Journal of Environmental Chemical Engineering 12 (2024) 112530

Contents lists available at ScienceDirect

Journal of Environmental Chemical Engineering

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

Predictive modeling of *Enterococcus sp.* removal with limited data from different advanced oxidation processes: A machine learning approach

Pavel Pascacio [a,*], David J. Vicente [a], Fernando Salazar [a,c], Sonia Guerra-Rodríguez [b], Jorge Rodríguez-Chueca [a,b]

[a] International Centre for Numerical Methods in Engineering (CIMNE), Barcelona 08034, Spain
[b] Department of Industrial Chemical & Environmental Engineering, Escuela Técnica Superior de Ingenieros Industriales, Universidad Politécnica de Madrid, Madrid 28006, Spain
[c] Flumen Research Institute, Universitat Politècnica de Catalunya (UPC), Barcelona 08034, Spain

ARTICLE INFO

Keywords:
Advanced Oxidation Processes
Random Forest
Wastewater Treatment
Enterococcus sp.
Data Partition
Machine Learning

ABSTRACT

The removal of contaminants through Advanced Oxidation Processes (AOPs) is a complex task that demands the simultaneous consideration of multiple operating parameters, such as type and concentration of oxidant and catalyst, type and intensity of radiation, composition of aqueous matrix, etc. Designing efficient AOPs often requires expensive and time-consuming laboratory experiments. To improve this process, this study proposes a Machine Learning approach based on a Random Forest (RF) model, to predict *Enterococcus sp.* concentration in wastewater treated with various AOPs, even when dealing with limited data. To assess our approach under diverse conditions, a data partitioning methodology is used to categorize the different AOPs into three distinct study cases of increasing complexity, from Case I to Case III. The evaluation of the RF model's performance, combined with the data partitioning methodology, demonstrated its usefulness in predicting missing or additional disinfection values at any instant during the AOPs. Specifically, in Case I, the model excels at generalizing predictions across various AOP *treatments*, followed by Case II and III, which achieve Root Mean Squared Error (RMSE) values below or comparable to the average RMSE of Case I (0.72) in 8 out of 15 and 2 out of 4 *treatments*, respectively. Moreover, the effects of imbalanced data on model performance are discussed. This highlights the potential of our approach to assess AOPs performance and facilitate the design of new experiments of the same *treatment* type without the need for additional laboratory trials, even in challenging conditions.

1. Introduction and background

Conventional disinfection treatments, namely chlorination or ozonation, have significant drawbacks such as the formation of disinfection by-products (DBPs) or the lack of residual effect [1–3]. For this reason, in recent years Advanced Oxidation Processes (AOPs) have emerged as an alternative series, although not all the AOPs ensure the non-formation of DBPs [4,5]. These processes can achieve the simultaneous removal of different types of contamination, simplifying the wastewater reclamation process and potentially reducing the time and cost of the process [6]. There is a broad range of AOPs, but all of them are based on the *in-situ* generation of highly reactive species [7,8] with enough oxidation capacity to degrade biological and organic pollutants. The most common oxidizing species generated during AOPs could be hydroxyl radicals (1.8–2.7 V), sulfate radicals (2.5–3.1 V), singlet oxygen (2.2 V), superoxide radical (2.4 V) or ozone (O_3) (2.1 V) [9]. Some of the most studied AOPs are based on the use of UV radiation. These photo-assisted AOPs can combine or not the use of different oxidants, namely hydrogen peroxide (H_2O_2), peroxymonosulfate (PMS), O_3, etc. [8,10,11], with different catalysts in homogeneous or heterogeneous phase in order to increase the kinetic removal of the pollutants [12,13]. Although many AOPs have been shown to be able to inactivate microorganisms and degrade a large variety of pollutants, under the conditions evaluated, they are not robust treatments, as the radical formation mechanisms depend on specific process parameters and can be greatly affected by system design and water quality [14,15]. Indeed, AOPs are highly dependent on water composition, as the presence of inorganic ions (e.g., Cl^-, CO_3^{2-},NO_3^-) and organic matter (e.g., humic acid, formic acid, citric acid) can promote or hinder treatment in actual wastewaters [16].

* Corresponding author.
E-mail address: ppascacio@cimne.upc.edu (P. Pascacio).

https://doi.org/10.1016/j.jece.2024.112530
Received 4 December 2023; Received in revised form 14 March 2024; Accepted 16 March 2024
Available online 19 March 2024

Table 1
Summary of state-of-the-art (SOTA) studies.

Work Ref.	Type of AOP	Contaminant	ML algorithm	Inputs	Outputs
Jiang et al. [22]	Photocatalyst (*Cat:* TiO_2)	78 Organic compounds	ANN	*Cat:* dosage; *Cont:* type, ini conc.; *Other:* pH, UV-C intensity	-log(*k*)
Smaali et al. [24]	Photocatalyst (*Oxid:* PPS; *Cat:* TiO_2, ZnO)	DFC	ANN + GA	*Oxid:* init conc.; *Cont:* init conc.; *Other:* pH	Y%
Zhang et al. [27]	CWAO, CWOP, WEO	COD, TOC	ANN	*Cat:* dosage; *Other:* pH, temp, press, time	Removal (time)
Cüce and Özçelik [21]	Fentonm Photofenton (*Oxid:* H_2O_2; *Cat:* Fe^{2+}))	COD	FF-NN,RBF-NN, SVM	*Oxid:* dosage; *Cat:* dosage; *Oxid/Cat* ratio; *Other:* pH, UV-C (# lamps, intensity), time	Removal (time)
Navidpour et al. [23]	Photocatalyst (*Oxid:* APS, NAPS, PMS; *Cat:*18 types)	PFOA	AdaBoost, GBR, RF	*Cont:* init conc.; *Other:* pH, UV-C (intensity, wavelength), temp, time	Removal (time)
Le et al. [20]	Fenton (Oxid:H_2O_2)	RB19	RF	*Oxid:* dosage; *Other:* pH, temp, time	Removal (time)
Sun et al. [26]	EO	2, 4-DCP	MLR, ANN, RF, SVM, XGBoost	*Cont:* init conc., 15 additional properties (quatum descriptors); *Other:* electrodes (type, area), current density, electrode gap, reactor volume, electrolyt (type, init conc., temp, pH)	*k*
Wang et al. [39]	Photo-Fenton with ultrasound	TC	ANN, GBR, KNN, XGBoost	*Cat:* dosage; *Other:* Fc, temp, time, ultrasound time, rpm	*k*
Zhang et al. [28]	Fe-based catalyst PMS activation	Organic compounds (number not specified)	ANN, RF, SVM	*Oxid:* dosage; *Cat:* dosage, 17additional properties; *Cont:* init conc.,	*k*
This work	diverse AOPs	*Enterococcus sp.*	RF	*Oxid:* dosage, type; *Cat:* dosage, type; *Other:* UVA (yes/no), time, type of aqueous matrix	Removal (time)

Notes:
- Type of AOP – Oxid: oxidant agent; Cat: catalyst; APS: mmonium persulfate; CWAO: catalytic wet air oxidation process; CWOP: catalytic wet peroxide oxidation process; EO: electrochemical oxidation; NaPS: sodium persulfate; PPS: sodium persulfate; TiO_2: titanium dioxide; WEO: wet electrocatalytic oxidation process; ZnO: zinc oxide
- Contaminant – 2,4-DCP: 2,4-dichlorophenol; COD: Chemical oxygen demand; DFC: Diclofenac; PFOA: Perfluorooctanoic acid; RB19: reactive blue 19; TC: tetracycline; TOC: total organic carbon
- ML algorithms – AdaBoost: Adaptive Boosting; ANN: Artificial Neural Network; FF-NN: Feed Forward Neural Network; GA: Genetic Algorithm; GBR: Gradient Boosted Regressor; MLR: Multiple Linear Regressor; RBF-NN: Radial Basis Function; Neural Network; RF: Random Forest; SVM: Support Vector Machine; XGBoost: Extreme Gradient Boost
- Inputs – Oxid: oxidant agent-based parameters; Cat: Catalyst-based parameters models); Cont: Contaminant-based parameters; Other: other inputs not included in the former categories; Fc: ferrocene nanoparticles amount; init. conc.: initial concentration, ph: initial ph; temp: reaction temperature; press: oxygen partial pressure; rpm: revolutions per minute; t: reaction time; UV: ultraviolet
- Output – k: pseudo-first order rate constant; Y%: percentage degradation ratio of contaminant "Y"; Removal (time): Contaminant removal over time.

Therefore, AOPs may involve many variables, including radical generation factors, namely type and concentration of oxidant concentration, radiation presence/type/intensity, type and concentration of catalyst, the composition of the wastewater, or pollutant characteristics. Experimentally finding optimal treatment conditions is costly and time-consuming, considering the numerous involved variables and time needed for analyses. To address this, treatment efficiency prediction methods are proposed, with growing popularity of artificial intelligence, particularly Machine Learning (ML). ML involves developing programs that learn from data, making it useful for scenarios with non-linear relationships or missing information in datasets [17,18].

In ML, it is possible to differentiate between supervised and unsupervised learning, with supervised learning being the technique used to infer a function from training data. The training data consists of pairs of data: on the one hand, the input data (or independent variables) are available and, on the other hand, the results to be predicted by the model (dependent variables) are entered. Supervised learning algorithms are often applied to the processing of previously labelled data to predict either the values of a continuous set through regression or the category of a discrete set through classification [17]. Once the algorithm has been trained, the goal is to have a model capable of generalising the identified relationships to make predictions on new data. The performance of prediction models is affected by many factors, such as the number of available data, the particular characteristics of these data, the number of iterations of the training process, the pre-treatment of the data or the bias derived from the selection of one or another algorithm [19].

To achieve reliable performance of the models obtained, experience during model building is necessary to allow understanding of the behaviour of the data as well as the nature of the experimental process. Therefore, the selection of appropriate tools for ML model training, testing and validation is a challenging task [19].

In recent years, there has been a significant increase in the number of articles related to the use of ML-based approaches in the context of AOPs. Some of the more relevant articles are summarized in Table 1, providing an overview of those findings through five key dimensions: type of AOP, target contaminant, ML algorithm, inputs, and outputs selected in each research. However, it is important to note that a wide diversity along these dimensions characterizes the findings from one research to another, such that the diversity along dimensions does not allow a straightforward comparison among the reported results in the studies. For example, AOPs are used based on various methods, including Fenton reactions [20], photo-Fenton [21], photo-catalysis [22–25] or electro oxidation-based reactions [26,27], among others.

Regarding the type of contaminants, the studies collected in Table 1 have targeted organic compounds such as pesticides [25], PFOAs [23], as well as other compounds such as phenols, antibiotics, and dyes [20, 26,28]. Despite the diversity of contaminants examined, it is notable the absence of studies focusing on the inactivation of pathogenic microorganisms (e.g., *E. coli*, *Enterococcus sp.*, etc.) through AOPs.

