which have not been described in the figure or table 7 (Supplementary material). For example: calcium (x5) in node 1, sodium (x15) in node 2, total alkalimetric titer (x7) in node 3, sulfate (x14) in node 4, manganese (x17) in node 5.... etc.

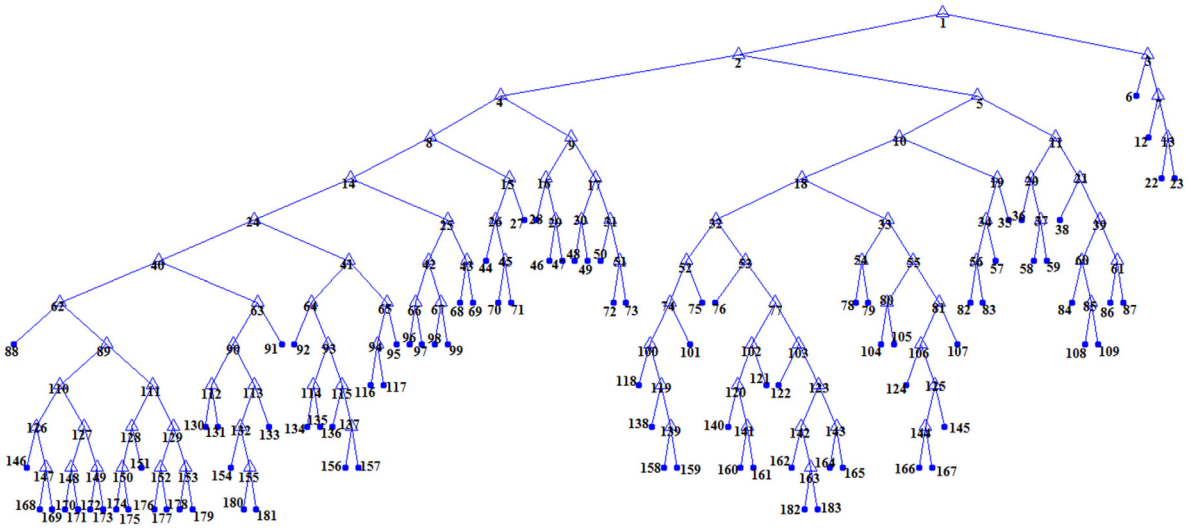


Fig. 4. The decision tree for organic matter in water.

There are very few works dealing with the study of the relationship of physicochemical parameters with organic matter, because the dynamic state of organic matter makes this study difficult because it constantly undergoes cycles of biological synthesis, metabolism and decomposition (Duursma and Dawson, 1981). The following interactions of organic matter were investigated (Fourier Transform Infrared Spectroscopy (FTIR) and Nuclear Magnetic Resonance (NMR) ... etc.), with metals (Hatira et al., 1990; Stevenson, 1977), with oxidants (Bergelin et al., 2000; Stevenson, 1977), coagulant metals (Lefebvre and Legube, 1990; Michot et al., 2005; Rezeg and Achour, 2004; Semmens and Ayers, 1985), especially aluminum sulfate.

However, the influence of the parameters of physicochemical analyses on organic matter was not confirmed; only non-linear relationships between some physicochemical parameters of water was shown, for example:

- Natural compounds in water, e.g. sulfate ions (SO_4^{2-}), are linked to essential cations: calcium, magnesium, and sodium. Apart from lead, barium, and strontium, most sulfates are soluble in water. They can nevertheless be reduced to sulfates, volatilized in the air into hydrogen sulfide (H_2S), precipitated an insoluble salt, or assimilated by living organisms (Graindorge and Landot, 2018).
- Calcium component affects water hardness, and it is generally the dominant element in drinking water. It exists as bicarbonate and in small amounts. Calcium-laden waters are complicated, and weakly laden ones are soft (Degremont, 2005). The presence of Ca^{2+} in water is mainly linked to two natural origins: either the dissolution of carbonate formations (CaCO3) or the dissolution of formations gypsum ($CaSO_4$) (Debieche, 2002).
- The presence of bicarbonates in water is due to the dissolution of carbonate formations (cipolin, limestone) by waters loaded with carbon dioxide (Debieche, 2002), according to the following reaction:

$$CaCO_3 + H_2O + CO_2 \longrightarrow 2HCO_3^- + Ca^{2+}$$

- Magnesium is one of the most common elements found in nature. Its content depends on the composition of the sedimentary rocks encountered. Thus it constitutes a significant element of water hardness (Rodier et al., 1975). From certain levels, magnesium gives water an unpleasant taste. It is found in water in the same forms as calcium, has the same drawbacks, and makes water hard (Degremont, 2005).
- The hardness, or hydrotimetric titer (TH), of water, essentially corresponds to the presence of calcium and magnesium salts. It is directly linked to the geological nature of the land crossed (Graindorge and Landot, 2018).
- In addition, the ion balance corresponds to:
  The sum of cations = the sum of anions

By considering these points, it can be concluded that almost all the physicochemical parameters have an influence on the organic matter by nonlinear relation, and this was confirmed by this study using the decision tree (intelligence artificial). Thus, the decision tree is very effective in understanding the behavior and influence of physicochemical parameters on organic matter.

On the other hand, these results also showed that artificial intelligence (the decision tree) can replace physicochemical characterization (FTIR and NMR ... etc.) to study the intercations of organic matter with other settings.

Table 3
Performances of the DT_Bag model tested.

Method	Min Leaf Size	Surrogate	Min Parent Size	Number of Learng Cycles	Max Number Splits	Number of node	Coefficients of correlation			RMSE		
							APP	VAL	ALL	APP	VAL	ALL
DT_Bag	1	ALL	2	47	99	199	1	0.9949	0.9991	0.0089	0.3717	0.1664

