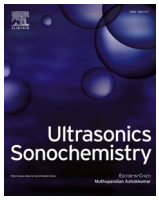
Ultrasonics Sonochemistry 110 (2024) 107032



Contents lists available at ScienceDirect

Ultrasonics Sonochemistry

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





Machine learning model to predict rate constants for sonochemical degradation of organic pollutants

Iseul Na [a,b], Taeho Kim [c], Pengpeng Qiu [d], Younggyu Son [a,b,*]

[a] Department of Environmental Engineering, Kumoh National Institute of Technology, Gumi 39177, Republic of Korea
[b] Department of Energy Engineering Convergence, Kumoh National Institute of Technology, Gumi 39177, Republic of Korea
[c] AI Lab, ROSIS IT, Seoul 07547, Republic of Korea
[d] State Key Laboratory for Modification of Chemical Fibers and Polymer Materials, Institute of Functional Materials, College of Materials Science and Engineering, Donghua University, Shanghai 201620, PR China

ARTICLE INFO

Keywords:
Machine learning
Ultrasound
Cavitation
Sonochemical degradation
Reaction rates

ABSTRACT

In this study, machine learning (ML) algorithms were employed to predict the pseudo-1st-order reaction rate constants for the sonochemical degradation of aqueous organic pollutants under various conditions. A total of 618 sets of data, including ultrasonic, solution, and pollutant characteristics, were collected from 89 previous studies. Considering the difference between the electrical power (P_{ele}) and calorimetric power (P_{cal}), the collected data were divided into two groups: data with P_{ele} and data with P_{cal}. Eight input variables, including frequency, power density, pH, temperature, initial concentration, solubility, vapor pressure, and octanol–water partition coefficient (K_{ow}), and one target variable of the degradation rate constant, were selected for ML. Statistical analysis was conducted, and outliers were determined separately for the two groups. ML models, including random forest (RF), extreme gradient boosting (XGB), and light gradient boosting machine (LGB), were used to predict the pseudo-1st-order reaction rate constants for the removal of aqueous pollutants. The prediction performance of the ML models was evaluated using different metrics, including the root mean squared error (RMSE), mean absolute error (MAE), and R squared (R^2). A significantly higher prediction performance was obtained using data without outliers and augmented data. Consequently, all the applied ML models could be used to predict the sonochemical degradation of aqueous pollutants, and the XGB model showed the highest accuracy in predicting the rate constants. In addition, the power density and frequency were the most influential factors among the eight input variables in prediction with the Shapley additive explanation (SHAP) values method. The degradation rate constants of the two pollutants over a wide frequency range (20–1,000 kHz) were predicted using the trained ML model (XGB) and the prediction results were analyzed.

1. Introduction

With rapid advancements in computer processing capabilities and algorithmic methodologies, effective machine learning (ML) and deep learning (DL) algorithms have been developed and applied in numerous domains, such as healthcare, finance, autonomous driving, and sound analysis [1–7]. Environmental engineering researchers have begun applying these techniques for data analytics. Since the late 2010 s, extensive ML research has been conducted, with most studies focusing on air pollution monitoring and water quality management [8–12].

Over the last two decades, advanced oxidation processes (AOPs) have been widely investigated for the removal of recalcitrant and emerging pollutants using electromagnetic and mechanical wave energies and a variety of oxidizing chemicals and catalysts [13–18]. Recent advancements in ML have led to the development of AOPs for investigating various process variables from new perspectives and predicting pollutant removal [19–29]. The ML-based approach is expected to streamline experimental processes by reducing time and obviating extensive chemical- and labor-based experiments [19,20]. By analyzing and training experimental data collected from previous research, pollutant degradation (removal efficiencies and reaction rate constants) can be predicted using a well-trained ML algorithm. Recently, researchers have reported the application of ML techniques to ultrasound-based AOPs. Zhou et al. predicted pseudo-1st-order reaction rate constants for the removal of pharmaceuticals in sonoelectrochemical processes using various ML models. They selected 15 input variables,

* Corresponding author.
E-mail address: yson@kumoh.ac.kr (Y. Son).