The selection and diversity of ML algorithms in each study have a great dependency on the specific goals of the studies, as well as other factors such as input data available, output requirements, type of contaminant, and data dimensionality. The implemented ML algorithms include Artificial Neural Network (ANN), Bayesian Network (BN),

Decision Tree (DT), Gradient Boosting Decision Tree (GB), *k*-Nearest Neighbours (*k*-NN), Multiple Linear Regressor (MLR), RF, Response Surface Method (RSM), Support Vector Machine (SVM), and Extreme Gradient Boost (XGBoost). Among the various ML algorithms available, we have chosen the RF algorithm in our study due its advantages such as low computational complexity, robust predictive capabilities for unseen data (generalization), and automatic feature importance selection during training [29]. Furthermore, RF has been successfully implemented in studies related to AOP [20,23,26] and other fields [29–32].

Similarly, the selection of inputs in each study is highly influenced by the research goals, data availability, and specific dimensions such as the type of AOP, ML algorithm, and contaminant, as evidenced by the diverse range of inputs presented in Table 1 for each study.

The last dimension corresponds to the outputs, which classifies the prediction conducted into two categories. The first one comprises studies in which the output is the predicted organic contaminant removal or microbial inactivation at any instant of the process. For instance, "reaction time" - or other analogous time parameter – is introduced as an input variable of the model [21,23,27]. The second category refers to studies in which the output is the reaction kinetic constant, typically represented as the pseudo-first order constant k, whereby no values of the degradation/inactivation process throughout time is considered [22,24,26,28].

Expanding the scope of our search, regardless of the wastewater treatment used, no studies employing predictive ML models for the inactivation of our target pollutant, *Enterococcus sp.*, were identified. Only by further extending the search to approaches not based on AOP and considering other pathogens, we find a limited number of papers that apply ML techniques to predict their inactivation. The target pathogens of the referred works are the following: faecal or total coliforms [33–35], *Pseudomonas sp.* [36], *Bacillus subtilis*, or *Staphylococcus aureus* [37]. *Enterococcus sp.* is one of the best indicators of faecal contamination, appearing in high concentration in all urban wastewater, and its selection as the target contaminant to test the efficiency of disinfection treatments is preferable over other indicator pathogens such as *E. coli*, mainly due to its greater resistance to treatments as a result of the different composition of the cell membrane of Gram-positive bacteria compared to Gram-negative ones [38].

To the best of our knowledge, there are currently no published research papers that specifically address the prediction of disinfection results using multiple AOPs and when faced with limited data samples. In this work, we present a Random Forest (RF) model to predict *Enterococcus sp.* concentration in wastewater, and in conjunction with a data partitioning methodology based on study cases, we address the challenge of limited data samples and enhance model generalisation by leveraging various AOPs. This article, therefore, focuses on the following key aspects: i) assessing the potential of the RF models to predict the disinfection level of *Enterococcus sp.* concentration through different AOPs; ii) investigating the impact of sample size and the diversity of AOPs on the performance of the models, and iii) evaluating the performance of the former RF models under three distinct study cases, each varying in prediction complexity to mimic diverse real-world laboratory needs and challenges encountered during the AOP design.

2. Materials and methods

2.1. Description of experimental raw data

The raw data used in this study corresponds to the disinfection results of *Enterococcus sp.* based on the application of different AOPs at laboratory scale, which have been partially published by Guerra-Rodríguez et al. [10,40] with other purposes. The raw data contains a total of 580 samples, corresponding to 88 different experiments in which concentration values of this biological pollutant were taken over different time periods for each experiment. The parameters or variables that defined each experiment were the following: type and concentration of oxidant (i.e., non-oxidant, peroxymonosulfate (PMS), H_2O_2, or sodium sulfite); type and concentration of catalyst (i.e., non-catalyst, $FeSO_4$ or Fe(III)-citrate); type of aqueous matrix (Distilled Water (DW), Saline Water (SW) or Simulated Wastewater (SWW)), and presence or absence of UV radiation. A detailed overview of the distribution of raw data samples, organized by experiments' parameters and their interconnections, is presented in the Sankey plot shown in Fig. 3. This visualization, along with the percentage and number of samples indicated in each group, allows us to summarize and understand the dimensions of the experimental dataset and its correlation with key parameters. Additionally, a summary of additional parameters used in the experiments, such as concentrations of oxidant and catalyst, grouped by treatments is provided in the Supplementary material (Table S1). The experimental raw data to this article is available on request.

2.2. Study cases

In our study, we propose three distinct study cases to achieve our primary objectives: i) assessing the potential of the RF models to predict the disinfection level of *Enterococcus sp.* concentration through different AOPs; ii) investigating the impact of sample size and the diversity of AOPs on the performance of the models, and iii) evaluating the performance of the former RF models under three distinct study cases, each varying in prediction complexity to mimic diverse real-world laboratory needs and challenges encountered during the AOP design. Each of these study cases is built upon the experimental data detailed in Section 2.1, which have been organized using a data partition methodology. To clarify our data organization, we divided the data samples based on treatments and experiments. On one hand, a *treatment* is defined as an AOP with a unique combination of parameters (e.g., aqueous matrix, radiation, oxidant, and catalyst), which may include various concentrations of these parameters. On the other hand, an *experiment* refers to a specific case of *treatment* with identical parameter configuration (identical level of concentrations). According to these definitions, the aim and description of each case are as follows:

Case I. *"Randomly data division (General prediction case)":* Case I, aims to predict the level of disinfection in the *treatments* to provide missing or additional values at any instant of the process. The ML model is trained and tested with 75% and 25% of samples randomly extracted from all the experimental data, respectively. With this approach, most of the training samples will share high-similarity conditions with the ones to be predicted. The completeness of missing values or adding new values in the *experiments* can be useful for more accurate computation of process parameters (e.g., kinetic constant) and provide information for decision-making during the process (e.g., to determine where the process must be completed).

Case II. *"Unseen experiments (Prediction based on different experiments)":* Case II, aims to predict the complete disinfection behaviour of individual *experiments* – i.e., all the levels of disinfection throughout the time of an *experiment*. The ML model is trained with data from previously conducted *treatments* where at least one *treatment* shares similarities in the type of oxidant, catalyst, radiation, system design, and type of aqueous matrix, but not in their concentrations compared to the ones being predicted. The predictions made in this case could be valuable for determining the optimal concentrations of oxidants and catalysts and predicting their effectiveness without the need for new *experiments*, as long as similar *treatments* have already been conducted.

Case III. *"Unseen treatments (Prediction based on completely different treatments)":* Case III aims to predict the complete disinfection behaviour of new *treatments*. The ML model is trained with data from previously conducted *treatments* that do not resemble the new ones being predicted. For instance, a *treatment* being predicted shares samples with the same type of oxidant, radiation, and aqueous matrix, with at least one *treatment* contained in the set of previous *treatments*, but its type of

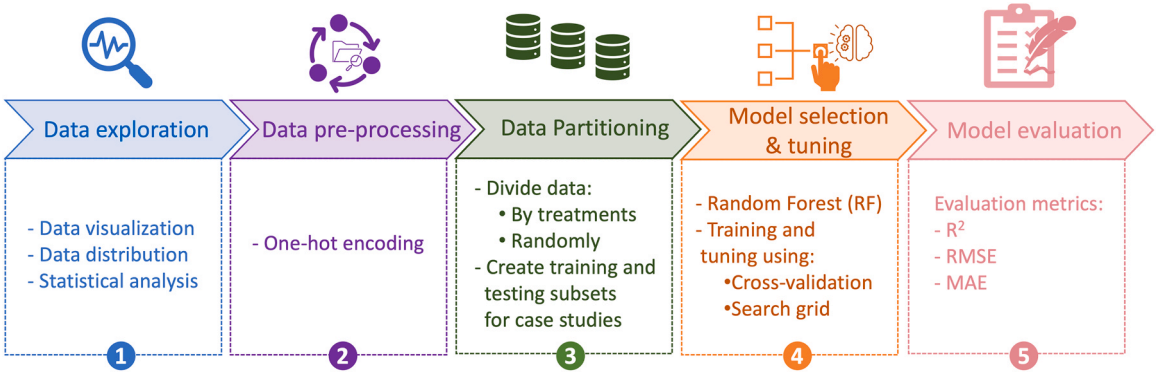


Fig. 1. Flowchart of methodology implemented in this study.

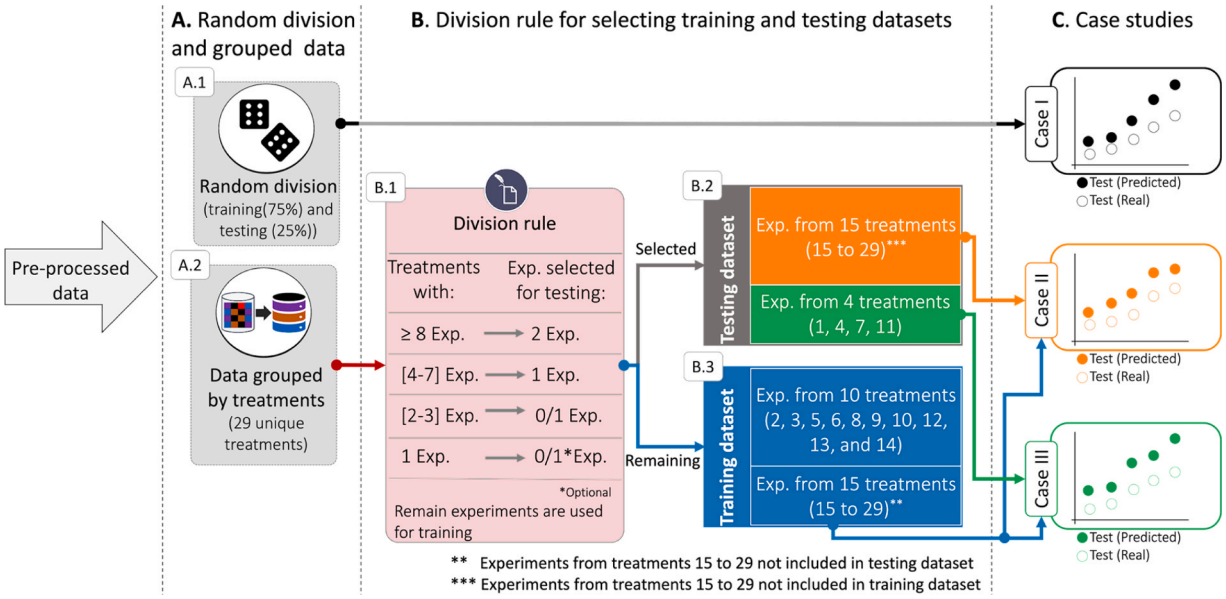


Fig. 2. Workflow diagram of the data partitioning methodology.

catalyst varies. This case presents one of the most challenging scenarios, as only a few similarities are found between the new *treatment* and a single *treatment* of the set of previous *treatments* used for prediction. Nonetheless, this approach enables us to explore new combinations of parameters and concentrations to design innovative *experiments*.

In the following sections, the methodology, and software and libraries used to evaluate the study cases are described.