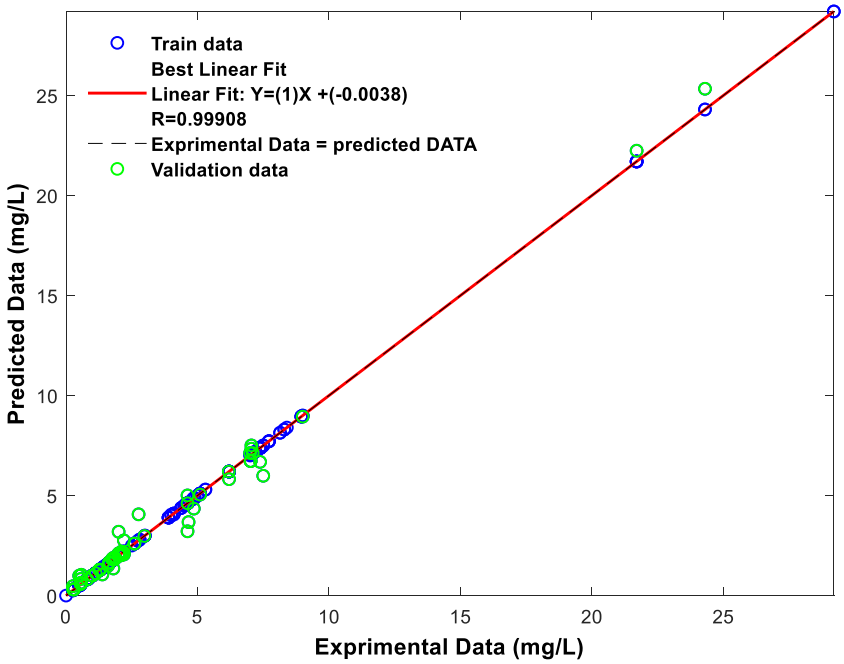


Fig. 5. Comparison between experimental and predicted values.

3.2. Bootstrap aggregated decision tree

In this part, to improve the result of the decision tree, it is coupled with the Bootstrap Aggregated technique based on the optimization of the number of data resample in the interval [10:50] given the size from our database.

Table 3 shows the results of the DT_Bag model. It displays the correlation coefficients and RMSE for training and validation and the parameters Number of Learning Cycles (the number of data resampling), Min Leaf Size, and Max Number of Splits. The model of the DT_Bag (Table 3) method showed more efficiency than the DT method in predicting organic matter. Due to that, the correlation coefficients of the DT_Bag method in all three phases were very high and also their RMSEs were very low (close to zero) compared to the DT method. Indeed, the correlation coefficients and RMSE of the DT_Bag method were ($R = 1$, RMSE $= 0.0089$) in the learning phase, ($R = 0.9949$, RMSE $= 0.3717$) in the validation phase and finally ($R = 0.9991$, RMSE $= 0.1664$) in all data. On the other hand, the correlation and RMSE coefficients of the DT method were $R = 0.9908$ and RMSE $= 0.5381$ in the learning phase, $R = 0.9907$ and RMSE $= 0.5690$ in the validation phase, and finally $R = 0.9905$ and RMSE $= 0.5444$ in all the data. These results (R and RMSE) show that the DT model has been improved by technique Bag. Knowing that the result of the DT_Bag method was obtained under the following conditions: Min Leaf Size $= 1$, Min Parent Size $= 2$, Num Learng Cycles $= 47$, and Max Num Splits $= 99$.

Fig. 5 represents the result of the DT_Bag model graphically (the predicted values according to the experimental values).

Fig. 6 represents the decision tree of organic matter by the DT_Bag method technique that is characterized by 199 nodes, the height equal to 14 (the root (node 1) up to node 198 or 199) and by 198 edges. This decision tree (Fig. 6) has also been described in table 7 (Supplementary material).

After improvement of the previous results by the DT_Bag method and by comparison of the results of table 7 (Supplementary material) for the two methods (DT and DT_Bag), we notice that calcium (x5) remained the most important parameter in DT_Bag (Table 7. Supplementary material), because it was distinguished in the 1 st node (root). In addition, the same previous physicochemical parameters of the decision tree model were found in the DT_Bag model, with the exception of aluminum (x19) and nitrate (x12), which became a significant parameter in this model (Table 7. Supplementary material). On the other hand, nitrogen dioxide (x10) became an indistinguishable parameter with

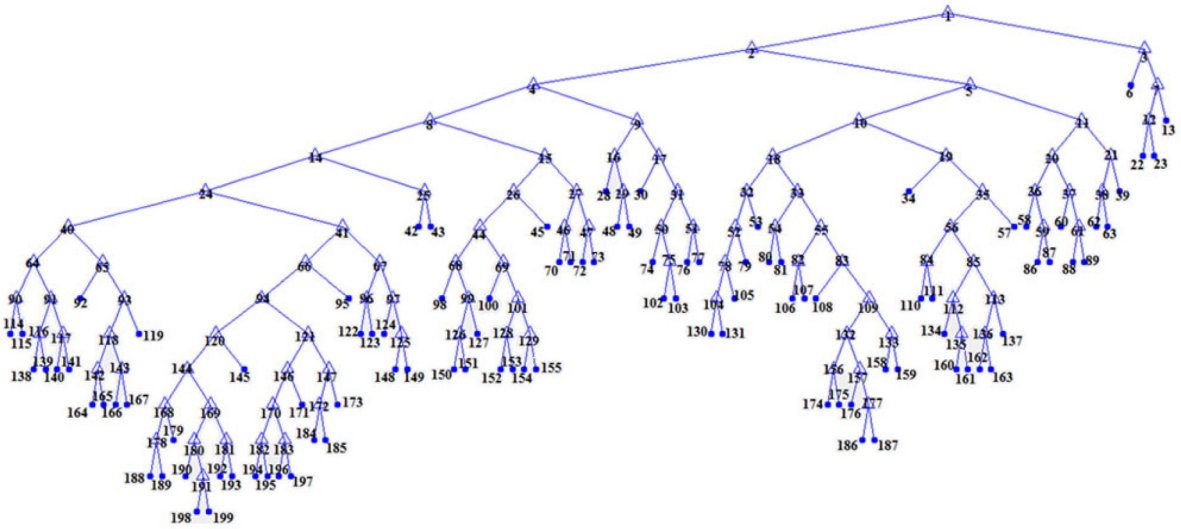


Fig. 6. The decision tree for organic matter in water by the Bag technique.

Table 4
Performances of the DT_Lsboost model tested.

Method	Min Leaf Size	Learn-Rate	Surrogate	Min Parent Size	Number of Learng Cycles	Max Number Splits	Number of node	Coefficients of correlation			RMSE		
								APP	VAL	ALL	APP	VAL	ALL
DT_Lsboost	1	1	ALL	2	46	72	145	0.9992	0.9994	0.9994	0.1209	0.1594	0.1295

ammonium (x11) and phosphate (x13) which was not described in the figure or table 7 (Supplementary material). On the other hand, Fig. 6 and table 7 again (Supplementary material) confirmed the nonlinear relationships between the physicochemical parameters of water and organic matter.

3.3. Least-squares boosting tree

In this part, to improve the decision tree result, it is coupled with Lsboot technique on optimization Learning Cycles in the interval [10:50] given their size of our database.

