water

Article

A Non-Hybrid Data-Driven Fuzzy Inference System for Coagulant Dosage in Drinking Water Treatment Plant: Machine-Learning for Accurate Real-Time Prediction

Adriano Bressane [1,2,*], Ana Paula Garcia Goulart [2], Carrie Peres Melo [1], Isadora Gurjon Gomes [2], Anna Isabel Silva Loureiro [1], Rogério Galante Negri [2,3], Rodrigo Moruzzi [1,2], Adriano Gonçalves dos Reis [1,2], Jorge Kennety Silva Formiga [2], Gustavo Henrique Ribeiro da Silva [1] and Ricardo Fernandes Thomé [1]

[1] Civil and Environmental Engineering Graduate Program, College of Engineering, São Paulo State University, 14-01 Eng. Luiz E.C. Coube Avenue, Bauru 17033-360, Brazil
[2] Environmental Engineering Department, Institute of Science and Technology, São Paulo State University, 500 Altino Bondensan Road, São José dos Campos 12245-000, Brazil
[3] Natural Disasters Graduate Program, Brazilian Center for Early Warning and Monitoring for Natural Disasters, 500 Altino Bondensan Road, São José dos Campos 12245-000, Brazil
* Correspondence: adriano.bressane@unesp.br

check for updates

Citation: Bressane, A.; Goulart, A.P.G.; Melo, C.P.; Gomes, I.G.; Loureiro, A.I.S.; Negri, R.G.; Moruzzi, R.; Reis, A.G.d.; Formiga, J.K.S.; da Silva, G.H.R.; et al. A Non-Hybrid Data-Driven Fuzzy Inference System for Coagulant Dosage in Drinking Water Treatment Plant: Machine-Learning for Accurate Real-Time Prediction. *Water* **2023**, *15*, 1126. https://doi.org/10.3390/w15061126

Academic Editor: Andrea G. Capodaglio

Received: 16 November 2022
Revised: 9 March 2023
Accepted: 13 March 2023
Published: 15 March 2023

Abstract: Coagulation is the most sensitive step in drinking water treatment. Underdosing may not yield the required water quality, whereas overdosing may result in higher costs and excess sludge. Traditionally, the coagulant dosage is set based on bath experiments performed manually, known as jar tests. Therefore, this test does not allow real-time dosing control, and its accuracy is subject to operator experience. Alternatively, solutions based on machine learning (ML) have been evaluated as computer-aided alternatives. Despite these advances, there is open debate on the most suitable ML method applied to the coagulation process, capable of the most highly accurate prediction. This study addresses this gap, where a comparative analysis between ML methods was performed. As a research hypothesis, a data-driven (D^2) fuzzy inference system (FIS) should provide the best performance due to its ability to deal with uncertainties inherent to complex processes. Although ML methods have been widely investigated, only a few studies report hybrid neuro-fuzzy systems applied to coagulation. Thus, to the best of our knowledge, this is the first study thus far to address the accuracy of this non-hybrid data-driven FIS (D^2FIS) for such an application. The D^2FIS provided the smallest error (0.69 mg/L), overcoming the adaptive neuro-fuzzy inference system (1.09), cascade-correlation network (1.18), gene expression programming (1.15), polynomial neural network (1.20), probabilistic network (1.17), random forest (1.26), radial basis function network (1.28), stochastic gradient tree boost (1.25), and support vector machine (1.17). This finding points to the D^2FIS as a promising alternative tool for accurate real-time coagulant dosage in drinking water treatment. In conclusion, the D^2FIS can help WTPs to reduce operating costs, prevent errors associated with manual processes and operator experience, and standardize the efficacy with real-time and highly accurate predictions, and enhance safety for the water industry. Moreover, the evidence from this study can assist in filling the gap with the most suitable ML method and identifying a promising alternative for computer-aided coagulant dosing. For further advances, future studies should address the potential of the D^2FIS for the control and optimization of other unit operations in drinking water treatment.

Keywords: coagulant dosage; fuzzy; machine learning; water treatment

1. Introduction

To remove contaminants, such as suspended solids, colloidal material, and microorganisms, coagulation is among the primary processes for the physical–chemical treatment of drinking water [1,2]. Jar tests are commonly used to determine the best dose of coagulant in drinking water treatment plants (WTPs) [3,4]. Considering the quality of raw water, the test simulates the coagulation step under laboratory conditions. Although this test has been used for many years, improving both its accuracy and response speed with respect to water quality changes, it remains very challenging [5]. Jar test experiments are manually performed and, hence, were not conceived for real-time decision-making. Additionally, coagulant dosing can become complex when raw water quality changes rapidly and substantially [6], particularly due to the critical influence of the potential of hydrogen (pH), turbidity, and color, among other properties of contaminants and hydraulic conditions, on coagulation performance [7–9]. Therefore, the jar test is not feasible for real-time adjustments [4,10].

On the other hand, reducing operating costs and improving efficacy in water treatment are some of the main challenges in the water sector, which also faces natural water degradation and strict standards and regulations. Therefore, the study and application of data-driven and real-time technologies, such as machine learning (ML), are essential to reduce costs and enhance water safety for the water industry [6,11]. However, the use of alternatives based upon mechanistic models for the coagulation process is a difficult task as it is a complex system in which there are uncertainties as interactions between the mechanisms of transfer and kinetics are not yet deeply understood [9,12].

In several areas of knowledge, empirical models using ML methods have been evaluated with a good ability to model complex non-linear problems [13]. Among the advantages of this computer-aided alternative, the prevention of errors associated with the human operator and the reduction in response times can be highlighted [10]. Another favorable factor is that the development of solutions based on ML only requires the availability of historical databases, which, in the case of drinking WTPs, are usually stored in sufficient quantities for this alternative [14]. Thus, applications based on methods such as artificial neural networks (ANNs) have become increasingly popular [12]. However, even with continual progress in research, highlighted among the most recent studies by Pandilov and Stojkov [6], Najafzadeh and Zeinolabedini [15], Ju et al. [16], Zhang et al. [12], Ghasemi et al. [8], Wang et al. [2], Narges et al. [5], and Zhu et al. [9], the results achieved on computer-aided coagulant dosing have not yet led to the replacement of the jar test, which is still widely performed in drinking WTPs [8]. Therefore, additional studies are still needed to strengthen the evidence that makes it possible to reduce the dependence on bath experiments, enabling more accurate predictions in real-time [2].

Several ML methods have been evaluated for coagulant dosing, with emphasis on different ANN architectures, such as the Levenberg–Marquardt neural network [17], inverse neural network [18], generalized regression neural network [19], adaptive neuro-fuzzy inference system [5,6], dynamic evolving neural-fuzzy system [20], radial basis function [2,10,11], multilayer perceptron [4,10,21], genetic algorithm-enhanced artificial neural network [12], variable structure neural network [8], and backpropagation neural network [9]. Other tested ML methods include the linear regression model [22], k-nearest neighbors [23], fuzzy linear and non-linear regression models [10], k-means clustering [11], and random forest [2].

Despite advances in recent years, there are still gaps in terms of the best method of ML applied to coagulation control. We hypothesize that a non-hybrid data-driven fuzzy inference system (D^{2}FIS), introduced in 2022, should provide the highest accuracy due to its ability to deal with intrinsic coagulation uncertainties that are not fully controlled during the WTP operation. To the best of our knowledge, this is the first study to date to assess the performance of this D^{2}FIS in predicting coagulant dosage.

The theory of fuzzy sets was introduced by Lotfi Zadeh to address the uncertainties that arise in complex systems [24]. To this end, inference systems based on fuzzy artificial intelligence with non-linear functions and soft boundaries allow a gradual transition between intervals and degrees of truth, admitting partial membership in more than one set of linguistic values [25].