2.3. Methodology

We proposed a five-step methodology to explore and analyse the experimental raw data, select the best ML model's hyperparameters for predicting the concentration of contaminant (i.e., *Enterococcus sp.*), and evaluate its performance. The first step corresponds to data exploration; the second is dedicated to data pre-processing; the third involves the implementation of a data partitioning methodology to build the training and testing datasets according to the case studies defined previously; the fourth focuses on the selection of the model, training the model and refinement of its hyperparameters, and the last step evaluates the model's performance. It should be noted that the data partitioning methodology used in our study cases plays a crucial role in exploring the real-world application of the proposed ML model and effectively addressing the challenge of limited available data. Fig. 1 presents the flowchart of our proposed methodology, and the following sections describe each of its steps in detail.

2.3.1. Step 1: Data exploration

Given the diversity of experiments conducted to retrieve raw data, data exploration becomes essential for understanding its main characteristics [41]. This initial step allows us to assess its dimensionality, diversity, distribution, and completeness. Furthermore, it helps us identify any misplaced, nonessential, or duplicated values in the dataset that could impact subsequent analysis. To this end, we used a Sankey diagram to graphically evaluate the dataset's distribution and dimensionality. Additionally, we utilized box and bar plots to identify missing and nonessential values considering the reduction of the sample size in the dataset's features.

2.3.2. Step 2: Data pre-processing

Data pre-processing aims to address the identified disadvantages of raw data, ensuring the dataset has an appropriate structure, data type, and completeness for feeding and training the ML model. The experimental study conducted in the AOP encompassed the investigation of various types of catalysts and oxidants, along with their concentrations, type of aqueous matrix, and radiation conditions. This results in in both categorical (e.g., name of the oxidants and catalysts) and numerical (e.g., time, oxidants' concentration) data. While many existing ML techniques can handle numerical data in their algorithms, most of them are unable to directly work with categorical data [42]. Therefore, to

effectively apply ML, it is necessary to encode/transform categorical data into a numerical format [43]. Among the most relevant encoding data techniques, one-hot encoding stands out as one of the most popular approaches [44]. Consequently, in our data pre-processing step, we utilized the one-hot encoding technique provided by the Scikit-Learn library [42] to encode the categorical data. Additionally, we performed a data-cleaning process to remove rows of data that contain a fixed output (i.e., rows at experimental time zero), since these features provide a constant target value (i.e., zero level of disinfection) that is expected at the beginning of the AOP treatments and are not necessary for prediction purposes.

2.3.3. Step 3: Data partitioning methodology for the case studies

To explore the feasibility and implement the ML model for each study cases described in Section 2.2, a thoughtful partition of the pre-processed data into training and evaluation datasets is crucial. To maximize the use of experimental data available for each specific study case, we propose a data partitioning methodology. This methodology is based on a division rule that chooses particular *experiments* from the *treatments* saved to build the datasets. Fig. 2 illustrates the workflow diagram of the data partitioning methodology suggested. The workflow is summarized as follows:

- ***Phase A:*** This phase is fed by the pre-processed data and includes two main tasks. The first task involves the random data division, visually represented in Fig. 2 (A.1). In this task, the data is divided into two sets: training (75%) and testing (25%) for use in Case I. The second task, depicted in Fig. 2 (A.2), groups the pre-processed data by unique *treatments*. A *treatment* is considered unique when there are no other *treatments* with the same combination of the type of aqueous matrix, radiation, oxidant, and catalyst. Within a unique *treatment*, one or more *experiments* may be available, which contain a diverse initial concentration of oxidant and/or catalyst. In the second task, the 88 *experiments* were classified into 29 unique *treatments*, which fed the ***Phase B***.
- ***Phase B:*** In this phase, we introduce a division rule-based procedure illustrated in Fig. 2 (B.1) for creating both the testing dataset (see Fig. 2 (B.2)) and the training dataset (see Fig. 2 (B.3)). These datasets are intended for use in Cases II and III. The division rule procedure takes as input the 29 unique *treatments* and counts the number of *experiments* within each of them to select the *experiments* to be included in the testing and training datasets based on a set of rules. The division rules are as follows:

1. If the treatment contains ≥ 8 *experiments*, randomly select 2 *experiments* for testing (Case II);
2. If the *treatment* contains [4 − 7] *experiments* randomly select 1 *experiment* for testing (Case II);
3. If the *treatment* contains [2−3] *experiments*, randomly select 0/1 *experiment* for testing (Case II), and
4. If the *treatment* contains 1 *experiment*, randomly select 0/1 *experiment* for testing (Case III).

Rules 3 and 4 allow for the optional selection of an *experiment* to be included in the testing dataset to maintain a 7:3 ratio for training and testing in the selection process. *Experiments* not included in the testing dataset are included in the training dataset. As a result, we have a testing dataset containing testing data for Cases II and III. For Case II, the testing dataset includes *experiments* from 15 *treatments* (15 to 19) that meet the criteria rules 1, 2, and 3. For Case III, the testing dataset includes *experiments* from 4 *treatments* (1, 4, 7, and 11) that fulfil rule 4.

Regarding the training dataset, which is used for training Cases II and III, it includes the remaining *experiments* not included in the selection made by rules 1, 2, 3, and 4. Specifically, it includes *experiments* from 15 *treatments* (15 to 19), which also have counterparts in the testing dataset, and *experiments* from 10 *treatments* (2, 3, 5, 6, 8, 9, 10, 12, 13, and, 14)

- ***Phase C:*** In this phase, we link the datasets created during ***Phase A*** and ***B*** with the case studies, as illustrated in Fig. 2 (C).

For Case I, we utilize both the training dataset (comprising 75%) and the testing dataset (comprising 25%), represented by the black line in Fig. 2. The training dataset is used as input for training the RF model, while the testing dataset is employed to evaluate its performance.

For Cases II and III, the whole training dataset (see Fig. 2 (B.3)) is used for training a unique RF model. Nevertheless, the performance of the RF model in Cases II and III is evaluated independently. In Case II, the testing dataset encompasses *experiments* from 15 *treatments*, denoted by the orange block in Fig. 2 (B.2) was used to evaluate the RF model performance. Meanwhile, in Case III, the testing dataset comprises *experiments* from 4 *treatments*, as represented by the green block in Fig. 2 (B.2).

A detailed description of the tuning and evaluation procedure of the RF models used in the case studies is presented in Sections 2.3.4 and 2.3.5.

2.3.4. Step 4: Model selection and tuning

The fourth step focuses on the selection of a suitable ML model for predicting contaminant concentration under the described conditions. We chose the RF algorithm, which is one of the more often used in different fields, due to its advantages in both performance and implementation. Its main features include low computational complexity, strong predictive capabilities for unseen data (generalization) and automatic feature importance selection during training [29,45,46]. Its ability to deliver acceptable performance even with default hyperparameters, coupled with a straightforward hyperparameter setting, contributes to its popularity [30]. Furthermore, RF has been successfully implemented in studies related to AOP [20,23,26] and other fields [29–32]. The RF method is a natural evolution of the decision tree method that seeks to eliminate the tendency of the latter to overfit. Specifically, the RF is an ensemble method, composed of multiple DTs, in which each tree is slightly different from the others. Each tree can do a relatively good prediction job but is likely to overfit a part of the data [47]. By constructing many trees, it is possible to reduce the amount of overfitting by averaging their results. Therefore, in this method, each of the trees takes a random sample of the input data to generate a prediction result independently [17]. The predictions are then averaged when the data are quantitative or used for a vote for qualitative data [18]. The number of trees generated in this method is a parameter to be determined and can be several tens or even hundreds.

The optimal hyperparameters for the RF model were determined using a grid search and cross-validation techniques. In this process, we systematically explored a predefined range of hyperparameter values to identify the combination that yielded the best model performance. The parameters included in the grid search were: *number of estimators* (10, 50, 100, 200), *maximum number of features* (3, 5, 7), and *minimum number of sample leafs* (1, 3, 5, 7). Cross-validation was employed to assess the model's performance with each set of hyperparameters. This technique involved dividing the dataset into multiple subsets, or folds, and training and evaluating the model on different combinations of training and validation sets. The number of folds used was 5. It should be noted that combining grid search and cross-validation techniques not only helps to select the best hyperparameters but also guarantees, through cross-validation, that the results remain independent of the partitioning between training and test data. For a comprehensive overview of the results obtained from the combination of hyperparameters, please refer to Supplementary material (Tables S2 and S3).

2.3.5. Step 5: Model evaluation

Within the main evaluation metrics used in regression algorithms to evaluate the performance and quality the RMSE, Mean Absolute Error (MAE), and coefficient of determination (R^2) are the most used [48]. Therefore, we use them in our RF model evaluation. Additionally, to

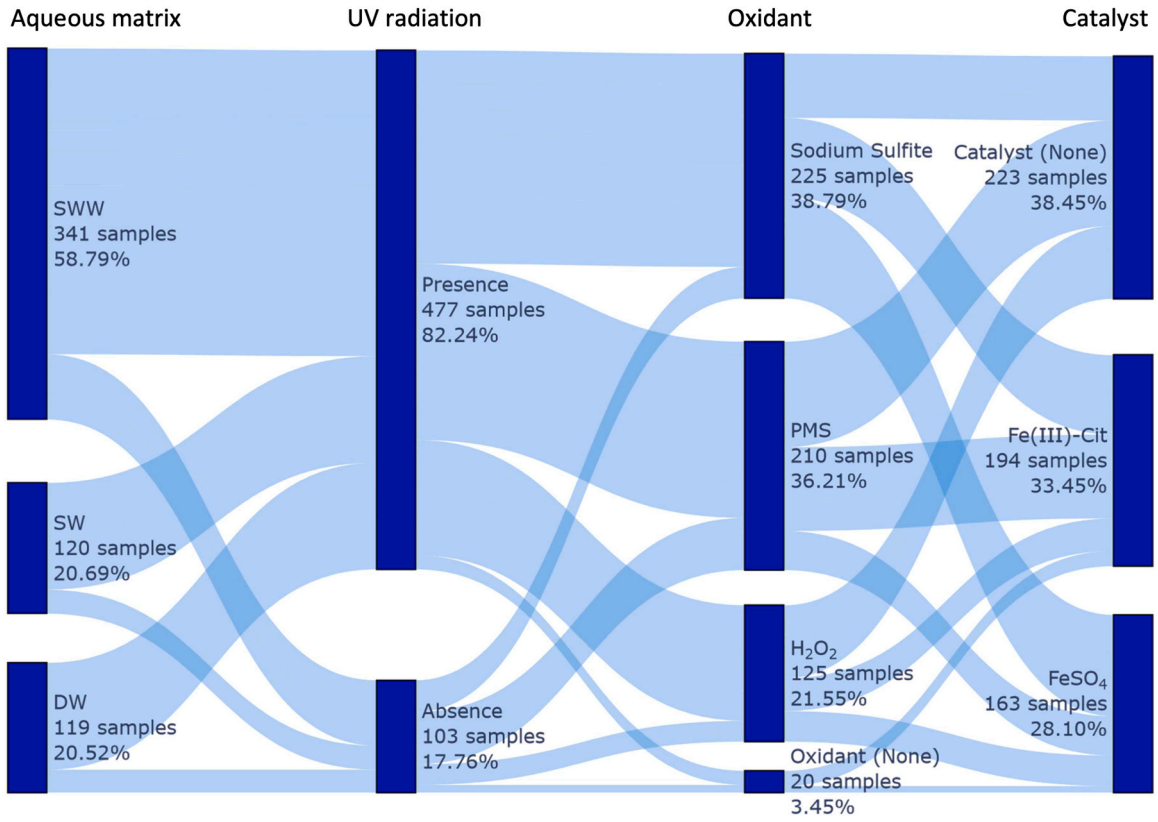
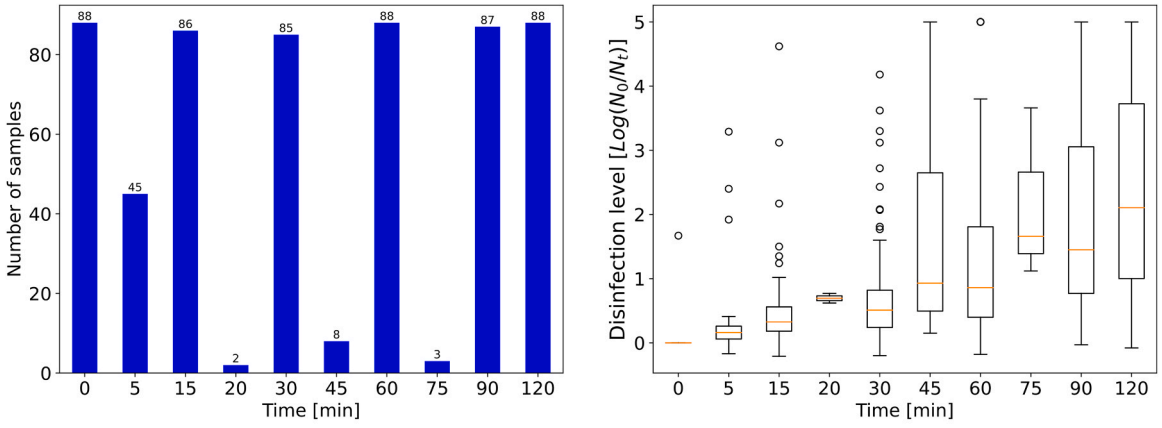


