Applied Computing and Intelligence

https://www.aimspress.com/journal/aci

Applied Computing and Intelligence
5(2): 264–285.
DOI: 10.3934/aci.2025015
Received: 25 May 2025
Revised: 13 October 2025
Accepted: 29 October 2025
Published: 04 November 2025

*Research article*

# Multivariate prediction of total organic carbon in river water using random forest and deep learning regression algorithms

**Eric Kipkirui Kemei[1,*], Kristof Van Laerhoven[2], Nancy Wangeci Karuri[1] and Robert Kimutai Tewo[1]**

[1] Department of Chemical Engineering, Dedan Kimathi University of Technology, Private bag, 10143 Nyeri, Kenya
[2] Department of Ubiquitous Computing, University of Siegen, H-A 8110, Holderlin Str., Siegen 57076, Germany

* **Correspondence:** Email: eric.kemei@dkut.ac.ke; Tel +254720286135.

Academic Editor: Azlan Ismail

**Abstract:** Total organic carbon (TOC) is used to determine the total amount of organic compounds in water. It has been used to indicate water purity levels for industrial water for decades. Analyzing TOC in water is often time-consuming and an expensive activity, requiring the use of multiple high-precision sensors. Our main aim of this study was to compare the use of the Random Forest (RF) algorithm, one-dimensional convolutional neural network (1D-CNN), and multilayer perceptron (MLP) in predicting TOC using water quality parameters selected based on their strength of correlation. RF was chosen because it can model complex interactions between the various parameters and is resistant to overfitting. 1D-CNN can handle local dependencies and spatial relationships between input features whereas MLP handles independent numerical features in the dataset. The dataset was obtained from analysis done in Puget Sound marine waters around Seattle King County in the USA at the Duwamish River at three sampling locations. Learning curve analysis demonstrated that the dataset size was sufficient for stable training and generalization, while five-fold cross-validation confirmed consistent model performance across data splits. The effects of wet and dry seasons on the parameter levels were done and their impact on RF model accuracy was assessed. The selection of parameters based on Gini importance ranking was done to evaluate their effect on the accuracy of the RF model. Our results indicated that the prediction of TOC using RF regression was the most accurate with a coefficient of determination ($R^2$) of 0.732. The 1D-CNN and MLP had $R^2$ of 0.714 and 0.638, respectively. The mean absolute error (MAE) for RF, 1D-CNN, and MLP were 0.120 mg/L, 0.244 mg/L and 0.270 mg/L, respectively. It was concluded that the RF algorithm would be more feasible in predicting TOC in river water than the two deep learning methods.

**Keywords:** random forests; convolutional neural network; multilayer perceptron; regression; machine learning; total organic carbon

## 1. Introduction

Access to clean water is important for our ever-growing population in the world today. About 0.3% of the water resources is fit for consumption [1,2]. Water shortages exist in many regions, with more than one billion people lacking adequate drinking water. The annual global water demand stands at 4600 $km^3$ [3]. Rivers act as sources of drinking water and are the basic elements in sustainable development especially in industrial and agricultural activities. Activities such as the use of chemical fertilizers, animal husbandry, mining, and combustion of fossil fuels affect the quality of all natural water bodies the most [4].

Total organic carbon (TOC) is important in determining the quality of water. It is vital for ecosystem properties and water quality for human use [5]. TOC also indicates the degree of mineralization of toxic organic pollutants in the water. It is critical in determining the degree of biodegradation and purification of water. It measures the amount of carbon present in the water as particulate and dissolved organic molecules. It is the sum of two fractions: (i) The dissolved organic carbon (DOC), and (ii) the particulate organic carbon. TOC is a potential alternative to both the chemical oxygen demand (COD) and the biochemical oxygen demand ($BOD_5$) tests and is more precise than the COD test in measuring water quality [6]. Acquiring accurate and reliable TOC measurement is a labor-intensive, expensive, and time-consuming laboratory activity [7]. It requires catalytic oxidation with temperatures of up to 680ºC. To measure TOC, there is a need to invest in a total inorganic carbon (TIC) removal device to give a more accurate measurement. To produce reliable results, manufacturers of TOC analyzers occasionally have to specify TIC limits required in samples. This poses a challenge when measuring low levels of TOC when high levels of TIC are present in water because it can lead to over or under-reporting of TOC. Accurate prediction of TOC using machine learning techniques using easily measurable water quality parameters can help address this challenge.

Several researchers have explored the use of different ML techniques in predicting TOC in natural streams. The researchers in [8–10] explored the application of Artificial Neural Network (ANN), kernel extreme machine learning, and extreme machine learning models with different activation functions to estimate TOC levels in rivers. It was found that Kernel-based extreme learning machine was more accurate tool to predict TOC concentration in river water than the other methods. Neural networks mimic the human brain by using neurons to solve regression problems. Alizadeh et al. [11] studied the use of ANN to model the relation between wireline logs and TOC content in source rocks. ANN gave more precise results compared to a method that uses porosity and resistivity logs in determining TOC content in source rocks. Asgari Nezhad et al. [12] used multilayer perceptron neural network (MLP) and Sequential Gaussian Simulation (SGS) for estimation and simulation of TOC geochemical parameters. According to their research, the MLP model was more accurate than the SGS model. Back Propagation Neural Network (BPNN) machine learning algorithm has also been used to predict TOC levels. It contains an input layer, a hidden layer, and an output layer. The processing units in one layer form a connection with those in another layer, whereas the processing units in the same layer do not form a connection [13]. It is one of the simplest of all the machine learning algorithms because it is easy to train and can be established quickly [13]. However, the main drawback of BPNN

is that it easily becomes stuck in local minima during optimization, which greatly reduces its reliability when predicting TOC [13].

To overcome the limitation of BPNN of low convergence rate and instability, some investigators explored the use of support vector machines (SVM) to predict TOC. SVM is a supervised learning tool used for classification and regression tasks. It uses mathematical relations to determine the optimal hyperplane that separates data points of different classes using maximum margin technique [13]. Tan et al. [14] investigated the use of SVM regression for TOC content prediction from wireline logs in organic shale. Their study revealed that SVM technology is a powerful tool for TOC prediction and is more effective and applicable than a single empirical model, ΔlogR, and some network methods. A study carried out to determine TOC using SVM and well logs showed similar results [15]. Bolandi et al. [16] also concluded that SVM is better than ANN in predicting TOC while analyzing the organic richness of source rocks [17]. Rui et al. [18] studied TOC content prediction based on SVM with particle swarm optimization. The SVM method was more accurate in the prediction of TOC compared to ΔlogR and multilayer perceptron based on the $R^2$. SVM have been shown to obtain a better prediction results for low-dimensional data comprising a small number of samples [13]. However, when the input data is noisy and large its efficiency is reduced [19]. On the other hand, SVM does not have a feature abstraction capability like multi-hidden-layer neural networks, resulting in a weaker prediction effect on engineering problems with certain ambiguities such as well logging evaluation [13].

One dimensional convolutional neural network (1D-CNN) is a powerful regression tool because it can extract patterns from sequential data inputs and reduce noise through its convolutional layer [20]. Asante-Okyere et al. [21] developed a CNN model for TOC prediction based on mineralogy and geophysical well log data. In the study, the CNN model based on mineralogical data logs had better accuracy than those using geophysical well logs with $R^2$ of 0.86 and 0.748, respectively. This means that the choice of log input data into the CNN models affects the accuracy of the results. In another study, while conducting multiple linear regression for predicting BOD in river water, it was also noted that the choice of input variables affects the accuracy of the model during regression [22].

Random forest (RF) is a classic algorithm derived from decision tree theory and ensemble learning theory that contains a collection of decision trees. There is no relationship between each decision tree; therefore, each tree is independently modeled. Emphasis is placed on the randomness [23], and when modeling, samples are picked from the training set at random during each training iteration [23]. Because they exert the strength of ensemble learning through random means, RF methods have a high prediction accuracy and a strong generalization ability [24]. They can also handle large amounts of data in training datasets by building more decision tree models, and they do not require data normalization and are adaptable to data with unbalanced distributions [13]. Sun et al. [17] compared the use of RF, SVM and XGBoost to predict TOC Content in Organic-rich shale and found that RF had improved accuracy over SVM and XGBoost techniques.

These ML techniques have different benefits and drawbacks that depend on their applications. Their development and use in natural rivers and streams need to be assessed. In this study, we present models for predicting TOC in Duwamish River water that use RF, 1D-CNN, and MLP machine learning algorithms. These three algorithms were chosen because each have their own strengths to handle the kind of dataset available for this study. RF is robust to noise and is less sensitive to outliers in the dataset. It also has feature importance analysis that evaluates the parameters that matter most in the regression task. 1D-CNN is efficient at local pattern detection, and captures features like trends, spikes, and periodicity using filters. MLP, on other hand, is a good baseline model that is simple to implement. It is good as a benchmarking algorithm to analyze and evaluate neural networks. We

compare the accuracy and efficiency of the three algorithms in predicting TOC. The information herein provides an important foundation for the development of natural stream TOC prediction models using ML.