https://doi.org/10.1016/j.ultsonch.2024.107032
Received 13 June 2024; Received in revised form 2 August 2024; Accepted 15 August 2024
Available online 21 August 2024
1350-4177/© 2024 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

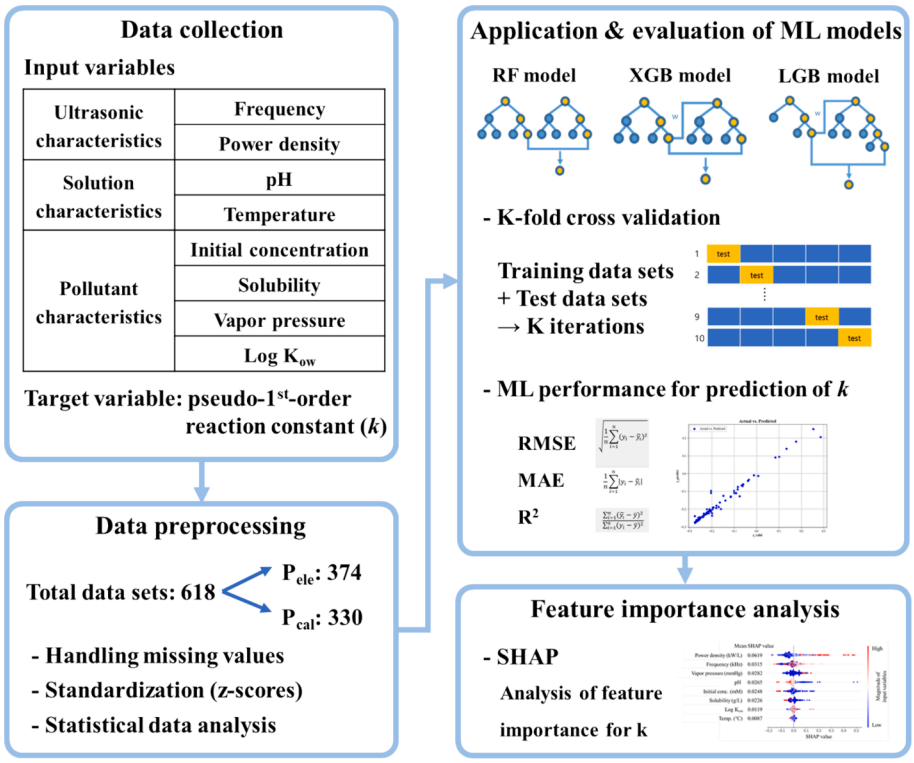


Fig. 1. Flow chart for machine learning using data from previous research on sonochemical processes.

including six electro-oxidation features, two ultrasonic features, four pollutant features, and three environmental features [21]. A research group investigated ML models to predict the tetracycline removal rate in sonophotocatalytic processes using various catalysts. They developed ML algorithms with tetracycline concentration, catalyst dose, irradiation time, light power, ultrasound on/off, and additive oxidant concentration as input variables [22,23,30]. Glienke et al. obtained experimental data for the degradation of 32 phenol derivatives using an 860 kHz sonoreactor and analyzed the data using various ML models [31]. All the above-mentioned sonochemists used only their own experimental data, relatively small numbers of datasets compared to those in previous AOPs research, for ML training and evaluating ML performance.

This study aims to investigate the applicability of ML to sonochemistry using datasets collected from previous studies. Three well-known ML models, random forest (RF), extreme gradient boosting (XGB), and light gradient boosting machine (LGB), were used to predict the pseudo-1st-order reaction rate constants of sonochemical pollutant degradation. A total of 618 datasets from 89 previous research papers, providing a comprehensive understanding of sonochemical degradation, were used for the training of the ML models. Eight input variables, including frequency, power density, pH, temperature, initial concentration, solubility, vapor pressure, and octanol–water partition coefficient (K_{ow}), and one target variable of the degradation rate, constant were considered. Different metrics, including the root mean squared error (RMSE), mean absolute error (MAE), and R squared (R^2), were analyzed and compared to evaluate the performance of the ML models. In addition, the rate constants for the removal of two selected pollutants (naphthol blue black and phenol), which are the most frequently reported pollutants in our collected data, under various frequency conditions (20–1,000 kHz) were predicted using the trained ML model.