Fig. 3. Distribution of raw data samples grouped by type of aqueous matrix, UV radiation, oxidant, and catalyst.



a) Number of samples vs Time

b) Disinfection level vs Time

Fig. 4. Representation of raw data considering the number of samples (Bar plot) and disinfection level (Boxplot) at each recorded time.

evaluate the Confidence Interval (CI) of outcomes by case, the Mean Error (ME) metrics is considered. Those metrics are mathematically expressed by the Eqs. (1), (2), (3) and (4), respectively.

$$RMSE(y,\hat{y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \quad (1)$$

$$MAE(y,\hat{y}) = \frac{1}{n}\sum_{i=1}^{n}|(y_i-\hat{y}_i)| \quad (2)$$

$$R^2(y,\hat{y}) = 1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)}{\sum_{i=1}^{n}(y_i-\bar{y})} \quad (3)$$

$$ME(y,\hat{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i) \quad (4)$$

where n denotes the total of samples, y_i is the true value for the i^{th} sample, $\hat{y}_i$ is the prediction value of the i^{th} sample, and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

2.4. Software and libraries

In our study, we utilized the Python 3.7.4 programming language along with a set of libraries to carry out various tasks. Specifically, we utilized Scikit-learn [42] for implementing GridSearchCV, the RF model, pre-processing, data splitting, and model evaluation. Furthermore, for

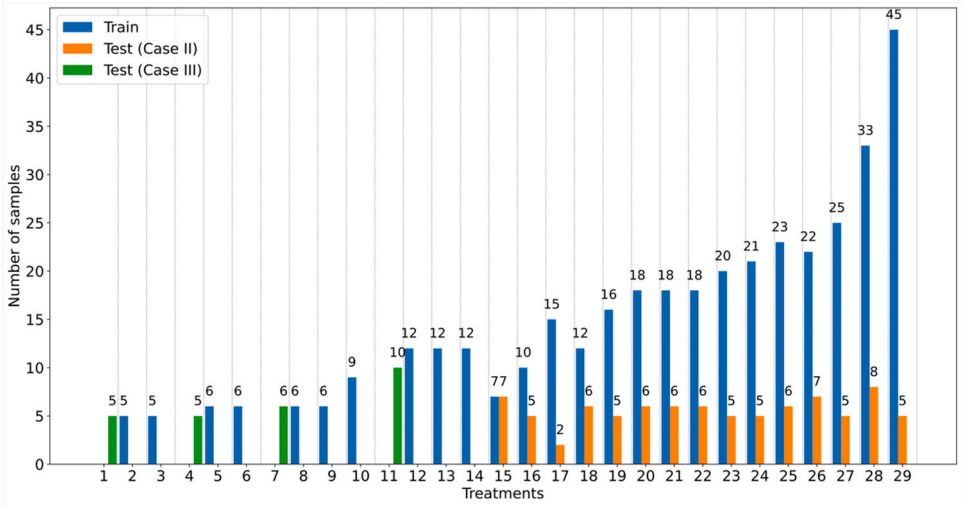


Fig. 5. Distribution of processed data (training and testing) used for Case II and III, divided by *treatment* and sample count.

Table 2
Summary of optimal hyperparameters used for the RF model in Case I and Cases II and III.

RF model	Max. Features	Min. Sample leaf	Num. Estimators	R^2 (train)
Case I	5	1	200	0.96
Cases II and III	7	3	200	0.8

data visualization, analysis & manipulation, and mathematical operations, we used Plotly [49], Pandas [50], and Numpy [51] libraries, respectively.

3. Results and discussion

In this section, we present and analyse the sample size and the diversity of the *treatments* in both raw and processed data to identify their main characteristics, which are useful for the correct implementation and evaluation of the RF model using the data partitioning methodology. Additionally, we present and discuss the main results of evaluating the three case studies described in Section 2.2 and provide insights into the implementation of the RF model to handle reduced sample sizes and diverse AOP treatments.

3.1. Exploratory analysis and description of the raw and processed data

To proceed with the data partitioning methodology of the experimental data, categorization and analysis of the parameters involved in the experimental phase are essential.

From left to right, the Sankey plot in Fig. 3 shows the type of aqueous matrix, UV radiation, oxidant, and catalyst considered in all the *experiments* performed. The raw data samples were grouped based on these experimental conditions, and the percentage and number of samples in each group are indicated alongside. As can be observed from Fig. 3, the experimental phase involved three aqueous matrices: "SWW" was the most prevalent (58.79%) and in a very lower percentage the "SW" and "DW" with a 20.69% and 20.52%, respectively. The majority of *experiments* were conducted in the "presence" of UV radiation (82.24%) while a smaller percentage occurred in its "absence" (17.76%). There were four oxidation conditions considered: "sulfite" and "PMS" were the most common at 38.79% and 36.21%, respectively, followed by "H_2O_2" (21.55%), and with 3.45% of *experiments* conducted without any oxidant. Regarding catalyst usage, two out of three conditions were

Table 3
Inputs and output considered for the ML model.

Parameters	Description
Type	Type of Oxidant is a categorical value with three possible options: - PMS - Sodium sulfite - H_2O_2
Concentration	Concentration of Oxidant is a numerical value that indicates the initial concentration of the type oxidant used
Type	Type of Catalyst is a categorical value with three possible options: - Not catalyst - Fe(III)-Cit - $FeSO_4$
Concentration	Concentration of catalyst is a numerical value that indicates the initial concentration of the type catalyst used
Type	Type of Aqueous matrix is a categorical parameter with three options: - DW - SWW - SW
Type	Type of UV radiation is a categorical parameter with two options: - Prescence - Absence
Collection	Time collection is a numerical value that refers to the moment when experimental data is collected
Level of disinfection	Level of contaminant disinfection is a numerical value that indicates the level of *Enterococcus sp.* disinfection at each time collection

nearly equally distributed, with 38.45% and 33.45% for *experiments* without catalyst and "Fe(III)-Cit", respectively, followed by "$FeSO_4$" at 28.10%.

Similarly, Fig. 4 provides a representation of raw data, offering insights into both the distribution of raw data samples collected at each time interval (Fig. 4(a)) and the

the variation in disinfection levels ($Log(N_0/N_t)$) over time (Fig. 4(b)). From the bar plot (Fig. 4(a)), we can observe a relatively consistent

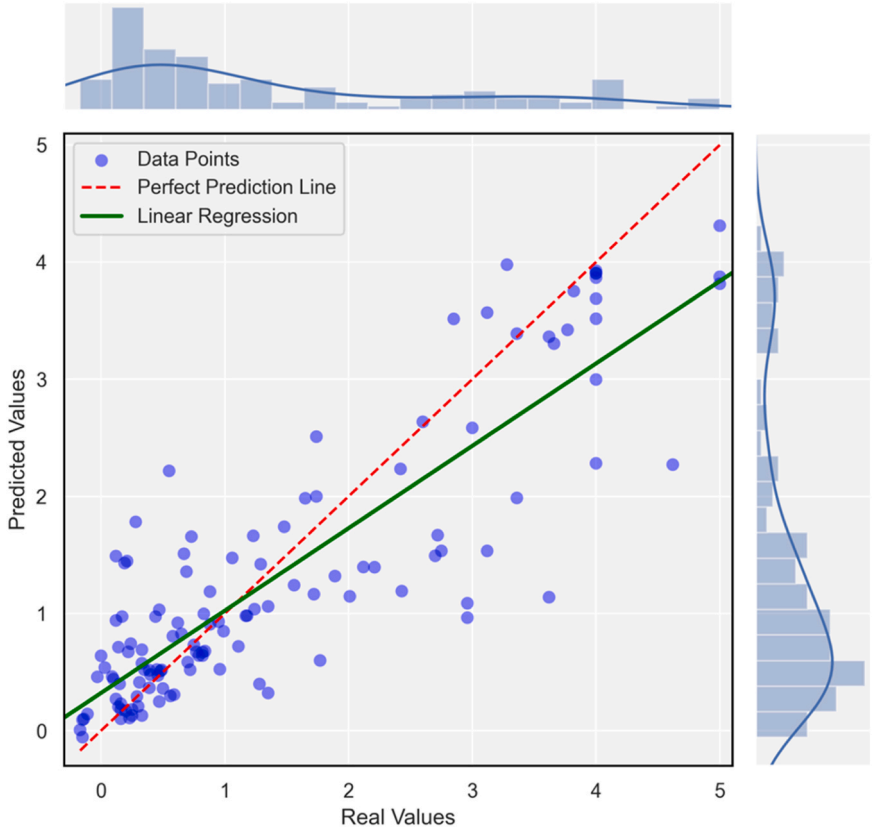


Fig. 6. Performance evaluation of RF model in Case I (testing dataset): Predicted vs. real values. Dotted red diagonal: Perfect prediction line. Upper side: Distribution of predicted values. Right side: Distribution of real values. Green line: Linear regression line.

distribution of samples across most time intervals, except for time points 5, 20, 45, and 75, which show fewer collected samples but can be compensated by the large number of samples of near time intervals. Regarding the box plot (Fig. 4(b)), there is a notable trend of increasing dispersion in disinfection levels as the time intervals increase, except at time zero.