Table 4 shows the results of the DT_Lsboost model. It shows the correlation coefficients and RMSE for training, validation, and all data and the parameters Number of Learning Cycles, Learn Rate, Min Leaf Size, and Max Num Splits. The Lsboot method model in Table 4 was more efficient than the DT method in predicting gross organic matter in the correlation coefficient and the RMSE statistical error. Due to that, the correlation coefficients of the DT_Lsboost method in all three phases were very high and also their RMSEs were very low (close to zero) compared to the DT method. Indeed, the correlation coefficients and RMSE of the DT_Lsboost method were ($R = 1$, RMSE $= 0.00047$) in the learning phase, ($R = 0.9815$, RMSE $= 0.0835$) in the test phase, ($R = 0.984$, RMSE $= 0.0547$) in the validation phase and finally ($R = 0.9959$, RMSE $= 0.0389$) in all the data. On the other hand, the correlation and RMSE coefficients of the DT method were $R = 0.9908$ and RMSE $= 0.5381$ in the learning phase, $R = 0.9907$ and RMSE $= 0.5690$ in the validation phase, and finally $R = 0.9905$ and RMSE $= 0.5444$ in all the data. These results (R and RMSE) show that the DT model was improved by the Lsboost technique. It should be noted that the result of the DT_Lsboost method was obtained under the following conditions: Min Leaf Size $= 1$, Min Parent Size $= 2$, Num Learng Cycles=46, Learn Rate=1, and Max Num Splits $= 72$.

Fig. 7 represents the result of the DT_Lsboost model graphically (the predicted values according to the experimental values).

Fig. 8 shows the decision tree of organic matter by the DT_Lsboost method technique that is characterized by 145 nodes, the height equal to 10 (the root (node 1) up to node 144 or 145) and by 144 edges. This decision tree (Fig. 8) was also described in table 7 (Supplementary material). After having improved the previous results of the DT model by the Lsboost technique and by comparing the results of table 7 (Supplementary material) for the two methods (DT and DT_Lsboost), we notice that the same previous physicochemical parameters of the tree model decisions were found in the DT_Lsboost model, with the exception of ammonium (x11), nitrates (x12), which became a significant parameter in this model (Fig. 8). Especially the ammonium (x11) which became the most important parameter in this model (table 7.

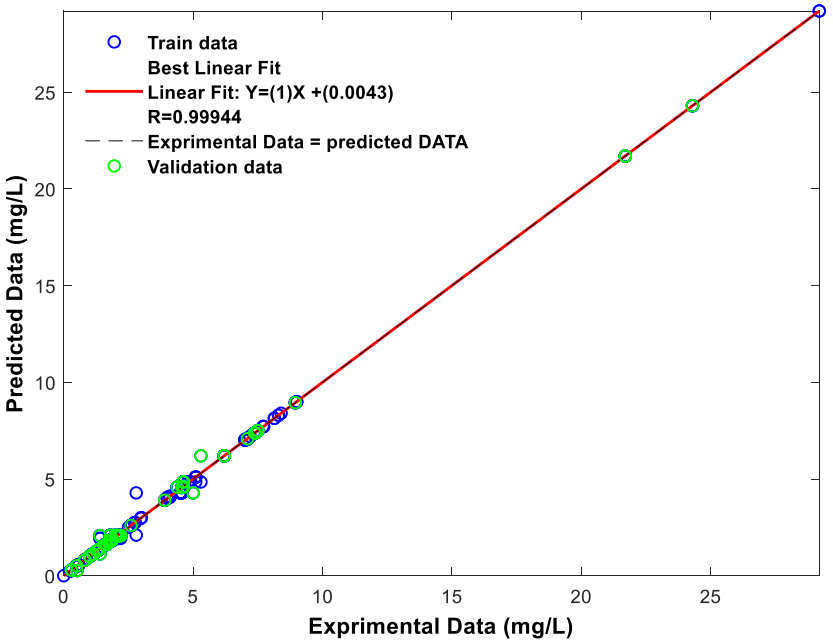


Fig. 7. Comparison between experimental and predicted values.

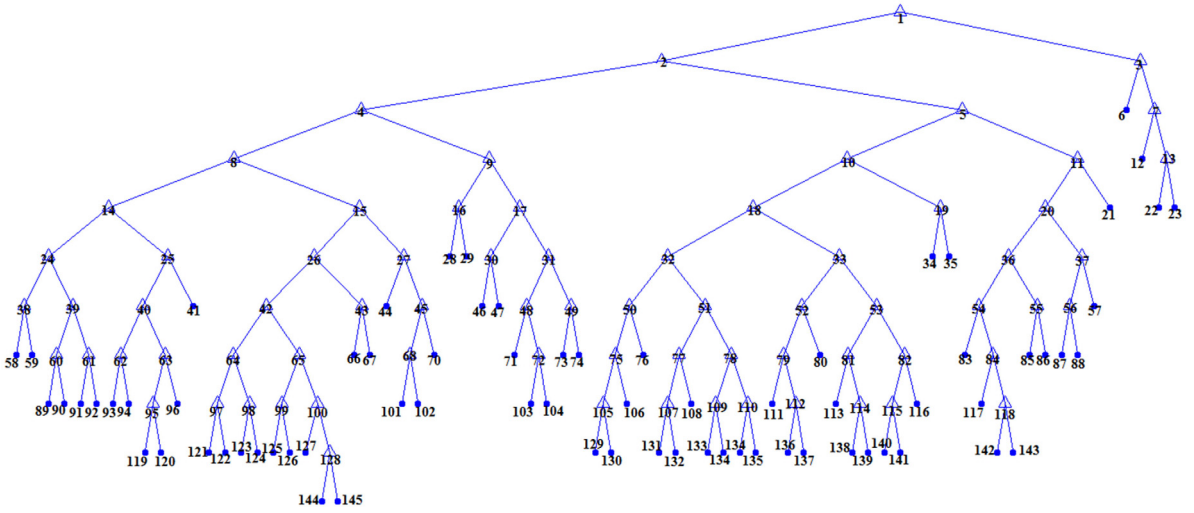


Fig. 8. The decision tree for organic matter in raw water by the Lsboot method.

Supplementary material), because it was distinguished in the 1 st node (root). On the other hand, nitrogen dioxide (x10) and sodium (x15) became insignificant parameters with aluminum (x19) and phosphate (x13), which were not described in the figure or table 7 (Supplementary material). On the other hand, Fig. 8 and table 7 (Supplementary material) again confirmed the nonlinear relationships between the physicochemical parameters of water and organic matter.

3.4. The performance evaluation of the developed models

A prediction was also made to test the accuracy of the models developed. The idea was to test experimental data of organic matter, which was not exploited during the training of our network; this database was built in 2020. The results are presented in Table 5 by correlation coefficient and RMSE and are also shown graphically in Fig. 9.