The development of FISs that use data-oriented methods for regression tasks occurred relatively recently, but they have already become one of the most popular approaches in several areas [26]. Among the environmental applications reported in the literature, FISs have been developed to support participatory planning [27,28] and impact assessment [29–31], pattern recognition [32,33], and land reclamation [34].

In addition, this study should contribute to a better understanding of the following questions: (i) which water quality parameters are most important for accurate coagulant dosing? and (ii) how do variations in these parameters affect coagulant dosing in real time?

2. Methods

The dataset used in this study was derived from the drinking water treatment plant Dr. Armando Pannunzio (WTP Cerrado) at Sorocaba, a city with a territorial area of 449.87 km^2 and 695,000 inhabitants (1304.18 inhab/km^2), one of the most important economic and technological hubs of São Paulo State [35], in southwest Brazil (Figure 1).

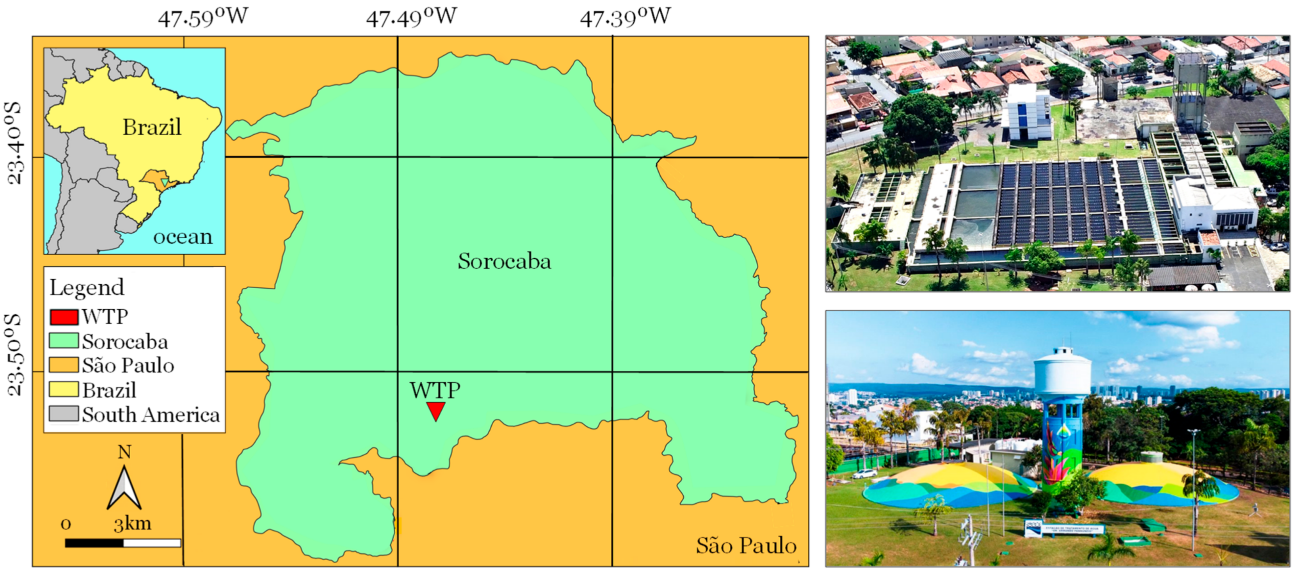


Figure 1. Drinking water treatment plant—WTP Cerrado at Sorocaba city, São Paulo State, southwest, Brazil. Source: modified from Santinon [36].

The WTP Cerrado treats 2.2 m^3/s of water via conventional treatment (coagulation–flocculation–sedimentation–filtration) using coagulant polyaluminum chloride (PAC) within the dose range of 30 to 40 mg/L (Figure 2).

A one-year database (January to December 2021) of quasi-daily tests (n = 291) was used, monitoring PAC and raw water quality indicators (pH, color, turbidity, fluoride, and chloride) (Table 1 and Figure 3).

Table 1. Database with quality indicator parameters of raw water and PAC.

	pH (PAN) *	Color (HU)	Turbidity (NTU)	Fluoride (mg/L)	Chloride (mg/L)	PAC (mg/L)
Average	6.76	2.01	0.25	0.69	1.84	32.0
Median	6.80	2.00	0.20	0.07	1.90	32.0
St. Deviation	0.11	1.18	0.16	0.03	0.24	1.84
Minimum	6.40	0.00	0.03	0.60	0.90	30.0
Maximum	7.00	7.00	0.78	0.76	2.70	40.0
Asymmetry	−0.25	1.14	1.29	−0.31	−0.63	1.40
Kurtosis	0.26	2.09	1.32	−0.03	1.68	2.39
Normality (p) **	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001

Note: * Values based on the proton activity mean (PAN); ** Shapiro–Wilk test.

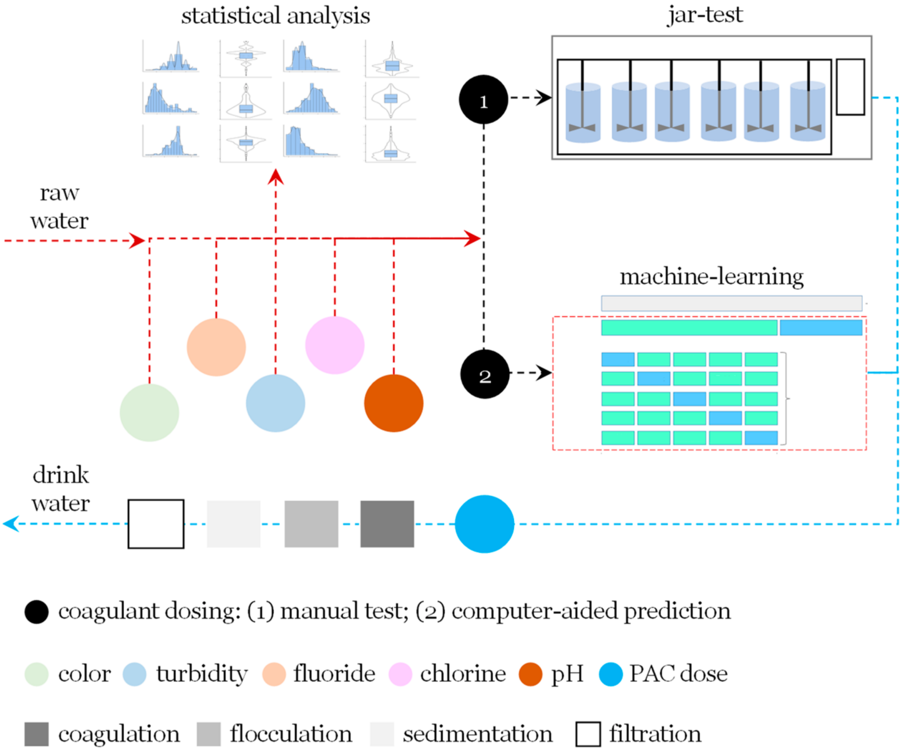

statistical analysis
jar-test
raw water
1
2
machine-learning
drink water
● coagulant dosing: (1) manual test; (2) computer-aided prediction
● color ● turbidity ● fluoride ● chlorine ● pH ● PAC dose
■ coagulation ■ flocculation ■ sedimentation □ filtration

Figure 2. WTP via conventional treatment (coagulation–flocculation–sedimentation–filtration) with manual or computer-aided coagulant dosing.