In this study, we aim to (i) develop predictive tools using RF, 1D-CNN, and MLP machine learning techniques for measuring TOC; (ii) assess the accuracy of RF, 1D-CNN, and MLP in predicting TOC; and (iii) assess the efficiency of RF, ID-CNN, and MLP in predicting TOC. The water quality parameter logs were taken from Duwamish River in Puget Sound along the northwestern coast of the U.S. state of Washington. There have been limited studies on the development of TOC prediction models using ML in natural streams.

## 2. Materials and methods

### *2.1. Study area and dataset description*

The dataset used for this study contained water quality parameters analyzed continuously at three sampling locations in the Duwamish River basin, including South Park Bridge, Harbor Island Marina, and East Waterway in the Puget Sound, USA. The three locations were designated in the dataset as LTMU03, LTKE03, and HNFD01, respectively. The three were chosen out of 191 sample locations for the analysis because they had the most complete datasets. The dataset contained water quality parameter logs collected and updated continuously from 21$^{st}$ October 1965 up to 20$^{th}$ March 2023. A sample preview of seven columns and the corresponding first five rows of the dataset is shown in Table 1.

**Table 1.** A preview of the dataset (source: https://catalog.data.gov/dataset/water-quality).

| Collection date and time | Depth (m) | Locator | Parameter | Value | Units | Analysis method |
|---|---|---|---|---|---|---|
| 9/21/2015 10:29 | 54.2 | JSUR01 | Temperature | 13 | deg C | CTD |
| 6/17/2015 10:43 | 1 | NSAJ02 | Silica | 0.951 | mg/L | Whitledge 1981 |
| 6/7/2016 10:13 | 178 | LSNT01 | Light Intensity (PAR) | 0 | umol/sm$^2$ | CTD |
| 8/15/2016 13:22 | 1.5 | LTUM03 | Salinity | 17.9 | PSS | SM2520-B |
| 5/18/2015 9:22 | 196 | KSBP01 | Chlorophyll, Field | 0.75 | ug/L | CTD |

From Table 1, the locator is a feature that assigns a distinctive identification of the area and particular locations where the samples are drawn from. The total number of unique locators are 191. The parameter column contains a total of 57 quality parameters that were measured at different locations, as per the standard procedures shown in the analysis methods column.

### *2.2. Dataset size suitability and model validation*

Random Forest, MLP, and 1D-CNN models were validated using learning curve analysis and k-fold cross-validation to assess their predictive stability and to determine the sufficiency of the dataset size for the regression task. Learning curves were generated by progressively increasing the training data from 10% to 100% of the dataset and obtaining the corresponding training and validation $R^2$ values. For this analysis, a 5-fold cross-validation procedure was applied. The dataset was partitioned into five equal subsets, with each fold serving once as the validation set, while the remaining folds were used for training. For each fold, the $R^2$, mean squared error (MSE), and mean absolute error (MAE) were computed, and their mean values and standard deviation reported to represent each model performance.

*2.3. Data preprocessing*

The variables in the parameter column were pivoted and filtered into their respective columns using the collect date, depth, and locator groups. They were then filled with their respective values in the 'Value' column. Where there was more than one value at a cell intersection, the mean was evaluated and filled up. The distribution of the parameter values in the 191 sample locations was analyzed to determine the locations that had the most sample collection campaigns with the least number of missing data points. The missing data points were filled using the k- nearest neighbors imputation algorithm. The common outliers among the parameters were identified and removed from the dataset using elliptic envelope method with a 95% confidence level.

The Pearson correlation coefficient, *r*, was used to evaluate the coefficient of association between the parameters with TOC in the dataset. The guidelines given by Schober and colleagues was used to determine the direction and strength of the linear relationship [25]. The parameters that depicted moderate to strong positive or negative correlation was used for the regression task.

For statistical analysis, and to better understand how the parameters were affected by the seasons, the data was split into the dry and wet season category, as per the dates illustrated in Figure 1.


**Figure 1.** Dry and wet seasons in Seattle: https://weatherspark.com/y/913/Average-Weather-in-Seattle-Washington-United-States-Year-Round.

This shows the distinction between the dry and wet seasons and chance of precipitation occurrence throughout the year at the sampling location. From Figure 1, a wet season is from mid-October to April, and November is the wettest month (whereas the dry season is from May to mid-October), while August is the driest month.

Analysis of Variance (ANOVA) was used to evaluate the level of significance of the mean values of the parameters based on the dry and the wet seasons. The permutations test was also used to test the significance levels of the parameters at the two seasons. The Pearson correlation coefficient between the parameters in the two seasons was also analyzed for comparison.

*2.4. Random forest regression*

Random forest regression algorithm was used to predict TOC using water quality parameters that had moderate to strong negative or positive Pearson correlation coefficient, *r* of greater than 0.36. The data split for the ML algorithm was kept at 80% and 20% for the training and testing set, respectively.

The sample sets were created from the random sample using the bootstrap ensemble aggregation method. The Gini impurity factor was used to rank the features and better understand variable contribution to the regression algorithm. The MAE, mean squared error (MSE), and the $R^2$ were used to determine the model accuracy.

The depth and the number of decision trees were optimized to achieve the highest prediction accuracy of the regression model. The depth and the number of decision trees were varied through permutations using python programming algorithm on the values shown in Table 2.

**Table 2.** Decision tree hyperparameter variation.

| Hyperparameters | Variation levels | | | | | | |
|---|---|---|---|---|---|---|---|
| Number of decision Trees | 50 | 100 | 150 | 200 | 250 | 300 | 350 |
| Depth of decision trees | 5 | 10 | 15 | 20 | 25 | 30 | 35 |

For the RF model, hyperparameters, including the number of estimators (number of decision trees) and maximum tree depth, were optimized using Grid Search in the sklearn learn library in python. The algorithm would first pick the number of decision trees and the depth as varied and then run the regression task. It would then evaluate the accuracy of the pair using performance metrics defined. Finally, after running all the pairs, the algorithm would pick the best performing pair for the regression task. The model parameters used for RF algorithm are those that exhibited good correlation with TOC. The parameters include depth, light transmissivity, dissolved organic carbon, date, total suspended solids, silica content, orthophosphate phosphorous content, salinity, and density.

### *2.5. 1D-CNN regression algorithm*

As a basis of precision comparison between ML algorithms, other deep learning models (1D-CNN and MLP) were used to predict TOC using the same water quality parameters used in the RF algorithm. 1D-CNN contains the input layer, the convolution layer, the pooling layer, the flatten layer, the fully connected (dense) layer, and the output layer.

Nine input water quality parameters, including DOC, sampling date, depth, density, silica, orthophosphate phosphorous, light transmissivity, date, and total suspended solids, were used in the 1D-CNN regression task. The model parameters were selected on the basis of their correlation strength with TOC. The number of convolutional layers were varied from 1 to 3 to determine the optimal number of layers required to yield higher accuracy in TOC prediction. The number of filters and the kernel size hyperparameters in the convolutional layer were also optimized. The MSE, MAE, and the $R^2$ were used as performance metrics. The number of filters were varied from a minimum value of 32 to a maximum value of 512 with a step interval of 32. The values of kernel sizes were 3, 5, and 7. The Rectified linear unit was used as the activation function in the convolutional layer to introduce the property of nonlinearity and reduce the dimensionality of the feature maps produced by the convolutional layer.

The max pooling method was used to select the maximum value from each patch of the feature map. The data was then fed into the flattened layer to convert into a one-dimensional array ready to be fed into the fully connected (dense) layer. The number of dense layers and neurons in each layer were optimized as follows: The number of dense layers were varied from 1 to 10, and the neurons in each layer varied from a minimum value of 32 to a maximum of 512 using a step interval of 32. There was only one output layer that was the predicted TOC from the parameters fed at the input. The dataset was set into a validation split of 20% placed at random. In this study, the number of epochs during the deep learning process was also optimized. This was done by setting the maximum number of epochs to 100,

using the early stopping criterion, and setting the patience to 5. Minimum validation loss at every epoch was evaluated and iterated until the minimum values were achieved. The optimal number of passes through the training dataset that leads to the best performance of the model on unseen data was evaluated and optimized using the validation dataset.

### *2.6. The multilayer perceptron regression algorithm*

Multilayer perceptron regression was used to train the dataset to obtain a desired output in the regression operation. It contains an input layer, hidden layers, and the output layer. In this study, the input into the model was the water quality parameters similar to those used in the RF and the 1D-CNN algorithms, and the output was the predicted TOC. The nine neurons from the input layer were fed into the hidden layer where the activation function (ReLU) was applied. The number of hidden layers and the neurons in each layer were optimized using the MAE as the performance metric for the model. The number of training epochs were also optimized using the early-stopping criterion while setting the patience to 5.