Analysing the combination of parameters in Fig. 3, we observe that the *treatments* primarily involve the use of a "simulated" aqueous matrix, the presence of radiation, and the utilization of oxidants such as "sodium sulfite" or "PMS". Moreover, *treatments* without a catalyst or with the "Fe (III)-Cit" catalyst constitute the majority of the *treatments*. This imbalance in sample sizes makes it challenging to predict *treatments* based on a DW or SW aqueous matrix, the absence of radiation, and the presence of other oxidants and catalysts, which together represent a minority of the samples.

Therefore, we apply a thoughtful division of data, as described in Section 2.3.3, to exploit the diversity of the overall *treatments* and evaluate the real study case proposed. Before data division, based on the results of Fig. 4(b), the samples on time zero are removed, since they represent a fixed level of disinfection (zero $Log(N_0/N_t)$). In the data partitioning methodology, a critical challenge was formulating partition rules for Cases II and III, requiring the careful mapping of experimental conditions. This was crucial to ensure unbiased *experiment* selection for training in each *treatment* and maintain a balanced sample distribution for testing across all *treatments*, particularly considering the limited number of experiments per treatment.

In Case I, our objective is to provide missing or additional values for treatments rather than predicting complete treatments. To achieve that, we randomly divided the overall treatments samples into training (75%) and testing (25%) datasets, resulting in 369 and 123 samples of processed data, respectively. On the contrary, in Cases II and III, our focus shifts to predicting unseen experiments and treatment, respectively.

Fig. 5 illustrates the distribution of processed data used for training and testing the RF model in Cases II and III, categorized by *treatment* and the number of samples in accordance with the division rule proposed (see flow diagram in Fig. 2). The x-axis represents each *treatment*, while the y-axis indicates the number of samples in each *treatment* group. The blue, orange, and green bars correspond to the training and testing data for Cases II and III, respectively.

In the training and evaluation of the performance of the RF models, we considered seven input parameters: the type and concentration of oxidant, the type and concentration of catalyst, the aqueous matrix type, radiation type, time collection, and one output parameter —the level of disinfection of the pollutant (*Enterococcus sp.*). Table 3 provides a list and description of each of these parameters. Moreover, a detailed summary of each *treatment* used is provided in the Supplementary material (Table S1).

3.2. Evaluated cases

In Section 2.3, we described the data partitioning (training and testing) for Cases I to III, the modelling process, and hyperparameter tuning to assess the performance of RF models. As a result of hyperparameter tuning of the RF model using grid search and cross-validation, optimal hyperparameters were obtained for each case. Specifically, for Case I, we implemented an RF model with the following hyperparameter values: 200 for the number of estimators, 5 for maximum features, and 1 for the minimum sample leaf. The RF performance on the overall training data (without cross-validation) for Case I yielded an R^2 value of 0.96, indicating strong predictive capability. In Cases II and III, we implemented RF models with the following hyperparameter values: 200 for the number of estimators, 7 for maximum features, and 3 for the minimum sample leaf. The RF performance on the overall training data (without cross-validation) for Cases II and III resulted in an R^2 value of

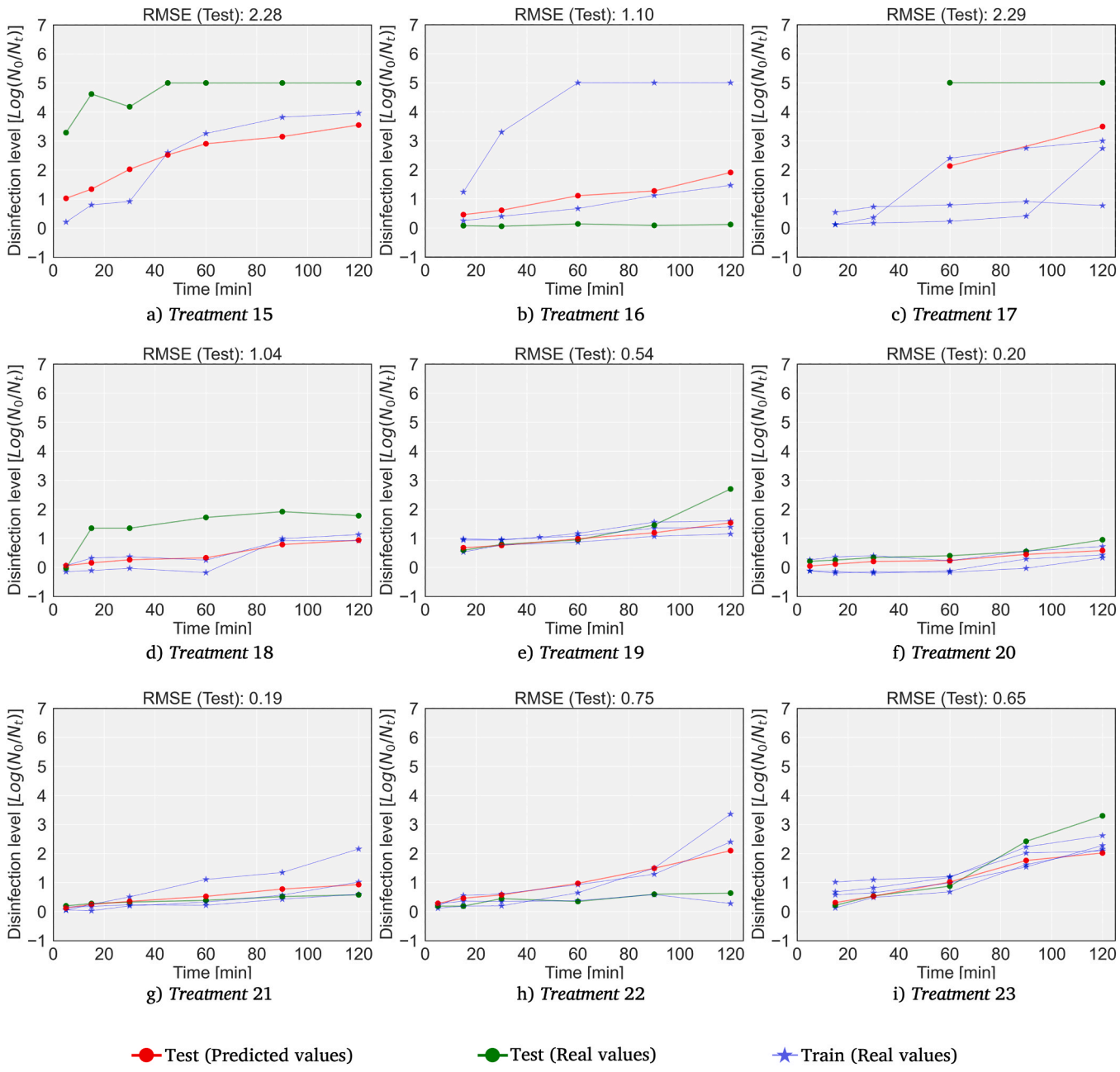


Fig. 7. Performance of RF model for *treatments* 15 to 29 in Case II. Performance of RF model for *treatments* 15 to 29 in Case II.

0.8, suggesting a good predictive performance though slightly lower compared to Case I. Table 2 summarizes the optimal hyperparameters employed for the RF model in Case I, and Cases II and III. For a comprehensive overview of the hyperparameter combinations used in hyperparameters tuning process along with corresponding coefficient of determination (R^2) performance metric, please consult the Supplementary material (Tables S2 and Table S3). The distribution of level of disinfection used in the training data for Case I and Cases II and III is presented in Figure S1 and Figure S2 of the Supplementary material.

Fig. 6 shows the results for Case I. The scatter plot within Fig. 6 offers a comparison between actual and predicted values of the level of disinfection predicted by our RF model, providing an insightful assessment of the RF model's performance for individual samples. Moreover, Fig. 6 features histogram plots on the top and right sides of scatter plot, providing insights into the distribution of both real and predicted values, respectively. A linear regression line (depicted in green) to indicate the overall trend of scatter points. Additionally, a perfect prediction line (red dashed line) aids in assessing the alignment of the scatter points with the ideal prediction scenario.

Specifically, we notice a dense cluster of points in the lower value range (<2), closely aligned along the diagonal dashed line. This pattern is similarly reflected in the distribution of real and predicted values, as evident on the histogram plots positioned on the top and right sides of the scatter plot. The clustering of points in the lower range indicates the usefulness of the RF for predicting lower values. However, as the actual values increase, the points on the plot become more dispersed, suggesting a decrease in prediction accuracy. Nevertheless, many points still remain relatively close to the perfect prediction line, as can be seen around values equal to 4, which indicates that while the model's performance might decrease as values grow, it consistently provides moderately accurate predictions across the entire spectrum of values. Moreover, we can notice that a high values range, the scatter points are

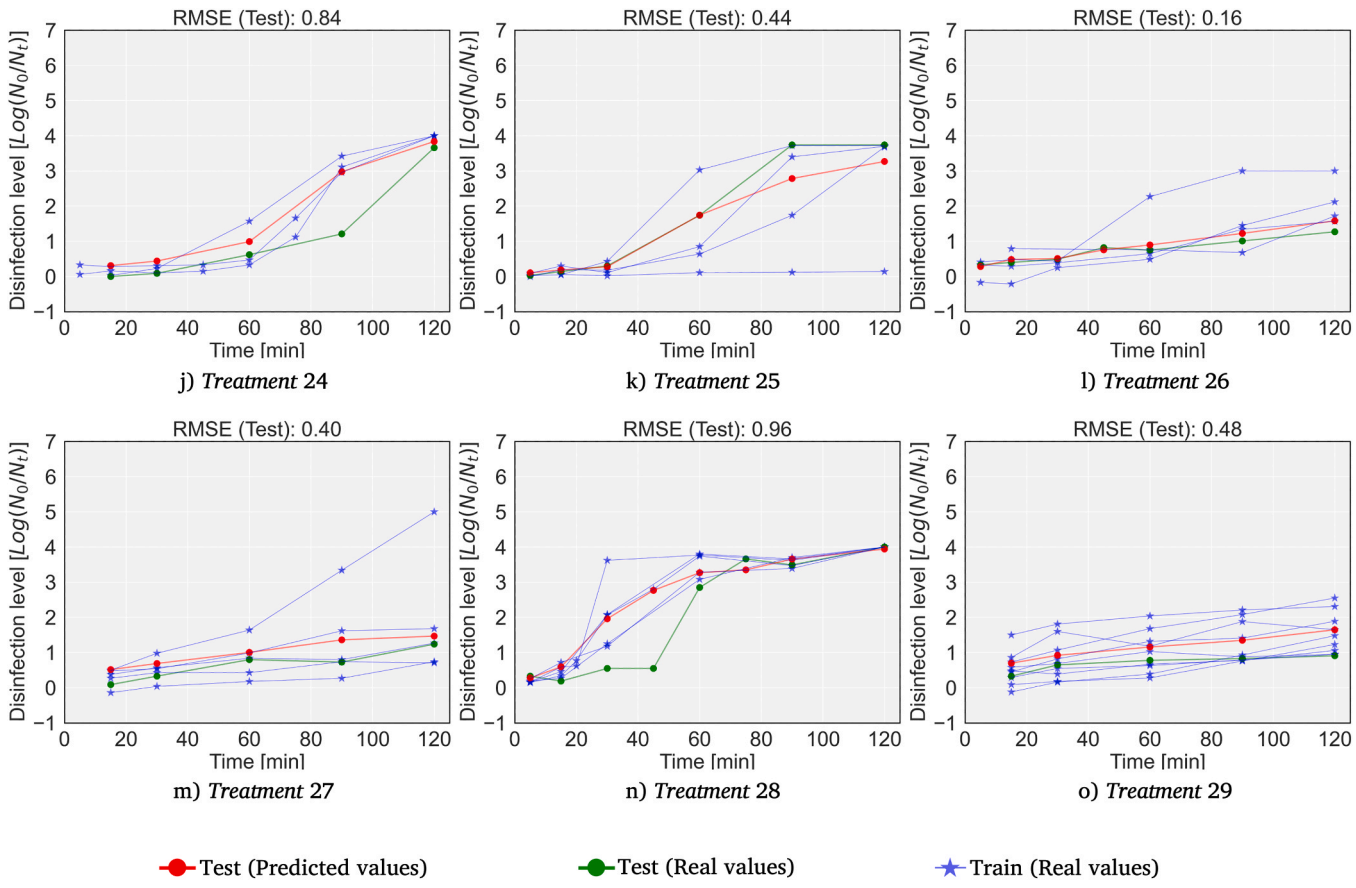


Fig. 7. (*continued*).