The results obtained show a high efficiency for all the used models—given their very high correlation coefficient and low RMSE value. But remains the best model DT_Lsboost the result obtained, due to the correlation coefficient $R = 0.9939$ and RMSE $= 0.2487$ which were better than the others. Following the DT_Bag model with $R = 0.9919$ and RMSE $= 0.2903$.

Table 5
Performance comparison between test prediction Models.

Model	*R*	*RMSE* (mg/L)
DT	0.9758	0.5534
DT_Bag	0.9919	0.2903
DT_Lsboost	**0.9939**	**0.2487**

Then, DT model with $R = 0.9758$ and RMSE = 0.5534. Despite the correlation coefficient and the RMSE of DT model were a little less than the other models, but still had an outstanding and acceptable result.

3.5. Comparison of models for predicting organic matter concentrations

To choose the best models in each part, all data for DT, DT_Bag, DT_Lsboost, were compared (Table 6). The results presented in Table 6 show correlation coefficients, coefficients of determination, adjustment coefficients, and statistical indicators obtained (RMSE, MSE, and MAE) for all models and testing the models.

From Table 6, it can be seen that all the models are efficient via their correlation coefficients, coefficients of determination, and very high adjustments, and shallow statistical indicators (RMSE, MSE, and MAE) for all models.

To have the best model among these models, a comparison was made for each phase in terms of R, R^2And R^2_{adj} and also in terms of statistical indicators (RMSE, MSE, and MAE):

The training phase: in this phase, it is found that the DT_Bag model was the best. Because their correlation coefficient, coefficient of determination, coefficient of adjustment were 1; followed by the DT_Lsboost model, which was almost identical to the DT_Bag model with correlation coefficient ($R = 0.9995$), coefficient determination ($R^2 = 0.999$), and adjustment coefficient ($R^2_{adj} = 0.9995$). And finally the DT model with correlation coefficient ($R = 0.9908$), determination of the coefficient ($R^2 = 0.9817$) and adjustment coefficient ($R^2_{adj} = 0.9806$)

On the other hand, in terms of statistical indicators, it is found that the DT_Bag model has very low values compared to the other models: RMSE = 0.0089, MSE = 0.0001, and MAE = 0.0025. This means that in this phase, the DT_Bag model has been well trained. But it should also be noted that the DT_Lsboost model was found to have statistical indicators slightly superior to the DT_Bag (RMSE = 0.1209, MSE = 0.0146, and MAE = 0.0407). It can therefore also be concluded that the DT_Lsboost model has been well trained.

The validation phase: in this phase, it is noticed that the DT_Lsboost model gave us impeccable results with a better correlation coefficient ($R = 0.9992$), determination coefficient ($R^2 = 0.9985$), and adjustment coefficient ($R^2_{adj} = 0.9980$). The same has been observed in terms of statistical indicators. It can be seen that the DT_Lsboost model has very low values of RMSE = 0.1594, MSE = 0.0254 and MAE = 0.0621. Then followed by DT_Bag and DT, respectively.

Note that the DT_Bag model validation results are also an extraordinary performance result with ($R = 0.3717$, $R^2 = 0.1382$, $R^2_{adj} = 0.1796$, RMSE = 0.1664, MSE = 0.0277 and MAE = 0.0379). The all data phase: in this phase, it is noticed that the DT_Lsboost model has a better correlation coefficient ($R = 0.9994$), determination coefficient ($R^2 = 0.9989$) and adjustment coefficient ($R^2_{adj} = 0.9978$). Followed by the DT_Bag model which does not present a big difference between them in terms of correlation coefficient ($R = 0.9991$), of determination coefficient ($R^2 = 0.9982$) and of adjustment coefficient ($R^2_{adj} = 0.9981$). Followed by DT. The same was recorded according to the statistical indicators and loss error. Because, the statistical indicators and loss error of model DT_Lsboost were: RMSE = 0.1295, MSE = 0.0168, MAE = 0.0450 and loss= 0.0168. And also the DT_Bag model were as follows: RMSE = 0.1664, MSE = 0.0277, MAE = 0.0379 and loss= 0.0277.

These results were also confirmed by model test data, this data showed much more efficiency of DT_Lsboost ($R = 0.9939$, $R^2 = 0.9878$, $R^2_{adj} = 0.9777$, RMSE = 0.2487, MSE = 0.0618 and MAE = 0.0746). Followed by DT_Bag which was almost equal with DT_Lsboost ($R = 0.9919$, $R^2 = 0.9838$, $R^2_{adj} = 0.9703$, RMSE = 0.2903, MSE = 0.0843 and MAE = 0.0859).

It should also be noted that the DT model remains very efficient and acceptable ($R = 0.9758$, $R^2 = 0.9522$, $R^2_{adj} = 0.9124$, RMSE = 0.5534, MSE = 0.3062 and MAE = 0.3648).

In the light of these results, it can be seen that the performance of the DT model has been dramatically improved by the Lsboost technique and by Bag.

Fig. 10 again confirms the effectiveness of the models which were coupled with Lsbbost and Bag.

3.6. Analysis of residues

To find the best model, model residues are analyzed. This method is an efficient way to detect the optimal performance of the models, and it consists in measuring the absolute or relative error of each experimental value with the predicted value [63]. Fig. 11 shows the residual values associated with the model defined by the models (DT, DT_Bag, and DT_Lsboost) as a function of the estimated values. This figure shows that the residuals obtained by the DT_Lsboost and DT_Bag methods were less dispersed (close to zero) than those obtained by the DT model in the learning phase, which