## 3. Results and discussion

### *3.1. Dataset preprocessing*

The dataset was analyzed for completeness for the three sampling locations, and the results are shown in Table 3. The parameters with less than ten datapoints of the possible 1187 were excluded from the regression task while those with more than 74% data completeness and with moderate to strong negative or positive correlation coefficient, $r$ of greater than 0.36 with TOC were used as input regression parameters.

**Table 3.** Dataset summary.

| Parameter | Number of data points | Percentage missing datapoints |
|---|---|---|
| Depth | 1187 | 0 |
| Total Organic Carbon | 1187 | 0 |
| Density | 1187 | 0 |
| Dissolved Organic | 1185 | 0.16 |
| Light Transmittance | 1143 | 3.70 |
| Orthophosphate Phosphorus | 871 | 26.62 |
| Silica | 1181 | 0.50 |
| TSS | 1177 | 0.84 |
| Salinity | 1187 | 0 |
| Date and Time | 1187 | 0 |

Orthophosphate phosphorus had the highest number of missing data points at 26.62%. The rest of the parameters had less than 4% out of the possible 1187 data missing. The missing data points were imputed using k-nearest neighbors imputation algorithm from the sklearn impute library in python by setting the value of $k$ as 5. Common outliers among the parameters were identified using elliptic envelop method with a 95% confidence level. Forty-two outliers were identified and removed from the dataset, resulting in 1145 complete sets of data.

*3.2. Dataset size suitability and model validation*

To analyze the suitablity of the dataset size for regression task using the three models, learning curves and k-fold cross validation analysis were conducted. Figure 2 shows the learning curve for RF, 1D-CNN, and MLP.

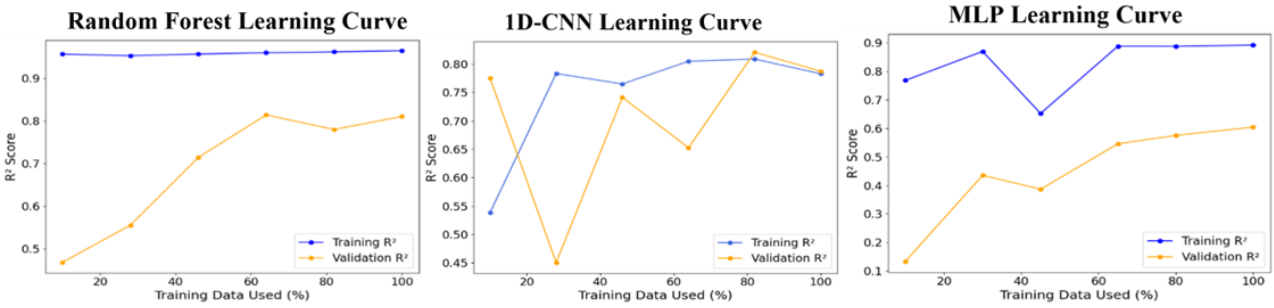


**Figure 2.** Learning curves for Random Forest, 1D-CNN, and MLP models showing stable convergence of training and validation.

From Figure 2, the learning curves for RF, 1D-CNN, and MLP models demonstrate that the dataset size used was large enough to enable stable model training and meaningful generalization. For RF curves validation $R^2$ values rises steadily converging near 0.81 at full training data usage while maintaining a high training $R^2$ of about 0.96. This implies that RF has a strong predictive ability with marginal overfitting. The 1D-CNN demonstrated the most balanced performance, with training and validation $R^2$ values ranging between 0.78 and 0.82 at higher data fractions, indicating that the model effectively captured nonlinear dependencies without significant variance errors. The MLP curve showed validation $R^2$ gradually approached the training $R^2$ as the training fraction increased, stabilizing at 0.6 with a training $R^2$ of 0.9, implying adequate learning capacity. Collectively, the consistent convergence of validation scores across all models as the training fraction increased suggested that the dataset size of 1145 data points was adequate for model generalization, with only minor gains expected from additional data. This agreed with the works of the researchers in [26] who suggested that for machine learning models, the minimum dataset size should be >1000 for better accuracy. Among the three models, RF and 1D-CNN displayed superior generalization compared to the MLP, underscoring their suitability for capturing complex relationships within the dataset. Table 4 shows the mean values of the performance metrics of the three models using 5-fold cross validation analysis to assess the model's stability.

**Table 4.** Mean performance metrics of 5-fold cross validation analysis for RF, 1D-CNN, and MLP.

| Performance Metric | RF | 1D-CNN | MLP |
|---|---|---|---|
| Test $R^2$ | 0.771 ± 0.031 | 0.736 ± 0.036 | 0.757 ± 0.039 |
| MSE | 0.096 ± 0.020 | 0.109 ± 0.022 | 0.101 ± 0.022 |
| MAE | 0.200 ± 0.017 | 0.226 ± 0.019 | 0.216 ± 0.018 |

The 5-fold cross-validation analysis provided a robust measure of model stability. The RF model achieved the highest predictive performance overall due to its ability to handle nonlinear relationships and feature interactions efficiently without requiring extensive parameter tuning. From the test $R^2$ values, the MLP model slightly outperformed the 1D-CNN, which could be attributed to the relatively small and tabular nature of the dataset, where fully connected networks tended to generalize more effectively than convolutional architectures that were optimized for sequential or spatially correlated

data. The low standard deviations (< 0.04) across folds indicated stable performance and minimal sensitivity to data partitioning, thus indicating the adequacy and representativeness of the dataset. Collectively, the convergence of the learning curves and the consistency of cross-validation results suggested that the dataset size used was robust for model training and was reliable with limited benefits expected from additional data and could be used effectively for predictions of total organic carbon concentrations in river water.

### *3.3. Parametric correlation*

The strength of the relationships between TOC and other water parameters were determined. The scatter plots showing parameters with positive and negative correlation with TOC are presented in Figures 3 and 4, respectively, and the pair-wise correlation matrix is shown in Table 5. The scatter plot of TOC with DOC shows that the data points were closely clustered around the regression line whose trend indicated a positive linear relationship. The correlation matrix showed that the value of *r* for the TOC- DOC pair was 0.852, indicating a strong positive linear relationship. The strong correlation was expected because DOC is the soluble portion of TOC. The other parameters that had a positive linear correlation with TOC were silica and total suspended solids, and those exhibiting negative linear correlation were orthophosphate phosphorus, density, light transmissivity, salinity, and depth. The relationships between TOC and these water parameters were linear but exhibited a significant amount of scatter. The directionality and strength of correlation was supported by the Pearson correlation coefficients in Table 5. Based on the guidelines of Schober et al., [25] all the water parameters were moderately to strongly correlated with TOC ($|r|>0.40$). The exceptions were depth and total suspended solids, having $0.36<|r|<0.40$, which was within the window of a linear relationship [27]. Parametric correlations suggested that a linear model may have been used to fit the data to TOC.

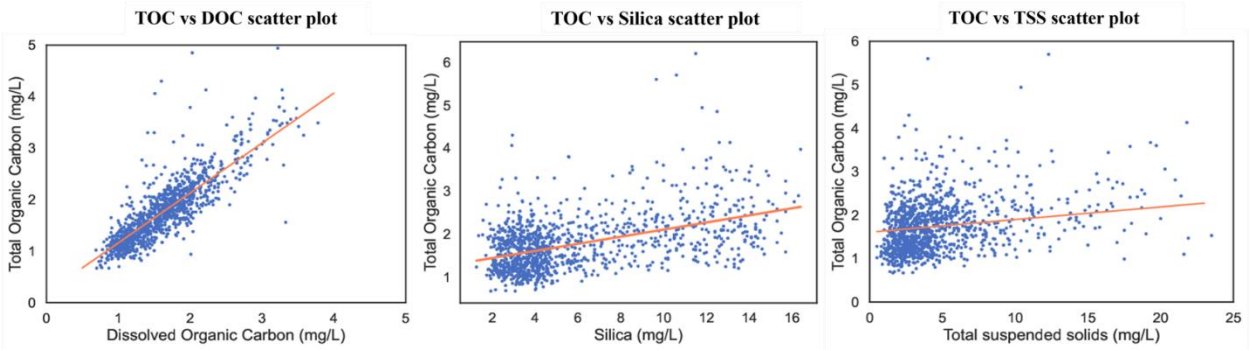


**Figure 3.** Scatter plots of TOC versus water quality parameters with positive linear correlation: DOC, silica, and Total suspended solids.