more frequent under the perfect prediction line indicating underestimation. This is more clearly noticeable for the overall trend of scatter points represented by the linear regression line.

The variability in RF performance for predicting low and high-range disinfection values is linked to the dispersion of sample values in the raw data, of which 75% was randomly used to train the RF model. This dispersion is illustrated in the box plots of Fig. 4(b). Predominantly, lower disinfection values are present in the first 60 minutes of the experiments, as indicated by the median (orange bar in Fig. 4(b)). Values between 2-4 are found in the upper quartile and whisker, while values above 4 are only present in the upper whisker and maximum values. Analysing the box plots, we infer that the RF model was trained with a higher proportion of low values compared to higher ones, contributing to its enhanced performance in predicting lower disinfection values. One of the main challenges in enhancing the performance of the RF model is the scarcity of experimental data with high disinfection values. While focusing on lower values improves predictions within that range, the scarcity representation of higher values poses a limitation on accurately predicting such values.

Fig. 7 presents the results for Case II, where the outcomes are plotted based on *treatments*, ranging from 15 to 29. This representation allows us to conduct an independent analysis of the performance of the RF model. Furthermore, it enables a graphical correlation of the contributions from the portion of data used to train the RF model, which belongs to the same *treatment*, with respect to the prediction accuracy. The real values of the training data are depicted in blue, while green and red dots connected by a line respectively represent the real and predicted values of the testing data for the corresponding *treatment*. In each plot, we present the performance of the RF model for each *treatment* using the RMSE metric.

By examining the plots in Fig. 7 along with the corresponding RMSE values displayed within each, we observe a consistent trend of decreasing RMSE as the number of experiments included in the training (represented in blue) increases for each *treatment*. This reduction correlates directly with the number of samples used in training, as demonstrated in Fig. 5 through the blue bars for *treatments* 15 to 29. However, there are exceptions, such as *treatments* 20, 21, and 26 (see Fig. 7(f), (g), and (l)), which exhibit lower RMSE values despite having fewer *experiments* in the training set. This positive phenomenon may be associated with the influence of our data partitioning methodology, which considers *experiments* belonging to the overall *treatments*, except *treatments* 1, 4, 7, and 11 used for evaluating Case III, in the training of the RF model. Specifically, *treatments* 20 (DW-Presence-Sodium Sulfite-Fe(III)-Cit), 21 (DW-Presence-Sodium Sultife-None), and 26 (SWW-Presence-Sodium Sulfite-Fe(III)-cit) share one or more parameters as type aqueous matrix, UV radiation, Oxidant, and catalyst. These parameters are frequently used in the *treatments* with a higher number of *experiments* in the training dataset, such as SWW, Presence, Sodium Sulfite and Fe(III)-Cit, respectively.

Fig. 8 presents the results for Case III, which encompasses *treatments* 1, 4, 7, and 11. In contrast to the plots in Fig. 7, these plots do not include information related to training data, as Case III focuses on the prediction of new *treatments* without utilizing experimental data from these *treatments* as part of the training set. From Fig. 8, we note that despite the absence of experiments related to the *treatments* used for prediction during training, our RF model performs well on *treatments* 4 and 11 (see Fig. 8(b) and (d)) achieving low RMSE values of 0.40 and 0.54, respectively. These values are comparable to those obtained in the predictions for Case II. However, for *treatments* 1 and 7 (as depicted in Fig. 8(a) and (c)), the prediction accuracy of our RF model decreases significantly, with RMSE values of 1.85 and 1.93, respectively.

To sum up, Table 4 presents a comprehensive overview of the performance evaluation results for our proposed RF model across all three

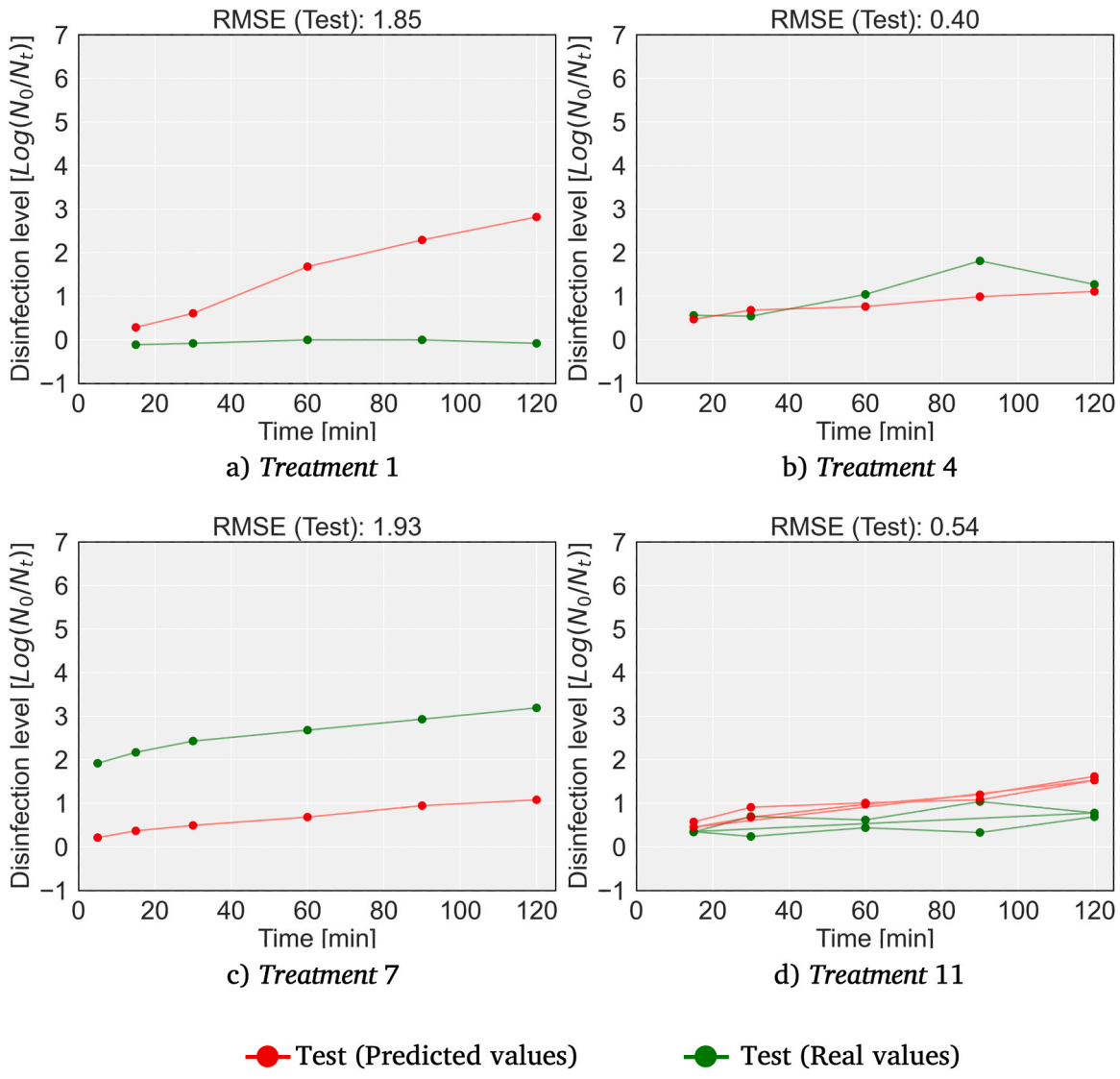


Fig. 8. Performance of RF model for *treatments* 1, 4, 7, and 11 in Case III. Plot of *Treatment* 11 contain two experiments, one denoted by 'x' dots for distinction.

study cases. Specifically, the Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and 95% confidence intervals (CI) for Mean Error (ME) as the key evaluation metrics. In Case I, all *treatments* are evaluated together, whereas in Cases II and III, individual *treatments* are assessed separately using RMSE and MAE metrics. Additionally, Table 4 provides the average performance results for Cases II and III, offering a comprehensive insight into the model's application in diverse scenarios.

Based on the average evaluation metrics presented in Table 4 for each study case, we noticed that our proposed RF model and data partitioning methodology demonstrate good performance when applied to Case I. Subsequently, their performance is slightly lower in Case II, and in Case III is the least favourable, as indicated by the RMSE values of 0.72, 0.82, and 1.18, respectively. Furthermore, our approach in Cases II and III achieved RMSE values below or similar to the RMSE average of Case–I in 8 out of 15 and 2 out of 4 *treatments*, respectively. This result underlines the potential utility of our approach to assist in the evaluation of the performance of AOP *treatments* and the design of new *experiments* of the same *treatment* type without the need for additional laboratory trials. Moreover, consistent with the results obtained from the RMSE, the average MAE values of 0.49, 0.7, and 1.07 for Cases I, II, and III, respectively, support the performance trends observed. The 95% confidence intervals further reinforce these findings. In Case I, a narrow 95% CI [-0.02, 0.22] confirms the model's strong predictive usefulness. In contrast, Case II, characterized by a wider 95% CI [-0.08, 0.33], indicates greater uncertainty than Case I, offering insight into the decrement of predictive performance. Case III is underscored by a wide 95% CI [-0.47, 0.52], emphasizing significant uncertainty.

It should be noted that non-machine learning models are limited to Case I and cannot be extended to Cases II and III, as they are designed for prediction based on different experiments and entirely different treatments, respectively. In Case I, a single RF model demonstrated adaptability across diverse scenarios, unlike non-machine learning approaches that require specific parameter settings for each experiment. This highlights the inherent adaptability of machine learning, eliminating the need for tailored settings in various experiments.