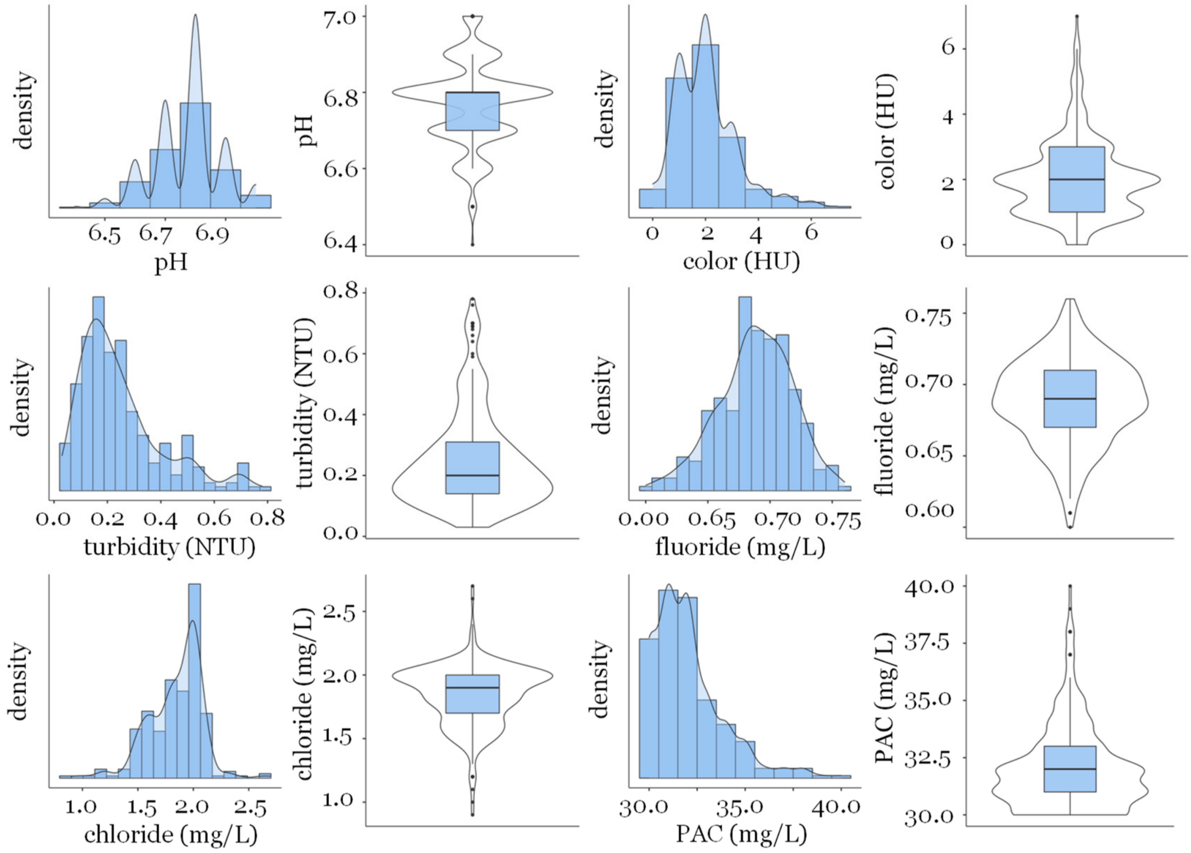


Figure 3. Exploratory analysis of quality indicator parameters of raw water and PAC.

As an artificial intelligence method specifically developed for the D^{2}FIS, the Wang–Mendel algorithm ('wm') was adopted in the present study. A non-hybrid ML method based on this algorithm was made available by Guillaume et al. [37] in the package 'FisPro' in the R programming language, which was used in our research. The 'wm' is a method of inducing IF-THEN fuzzy rules within the general architecture of the inference system [38]. In this study, the 'wm' method was associated with the fuzzy hierarchical partition strategy ('hfp'), resulting in an inference system with high completeness and robustness, which is able to achieve better forecast accuracy.

This D^{2}FIS is a rule-based inference system capable of extracting knowledge from the raw data and, at the same time, preserving the interpretability of the resulting model [39]. The modeling process of the fuzzy inference system includes the following main steps [40]: (i) fuzzification: the input space partitioning of the predictors occurs using fuzzy membership functions, which model the linguistic values of each variable, using the 'hfp', for instance; (ii) inference: an induction technique, such as the 'wm' method, in which logical operators of conjunction (minimum, product, or Lukasiewicz) and disjunction (sum or maximum) are applied to build relational propositions (IF-THEN rules); and (iii) defuzzification: the fuzzy output is converted to a crisp value (Figure 4).

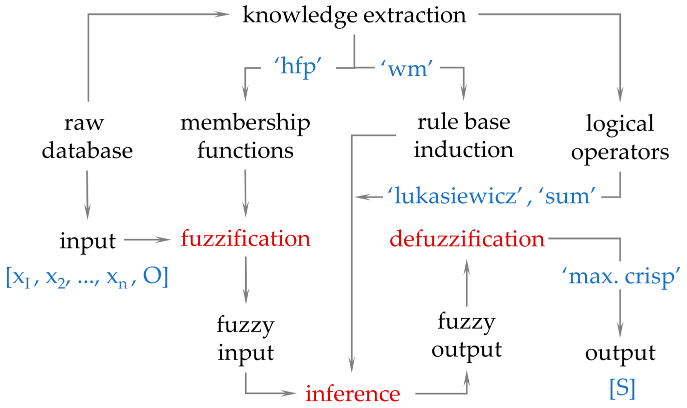


Figure 4. Modeling process of the data-driven fuzzy inference system (D^{2}FIS), considering the water quality parameters as input and the coagulant dosage as output.

To test the research hypothesis, the accuracy of this D^{2}FIS was compared to that obtained by some of the primary methods applicable to prediction tasks: adaptive neuro-fuzzy inference system (ANFIS), cascade-correlation network (CCN), gene expression programming (GEP), polynomial neural network (GMDH), multilayer perceptron network (MLP), probabilistic network (PNN), radial basis function network (RBFN), random forest (RF), stochastic gradient tree boost (SGT), and support vector machine (SVM). As a standard way to measure the performance of a model in predicting quantitative data, the root mean square error (*RMSE*), the most common metric for comparing models [41], was calculated to analyze the coagulant dosing accuracy (Equation (1)):

$$RMSE = \sqrt{\frac{1}{n}\sum_i^n (S_i - O_i)^2} \quad (1)$$

where O_i are the observations, S_i are the predicted values of the dose of the coagulant, and n is the number of observations.

Each ML method has a different set of parameters according to its operating mechanism, making it impossible to adopt the same values. Therefore, during machine learning, a grid search method was used, which identifies the best parametrization values, that is, the values capable of providing the highest accuracy for each ML method [42]. Considering different combinations of parameters, the setting that minimized the RMSE was determined based on 5-fold cross-validation (Figure 5), using 70% of the dataset for the learning process and 30% for validation testing, a quite common split used to train and evaluate the performance of regression models [43].

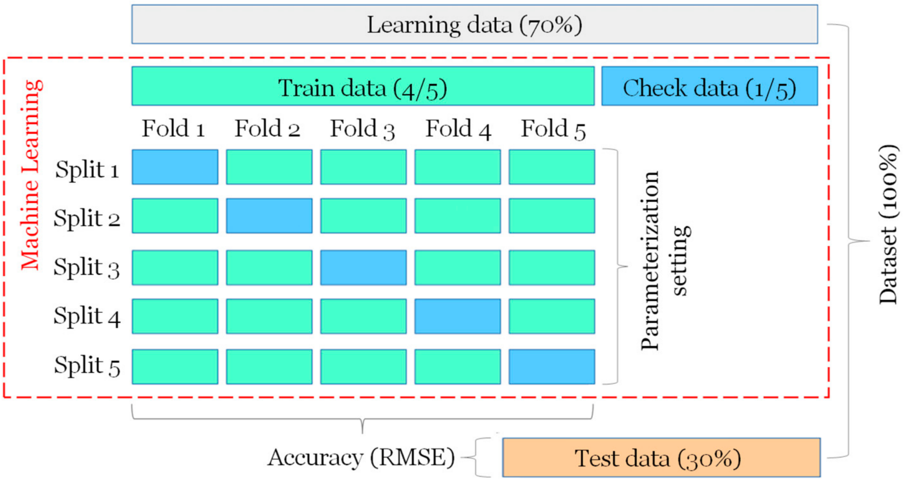


Figure 5. Determining parameterization settings based on 5-fold cross-validation. Source: modified from Scikit-learn developers [44].