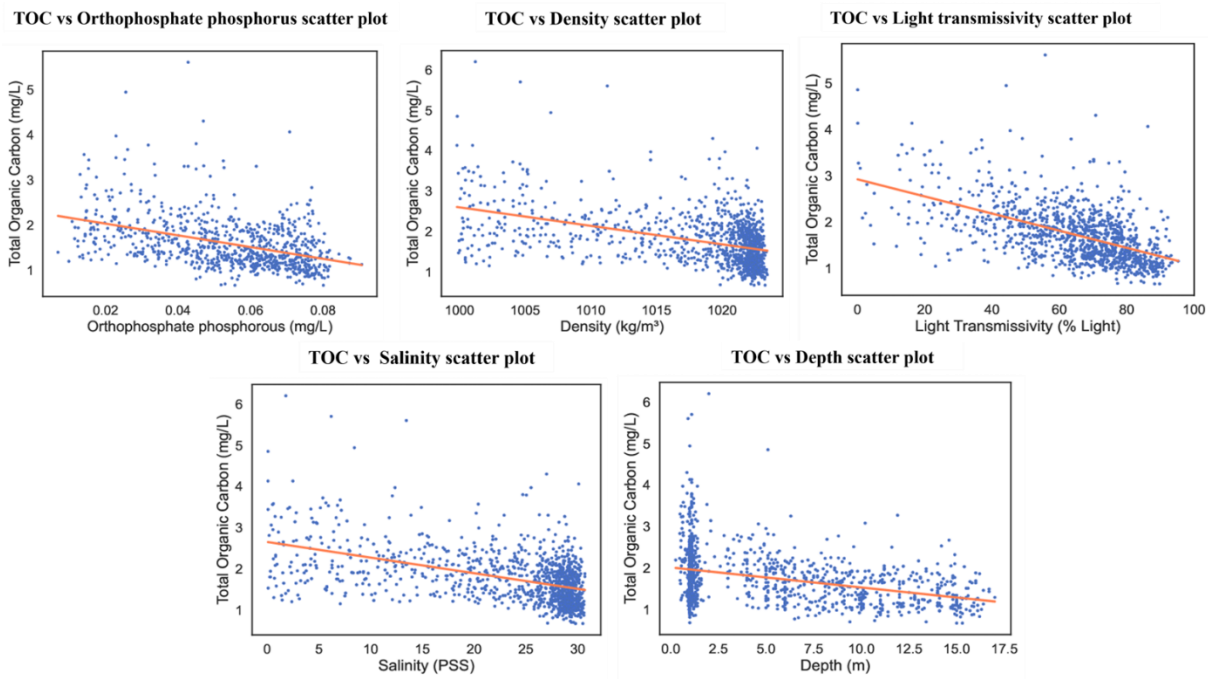


**Figure 4.** Scatter plots of TOC versus water quality parameters with a negative linear correlation: Orthophosphate phosphorus, density, light transmissivity, salinity, and depth.

**Table 5.** Pearson correlation matrix.

| Parameter | Depth | Density | DOC | Light Transmissivity | Orthophosphate Phosphorus | Silica | TOC | TSS | Salinity | Date |
|---|---|---|---|---|---|---|---|---|---|---|
| Depth | 1.000 | | | | | | | | | |
| Density | 0.505 | 1.000 | | | | | | | | |
| DOC | -0.311 | -0.429 | 1.000 | | | | | | | |
| Light Transmissivity | 0.444 | 0.558 | -0.483 | 1.000 | | | | | | |
| Orthophosphate Phosphorus | 0.499 | 0.799 | -0.351 | 0.550 | 1.000 | | | | | |
| Silica | -0.534 | -0.850 | 0.455 | -0.574 | -0.736 | 1.000 | | | | |
| TOC | -0.370 | -0.456 | 0.852 | -0.553 | -0.423 | 0.472 | 1.000 | | | |
| TSS | -0.114 | -0.296 | 0.260 | -0.554 | -0.257 | 0.242 | 0.361 | 1.000 | | |
| Salinity | 0.534 | 0.977 | -0.465 | 0.601 | 0.814 | -0.917 | -0.492 | -0.306 | 1.000 | |
| Date | 0.066 | -0.006 | -0.483 | 0.278 | -0.065 | -0.094 | -0.483 | -0.106 | 0.045 | 1.000 |

Time may influence water quality parameters through seasonal changes in precipitation and temperature. The influence of time on water quality parameters was examined graphically through time series analyses. The various dates and time were converted into equivalent time stamps to enable the scatter plots to be plotted over the time scale. Figure 5 shows time series plots for orthophosphate, silica, density, and salinity taken from the year 2005 to the year 2022 in the river. The four parameters showed the strongest variation with time (Supporting Figure 1). In Figure 5, orthophosphate phosphorus had a clear periodic pattern with consistent peaks and troughs. This meant that the levels of orthophosphate phosphorus were influenced by seasonal changes. Silica levels had similar patterns to those of orthophosphate phosphorus but comparable periodicity. Density exhibited sinusoidal trend in the upper range values similar to that of salinity. This was in agreement with the strong correlation between density and salinity in Table 5 and the work of others [28]. Seasonal or periodic changes of

orthophosphate phosphorus, silica, depth, and salinity with time did not seem to significantly influence TOC. The target variable (TOC) did not portray a periodic trend, whereas the input variables exhibited periodic and non-periodic trends.

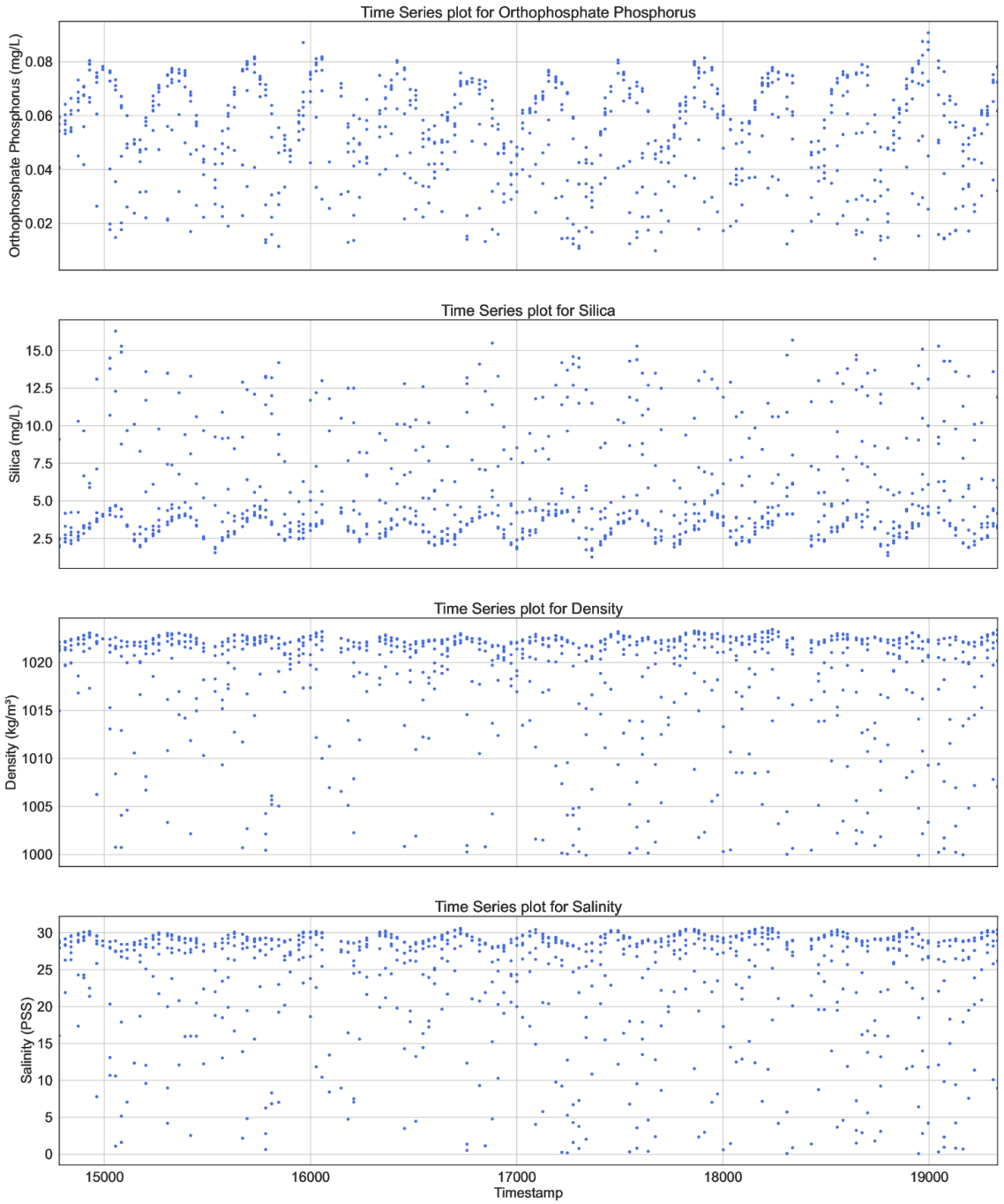


**Figure 5.** Time series scatter plots for parameters with periodic pattern: Orthophosphate phosphorous, silica, density, and salinity.