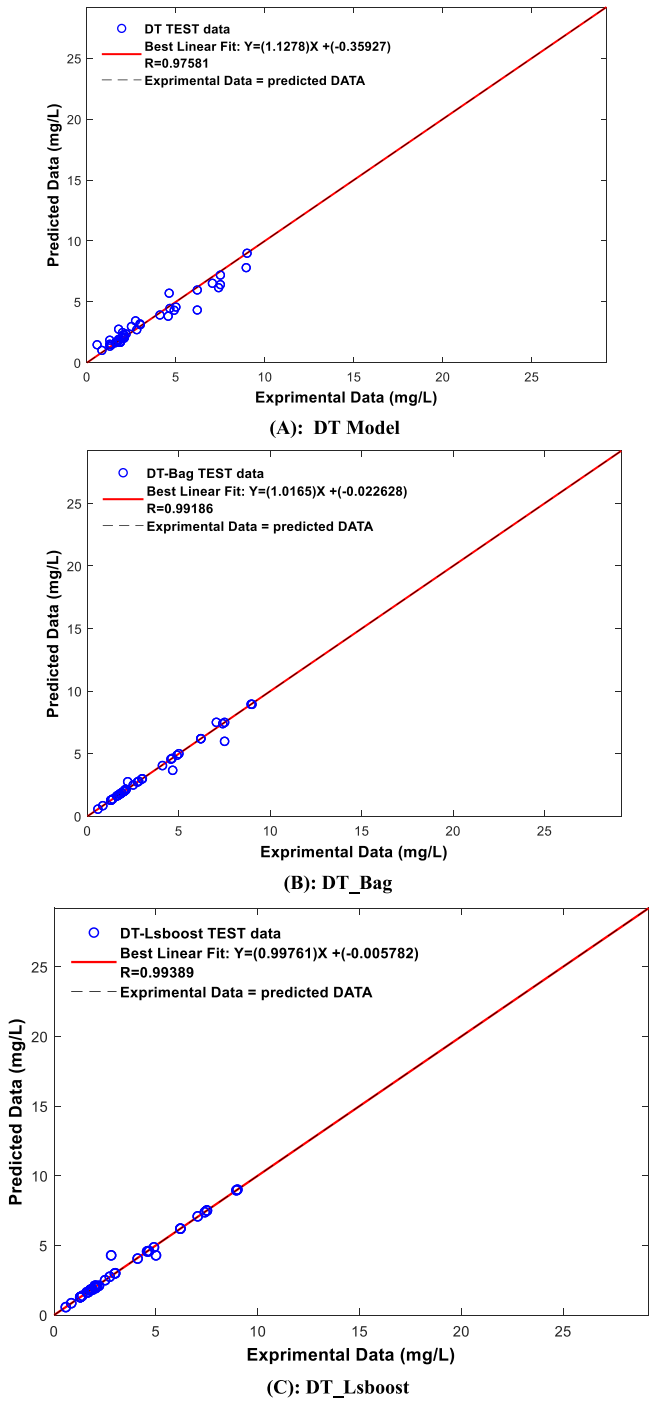


Fig. 9. Comparison between experimental and predicted for test models.

shows that the Bag and Lsboost techniques have improved the performance of the DT model. On the other hand, at the validation stage, it is clear that the DT_Lsboost model was less dispersed than the DT_Bag model.

Not to mention the number of parameters because the DT_Bag model works with 199. In contrast, the DT_Lsboost model works with 145 parameters and with very low loss value 0.0168.

In light of these results, it can be seen that Lsboost technology has dramatically improved the DT model. This result is also confirmed by the results of tests on models (interpolation). It should also be noted that the Bag technique was also

Table 6
Performance comparison between prediction Models.

Models	Train	Val	ALL	Train	Val	ALL	Model parameter		Test models	
	$R/R^2/R^2_{adj}$			RMSE/MSE/ MAE			Regression error (Loss)	Number of parameter	$R/R^2/R^2_{adj}$	RMSE/MSE/ MAE
	0.9908	0.9907	0.9905	0.5381	0.5690	0.5444			0.9758	0.5534
DT	0.9817	0.9814	0.9785	0.2895	0.3238	0.2964	0.2964	183	0.9522	0.3062
	0.9806	0.9761	0.9775	0.3092	0.3915	0.3257			0.9124	0.3648
	1.0000	0.9949	0.9991	0.0089	0.3717	0.1664			0.9919	0.2903
DT_Bag	1.0000	0.9898	0.9982	0.0001	0.1382	0.0277	0.0277	199	0.9838	0.0843
	1.0000	0.9869	0.9981	0.0025	0.1796	0.0379			0.9703	0.0859
	0.9995	**0.9992**	**0.9994**	**0.1209**	**0.1594**	**0.1295**			**0.9939**	**0.2487**
DT_Lsboost	**0.9990**	**0.9985**	**0.9989**	**0.0146**	**0.0254**	**0.0168**	**0.0168**	**145**	**0.9878**	**0.0618**
	0.9989	**0.9980**	**0.9988**	**0.0407**	**0.0621**	**0.0450**			**0.9777**	**0.0746**

very effective in optimizing the DT model, as the results obtained with this technique were very close to the DT_Lsboost model.

3.7. Comparison of models for predicting organic matter concentrations with other models

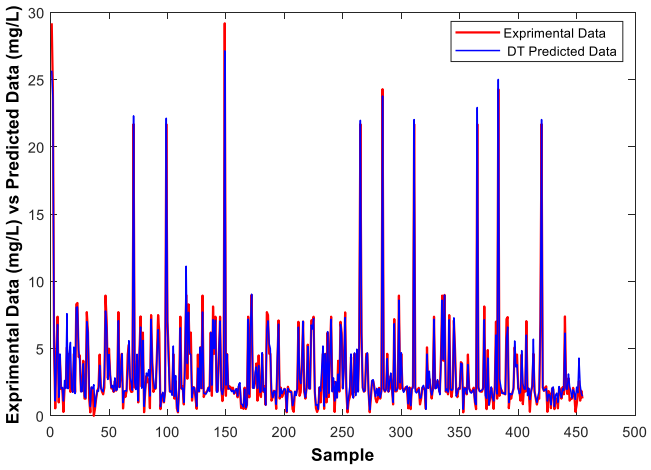
Only few studies dealt with the prediction of the organic matter of water as a function of the physicochemical parameters, for example:

- A study was carried out on the prediction of the organic carbon content in the terminal quaternary deposits of the Alboran Sea, using two modeling tools: multiple linear regression and artificial neural networks in the form of multilayer Perceptron type (MLP). The results obtained by the two models are somewhat acceptable. Because, results of neural networks demonstrated a significant learning and prediction capacity for organic carbon contents with a coefficient of determination of 0.96 to 0.99 and a very low maximum square error of 0, 00009 to 0.0003 for all four databases. For multiple linear regression, the results were less significant with a coefficient of determination of 0.86 for the entire database (El (Hmaidi et al., 2013)).
- Another study used artificial neural networks (ANNs) to estimate the daily BOD at the entrance to biochemical wastewater treatment plants. The plant-wide dataset (364 daily records for the year 2005) was obtained from a local wastewater treatment plant. Various combinations of daily water quality data including chemical oxygen demand (COD), water flow (Qw), suspended solids (SS), total nitrogen (N) and Total phosphorus (P) were used as inputs to the ANN to assess the degree of effect of each of these variables on daily intake BOD. The results of the ANN model were compared with the Multiple Linear Regression (MLR) model. The mean squared error, mean absolute relative error, and coefficient of determination statistics were used as benchmarks for evaluating model performances. The ANN technique whose inputs were COD, Qw, SS, N, and P gave root mean square errors of 708.01, mean absolute relative errors of 10.03%, and a coefficient of determination of 0.919, respectively. Based on the comparisons, it was found that the ANN model can be successfully used to estimate daily input BOD to biochemical wastewater treatment plants (Dogan et al., 2008);