To analyze which raw water quality parameters are most important for accurate coagulant dosing and how their variations affect real-time prediction, the 'Explain' and 'ICE' (individual conditional expectation) widgets were used. Both widgets are available in the Orange data mining software (version 3.34).

3. Results and Discussion

The performance of the ML methods is presented in Table 2, where the accuracy (RMSE) based on the testing data varies significantly between 1.28 (RBFN) and 0.86 mg/L (D^2FIS).

Table 2. Overall accuracy of each ML method based on the RMSE.

ML Method	Parameterization Setting	RMSE (mg/L)	
		Train	Test
ANFIS	Model: subtractive clustering; radii: 0.7: functions: 2; pre-overfitting epochs: 96; optimization method: hybrid.	0.99	1.09
CCN	Kernel: gaussian; candidates: 10^2; epochs: 10^3; neurons range: [0–10^3]; overfitting control: cross-validation.	1.10	1.18
D^2FIS	Model: 'wm'; conjunction: Lukasiewicz; disjunction: sum; functions: 6; grid: 150 'hfp'; defuzzification: maximum crisp.	0.44	0.86
GEP	Population: 50; maximum tries: 10^4; genes: 4; gene head length: 10; generations without improvement: 10^3.	1.10	1.15
GMDH	Layer: 20; polynomial order: 16; neurons per layer: 20; function: linear; connections: to previous layer.	1.17	1.20
MLP	Layer: 3; hidden layer function: smooth; output layer function: linear; train: scaled conjugate gradient.	1.11	1.17
PNN	Kernel: gaussian; steps: 20; sigma: each variable [10^{-4}–10]; prior probability: frequency distribution.	0.71	1.17
RBFN	Neurons: 10^3; radius: [10^{-2}–10^3]; population size: 200; generations: 20; generation flat: 5.	0.96	1.28
RF	Number of trees: 860; number of attributes considered at each split: 35; limit depth of individual trees: 15.	0.48	1.26
SGT	Trees: [10^3–10^2]; depth: 10; minimum size node: 5; shrink factor: auto; prune: minimum absolute error, smooth: 5.	0.56	1.25
SVM	Kernel: RBF; model: epsilon-SVR; optimize: minimize total error; stopping criteria: 10^{-3}.	1.12	1.17

In general, although some ML algorithms stand out for their high performance in specific applications, it is essential to note that task accuracy is also highly associated with data behavior [45]. Therefore, comparing several ML methods is important to verify the best alternative applicable to each case [46].

Analyzing Table 2, the results can be organized into three groups based on the performance of the ML methods during the tests. In the first group, with low performance (RMSE equal to or greater than 1.20 mg/L), are the GMDH, SGT, RBFN, and RF methods. Wang et al. [2] proposed a method for optimizing the coagulation process during drinking water treatment using distinct ML approaches, including the RBFN method. Although it delivers better performance compared to multiple linear regression models, the RBFN was outperformed by the random forest algorithm. In the present study, the SGT achieved the second-worst accuracy with 1.25 RMSE. This algorithm develops a sequential training through which the decision trees grow in series. In this way, a tree is built to correct the errors of the previous one (boosting), which generally provides superior performance unless there is influence from noisy data [32,47].

In the second group, with intermediate performance (RMSE from 1.15 to 1.20 mg/L), were the CCN, GEP, MLP, PNN, and SVM. The CCN is a type of self-organizing neural network whose size and topology are determined by adding neurons to its architecture to guarantee improved learning over the training process [48]. Consequently, this algorithm may overfit the training data and lose generalization ability. In this situation, we used an overfitting control pruning strategy to minimize the cross-validation error. Despite this, the CCN's performance dropped from 1.10 in training to 1.18 RMSE during testing. Wadkar et al. [49] also evaluated the CCN method to predict coagulant dose. The authors indicated that beyond large amounts of training data, as required by most ANN-based approaches, the CCN method showed a sensitive/fragile relationship between the network's architecture and the prediction error rates. Consequently, this method may demand great attention concerning its parametrization.

In turn, the MLP enables non-linear mappings using activation functions based on the backward propagation of errors to adjust the ANN weight connections. Moreover, the network architecture of minimum training error during the ML process was considered to prevent model overfitting, delivering a 1.17 RMSE. Additionally, according to Jayaweera et al. [4], although the MLP method has been useful for predicting the optimum coagulant dosage for water treatment, the high computational cost and requirement of sufficient training data are the primary drawbacks.

Almost all analyzed ANNs achieved a similar performance, approximately 1.17 RMSE. To minimize misclassification, the PNN uses probability density functions to define complex decision boundaries, which generally improves its accuracy [32]. Zhang et al. [23] analyzed the performance of the SVM method applied to predict coagulant dosage in water treatment plants of distinct sizes and concluded that such a method performs better for large- and medium-sized water systems compared to small ones. Although it shares similarities with ANNs, the SVM shows a better ability to deal with high-dimensional data and is less prone to overfitting [50]. Despite this, the SVM also achieved only 1.17 RMSE.

Finally, with higher accuracy, the ANFIS and D^2FIS reached 1.09 and 0.86 RMSE over the test data, respectively. Despite acceptance among researchers, the ANFIS suffers from limitations, such as the curse of dimensionality computational expense, and it is not good at explaining how it reaches decisions [51], being overcome by the non-hybrid D^2FIS in the present study.

Of note, the D^2FIS error is relatively low given that PAC variation ranges from 30 to 40 mg/L, that is, less than ±0.9 mg/L in a variation range more than 10 times greater [30,40]. The occurrence of outliers makes the capability of the D^2FIS to handle data behaviors that are critical to other ML methods even more evident (Figure 6).

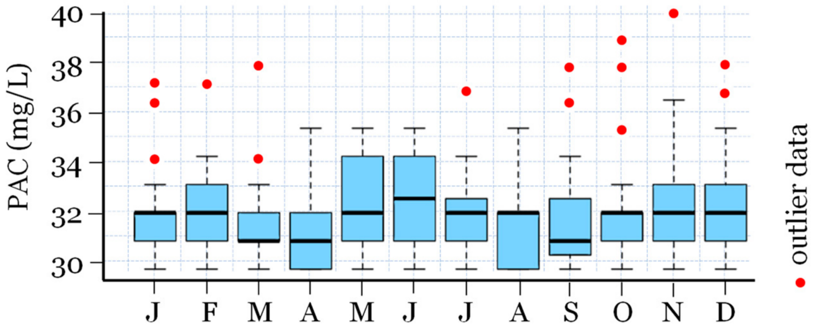


Figure 6. Data behaviors and occurrence of outliers related to coagulant dose throughout the year (January to December).

As outliers are data points that are significantly different from the rest of the dataset, they are usually removed before ML to ensure that the trained model generalizes well to the valid range of test inputs. However, in our study, the outliers were not removed because they are not abnormal observations that arise due to inconsistent data entry, erroneous measurements, etc. In other words, the outliers were maintained because they are a consequence of abrupt variations characteristic of the phenomenon under analysis. Identifying the occurrence of these outliers allowed us to understand why most ML methods showed a significant drop in accuracy during the test in contrast to the D^{2}FIS, which was able to better deal with this challenge.

Using the Wang–Mendel rule induction technique, fuzzification of the linguistic values of each variable was performed using triangular functions (Figure 7), which is one of the most widely accepted and used fuzzy membership functions [25].