Figure 6 illustrate the parameters that had no regular periodic oscillations. Light transmissivity exhibited an irregular trend with no cyclic pattern. This meant that seasonal changes through the years do not affect light transmissivity in the river water, and the changes in the levels of light transmissivity are influenced by factors other than seasonal variations. This is in agreement with findings by Nicolaus et al. [29].


**Figure 6.** Time series scatter plots for parameters without nonperiodic pattern: Light transmissivity, dissolved organic carbon, total organic carbon, and total suspended solids.

### *3.4. Statistical comparisons of the dry and the wet season*

Since the data set had parameters that showed periodicity in their values, a comparison was carried out for parameter values for the dry and wet seasons. A split of the dataset based on the wet and dry seasons resulted in datasets containing a total of 666 and 479 datapoints, respectively. A comparison of the Pearson correlation coefficient for the two seasons is shown in Figure 7. The Pearson correlation coefficient for the dry and wet season between water quality parameters and TOC were higher in the wet season than those in the dry season. The only exception was TSS. Therefore, the levels of

precipitation impacted the correlation of TOC with the water quality parameters. This agreed with the findings of Correll et al. [30].

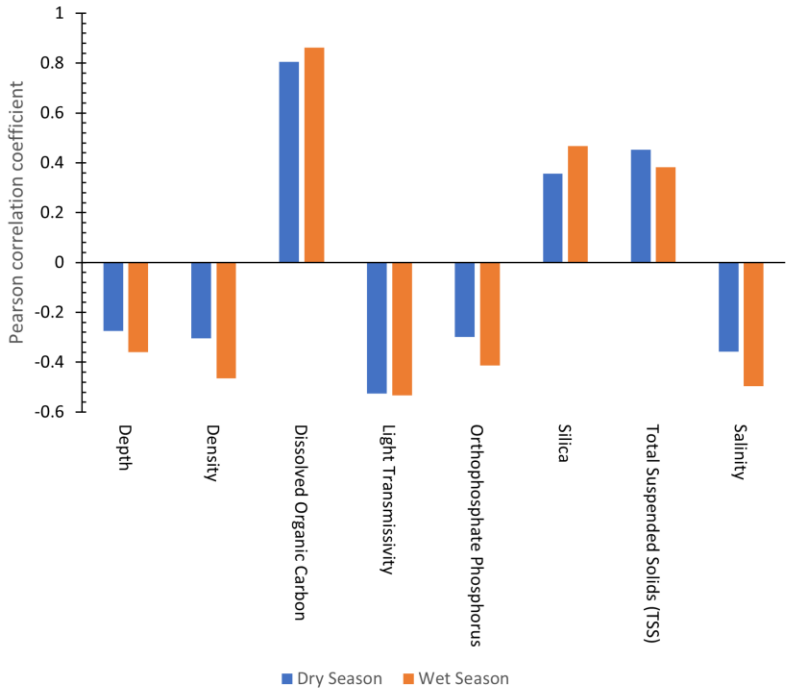


**Figure 7.** Pearson correlation coefficients between water quality parameters and TOC for the wet and dry seasons.

Analysis of variance (ANOVA) on the water quality parameters from the three sampling sites on Duwamish River for the dry and wet seasons was carried out. This is summarized in Table 6. The *p*-value indicated the probability of observing a difference in the means of the data during the wet and dry seasons, respectively, assuming that there was no true difference between the two seasons. At a 95% confidence level, density, DOC, orthophosphate phosphorus, silica, and salinity had significant differences in their mean values for the two seasons (dry and wet). This agreed with the findings of the time-series analysis that density, orthophosphate phosphorus, silica, and salinity showed periodic variations in their value. It was possible that seasonal variations in DOC variables were masked by their scatter. Light transmissivity, TOC, and TSS did not have a significant difference between the mean values of the two seasons. RF, 1D-CNN, and MLP machine learning algorithms were used for regression analysis of the relationship between TOC and other water quality parameters to evaluate their performance in predicting TOC values for the mixture of periodic and non-periodic data.

**Table 6.** Variance, mean, and *p*-values for the dry and wet seasons.

| Parameter | Units | Mean | | Variance | | p-Value |
|---|---|---|---|---|---|---|
| | | Dry | Wet | Dry | Wet | |
| 1. Density | $Kg/m^3$ | 1018.571 | 1017.721 | 30.488 | 51.771 | 0.028 |
| 2. Dissolved Organic Carbon | mg/L | 1.529 | 1.681 | 0.274 | 0.422 | 0.013 |
| 3. Light Transmissivity | % Light | 63.250 | 63.797 | 279.518 | 447.535 | 0.640 |
| 4. Orthophosphate Phosphorus | mg/L | 0.046 | 0.057 | 0.000 | 0.000 | 3.67E-16 |
| 5. Silica | mg/L | 5.032 | 6.756 | 11.680 | 14.969 | 6.33E-15 |
| 6. Total Organic Carbon | mg/L | 1.798 | 1.808 | 0.366 | 0.645 | 0.817 |
| 7. Total Suspended Solids | mg/L | 5.984 | 6.347 | 72.069 | 173.072 | 0.591 |
| 8. Salinity | PSS | 24.141 | 22.350 | 51.828 | 87.435 | 3.79E-04 |

### *3.5. RF regression*

RF regression was used to develop a predictive model of TOC from the other water quality parameters. RF hyperparameter values shown in Table 2 were optimized using a python programming machine learning algorithm, and the results are shown in Table 7. The optimal depth and number of decision trees were 15 and 150 for the dry season, respectively, and 35 and 50 for the wet season, respectively. When the two seasons were combined, the optimal depth and number of decision trees were 20 and 150, respectively. The $R^2$ was highest during the wet season at 0.788 and lowest during the dry season at 0.685. Higher accuracy during the wet season could be attributed to the higher Pearson correlation coefficient between TOC and water parameters for seven of nine parameters during the wet season compared to those of the dry season (Figure 7). When the data was taken for the sampling locations, the coefficient of determination, $R^2$, became 0.732. The scatter plot for observed versus predicted values of the RF regression model is shown in Figure 8.

**Table 7.** Random Forest regression results for dry season, wet season, and both seasons combined.

| Parameter | Dry season | Wet season | Seasons combined |
|---|---|---|---|
| Optimal depth of decision tree | 15 | 35 | 20 |
| Optimal number of decision trees | 150 | 50 | 150 |
| Mean squared Error | 0.082 | 0.083 | 0.111 |
| Root Mean Squared Error | 0.286 | 0.287 | 0.334 |
| Coefficient of Determination ($R^2$) | 0.685 | 0.788 | 0.732 |

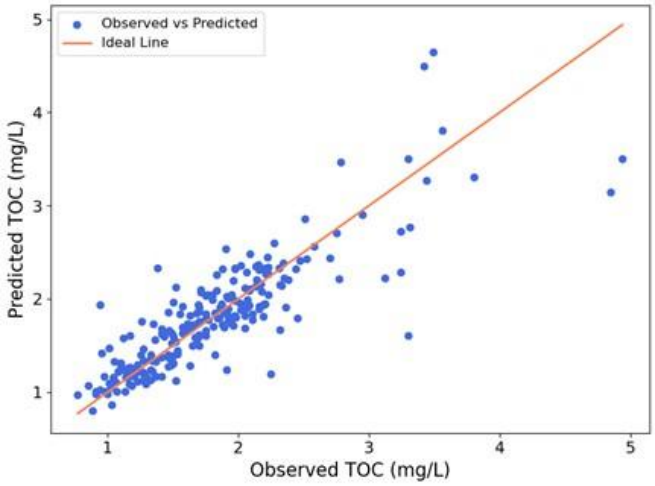


**Figure 8.** Scatter plots for predicted versus observed values for the RF TOC regression model: Considering DOC, light transmissivity, date, depth, silica, orthophosphate phosphorus, salinity, and density as input parameters.

Sampling date, density, light transmissivity, and total suspended solids were considerably cost effective and easy measurements to take. When these water quality parameters were used for the prediction of TOC, the coefficient of determination, $R^2$, changed from 0.732 to 0.600. The scatter plot is presented in Figure 9. Moreover, the more input variables there were, the better the model fit. However, in a setting where not all the water parameters were available, the model could be used to provide a reasonable estimate of TOC values.