3.3. Comparison of study with SOTA studies

To compare our study with respect to SOTA AOPs based on ML algorithms, we have considered nine relevant studies listed in Table 1. It should be noted that a direct comparison is not feasible, because these studies focused on AOPs, contaminants, inputs, and outputs which are not compatible with our dataset and research goals.

Furthermore, the performance of ML algorithms can vary significantly in different scenarios [23,52,53]. Therefore, to assess the advantages and differences of our study versus SOTA AOPs coupled with

Table 4
Performance evaluation of RF model across the three study cases.

Case	Treatment	Evaluation metrics			Treatment conditions			
		RMSE	MAE	ME [95% CI]	Aqueous matrix	UV radiation	Oxidant	Catalyst
I	All (1-29)	0.72	0.49	0.09 [-0.02, 0.22]	–	–	–	–
II	15	2.28	2.22	–	SW	Absence	PMS	None
	16	1.10	0.98	–	SWW	Absence	PMS	None
	17	2.29	2.19	–	SWW	Presence	PMS	None
	18	1.04	0.96	–	DW	Presence	Sulfite	$FeSO_4$
	19	0.54	0.32	–	SWW	Presence	Sulfite	None
	20	0.20	0.18	–	DW	Presence	Sulfite	Fe(III)-Cit
	21	0.19	0.15	–	DW	Presence	Sulfite	None
	22	0.75	0.58	–	SWW	Presence	H_2O_2	$FeSO_4$
	23	0.65	0.44	–	SWW	Presence	PMS	Fe(III)-Cit
	24	0.84	0.59	–	SW	Presence	H_2O_2	Fe(III)-Cit
	25	0.44	0.26	–	SWW	Presence	H_2O_2	None
	26	0.16	0.13	–	SWW	Presence	Sulfite	Fe(III)-Cit
	27	0.40	0.37	–	SWW	Presence	PMS	$FeSO_4$
	28	0.96	0.63	–	SW	Presence	PMS	Fe(III)-Cit
	29	0.48	0.46	–	SWW	Presence	Sulfite	$FeSO_4$
	Average (15-29)	0.82	0.70	0.11 [-0.08, 0.33]	–	–	–	–
III	1	1.85	1.59	–	SWW	Absence	H_2O_2	None
	4	0.40	0.30	–	SWW	Presence	None	$FeSO_4$
	7	1.93	1.92	–	SWW	Presence	None	Fe(III)-Cit
	11	0.54	0.45	–	SWW	Absence	Sulfite	$FeSO_4$
	Average (1, 4, 7, 11)	1.18	1.07	0.01 [-0.47, 0.52]	–	–	–	–

ML algorithms, we conducted a comparison of methodologies and goals for each of the cases we defined (Cases I, II and III).

In Case I, our main aim is to predict missing or additional data points at any given moment during a degradation or inactivation process within a specific *experiment*. This aligns with the goals of four out of the nine studies we have reviewed [20,21,23,27].

Specifically, they share the methodology of randomly splitting the dataset for ML implementation. However, it is important to note that two of these studies [21,27] present relevant differences: while we used a single predictive model for all 29 diverse *treatments*, both works designed independent predictive models for each of the AOPs. On one hand, this choice limits the scalability and practicality of their models for application to others AOPs. On the other hand, being specifically trained for individual AOPs leads to a higher accuracy ($R^2 > 0.9$). In this regard, our works is similar to Navidpour et al. [23], who implemented unique predictive models for all AOPs studied and considered similar input parameters to represent the non-linearity of the AOPs. As expected, this results in lower accuracy ($R^2 < 0.8$).

Regarding Cases II and III, none of the reviewed studies fully address the challenge of predicting various values over time for entirely new *experiments* or *treatments*. This predictive approach has precedence in other research fields, as exemplified by Salazar and Crookston [54]. It is important to note that five out of nine reviewed studies [22,24,26,28,39] share the main objectives of Cases II & III, which facilitates the optimization of parameters associated with the reaction conditions of a particular *treatment*. However, they all differ from our study in one significant aspect: the output of their models is the kinetic reaction constant "k" instead of a set of values representing the degradation process over time are predicted as a whole. Three of these works [22,24,39] clearly align with our Case II, as they aim to optimize parameter values for the same type of *treatment*. Nevertheless, they diverge from our methodological approach in that they exclusively rely on data from the same type of *treatment* to train the ML models. As a result, we argue that our models exhibit a higher degree of generalizability, as they are trained with diverse *treatments*, encompassing various catalysts and oxidants, as well as the presence or absence of UV irradiation.

Regarding the objectives of Case III, Sun et al. [26] and Zhang et al. [28], introduced different types of catalysts [26,28] and various oxidants [26] as input variables, thus including different *treatments* in the optimization target. Notably, these studies stand out from most previous works, including our own, due to its broader scope, i.e., the predictive capability for different contaminants. For that purpose, they incorporate different approaches to parameterize each contaminant: A Linear Solvation Energy Relation (LSER) model [28] and the "quantum chemical descriptors" methodology [26]. While these methodologies offer interesting possibilities for future research, their applicability is constrained to organic compounds or relatively small molecular clusters. This limitation arises from the complexity and size of organisms such as *Enterococcus sp.* and other bacterial pathogens, which are not covered by these methods.

4. Conclusions and future research avenues

In this article, we introduced a RF model to address the complex task of predicting *Enterococcus sp.* concentration in wastewater based on multiple AOPs when faced with limited data samples. The model's performance was assessed using a real-world wastewater database comprising wastewater samples treated with multiple AOPs and three distinct cases constructed through a data partitioning methodology. The prediction difficulty increases progressively from Case I to Case III. The results indicated that the RF model performs well in Case I, moderately in Case II, and less favourably in Case III, with average RMSE values of 0.72, 0.82, and 1.18, respectively. Furthermore, despite the complexity of Cases II and III, our approach is able to achieve RMSE values below or similar to the RMSE average of Case I in 8 out of 15 *treatments* for Case II and 2 out of 4 *treatments* for Case III. Moreover, our study objectives and methodology were compared with SOTA AOPs based on ML algorithms demonstrating its unique approach and innovation in predicting disinfection levels of a complex microorganism such as *Enterococcus sp.*

In conclusion, the RF model demonstrated its usefulness in predicting disinfection levels of missing or additional values at any instant during the AOPs. This is particularly evident in Case I, where the model excels at generalizing its predictions to various AOP treatments with a unique model. However, it is important to note a limitation in the model's accuracy when predicting higher disinfection levels. This decrease in accuracy is attributed to the scarcity of experimental data with high disinfection values in the training dataset. Moreover, it is also valuable in predicting the disinfection behaviour of completely unseen *experiments* based on data from identical *treatments* but different initial concentration levels of oxidant and/or catalyst, as demonstrated in Case

II. This capability allows us to design new *experiments* of the same *treatment* type without the need for additional expensive and time-consuming laboratory tests. Nevertheless, due to the limited number of samples available for each *treatment* and the diversity of treatments in our database, it is premature to draw conclusions about the potential utility of our approach to assist in the design of new *treatments*, as presented in Case III.

It is important to highlight that our proposed RF model provides a straightforward implementation, effective handling of non-linear relationships, and requires only a minimal number of training parameters. Therefore, in the case of having a more balanced distribution of samples and/or an increase in the number of samples across *treatments*, they can be easily added in the training phase to enhance the model performance.

As potential avenues for future research to address limitations in the model's accuracy due to imbalanced data in the training dataset, we propose apply data augmentation techniques to artificially increase the size of the high-value dataset for Case I. Moreover, we will explore the use of synthetic data generation methods to create new data and balance the number of samples across *treatments* in the training dataset for Case II and III.

CRediT authorship contribution statement

Jorge Rodríguez-Chueca: Writing – review & editing, Resources, Project administration, Investigation, Funding acquisition, Conceptualization. **Fernando Salazar:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization. **Sonia Guerra-Rodríguez:** Writing – original draft, Investigation, Data curation. **Pavel Pascacio:** Writing – original draft, Visualization, Software, Methodology, Formal analysis, Conceptualization. **David J. Vicente:** Writing – original draft, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of Competing Interest

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

Data Availability

Data will be made available on request.

Acknowledgment

The publication is part of Projects TED2021-129969A-C32 and TED2021-129969B-C33 funded by MCIN/AEI/ 10.13039/ 501100011033 and by "European Union NextGenerationEU/PRTR". Sonia Guerra-Rodríguez acknowledges the Universidad Politécnica de Madrid (UPM) for the financial support provided through the predoctoral contract granted within the "Programa Propio". Jorge Rodríguez-Chueca acknowledges Comunidad de Madrid by the pluriannual agreement with the Polytechnic University of Madrid in the line of action Programme of Excellence for University Teaching Staff (M190020074BJJRC). This work was also funded by the Spanish Ministry of Economy and Competitiveness through the "Severo Ochoa Programme for Centres of Excellence in R&D" (CEX2018-000797-S) and the Generalitat de Catalonia through the CERCA Program.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jece.2024.112530.