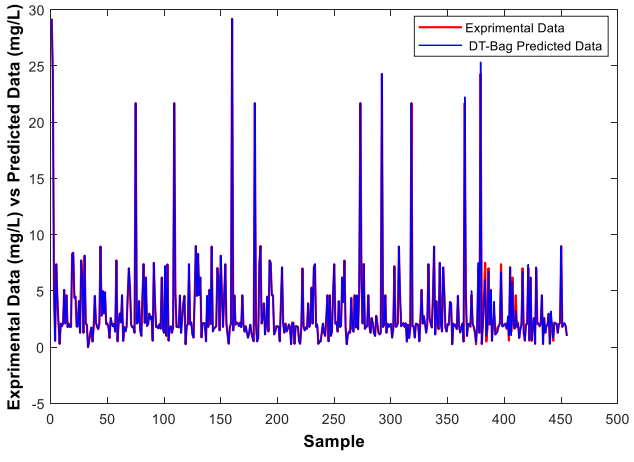
If we compare the above works to our models, namely DT, Bag_DT or Lsboost_DT, it can be concluded that our models produce very good results compared with those works either in terms of coefficient of determination or according to the statistical error, showing the efficiency of the models developed in the present study.

4. Conclusions

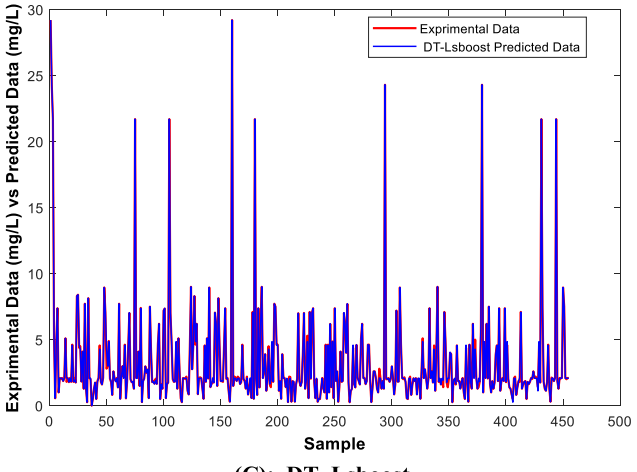
The present study aimed to model the organic matter of water by decision tree enhanced by two techniques which are bootstrap aggregates (Bag) and least-squares boosting (Lsboost). Comparison between DT_Bag and DT_Lsboost and the traditional decision tree (DT) revealed that the DT_Lsboost model gave the best performance (the mean root squared errors for the dataset was 0.1295), followed by the DT_Bag model, which gives performance not far from the DT_Lsboost model (the mean root squared errors for the dataset was 0.1664). The enhanced DT models have greater precision and can describe the concentration of organic matter more accurately with the single decision tree model (the mean root squared errors for the dataset was 0.5444). These results are also confirmed by the results of the model tests (interpolation data of 45 points) which gave the mean root squared errors of 0.2487 for DT_Lsboost, 0.2903 for DT_Bag and 0.5534 for DT. This shows that bootstrap aggregates (Bag) and least-squares boosting (Lsboost) dramatically improves the predictions with precision and robust the model of a single decision tree when applied to invisible data.



(A): DT Model



(B) : DT_Bag Model



(C): DT_Lsboost

Fig. 10. Relationship between experimental data and the predicted data of samples using DT, DT_Lsboost and DT_Bag modeling.

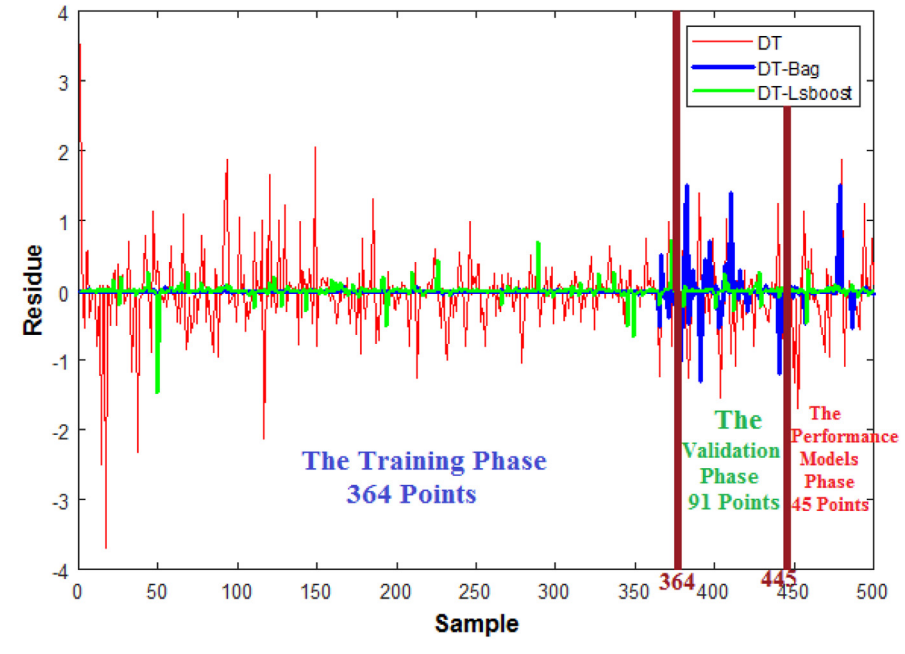


Fig. 11. Residues relating to the models established by the different models depending of estimated values.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.eti.2022.102419.