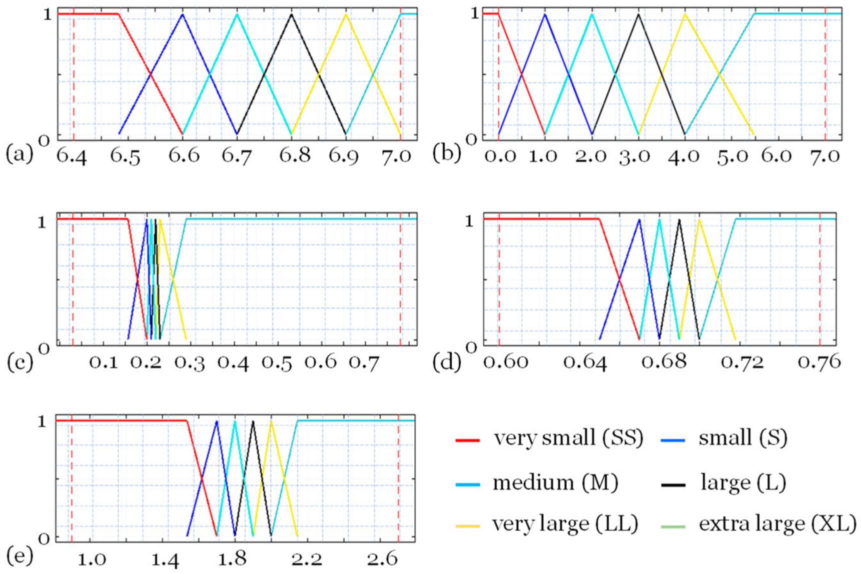


Figure 7. Fuzzification of the predictor variables in the input space using triangular membership functions and 6 linguistic values (SS, S, M, L, LL, and XL): (**a**) pH, (**b**) color (HU), (**c**) turbidity (NTU), (**d**) fluoride (mg/L), and (**e**) chloride (mg/L).

The input space of the predictor variables shown in Figure 7 was partitioned into linguistic values (SS, S, M, L, LL, and XL) based on fuzzy soft boundaries. Khameneh et al. [52] define a fuzzy soft boundary as a parameterization extension of the concept of a boundary in the classical sense. The properties associated with this extension allow a fuzzy model to make inferences based on partial degrees of certainty, which cannot be properly handled using traditional tools [53].

Considering these linguistic values and ranges, some examples of rules (R_i) generated by the D^{2}FIS during machine learning are as follows:

R_1: if pH is S and color is LL and turbidity is XL and fluoride is M and chloride is SS, then the dosage of coagulant (PAC) = 37 mg/L.
R_7: if pH is L and color is M and turbidity is XL and fluoride is SS and chloride is LL, then the dosage of coagulant (PAC) = 30 mg/L.

During parameterization in the ML process, the 'hfp' partitioning procedure provided the best fit of the data to the model. For the Lukasiewicz conjunction operator, six antecedent terms (linguistic values) were sufficient for the D^{2}FIS to decrease the RMSE close to 0.44 during training. Whereas some operators consider only the lowest membership in the disjunction step, the sum t-norm considers all membership values, which provides improved performance in the regression task [32].

In the present study, the coagulation process represents a complex system in which there are uncertainties as interactions between the mechanisms of transfer and kinetics are not yet deeply understood [9,12]. To address this challenge, ML helped identify patterns of association between predictor and response variables based on a data-driven system by extracting knowledge from the historical database. In particular, the data-driven fuzzy approach proved to be efficient in solving this complex and poorly understood problem, characterized by uncertainty due to imprecise knowledge. In other words, data-driven fuzzy allows for dealing with uncertainties to provide a powerful framework for computational reasoning [25].

As a limitation, classical ML methods were not designed to deal with uncertainty. Therefore, when the degree of uncertainty of the problem becomes significant, the solution provided by classical ML methods is not able to provide a solution with greater accuracy [53]. As verified in the experimental evidence of our study, such methods even reach good accuracy during training, but the predictive performance drops during the validation tests. To overcome this GAP, the approach proposed in our study demonstrated that data-driven fuzzy outperformed the ML methods used in previous studies.

It is worth mentioning the fuzzy approach was initially developed aiming at using knowledge-based systems without self-learning capability. Thus, to build a method capable of combining self-learning ability with the ability to deal with uncertainty, hybrid neuro-fuzzy systems were created, such as the ANFIS. The main novel point of the proposed D^{2}FIS method is that it is not a hybrid ML method, as has already been evaluated in previous studies. Thus, to the best of our knowledge, this is the first study thus far to address the accuracy of this non-hybrid D^{2}FIS for coagulant dosage.

From these results, computer-aided coagulant dosage can be highly accurately determined using the D^{2}FIS approach proposed in this study. As a practical implication, this alternative avoids errors associated with the WTP operator's experience; it can predict dosages accurately and in real time, saving operational resources, the acquisition and maintenance of equipment, and the consumption of raw material required by jar tests. Considering that the jar test can result in underdosing or overdosing, further verification from future studies will be needed to certify that the proposed method can predict the optimal dose of coagulant.

On the other hand, considering that, with a smaller margin of error, the proposed method can provide a dosage comparable to the dosage through a manual experiment, the data-driven fuzzy can help the drinking water treatment system to overcome critical limitations of the jar test, which are not conceived for real-time decision-making. As previously pointed out, coagulant dosing can become complex when raw water quality changes rapidly and substantially [6]. Thus, the jar test is not feasible for fast enough adjustments [4,10].

The results of the 'Explain' and 'ICE' widgets are shown in Figure 8, where the effects of each raw water quality parameter on coagulant dosing accuracy and the dependence of real-time prediction on variation in these parameters can be seen.

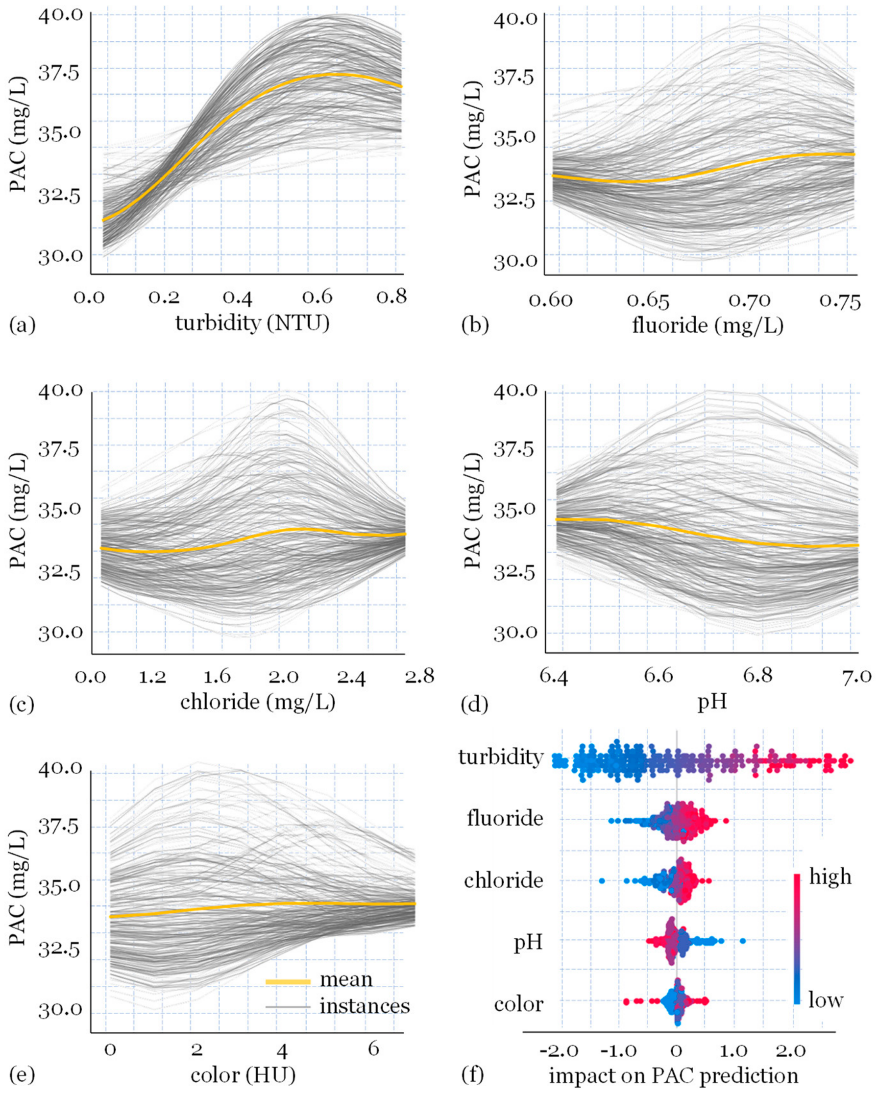
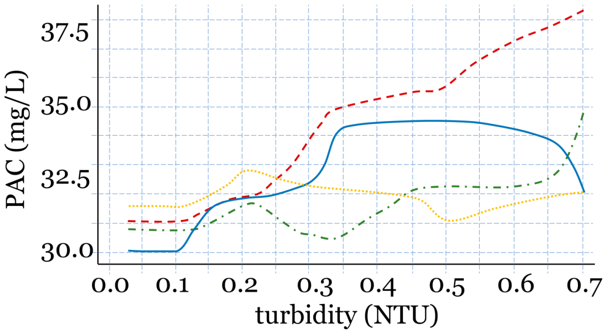