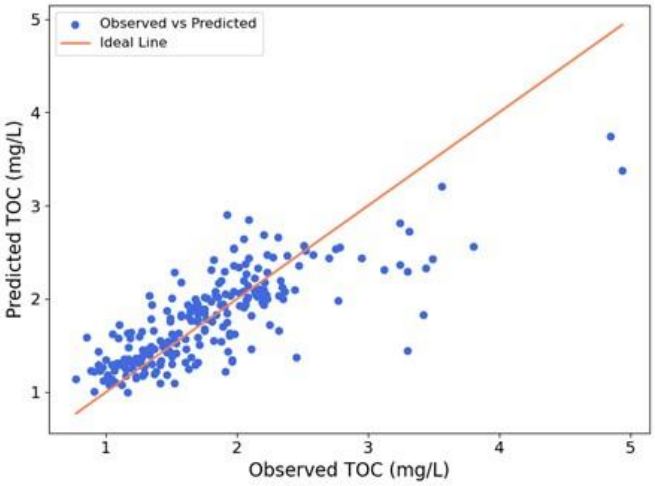


**Figure 9.** Scatter plots for predicted versus observed values for the RF TOC regression model considering date, density, light transmissivity, and total suspended solids as input parameters.

Gini importance was used to measure the potential of the variables in their contribution to the RF regression model. Table 8 shows the variables together with the sampling locations and their respective Gini importance. As expected, DOC accounted for the highest contributor to the prediction of TOC because it had the highest Gini factor of 0.779 in the RF regression algorithm. This agreed with scatter plots of TOC and DOC as well as the pair-wise Pearson correlation coefficient for the two water quality parameters. Interestingly, the three sample locations in the river had the least contribution to the prediction of TOC at < 0.2%. This meant that sampling locations in Duwamish River had the least effect on TOC prediction using the RF regression model. In fact, when the sample locations were excluded from the RF algorithm, the coefficient of determination ($R^2$) increased from 0.710 to 0.732.

**Table 8.** Gini importance factors.

| Feature | Gini Importance | | |
|---|---|---|---|
| | All Features | Excluding sample locations | Selected features |
| Dissolved Organic Carbon | 0.779 | 0.774 | - |
| Light Transmissivity | 0.038 | 0.038 | 0.133 |
| Date | 0.037 | 0.040 | 0.465 |
| Total suspended solids | 0.029 | 0.031 | 0.094 |
| Depth | 0.027 | 0.028 | - |
| Silica | 0.025 | 0.027 | - |
| Orthophosphate Phosphorous | 0.025 | 0.026 | - |
| Salinity | 0.018 | 0.019 | - |
| Density | 0.016 | 0.016 | 0.309 |
| Location LTKE03 | 0.002 | - | - |
| Location LTUM03 | 0.002 | - | - |
| Location HFND01 | 0.002 | - | - |

When sample location was excluded, DOC was significant in the prediction of TOC (Gini importance of 0.774) followed by the date. In the absence of sample location, the other water quality parameters had comparable Gini importance values. When date, density, light transmissivity, and total suspended solids were used as logs, sampling date had the highest Gini importance factor of 0.465. Since these four parameters were easily measurable, they could be used as a prediction estimate for

TOC prediction because they gave a coefficient of determination of 0.600. When necessary, and for higher accuracy, all seven water quality parameters could be used to predict TOC.

### *3.6. One-dimensional convolutional neural network regression*

1D-CNN was used to predict TOC using the same input parameters as those used in RF algorithm. The input parameters included; DOC, light transmissivity, date, TSS, depth, silica, orthophosphate phosphorous, salinity, and density. Table 9 shows the optimization results for 1D-CNN regression, which includes the number of convolutional layers and dense layers optimized. The test loss measures the difference between the predicted values and the true values. It is aggregated over all the samples in the test set to give the overall results. In this case, it was determined using the mean squared value. The1D-CNN prediction of TOC had a test loss, MAE, MSE, and $R^2$ of 0.114 mg/L, 0.244 mg/L, 0.119 mg/L, and 0.714, respectively. The number of convolutional layers and dense layers were 1 and 4, respectively. The number of neurons varied in each layer, as shown in Table 9. The scatter plot showing predicted versus observed values for 1D-CNN is shown in Figure 10.

**Table 9.** Optimization results for 1D-CNN

| Parameter | Value |
|---|---|
| Test loss | 0.114 |
| Mean absolute error, MAE (mg/L) | 0.244 |
| Mean squared error, MSE (mg/L) | 0.119 |
| Coefficient of determination, $R^2$ | 0.714 |
| Number of CNN layers | 1 |
| Number of filters in CNN | 128 |
| Kernel size in CNN | 5 |
| Number of dense layers | 4 |
| Neurons in dense layer 1 | 256 |
| Neurons in dense layer 2 | 64 |
| Neurons in dense layer 3 | 448 |
| Neurons in dense layer 4 | 512 |
| Number of epochs | 16 |

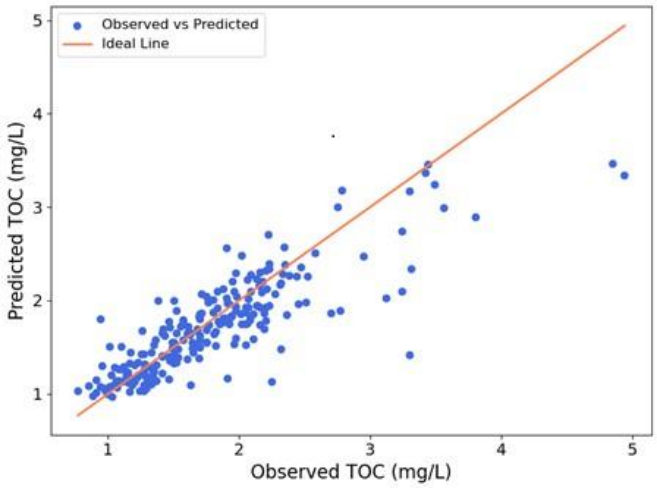


**Figure 10.** Scatter plots for predicted versus observed values for TOC prediction using 1D-CNN.

Figure 11 is a training and validation loss curve used to visualize how the number epochs affect the accuracy of TOC prediction using 1D-CNN architecture. The loss curve shows a rapid decrease in training and validation loss within the first five epochs. The model converged quickly within the five epochs, meaning that the model learned effectively. The two curves remained very close, indicating good generalization and that there was no major overfitting. Both losses were low and stable, meaning that underfitting did not arise. The optimal number of epochs for 1D-CNN were stopped at 16 to monitor the performance and prevent overfitting or underfitting during the training process. From the loss-curves, the 1D-CNN model learned quickly and had good prediction of the true values of TOC.

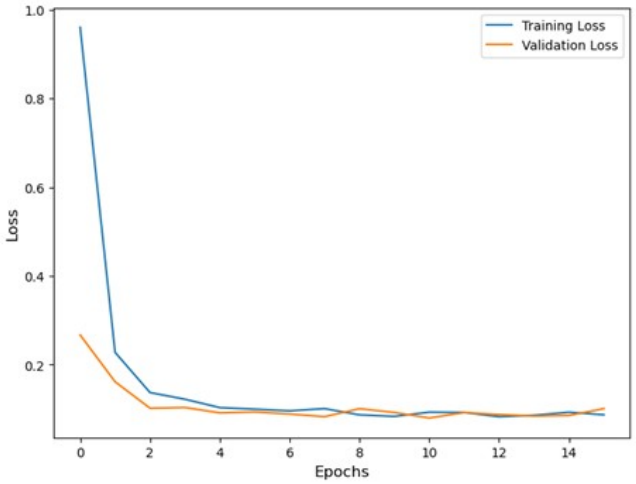


**Figure 11.** Training and validation loss curve for 1D-CNN.

*3.7. Multilayer perceptron regression*

A MLP regression algorithm was created to predict TOC using the seven water quality parameters mentioned in the previous sections. Table 10 shows the optimization results of the MLP regression algorithm. The test loss, MAE, MSE, and $R^2$ for the MLP regression algorithm were 0.151mg/L, 0.270 mg/L, 0.150 mg/L, and 0.638, respectively. The number of hidden layers were optimized to 5 with the number of neurons in each layer, as shown in Table 10. The scatter plot for the predicted versus the observed value is shown in Figure 12. During the training of the dataset, the epochs stabilized at 21 when the training loss was 0.155, as shown in Figure 13.