References

Ahmadi, M.A., Soleimani, R., Bahadori, A., 2014. A computational intelligence scheme for prediction equilibrium water dew point of natural gas in TEG dehydration systems. Fuel 137, 145–154.
Al-Anazi, A., Gates, I.D., 2010. A support vector machine algorithm to classify lithofacies and model permeability in heterogeneous reservoirs. Eng. Geol. 114, 267–277.
Alajmi, M.S., Almeshal, A.M., 2021. Least squares boosting ensemble and quantum-behaved particle swarm optimization for predicting the surface roughness in face milling process of aluminum material. Appl. Sci. 11 (2126), http://dx.doi.org/10.3390/app11052126.
Ashqar, H.I., Elhenawy, M., Rakha, H.A., Almannaa, M., House, L., 2021. Network and station-level bike-sharing system prediction: a an Francisco bay area case study. J. Intell. Transp. Syst. 1–11. http://dx.doi.org/10.1080/15472450.2021.1948412.
Badaoui, H.El., Abdallaoui, A., Manssouri, I., Lancelot, L., 2012. Elaboration de modèles mathématiques stochastiques pour la prédiction des teneurs en métaux lourds des eaux superficielles en utilisant les réseaux de neurones artificiels et la régression linéaire multiple. J. Hydrocarbons Mines Environ. Res. 3, 31–36.
Baptista, A.T.A., Coldebella, P.F., Cardines, P.H.F., Gomes, R.G., Vieira, M.F., Bergamasco, R., Vieira, A.M.S., 2015. Coagulation–flocculation process with ultrafiltered saline extract of *Moringa oleifera* for the treatment of surface water. Chem. Eng. J. 276, 166–173.
Barutçuoğlu, Z., Alpaydın, E., 2003. A comparison of model aggregation methods for regression. In: Artificial Neural Networks and Neural Information Processing—ICANN/ICONIP 2003. Springer, pp. 76–83.
Bauer, E., Kohavi, R., 1999. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. Mach. Learn. 36, 105–139.
Belsley, D.A., Kuh, E., Welsch, R.E., 1980. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. Wiley.
Bergelin, A., Van Hees, P.A.W., Wahlberg, O., Lundström, U.S., 2000. The acid–base properties of high and low molecular weight organic acids in soil solutions of podzolic soils. Geoderma 94, 223–235.
Bilali, A.El., Taleb, A., Brouziyne, Y., 2021. Groundwater quality forecasting using machine learning algorithms for irrigation purposes. Agricult. Water Manag. 245, 106625. http://dx.doi.org/10.1016/j.agwat.2020.106625.
Bousselma, A., Abdessemed, D., Tahraoui, H., Amrane, A., 2021. Artificial intelligence and mathematical modelling of the drying kinetics of pre-treated whole apricots. In: Kemija U Industriji: časopis Kemičara I Kemijskih inženjera Hrvatske. 70, pp. 651–667.
Breiman, L., 1996. Bagging predictors. Mach. Learn. 24, 123–140.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.
Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and Regression Trees. crc press, Boca Raton, Florida.
Dai, J., Xu, Q., 2013. Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification. Appl. Soft Comput. 13, 211–221.