Figure 8. Individual conditional expectation and the impact of each water quality parameter on the D^{2}FIS prediction: (**a**) PAC × turbidity; (**b**) PAC × fluoride; (**c**) PAC × chloride; (**d**) PAC × pH, (**e**) PAC × color; and (**f**) water quality parameters × PAC.

In Figure 8a, it can be seen that PAC increases sharply with the increase in turbidity up to the vicinity of 0.6 NTU, but after this value, the coagulant dosage starts to decrease. In turn, there is also a directly proportional relationship between PAC and the color (Figure 8e) but with a much smaller variation in the coagulant dosage compared to that caused by turbidity. In all other cases, there are multiple inflection points in the curve of the mean (yellow line) of PAC, suggesting that the increase or decrease in coagulant dosage due to fluoride (b), chloride (c), and pH (d) depends on the range of variation in these parameters. For example, the relationship between chloride and PAC is inverse up to a concentration of 1.2 mg/L, then it becomes directly proportional from this concentration up to 2.0 mg/L; between this last value and 2.6 mg/L, it becomes inverse again and then changes to an upward trend (Figure 8c). Figure 8f shows the increase in the prediction error if the relationship between the parameter and PAC is broken. In short, a higher deviation from the center of the graph means that the parameter has a bigger impact on the prediction. The blue color represents a lower parameter value, whereas the red color is a higher value. Thereby, red dots to the right and left of the center indicate that the parameter tends to vary directly or is inversely proportional to PAC, respectively. In turn, the values in the horizontal axis measure how much the parameter impacts the predicted value (S_i) from the average prediction $(\overline{S})$.

In line with the ICE analysis, turbidity was the parameter with the greatest impact on the coagulant dose prediction $(-2.2 < (\overline{S} - S_i) < 3.1)$, followed by fluoride $(-1.1 < (\overline{S} - S_i) < 0.9)$, chloride $(-1.3 < (\overline{S} - S_i) < 0.6)$, pH $(-0.3 < (\overline{S} - S_i) < 1.2)$, and color $(-0.9 < (\overline{S} - S_i) < 0.5)$, in descending order of importance. These findings can contribute to decision-making on monitoring raw water quality as the most important parameters should be prioritized in the case of limitations that make it impossible to evaluate all of them periodically. It is worth noting that these appointments are based on the curve of the mean (yellow line) of PAC, which allows for analyzing the expected effect of each parameter. If we take into account the interdependent relationships between the water quality parameters, the effect on PAC dosage becomes even more complex (Figure 9).



	color (HU)	fluoride (mg/L)	chloride (mg/L)	pH
- - - -	3.97	0.70	1.82	6.91
——	2.91	0.65	1.33	6.53
-·-·-	3.74	0.67	2.12	6.81
········	1.05	0.68	1.67	6.41

Figure 9. Impact on PAC prediction considering interdependent relationships between the water quality parameters.

Figure 9 shows some scenarios with randomly selected values for each parameter. In the first case (red line), it can be seen that for each range of turbidity, the relationship with PAC is directly proportional. In the other cases (blue, green, and yellow lines), however, this relationship changes significantly depending on the variation of the other parameters (color, fluoride, chloride, and pH). This shows the complexity of the dependence of the dose PAC on the raw water quality, which complicates the mathematical formulation of mechanistic models and makes even more evident the importance of machine learning as an alternative to cope with this complexity.

Based on this deeper understanding of the importance of each water quality parameter for dosing PAC, new machine learning was performed for the model with the highest performance in the previous analyzes (D^2FIS), which included the following improvements: (i) reducing the dimensionality of the inputs by suppressing the color variable due to its lower impact on prediction; it is expected that the saved computational effort can be used in an optimized way; (ii) fine-tuning the partitioning of the input space of the other variables by increasing the number of membership functions (mfs) proportionally to the importance of each parameter, so that turbidity increases to 10 mfs, fluoride to 9, chloride to 8, and pH to 7 (originally, they all had 6 mfs). The other parameterization settings were not changed. These adjustments did not significantly improve the training error (from 0.44 to 0.42 mg/L). However, the generalizability of the D^2FIS increased significantly, with the error decreasing from 0.86 to 0.69 mg/L, a 19.8% improvement in accuracy. Among the recent studies reported in the literature, Achite et al. [21] compared different ML models for predicting coagulant dosage in WTPs. As a result, a combination of the M5 tree and gorilla troops optimizer models achieved an RMSE of 1.17 over the test dataset, which was up to 7, 10, 22, and 35% more accurate than the random forest (1.26), artificial neural network (1.30), multivariate adaptive regression splines (1.50), and k-nearest neighbor (1.81), respectively. Narges et al. [5], in turn, reported that the ANFIS achieved an RMSE value of 1.83 and concluded that this model is an excellent approach for determining the best PAC doses in WTPs. Therefore, to the best of our knowledge, the results of the present study indicate that the non-hybrid D^{2}FIS (0.69 RMSE) can be considered the most promising alternative evaluated so far as a computer-assisted tool for real-time, highly accurate coagulant dosing in WTPs.

4. Conclusions

In this study, experiments were conducted to test and compare the accuracy of several different ML algorithms, namely a data-driven fuzzy inference system, cascade-correlation network, gene expression programming, polynomial neural network, multilayer perceptron network, probabilistic neural network, radial basis function network, stochastic gradient boosting, and support vector machine for coagulant dosing of a drinking water treatment plant.

As the main contributions from this comparative analysis, it is worth highlighting (i) filling the gap with the more suitable ML method applied to the coagulation process; (ii) identifying a promising alternative for computer-aided coagulant dosing; and (iii) stimulating further studies to assess the potential of the D^{2}FIS for the control and optimization of other unit operations in drinking water treatment. From these findings, it was possible to confirm the research hypothesis that the D^{2}FIS presented the highest accuracy due to its ability to deal with uncertainties inherent to complex processes.

By constituting a solution based on non-linear functions with soft boundaries, which allows the measurement of partial memberships (uncertainties), the D^{2}FIS affords the best generalization ability and provides a highly accurate prediction. In conclusion, the accuracy of the D^{2}FIS-based alternative (0.86 error) outperformed the other assessed ML algorithms, including the ensemble models (1.25), ANNs (1.20), and kernel-based methods (1.17) widely used in regression tasks. In addition, a deeper understanding of the importance of each water quality parameter for dosing PAC improved the accuracy of the D^{2}FIS by nearly 20%, reducing the RMSE in the tests to 0.69 mg/L.

In conclusion, the non-hybrid data-driven fuzzy inference system can be considered a promising alternative tool for real-time and highly accurate coagulant dosing in drinking water treatment. The outcomes indicate that the D^{2}FIS can help WTPs to reduce operating costs, prevent errors associated with manual processes and operator experience, standardize efficacy, and enhance safety for the water industry.