**Table 10.** Optimization results for MLP for TOC prediction.

| Parameter | Value |
|---|---|
| Test loss | 0.151 |
| Mean absolute error, MAE (mg/L) | 0.270 |
| Mean Squared Error, MSE (mg/L) | 0.150 |
| Coefficient of determination, $R^2$ | 0.638 |
| Number of hidden layers | 5 |
| Neurons in hidden layer 1 | 64 |
| Neurons in hidden layer 2 | 96 |
| Neurons in hidden layer 3 | 288 |
| Neurons in hidden layer 4 | 512 |
| Neurons in hidden layer 5 | 384 |
| Number of epochs | 21 |

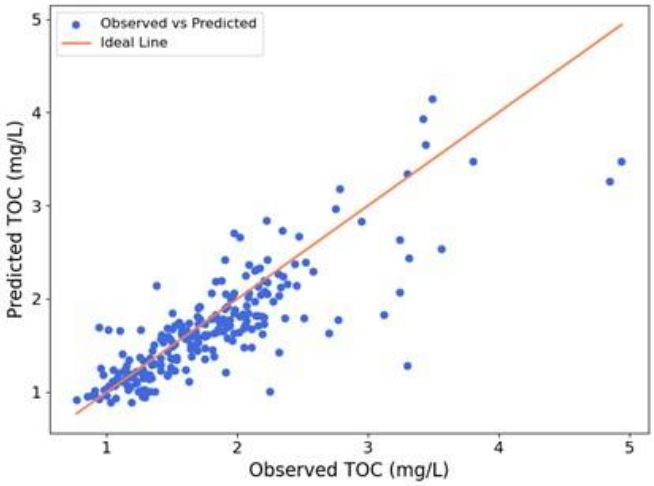


**Figure 12.** Scatter plots for predicted versus observed values for TOC prediction using MLP.

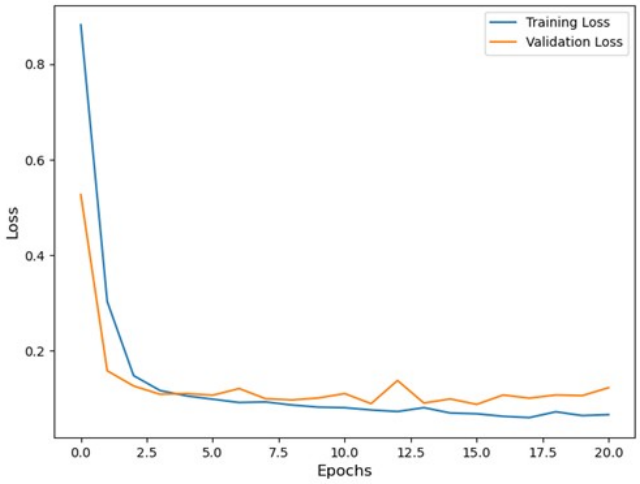


**Figure 13.** Training and validation loss curve for MLP.

An assessment of the predictive MLP model was done and compared to that of 1D-CNN. Figure 13 shows training and validation loss curve for the MLP regression algorithm. It had a similar trend to that of 1D-CNN. The training and validation loss dropped sharply within the first three epochs. The validation loss curve stabilized at epoch 5, but that of 1D-CNN fluctuated slightly at around epoch 12. This indicated mild overfitting. The training loss curve was more stable and consistently lower than the validation curve after 5 epochs. Both losses reached a stable low value and, thus, underfitting did not present a challenge for this algorithm. Both MLP and 1D-CNN provided good prediction and fast learning in modeling the data but ran the risk of overfitting [31]. They also lacked good interpretability, making it hard to extract meaningful insight from the filters or feature maps [20]. 1D-CNN requires large amount of labeled data for effective training, especially when the architecture is deep or the kernel size is large and, therefore, requires more computational resources [32].

### *3.8. Comparison of model performance in TOC prediction*

The performance of RF, 1D-CNN, and MLP in predicting TOC from the nine water quality parameters was compared. Table 11 shows a comparison of the three algorithms in predicting TOC. From the results, the RF algorithm had the highest accuracy for TOC prediction with a test $R^2$ of 0.732

and a trained $R^2$ of 0.969, indicating a strong learning capacity with minimal overfitting. In comparison, 1D-CNN had a test $R^2$ of 0.714 and a trained $R^2$ of 0.801, and MLP had a test $R^2$ of 0.638 and trained $R^2$ of 0.808, indicating a slightly lower training fit. The RF test $R^2$ of 0.732 was within the range as those in [33] where ML models in predicting TOC in river water were used. The accuracy of these models depends on the target variable and contextual features used [13]. These studies and the cross-validation results indicated that the predictive accuracy as wis realistic for site-specific monitoring tasks, and that the model errors (MAE/MSE) remained within practical limits for trend detection and decision support. RF provided the best fit for data and best captured variance in the data. It also had the least MAE of 0.120 mg/L compared to that of 1D-CNN and MLP, meaning that RF gave the most precise predictions in the regression process because the error between the observed and predicted values was the least. The MSE of 1D-CNN was the lowest (0.119 mg/L) among the three models, meaning that 1D-CNN performed better in minimizing large and small errors. Therefore, in terms of performance metrics, RF would be chosen over the two deep learning approaches (CNN and MLP) for TOC prediction in river water.

**Table 11.** Performance comparison of RF, 1D-CNN, and MLP regression algorithm for TOC prediction.

| Model | Test $R^2$ | Train $R^2$ | MAE (mg/L) | MSE (mg/L) |
|---|---|---|---|---|
| RF | 0.732 | 0.969 | 0.120 | 0.286 |
| 1D-CNN | 0.714 | 0.801 | 0.244 | 0.119 |
| MLP | 0.638 | 0.808 | 0.270 | 0.150 |

Deep learning is best suited for handling highly complex relationships in data and requires more computational resources and time [34]. It also requires data normalization, which can lead to loss of important data features in regression tasks. The RF algorithm, on the other hand, is easier to interpret and understand because it is possible to track down where the decisions are made during the regression task [23]. It is best suited to handle small datasets even with categorical features. RF can handle datasets without necessarily having to normalize or impute data and maintains accuracy for many multivariate regression tasks [35]. In this study, it was able to accurately handle input data containing parameters that exhibited both periodic and non-periodic time series trends without normalization. When the two deep learning algorithms were compared, 1D-CNN had a higher $R^2$ value and lower MSE and MAE values than the MLP regression algorithm in TOC prediction, indicating it was better at predicting the true value of TOC. The convolutional layer, which was not there in MLP, had the ability to capture spatial features and patterns in the dataset [20]. It is difficult to handle data sets that contain varying features and trends with MLP because it lacks a convolutional layer. Among the three models evaluated, the RF algorithm demonstrated superior predictive performance compared to the deep learning approaches (MLP and 1D-CNN). This could be attributed to the nature of the dataset, which was tabular and moderately sized, making RF particularly effective at capturing complex nonlinear relationships. RF's ensemble-based architecture aggregated multiple decision trees, thereby reducing model variance and enhancing robustness against overfitting. In contrast, the MLP and 1D-CNN models, although capable of learning intricate feature representations, were more sensitive to data distribution and architecture configuration. The consistency of RF's performance was further confirmed through five-fold cross-validation, where it achieved the highest mean $R^2$ ($0.771 \pm 0.031$) and the lowest error metrics (standard deviation) among the models. These results demonstrated RF's superior ability to generalize across data folds, demonstrating its reliability for regression tasks on moderately sized environmental datasets.

## 4. Conclusions

In this study, prediction of TOC through measurements of depth, density, DOC, light transmissivity, orthophosphate phosphorus, silica, TSS, salinity, and date was carried out using the RF algorithm, 1D-CNN, and MLP deep learning algorithms. Density, DOC, light transmissivity, orthophosphate phosphorus, silica, TSS, and salinity had a moderate to strong correlation with TOC. Seasonality, not location, influenced all the water quality parameters except for light transmissivity, TOC, and TSS. The RF regression algorithm had the highest prediction capability compared to the deep learning models. The RF regression model was robust in withstanding time series variations in some water quality parameters. Among the two deep learning models, 1D-CNN had better predictive ability than MLP. This was attributed to the presence of the convolutional layer, which can handle complex patterns in the dataset. RF is a powerful robust machine learning regression tool that can be applied in predicting TOC in the river water by carefully picking the input parameters. It can withstand parametric variations and is strong enough to handle noise in the datasets. This superior performance was further validated through five-fold cross-validation, which confirmed the model's consistency and minimized bias across data splits. Additionally, learning curve analysis indicated that the available dataset size was sufficient for model training and that the RF model achieved stable generalization without overfitting. The findings here provide novel predictive tools that would be useful in low resource settings or for resource planning when predicting TOC from other water quality parameters. In the future, researchers could explore advanced gradient boosting frameworks such as XGBoost and LightGBM to further enhance prediction accuracy and computational efficiency.

## Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

## Conflict of interest

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

## References

1. Z. Kılıç, The importance of water and conscious use of water, *Int. J. Hydro.*, **4** (2021), 239–241. https://doi.org/10.15406/ijh.2020.04.00250
2. G. Bennett, Status of drinking water supply and water stress levels in the African Great Lakes region: a time-series analysis from 1980 to 2020, *Water Supply*, **25** (2025), 228–239. https://doi.org/10.2166/ws.2025.017
3. A. Boretti, L. Rosa, Reassessing the projections of the world water development report, *Npj Clean Water*, **2** (2019), 15. https://doi.org/10.1038/s41545-019-0039-9
4. N. Akhtar, M. I. Syakir Ishak, S. A. Bhawani, K. Umar, Various natural and anthropogenic factors responsible for water quality degradation: a review, *Water*, **13** (2021), 2660. https://doi.org/10.3390/w13192660
5. L. McDonough, I. Santos, M. Andersen, D. O'Carroll, H. Rutlidge, K. Meredith, et al., Changes in global groundwater organic carbon driven by climate change and urbanization, *Nat. Commun.*, **11** (2020), 1279. https://doi.org/10.1038/s41467-020-14946-1

6. D. Dubber, N. F. Gray, Replacement of chemical oxygen demand (COD) with total organic carbon (TOC) for monitoring wastewater treatment performance to minimize disposal of toxic analytical waste, *J. Environ. Sci. Heal. A*, **45** (2010), 1595–1600. https://doi.org/10.1080/10934529.2010.506116
7. A. M. Handhal, A. M. Al-Abadi, H. E. Chafeet, M. J. Ismail, Prediction of total organic carbon at Rumaila oil field, Southern Iraq using conventional well logs and machine learning algorithms, *Mar. Petrol. Geol.*, **116** (2020), 104347. https://doi.org/10.1016/j.marpetgeo.2020.104347
8. E. Goz, M. Yuceer, E. Karadurmus, Total organic carbon prediction with artificial intelligence techniques, *Computer Aided Chemical Engineering*, **46** (2019), 889–894. https://doi.org/10.1016/B978-0-12-818634-3.50149-1
9. S. Kim, N. Maleki, M. Rezaie-Balf, V. Singh, M. Alizamir, N. Kim, et al., Assessment of the total organic carbon employing the different nature - inspired approaches in the Nakdong River, South Korea, *Environ. Monit. Assess.*, **193** (2021), 445. https://doi.org/10.1007/s10661-021-08907-4
10. I. S. Yeon, J. H. Kim, K. W. Jun, Application of artificial intelligence models in water quality forecasting, *Environ. Technol.*, **29** (2008), 625–631. https://doi.org/10.1080/09593330801984456
11. B. Alizadeh, K. Maroufi, M. H. Heidarifard, Estimating source rock parameters using wireline data: an example from Dezful Embayment, south west of Iran, *J. Petrol. Sci. Eng.*, **167** (2018), 857–868. https://doi.org/10.1016/j.petrol.2017.12.021
12. Y. Asgari Nezhad, A. Moradzadeh, M. R. Kamali, A new approach to evaluate organic geochemistry parameters by geostatistical methods: a case study from western Australia, *J. Petrol. Sci. Eng.*, **169** (2018), 813–824. https://doi.org/10.1016/j.petrol.2018.05.027
13. L. Zhu, X. Zhou, W. Liu, Z. Kong, Total organic carbon content logging prediction based on machine learning: a brief review, *Energy Geoscience*, **4** (2023), 100098. https://doi.org/10.1016/j.engeos.2022.03.001
14. M. Tan, X. Song, X. Yang, Q. Wu, Support-vector-regression machine technology for total organic carbon content prediction from wireline logs in organic shale: acomparative study, *J. Nat. Gas Sci. Eng.*, **26** (2015), 792–802. https://doi.org/10.1016/j.jngse.2015.07.008
15. X. Liu, Y. Lei, X. Luo, X. Wang, K. Chen, M. Cheng, et al., TOC determination of Zhangjiatan shale of Yanchang formation, Ordos Basin, China, using support vector regression and well logs, *Earth Sci. Inform.*, **14** (2021), 1033–1045. https://doi.org/10.1007/s12145-021-00607-4
16. V. Bolandi, A. Kadkhodaie, R. Farzi, Analyzing organic richness of source rocks from well log data by using SVM and ANN classifiers: a case study from the Kazhdumi formation , the Persian Gulf basin, off shore Iran, *J. Petrol. Sci. Eng.*, **151** (2017), 224–234. https://doi.org/10.1016/j.petrol.2017.01.003
17. J. Sun, W. Dang, F. Wang, H. Nie, X. Wei, P. Li, et al., Prediction of TOC content in organic-rich shale using machine learning algorithms: Comparative study of random forest, support vector machine, and XGBoost, *Energies*, **16** (2023), 4159. https://doi.org/10.3390/en16104159
18. J. Rui, H. Zhang, D. Zhang, F. Han, Q. Guo, Total organic carbon content prediction based on support-vector-regression machine with particle swarm optimization, *J. Petrol. Sci. Eng.*, **180** (2019), 699–706. https://doi.org/10.1016/j.petrol.2019.06.014
19. M. Sabzekar, S. Hasheminejad, Robust regression using support vector regressions, *Chaos Soliton. Fract.*, **144** (2021), 110738. https://doi.org/10.1016/j.chaos.2021.110738
20. A. O. Ige, M. Sibiya, State-of-the-art in 1d convolutional neural networks: a survey, *IEEE Access*, **12** (2024), 144082–144105. https://doi.org/10.1109/ACCESS.2024.3433513
21. S. Asante-okyere, Y. Yevenyo, S. Adjei, Improved total organic carbon convolutional neural network model based on mineralogy and geophysical well log data, *Unconventional Resources*, **1** (2021), 1–8. https://doi.org/10.1016/j.uncres.2021.04.001
22. I. K. Mutai, K. Van Laerhoven, N. W. Karuri, R. K. Tewo, Using multivariate linear regression for biochemical oxygen demand prediction in waste water, *Applied Computing and Intelligence*, **4** (2024), 125–137. https://doi.org/doi: 10.3934/aci.2024008

23. Y. Ao, H. Li, L. Zhu, S. Ali, Z. Yang, The linear random forest algorithm and its advantages in machine learning assisted logging regression modeling, *J. Petrol. Sci. Eng.*, **174** (2019), 776–789. https://doi.org/10.1016/j.petrol.2018.11.067
24. V. Rodriguez-galiano, M. Sanchez-castillo, M. Chica-olmo, M. Chica-rivas, Machine learning predictive models for mineral prospectivity : an evaluation of neural networks, random forest, regression trees and support vector machines, *Ore Geol. Rev.*, **71** (2015), 804–818. https://doi.org/10.1016/j.oregeorev.2015.01.001
25. P. Schober, C. Boer, L. A. Schwarte, Correlation coefficients: appropriate use and interpretation, *Anesth. Anal.*, **126** (2018), 1763–1768. https://doi.org/10.1213/ANE.0000000000002864
26. G. Peng, S. Sun, Z. Xu, J. Du, Y. Qin, S. Sharshir, et al., The effect of dataset size and the process of big data mining for investigating solar-thermal desalination by using machine learning, *Int. J. Heat Mass Tran.*, **236** (2025), 126365. https://doi.org/10.1016/j.ijheatmasstransfer.2024.126365
27. M. Fettweis, M. Schartau, X. Desmit, B. J. Lee, N. Terseleer, D. Van der Zande, et al., Organic matter composition of biomineral flocs and its influence on suspended particulate matter dynamics along a nearshore to offshore transect, *J. Geophys. Res.-Biogeo.*, **127** (2022), e2021JG006332. https://doi.org/10.1029/2021JG006332
28. H. Schmidt, S. Seitz, E. Hassel, H. Wolf, The density-salinity relation of standard seawater, *Ocean Sci.*, **14** (2018), 15–40. https://doi.org/10.5194/os-14-15-2018
29. M. Nicolaus, C. Petrich, S. R. Hudson, M. A. Granskog, Variability of light transmission through Arctic land-fast sea ice during spring, *Cryosphere*, **7** (2013), 977–986. https://doi.org/10.5194/tc-7-977-2013
30. D. L. Correll, T. E. Jordan, D. E. Weller, Effects of precipitation, air temperature, and land use on organic carbon discharges from rhode river watersheds, *Water Air Soil Poll.*, **128** (2001), 139–159. https://doi.org/10.1023/A:1010337623092
31. A. Al Bataineh, D. Kaur, S. Jalali, Multi-layer perceptron training optimization using nature inspired computing, *IEEE Access*, **10** (2022), 36963–36977. https://doi.org/10.1109/ACCESS.2022.3164669
32. E. Qazi, A. Almorjan, T. Zia, A one-dimensional convolutional neural network (1D-CNN) based deep learning system for network intrusion detection, *Appl. Sci.*, **12** (2022), 7986. https://doi.org/10.3390/app12167986
33. H. Oh, H. Park, J. Kim, B. Lee, J. Choi, J. Hur, Enhancing machine learning models for total organic carbon prediction by integrating geospatial parameters in river watersheds, *Sci. Total Environ.*, **943** (2024), 173743. https://doi.org/10.1016/j.scitotenv.2024.173743
34. A. L'Heureux, K. Grolinger, H. Elyamany, M. Capretz, Machine learning with big data: challenges and approaches, *IEEE Access*, **5** (2017), 7776–7797. https://doi.org/10.1109/ACCESS.2017.2696365
35. F. Tang, H. Ishwaran, Random forest missing data algorithms, *Stat. Anal. Data Min.*, **10** (2017), 363–377. https://doi.org/10.1002/sam.11348



AIMS Press