References

[1] Stuart W. Krasner, Howard S. Weinberg, Susan D. Richardson, Salvador J. Pastor, Russell Chinn, Michael J. Sclimenti, Gretchen D. Onstad, Alfred D. Thruston, Occurrence of a new generation of disinfection byproducts, Environ. Sci. Technol. 40 (23) (2006) 7175–7185.
[2] David L. Sedlak, Urs von Gunten, The chlorine dilemma, Science 331 (6013) (2011) 42–43.
[3] Guanghui Hua, David A. Reckhow, Comparison of disinfection byproduct formation from chlorine and alternative disinfectants, Water Res. 41 (8) (2007) 1667–1678.
[4] Juan Li, Zhong Zhang, Yingying Xiang, Jin Jiang, and, Ran Yin, Role of uv-based advanced oxidation processes on nom alteration and dbp formation in drinking water treatment: a state-of-the-art review, Chemosphere (2022) 136870.
[5] A.Ike Ikechukwu, Yunho Lee, Jin Hur, Impacts of advanced oxidation processes on disinfection byproducts from dissolved organic matter upon postchlor (am) ination: a critical review, Chem. Eng. J. 375 (2019) 121929.
[6] Informe sobre aguas residuales en España, n.d. http://www.un.org/es/comun/docs/?symbol=A/69/L.85.
[7] Monali Priyadarshini, Indrasis Das, Makarand M. Ghangrekar, Lee Blaney, Advanced oxidation processes: Performance, advantages, and scale-up of emerging technologies, J. Environ. Manag. 316 (2022) 115295.
[8] David B. Miklos, Christian Remy, Martin Jekel, Karl G. Linden, J.örg E. Drewes, Uwe Hübner, Evaluation of advanced oxidation processes for water and wastewater treatment–a critical review, Water Res. 139 (2018) 118–131.
[9] Xiaoguang Duan, Xu Zhou, Rupeng Wang, Shaobin Wang, Nan-qi Ren, ShihHsin Ho, et al., Advanced oxidation processes for water disinfection: Features, mechanisms and prospects, Chem. Eng. J. 409 (2021) 128207.
[10] Sonia Guerra-Rodríguez, Nerea Cediel, Encarnación Rodríguez, Jorge Rodriguez-Chueca, Photocatalytic activation of sulfite using fe (ii) and fe (iii) for enterococcus sp. inactivation in urban wastewater, Chem. Eng. J. 408 (2021) 127326.
[11] Jorge Rodríguez-Chueca, Sonia Guerra-Rodriguez, Julia M. Raez, Maria-Jose Lopez-Munoz, Encarnacion Rodriguez, Assessment of different iron species as activators of s2o82-and hso5-for inactivation of wild bacteria strains, Appl. Catal. B: Environ. 248 (2019) 54–61.
[12] Arjunan Babuponnusami, Karuppan Muthukumar, A review on fenton and improvements to the fenton process for wastewater treatment, J. Environ. Chem. Eng. 2 (1) (2014) 557–572.
[13] Jorge Rodríguez-Chueca, Carlos Amor, T.ânia Silva, Dionysios D. Dionysiou, Gianluca Li Puma, Marco S. Lucas, José A. Peres, Treatment of winery wastewater by sulphate radicals: Hso5-/transition metal/uv-a leds, Chem. Eng. J. 310 (2017) 473–483.
[14] L. Sbardella, I. Velo-Gala, J. Comas, I. Rodríguez-Roda Layret, A. Fenu, W. Gernjak, The impact of wastewater matrix on the degradation of pharmaceutically active compounds by oxidation processes including ultraviolet radiation and sulfate radicals, J. Hazard. Mater. 380 (2019) 120869.
[15] Jie Ma, Haiyan Li, Yongqi Yang, Xuening Li, Influence of water matrix species on persulfate oxidation of phenol: reaction kinetics and formation of undesired degradation byproducts, Water Sci. Technol. 2017 (2) (2018) 340–350.
[16] Ana R. Lado Ribeiro, Nuno F.F. Moreira, Gianluca Li Puma, and, Adrián M.T. Silva, Impact of water matrix on the removal of micropollutants by advanced oxidation technologies, Chem. Eng. J. 363 (2019) 155–173.
[17] Ruixing Huang, Chengxue Ma, Jun Ma, Xiaoliu Huangfu, Qiang He, Machine learning in natural and engineered water systems, Water Res. 205 (2021) 117666.
[18] Nawal Taoufik, Wafaa Boumya, Mounia Achak, Hamid Chennouk, Raf Dewil, Noureddine Barka, The state of art on the prediction of efficiency and modeling of the processes of pollutants removal based on machine learning, Sci. Total Environ. 807 (2022) 150554.
[19] Zaher Mundher Yaseen, An insight into machine learning models era in simulating soil, water bodies and adsorption heavy metals: Review, challenges and solutions, Chemosphere 277 (2021) 130126.
[20] Thi Thanh Nhi Le, Hai Bang Truong, Thi Hoa Le, Hoang Sinh Le, Thanh Tam Toan Tran, Tran Duc Manh, Quang Khieu Dinh, Xuan Cuong Nguyen, et al., Cu2o/fe3o4/uio-66 nanocomposite as an efficient fenton-like catalyst: Performance in organic pollutant degradation and influencing factors based machinelearning, Heliyon, 2023.
[21] H.üseyin Cüce, Duygu Özçelik, Application of machine learning (ML) and artificial intelligence (ai)-based tools for modelling and enhancing sustainable optimization of the classical/photo-fenton processes for the landfill leachate treatment, Sustainability 14 (18) (2022) 11261.
[22] Zhuoying Jiang, Jiajie Hu, Xijin Zhang, Yihang Zhao, Xudong Fan, Shifa Zhong, Huichun Zhang, Xiong Yu, A generalized predictive model for tio2–catalyzed photo-degradation rate constants of water contaminants through artificial neural network, Environ. Res. 187 (2020) 109697.
[23] Amir H. Navidpour, Ahmad Hosseinzadeh, Zhenguo Huang, Donghao Li, John L. Zhou, Application of machine learning algorithms in predicting the photocatalytic degradation of perfluorooctanoic acid, Catal. Rev. (2022) 1–26.
[24] Anfel Smaali, Mohammed Berkani, Fateh Merouane, Yasser Vasseghian, Noureddine Rahim, Meriem Kouachi, et al., Photocatalytic-persulfate-oxidation for diclofenac removal from aqueous solutions: modeling, optimization and biotoxicity test assessment, Chemosphere 266 (2021) 129158.
[25] Yasser Vasseghian, Mohammed Berkani, Fares Almomani, and, Elena-Niculina Dragoi, Data mining for pesticide decontamination using heterogeneous photocatalytic processes, Chemosphere 270 (2021) 129449.

[26] Ye Sun, Zhiyuan Zhao, Hailong Tong, Baiming Sun, Yanbiao Liu, Nanqi Ren, and Shijie You. Machine learning models for inverse design of the electrochemical oxidation process for water purification. *Environmental Science & Technology*, 2023.
[27] Cheng Zhang, Wenjing Sun, Huangzhao Wei, Chenglin Sun, Application of artificial intelligence for predicting reaction results in advanced oxidation processes, Environ. Technol. Innov. 23 (2021) 101550.
[28] Shu-Zhe Zhang, Shuo Chen, Hong Jiang, A new tool to predict the advanced oxidation process efficiency: Using machine learning methods to predict the degradation of organic pollutants with fe-carbon catalyst as a sample, Chem. Eng. Sci. 280 (2023) 119069.
[29] Joaquín Irazábal, Fernando Salazar, and, David J. Vicente, A methodology for calibrating parameters in discrete element models based on machine learning surrogates, Comput. Part. Mech. (2023) 1–17.
[30] D.J. Vicente, F. Salazar, S.R. López-Chacón, C. Soriano, J. Martin-Vide, Evaluation of different machine learning approaches for predicting high concentration episodes of ground-level ozone: A case study in catalonia, spain, Atmos. Pollut. Res. (2023) 101999.
[31] Fernando Salazar, Mohammad Amin Hariri-Ardebili, Coupling machine learning and stochastic finite element to evaluate heterogeneous concrete infrastructure, Eng. Struct. 260 (2022) 114190.
[32] Fernando Salazar, André Conde, Joaquín Irazábal, David J. Vicente, Anomaly detection in dam behaviour with machine learning classification models, Water 13 (17) (2021) 2387.
[33] Jacopo Foschi, Andrea Turolla, Manuela Antonelli, Soft sensor predictor of e. coli concentration based on conventional monitoring parameters for wastewater disinfection control, Water Res. 191 (2021) 116806.
[34] Narendra Khatri, Kamal Kishore Khatri, and, Abhishek Sharma, Artificial neural network modelling of faecal coliform removal in an intermittent cycle extended aeration system-sequential batch reactor based wastewater treatment plant, J. Water Process Eng. 37 (2020) 101477.
[35] Sofyan Sbahi, Naaila Ouazzani, Lahbib Latrach, Abdessamed Hejjaj, Laila Mandi, Predicting the concentration of total coliforms in treated rural domestic wastewater by multi-soil-layering (msl) technology using artificial neural networks, Ecotoxicol. Environ. Saf. 204 (2020) 111118.
[36] Esra' Bashayreh, Ahmad Manasrah, Shahnaz Alkhalil, Eman Abdelhafez, Estimation of water disinfection by using data mining, Ecol. Eng. Environ. Technol. 22 (2021).
[37] Efaq Ali Noman, Adel Ali Al-Gheethi, Radin Mohamed Radin Maya Saphira, Balkis A. Talip, Mohammed Al-Sahari, Norli Ismail, Mathematical prediction models for inactivation of antibiotic-resistant bacteria in kitchen wastewater by bimetallic bionanoparticles using machine learning with gene expression programming, J. Clean. Prod. 333 (2022) 130131.
[38] J. Rodríguez-Chueca, E. Barahona-García, V. Blanco-Gutiérrez, L. Isidoro-García, and, A.J. Dos santos García, Magnetic cofe2o4 ferrite for peroxymonosulfate activation for disinfection of wastewater, Chem. Eng. J. 398 (2020) 125606.
[39] Lijing Wang, Tianyi Yang, Xiangyu Xu, Guangya Zhang, Yunming Liu, Amin Ju, Gang Zhou, Bo Feng, Guangbo Che, Zhao Zhao, Acid groups decorated bimetal-organic catalyst for advanced oxidation technology at full ph range, J. Alloy. Compd. page 172370 (2023).
[40] Sonia Guerra-Rodriguez, Encarnacion Rodriguez, Javier Moreno-Andres, and, Jorge Rodriguez-Chueca, Effect of the water matrix and reactor configuration on enterococcus sp. inactivation by uv-a activated pms or h2o2, J. Water Process Eng. 47 (2022) 102740.
[41] Thérence Nibareke, Jalal Laassiri, Using big data-machine learning models for diabetes prediction and flight delays analytics, J. Big Data 7 (2020) 1–18.
[42] Ekaba Bisong, Ekaba Bisong, Introduction to scikit-learn, Build. Mach. Learn. Deep Learn. Models Google Cloud Platf.: A Compr. Guide Begin. (2019) 215–229.
[43] Kiran Maharana, Surajit Mondal, Bhushankumar Nemade, A review: Data pre-processing and data augmentation techniques, Glob. Transit. Proc. 3 (1) (2022) 91–99.
[44] John T. Hancock, Taghi M. Khoshgoftaar, Survey on categorical data for neural networks, J. Big Data 7 (1) (2020) 1–41.
[45] Andy Liaw, Matthew Wiener, et al., Classification and regression by randomforest, R. N. 2 (3) (2002) 18–22.
[46] Biau Gérard, Analysis of a random forests model, J. Mach. Learn. Res. 13 (2012) 1063–1095.
[47] Andreas C. Müller, Sarah Guido, Introduction to machine learning with Python: a guide for data scientists, " O'Reilly Media, Inc.", 2016.
[48] Kirill Kolodiazhnyi, Hands-On Machine Learning with C++: Build, train, and deploy end-to-end machine learning and deep learning pipelines, Packt Publishing Ltd, 2020.
[49] Plotly Technologies Inc, Collaborative data science, Plotly technologies inc, montreal, qc, 2015.
[50] Wes McKinney, et al., pandas: a foundational python library for data analysis and statistics, Python High. Perform. Sci. Comput. 14 (9) (2011) 1–9.
[51] Travis E. Oliphant, et al., Guide to numpy, volume 1, Trelgol Publishing USA, 2006.
[52] Ahmad Hosseinzadeh, John L. Zhou, Ali Altaee, Donghao Li, Machine learning modeling and analysis of biohydrogen production from wastewater by dark fermentation process, Bioresour. Technol. 343 (2022) 126111.
[53] Ahmad Hosseinzadeh, John L. Zhou, Ali Altaee, Mansour Baziar, and, Xiaowei Li, Modeling water flux in osmotic membrane bioreactor by adaptive network-based fuzzy inference system and artificial neural network, Bioresour. Technol. 310 (2020) 123391.
[54] Fernando Salazar, Brian M. Crookston, A performance comparison of machine learning algorithms for arced labyrinth spillways, Water 11 (3) (2019) 544.