Author Contributions: Conceptualization, A.B., J.K.S.F. and G.H.R.d.S.; data curation, R.G.N. and J.K.S.F.; formal analysis, A.B., A.P.G.G., I.G.G., A.I.S.L., R.G.N., R.M., C.P.M., R.F.T. and G.H.R.d.S.; funding acquisition, A.P.G.G. and G.H.R.d.S.; investigation, A.I.S.L., R.M., R.F.T. and A.G.d.R.; methodology, A.B., A.P.G.G., I.G.G., A.I.S.L., R.G.N., R.M., A.G.d.R., J.K.S.F. and C.P.M.; resources, A.B., A.G.d.R. and J.K.S.F.; software, A.P.G.G., I.G.G. and C.P.M.; supervision, A.B.; validation, R.M., C.P.M., R.F.T. and G.H.R.d.S.; writing—original draft, A.B., A.P.G.G., I.G.G., A.I.S.L., R.G.N., A.G.d.R., J.K.S.F., C.P.M. and G.H.R.d.S.; writing—review and editing, R.F.T. and R.M. All authors have read and agreed to the published version of the manuscript.

Funding: The São Paulo Research Foundation—FAPESP: grant number #2022/03675-8.

Data Availability Statement: The datasets generated and analyzed during the current study are available from the corresponding authors on reasonable request.

Conflicts of Interest: The authors declare that they have no known competing financial interest or personal relationship that could have appeared to influence the work reported in this paper.

References

1. Zhang, J.; Luo, Y. Multimodal control by variable-structure neural network modeling for coagulant dosing in water purification process. *Complexity* **2020**, *20*, 5392417. [CrossRef]
2. Wang, D.; Wu, J.; Deng, L.; Li, Z. A real-time optimization control method for coagulation process during drinking water treatment. *Nonlinear Dyn.* **2021**, *105*, 3271–3283. [CrossRef]
3. Menezes, F.C.; Fontes, R.M.; Oliveira-Esquerre, K.P.; Kalid, R. Application of uncertainty analysis of artificial neural networks for predicting coagulant and alkalizer dosages in a water treatment process. *Br. J. Chem. Eng.* **2018**, *35*, 1369–1381. [CrossRef]

4. Jayaweera, C.D.; Aziz, N. Development and comparison of extreme learning machine and multi-layer perceptron neural network models for predicting optimum coagulant dosage for water treatment. *J. Phys.* **2018**, *1123*, e012032. [CrossRef]
5. Narges, S.; Ghorban, A.; Hassan, K.; Mohammad, K. Prediction of the optimal dosage of coagulants in water treatment plants through developing models based on artificial neural network fuzzy inference system (ANFIS). *J. Environ. Health Sci. Eng.* **2021**, *19*, 1543–1553. [CrossRef]
6. Pandilov, Z.; Stojkov, M. Application of intelligent optimization tools in determination and control of dosing of flocculant in water treatment. *Int. J. Eng.* **2019**, *3*, 109–116.
7. Oliveira, A.S.; Lopes, V.S.; Coutinho Filho, U.; Moruzzi, R.B.; Oliveira, A.L. Neural network for fractal dimension evolution. *Water Sci. Technol.* **2018**, *78*, 795–802. [CrossRef]
8. Ghasemi, M.; Hasani Zonoozi, M.; Rezania, N.; Saadatpour, M. Predicting coagulation–flocculation process for turbidity removal from water using graphene oxide: A comparative study on ANN, SVR, ANFIS, and RSM models. *Environ. Sci. Pollut. Res.* **2022**, *29*, 72839–72852. [CrossRef] [PubMed]
9. Zhu, G.; Xiong, N.; Wang, C.; Zhongwu, L.; Hursthouse, A.S. Application of a new HMW framework derived ANN model for optimization of aquatic dissolved organic matter removal by coagulation. *Chemosphere* **2021**, *262*, 127723. [CrossRef]
10. Zangooei, Z.; Delnavaz, M.; Asadollahfardi, G. Prediction of coagulation and flocculation processes using ANN models and fuzzy regression. *Water Sci. Technol.* **2016**, *74*, 1296–1311. [CrossRef]
11. Kim, C.M.; Parnichkun, M. Prediction of settled water turbidity and optimal coagulant dosage in drinking water treatment plant using a hybrid model of k-means clustering and adaptive neuro-fuzzy inference system. *Appl. Water Sci.* **2017**, *7*, e3902. [CrossRef]
12. Zhang, Y.; Gao, X.; Smith, K.; Inial, G.; Liu, S.; Conil, L.B.; Pan, B. Integrating water quality and operation into prediction of water production in drinking water treatment plants by genetic algorithm enhanced artificial neural network. *Water Res.* **2019**, *164*, 114888. [CrossRef] [PubMed]
13. Kennedy, M.J.; Gandomi, A.H.; Miller, C.M. Coagulation modeling using artificial neural networks to predict both turbidity and DOM-PARAFAC component removal. *J. Environ. Chem. Eng.* **2015**, *3*, 2829–2838. [CrossRef]
14. Newhart, K.B.; Holloway, R.W.; Hering, A.S.; Cath, T.Y. Data-driven performance analyses of wastewater treatment plants: A review. *Water Res.* **2019**, *157*, 498–513. [CrossRef]
15. Najafzadeh, M.; Zeinolabedini, M. Prognostication of wastewater treatment plant performance using efficient soft computing models: An environmental evaluation. *Measurement* **2019**, *138*, 690–701. [CrossRef]
16. Ju, J.; Park, Y.; Choi, Y.; Lee, S. Comparison of statistical methods to predict fouling propensity of microfiltration membranes for drinking water treatment. *Desalination Water Treat.* **2019**, *143*, e716. [CrossRef]
17. Wu, G.D.; Lo, S.L. Predicting real-time coagulant dosage in water treatment by artificial neural networks and adaptive network-based fuzzy inference system. *Eng. Appl. Artif. Intell.* **2008**, *21*, 1189–1195. [CrossRef]
18. Robenson, A.; Shukor, S.A.; Aziz, N. Development of process inverse neural network model to determine the required alum dosage at Segama water treatment plant Sabah, Malaysia. *Comput. Aided Chem. Eng.* **2009**, *27*, 525–530.
19. Heddam, S.; Bermad, A.; Dechemi, N. Applications of radial-basis function and generalized regression neural networks for modeling of coagulant dosage in a drinking water-treatment plant: Comparative study. *J. Environ. Eng.* **2011**, *137*, 1209–1214. [CrossRef]
20. Heddam, S.; Dechemi, N. A new approach based on the dynamic evolving neural-fuzzy inference system (DENFIS) for modelling coagulant dosage (Dos): Case study of water treatment plant of Algeria. *Desalination Water Treat.* **2015**, *53*, 1045–1053.
21. Achite, M.; Farzin, S.; Elshaboury, N.; Valikhan Anaraki, M.; Amamra, M.; Toubal, A.K. Modeling the optimal dosage of coagulants in water treatment plants using various machine learning models. *Environ. Dev. Sustain.* **2022**, *65*, 1–27. [CrossRef]
22. Hernandez, H.; Lann, M.V. Development of a neural sensor for on-line prediction of coagulant dosage in a potable water treatment plant in the way of its diagnosis. *Iberamia Sbia* **2006**, *4140*, 249–257.
23. Zhang, K.; Achari, G.; Li, H.; Zargar, A.; Sadiq, R. Machine learning approaches to predict coagulant dosage in water treatment plants. *Int. J. Assur. Eng. Manag.* **2013**, *4*, 205–214. [CrossRef]
24. Zadeh, L.A. *Computing with Words*; Springer: Berlin/Heidelberg, Germany, 2012.
25. Barros, L.C.; Bassanezi, R.C.; Lodwick, W.A. *A First Course in Fuzzy Logic, Fuzzy Dynamical Systems, and Biomathematics*; Springer: Berlin/Heidelberg, Germany, 2017.
26. Zhang, J.; Qiu, H.; Li, X.; Niu, J.; Neyers, M.B.; Hu, X.; Phanikumar, M.S. Realtime nowcasting of microbiological water quality at recreational beaches: A wavelet and artificial neural network-based hybrid modeling approach. *Environ. Sci. Technol.* **2018**, *52*, 8446–8455. [CrossRef] [PubMed]
27. Mehryar, S.; Sliuzas, R.; Sharifi, A.; Reckien, D.; van Maarseveen, M. A structured participatory method to support policy option analysis in a social-ecological system. *J. Environ. Manag.* **2017**, *15*, 360–372. [CrossRef] [PubMed]
28. Bressane, A.; Biagolini, C.H.; Mochizuki, P.S.; Roveda, J.A.F.; Lourenço, R.W. Fuzzy-based methodological proposal for participatory diagnosis in linear parks management. *Ecol. Indic.* **2017**, *80*, 153–162. [CrossRef]
29. Caniani, D.; Labella, A.; Lioi, D.S.; Mancini, I.M.; Masi, S. Habitat ecological integrity and environmental impact assessment of anthropic activities: A GIS-based fuzzy logic model for sites of high biodiversity conservation interest. *Ecol. Indic.* **2016**, *31*, 238–249. [CrossRef]