Debieche, T.H., 2002. Evolution de la qualité des eaux (salinité, azote et métaux lourds) sous l'effet de la pollution saline, agricole et industrielle: application à la basse plaine de la seybouse nord-est algérien. Besançon.
Degremont, G., 2005. Mémento Technique de L'Eau, Tome 1, 10éme éd. Edit. Tec et doc..
Ding, Y.R., Cai, Y.J., Sun, P.D., Chen, B., 2014. The use of combined neural networks and genetic algorithms for prediction of river water quality. J. Appl. Res. Technol. 12, 493–499.
Dogan, E., Ates, A., Yilmaz, E.C., Eren, B., 2008. Application of artificial neural networks to estimate wastewater treatment plant inlet biochemical oxygen demand. Environ. Prog 27, 439–446. http://dx.doi.org/10.1002/ep.10295.
Dolling, O.R., Varas, E.A., 2002. Artificial neural networks for streamflow prediction. J. Hydraul. Res. 40, 547–554.
Duursma, E., Dawson, R., 1981. Marine Organic Chernistry. Elsevier Scientific. Publishing. Co, Amsterdam.
Faruk, D.Ö., 2010. A hybrid neural network and ARIMA model for water quality time series prediction. Eng. Appl. Artif. Intell. 23, 586–594.
Freund, Y., Schapire, R.E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. System Sci. 55, 119–139.
Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Statist. 118, 9–1232.
Galar, M., Fernández, A., Barrenechea, E., Bustince, H., Herrera, F., 2013. Dynamic classifier selection for one-vs-one strategy: avoiding non-competent classifiers. Pattern Recognit. 46, 3412–3424.
Graindorge, J., Landot, É., 2018. La QualitÉ de L'Eau Potable: Techniques Et Responsabilités, Territorial éditions.
Hastie, T., Tibshirani, r., Friedman, J., 2009. The Elements of Statistical Learning. Data Mining, Inference, and Prediction. Springer, 2008.
Hatira, A., Gallali, T., Rouiller, J., Guillet, B., 1990. Stabilité et solubilité des complexes formés entre le cuivre, le plomb, le zinc et les acides fulviques. Sci. Du Sol 28, 123–135.
Hmaidi, A.El., Badaoui, H.El., Abdallaoui, A., El Moumni, B., 2013. Application des réseaux de neurones artificiels de type PMC pour la prédiction des teneurs en carbone organique dans les dépôts du quaternaire terminal de la mer d'alboran. Eur. J. Sci. Res. 107, 400–413.
Ho, J.Y., Afan, H.A., El-Shafie, A.H., Koting, S.B., Mohd, N.S., Jaafar, W.Z.B., Sai, H.Lai., Malek, M.A., Ahmed, A.N., Mohtar, W.H.M.W., Elshorbagy, A., El-Shafie, A., 2019. Towards a time and cost effective approach to water quality index class prediction. J. Hydrol. 575, 148–165. http://dx.doi.org/10.1016/j.jhydrol.2019.05.016.
Hong, S.H., Lee, M.W., Lee, D.S., Park, J.M., 2007. Monitoring of sequencing batch reactor for nitrogen and phosphorus removal using neural networks. Biochem. Eng. J. 35, 365–370.
Jamin, D., 2010. Recherche du boson de higgs du modèle standard dans le canal de désintégration ZH-> nu nu bb sur le collisionneur tevatron dans l'expérience D0. développement d'une méthode d'étiquetage des jets de quark b avec des muons de basses impulsions transverses.
Kilian, L., Inoue, A., 2005. How useful is bagging in forecasting economic time series? A case study of US CPI inflation.
Kisi, O., Ay, M., 2014. Comparison of Mann–Kendall and innovative trend method for water quality parameters of the Kizilirmak river, Turkey. J. Hydrol. 513, 362–375.
LeChevallier, M.W., 1990. Coliform regrowth in drinking water: a review. J.-Am. Water Works Assoc. 82, 74–86.
Lee, C., Lee, G.G., 2006. Information gain and divergence-based feature selection for machine learning-based text categorization. Inf. Process. Manage. 42, 155–165.
Lefebvre, E., Legube, B., 1990. Coagulation par fe (III) de substances humiques extraites d'eaux de surface: Effet du pH et de la concentration en substances humiques. Water Res. 24, 591–606.
Libera, D.A., Sankarasubramanian, A., 2018. Multivariate bias corrections of mechanistic water quality model predictions. J. Hydrol. 564, 529–541. http://dx.doi.org/10.1016/j.jhydrol.2018.07.043.
Loh, W.-Y., 2002. Regression tress with unbiased variable selection and interaction detection. Statist. Sinica 36, 1–386.
Loh, W.-Y., Shih, Y.-S., 1997. Split selection methods for classification trees. Statist. Sinica 81, 5–840.
Lukoševičius, M., Jaeger, H., 2009. Reservoir computing approaches to recurrent neural network training. Comp. Sci. Rev. 3, 127–149.
Manssouri, I., Hmaidi, A.El., Manssouri, T.E., Moumni, B.El., 2014. Prediction levels of heavy metals (Zn, Cu and Mn) in current Holocene deposits of the eastern part of the mediterranean moroccan margin (Alboran sea). IOSR J. Comput. Eng. 16, 117–123.
Manssouri, I., Manssouri, M., Kihel, B.El., 2011. Fault detection by k-nn algorithm and mlp neural networks in a distillation column: comparative study. J. Inform. Intell. Knowl. 3 (201).
Melesse, A.M., Ahmad, S., McClain, M.E., Wang, X., Lim, Y.H., 2011. Suspended sediment load prediction of river systems: An artificial neural network approach. Agricult. Water Manag. 98, 855–866.
Mendes-Moreira, J., Soares, C., Jorge, A.M., Sousa, J.F.D., 2012. Ensemble approaches for regression: A survey. Acm Comput. Surv. (Csur) 45, 1–40.
Michot, L.J., Bihannic, I., Pelletier, M., Rinnert, E., Robert, J.-L., 2005. Hydration and swelling of synthetic Na-saponites: Influence of layer charge. Am. Mineral. 90, 166–172.
Noori, N., Kalin, L., Isik, S., 2020. Water quality prediction using SWAT-ANN coupled approach. J. Hydrol. 125220.
Panneerselvam, B., Muniraj, K., Pande, C., Ravichandran, N., 2021. Prediction and evaluation of groundwater characteristics using the radial basic model in semi-arid region, India. Int. J. Environ. Anal. Chem. 1–17. http://dx.doi.org/10.1080/03067319.2021.1873316.
Rajaee, T., Khani, S., Ravansalar, M., 2020. Artificial intelligence-based single and hybrid models for prediction of water quality in rivers: A review. Chemometr. Intell. Lab. Syst. 103978.
Rajaee, T., Mirbagheri, S.A., Zounemat-Kermani, M., Nourani, V., 2009. Daily suspended sediment concentration simulation using ANN and neuro-fuzzy models. Sci. Total Environ. 407, 4916–4927.
Rezeg, A., Achour, S., 2004. Incidence des groupements fonctionnels acides dans les mécanismes d'élimination de la matière organique par coagulation-floculation. séminaire international l'eau et le risque dans le contexte saharien 19.
Rodier, J., Geoffray, C., Rodi, L., 1975. L'analyse de l'eau: eaux naturelles, eaux résiduaires, eau de mer: chimie, physico-chimie. In: Bactériologie, Biologie. Dunod.
Rodier, J., Legube, B., Merlet, N., Brunet, R., 2009. L'analyse de l'eau. In: Eaux Naturelles, Eaux Résiduaires, Eau de Mer, 9e éd. Dunod.
Semmens, M.J., Ayers, K., 1985. Removal by coagulation of trace organics from Mississippi river water. J.-Am. Water Works Assoc. 77, 79–84.
Shah, M.I., Abunama, T., Javed, M.F., Bux, F., Aldrees, A., Tariq, M.A.U.R., Mosavi, A., 2021a. Modeling surface water quality using the adaptive neuro-fuzzy inference system aided by input optimization. Sustainability 13 (4576), http://dx.doi.org/10.3390/su13084576.
Shah, M.I., Javed, M.F., Alqahtani, A., Aldrees, A., 2021b. Environmental assessment based surface water quality prediction using hyper-parameter optimized machine learning models based on consistent big data. Process Saf. Environ. Protect. 151, 324–340. http://dx.doi.org/10.1016/j.psep.2021.05.026.
Shiri, N., Shiri, J., Yaseen, Z.M., Kim, S., Chung, I.-M., Nourani, V., Zounemat-Kermani, M., 2021. Development of artificial intelligence models for well groundwater quality simulation: Different modeling scenarios. PLoS One 16, e0251510. http://dx.doi.org/10.1371/journal.pone.0251510.
Stevenson, F.J., 1977. Nature of divalent transition metal complexes of humic acids as revealed by a modified potentiometric titration method. Soil Sci. 123, 10–17.
Tahraoui, H., Belhadj, A.E., Hamitouche, A.E., 2020. Prediction of the bicarbonate amount in drinking water in the region of médéa using artificial neural network modelling. Kemija U Industriji 69, 595–602. http://dx.doi.org/10.15255/KUI.2020.002.

Tahraoui, H., Belhadj, A.-E., Hamitouche, A., Bouhedda, M., Amrane, A., 2021a. Predicting the concentration of sulfate (SO42−) in drinking water using artificial neural networks: a case study: Médéa-Algeria. Desalin. Water Treat. 14.
Tahraoui, H., Belhadj, A.-E., Moula, N., Bouranene, S., Amrane, A., 2021b. Optimisation and prediction of the coagulant dose for the elimination of organic micropollutants based on turbidity. Kemija U Industriji: časopis Kemičara I Kemijskih inženjera Hrvatske 70, 675–691.
Teghidet, H., 2004. Contribution a L'étude de L'Entartrage Par Voie élèctrochimique. Béjaia. Université Abderrahmane Mira. Faculté des Sciences et des Sciences.
Yitzhaki, S., 1979. Relative deprivation and the gini coefficient. Q. J. Econ. 32, 1–324.
Yu, H., Rezaee, R., Wang, Z., Han, T., Zhang, Y., Arif, M., Johnson, L., 2017a. A new method for TOC estimation in tight shale gas reservoirs. Int. J. Coal Geol. 179, 269–277.
Yu, H., Wang, Z., Rezaee, R., Liu, X., Zhang, Y., Imokhe, O., 2017b. Fluid type identification in carbonate reservoir using advanced statistical analysis. In: SPE Oil and Gas India Conference and Exhibition. Society of Petroleum Engineers.
Zhang, J., 1999. Developing robust non-linear models through bootstrap aggregated neural networks. Neurocomputing 25, 93–113.
Zhang, Y., Xu, X., 2021. Predictions of the total crack length in solidification cracking through LSBoost. Metall. Mater. Trans. A 52, 985–1005. http://dx.doi.org/10.1007/s11661-020-06130-3.
Zheng, Z., 2006. Boosting and bagging of neural networks with applications to financial time series. In: Neural Network. The University of Chicago.