30. Bressane, A.; Silva, P.M.; Fiore, F.A.; Carra, T.A.; Ewbank, H.; De-carli, B.P.; Mota, M.T. Fuzzy-based computational intelligence to support screening decision in environmental impact assessment: A complementary tool for a case-by-case project appraisal. *Environ. Impact Assess. Rev.* **2020**, *85*, e106446. [CrossRef]
31. Bressane, A.; Fengler, F.H.; Roveda, J.A.F.; Roveda, S.R.M.M.; Martins, A.C.G. Arboreal identification supported by fuzzy modeling for trunk texture recognition. *Trends Appl. Comput. Math.* **2018**, *19*, 111–126. [CrossRef]
32. Zhang, H.; Sun, T.; Shao, D.; Yang, W. Fuzzy logic method for evaluating habitat suitability in an estuary affected by land reclamation. *Wetlands* **2016**, *36*, 19–30. [CrossRef]
33. Brazilian Institute of Geography and Statistics. Sorocaba City. 2022. Available online: https://www.ibge.gov.br/cidades-e-estados/sp/sorocaba.html (accessed on 11 September 2022).
34. Santinon, E. Drinking Water Treatment Plant Dr. Armando Pannunzio at Sorocaba City, São Paulo State, Brazil. 2022. Available online: https://noticias.sorocaba.sp.gov.br/saae-sorocaba-realiza-manutencao-preventiva-na-eta-cerrado-neste-domingo-21/ (accessed on 11 September 2022).
35. Guillaume, S.; Charnomordic, B.; Lablée, J.; Jones, H.; Desperben, L. FisPro: Fuzzy Inference System, Design and Optimization. R Package Version 1.1.1. 2022. Available online: https://CRAN.R-project.org/package=FisPro (accessed on 11 September 2022).
36. Alvarez-Estevez, D.; Moret-Bonillo, V. Revisiting the Wang–Mendel algorithm for fuzzy classification. *Expert. Syst.* **2018**, *35*, e12268. [CrossRef]
37. Zhai, Y.; Lv, Z.; Zhao, J.; Wang, W.; Leung, H. Data-driven inference modeling based on an on-line Wang-Mendel fuzzy approach. *Inf. Sci.* **2021**, *551*, 113–127. [CrossRef]
38. Bressane, A.; Gomes, I.G.; da Rosa, G.C.S.; Brandelik, C.C.M.; Silva, M.B.; Siminski, A.; Negri, R.G. Computer-aided classification of successional stage in subtropical Atlantic Forest: A proposal based on fuzzy artificial intelligence. *Environ. Monit. Assess.* **2023**, *195*, e184. [CrossRef]
39. Hodson, T.O. Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not. *Geosci. Model Dev.* **2022**, *15*, 5481–5487. [CrossRef]
40. Yu, T.; Zhu, H. Hyper-parameter optimization: A review of algorithms and applications. *Comput. Sci.* **2022**, *in press*.
41. Nguyen, Q.H.; Ly, H.B.; Ho, L.S.; Al-Ansari, N.; Le, H.V.; Tran, V.Q.; Pham, B.T. Influence of data splitting on performance of machine learning models in prediction of shear strength of soil. *Math. Probl. Eng.* **2021**, *2021*, 4832864. [CrossRef]
42. Scikit-Learn Developers. Cross-Validation: Evaluating Performance. 2022. Available online: https://scikit-learn.org/stable/modules/cross_validation.html (accessed on 11 September 2022).
43. Negri, R.G. *Pattern Recognition: A Directed Study*; Edgard Blucher: São Paulo, Brazil, 2021.
44. Bressane, A.; Spalding, M.; Zwirn, D.; Loureiro, A.I.S.; Bankole, A.O.; Negri, R.G.; Junior, I.D.B.; Formiga, J.K.S.; Medeiros, L.C.D.C.; Bortolozo, L.A.P.; et al. Fuzzy artificial intelligence—Based model proposal to forecast student performance and retention risk in engineering education: An alternative for handling with small data. *Sustainability* **2022**, *14*, 14071. [CrossRef]
45. Wei, Y.; Ding, J.; Yang, S.; Yang, X.; Wang, F. Comparisons of random forest and stochastic gradient treeboost algorithms for mapping soil electrical conductivity with multiple subsets using Landsat OLI and DEM/GIS-based data at a type oasis in Xinjiang, China. *Eur. J. Remote Sens.* **2021**, *54*, 158–181. [CrossRef]
46. Mohamed, S.M.; Mohamed, M.H.; Farghally, M.F. A new cascade-correlation growing deep learning neural network algorithm. *Algorithms* **2021**, *14*, 158. [CrossRef]
47. Wadkar, D.V.; Karale, R.S.; Wagh, M.P. Application of cascade feed forward neural network to predict coagulant dose. *J. Appl. Water Eng. Res.* **2022**, *10*, 87–100. [CrossRef]
48. Kalantar, B.; Pradhan, B.; Naghibi, S.A.; Alireza, M.; Mansor, S. Assessment of the effects of training data selection on the landslide susceptibility mapping: A comparison between support vector machine (SVM), logistic regression (LR) and artificial neural networks (ANN). *Geomat. Nat. Hazards Risk* **2018**, *9*, 49–69. [CrossRef]
49. Salleh, M.N.M.; Talpur, N.; Hussain, K. Adaptive neuro-fuzzy inference system: Overview, strengths, limitations, and solutions. In *International Conference on Data Mining and Big Data*; Springer: Cham, Switzerland, 2017; pp. 527–535.
50. Khameneh, A.Z.; Kiliçman, A.; Salleh, A.R. Fuzzy soft boundary. *Ann. Fuzzy Math. Inform.* **2014**, *8*, 687–703.
51. Hussain, S. On some properties of intuitionistic fuzzy soft boundary. *Commun. Fac. Sci. Univ. Ank. Ser. Math. Stat.* **2020**, *69*, 1033–1044. [CrossRef]
52. Ghodousian, A.; Naeeimi, M.; Babalhavaeji, A. Nonlinear optimization problem subjected to fuzzy relational equations defined by Dubois-Prade family of t-norms. *Comput. Ind. Eng.* **2018**, *119*, 167–180. [CrossRef]
53. Naresh, C.; Bose, P.S.C.; Rao, C.S.P. Artificial neural networks and adaptive neuro-fuzzy models for predicting WEDM machining responses of Nitinol alloy: Comparative study. *SN Appl. Sci.* **2020**, *2*, 314. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.