2. Experiment and analytical methodology

A brief flowchart for the ML using data from previous research on sonochemical processes in this study is shown in Fig. 1. A desktop computer for office (Intel i7-8700 CPU 3.20 GHz; RAM 16 GB) was used in this study.

2.1. Data collection

A literature search was conducted using bibliographic databases such as Google Scholar with five major keywords: “sonochemical,” ”degradation,” “ultrasound,” ”ultrasonic,” and “oxidation.” This search yielded 618 sets of data from 89 articles published since 2000 on the degradation of aqueous pollutants in sonochemical processes using only ultrasound, but not other AOPs. Each dataset included eight input variables, namely frequency, power density, pH, temperature, initial concentration, solubility, vapor pressure, and octanol–water partition coefficient (K_{ow}), and one target variable of the pseudo-1st-order reaction rate constant for the removal of pollutants was collected considering the ultrasonic, solution, and pollutant characteristics. Some rate constants were obtained from the figures of research papers using a high-resolution monitor (QHD 2560 × 1440) and PlotDigitizer software (https://plotdigitizer.com/) [32,33]. Some data on pollutant characteristics were obtained from well-known websites, such as PubChem (https://pubchem.ncbi.nlm.nih.gov/) and the Toxin and Toxin Target Database (T3DB, https://www.t3db.ca/). The collected data with a reference list was provided in the Supplementary Material. In spite of relatively larger numbers of datasets compared to those in previous research, more datasets should be added to enhance diversity in datasets [21–23,30,31]. In addition, it should be noted that the geometric effects were not considered in this study because most previous researchers rarely reported the details of the geometric conditions in their studies. Recently, it has been reported that small changes in the geometric conditions could result in significant differences in sonochemical activity [34–39].

2.2. Data preprocessing

In sonochemistry, information on the input power for ultrasound

Table 1
Model performance evaluation metrics.

Evaluation metrics	Mathematical expressions
Root Mean Squared Error (RMSE)	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$
Mean Absolute Error (MAE)	$\frac{1}{n}\sum_{i=1}^{n}\lvert y_i-\hat{y}_i\rvert$
R Squared (R^2)	$\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$

irradiation is provided in terms of electrical power (P_{ele}), which represents the energy consumed by the ultrasound generator, or calorimetric power (P_{cal}), which represents the heat energy induced by the absorption of ultrasound in a liquid medium. It has been reported that the difference between electrical power and calorimetric power is quite significant and P_{cal} typically ranges from 45 to 85 % of P_{ele} depending on the applied ultrasonic system [35,36,39]. Previous researchers provided either P_{cal} or P_{ele} in their studies. Thus, the collected datasets were divided into two categories [data with P_{ele} (374 sets) and data with P_{cal} (330 sets)] and analyzed separately. In addition, considering the size (or volume) difference of the sonoreactors, the power density (P_{ele} or P_{cal} / liquid volume) was used instead of P_{ele} or P_{cal}.

Some data, including temperature, pH, solubility, vapor pressure, and log K_{ow}, were not available, and missing data were replaced to obtain complete ML datasets. The temperature and pH were assumed as 25℃ and 6, respectively. Missing values of solubility, vapor pressure, and log K_{ow} were appropriately estimated based on related references and the reliable data we collected [40,41].

By statistically analyzing the datasets, outliers, whose values were higher or lower than three standard deviations of each variable were determined [42–44] and the total data were divided into two groups and analyzed using ML models: data including the outliers and data excluding the outliers. It has been reported that more accurate results can be obtained for predicting target values using data without outliers in ML algorithms [42,44–46].

As the final step in data preprocessing for ML training, the data were standardized to have a mean of zero and a standard deviation of one using z-scores, owing to the large differences in the magnitudes of the variables [47–50]. The preprocessed data were then partitioned into training and validation data using K-fold cross validation (K=10) to prevent overfitting of the ML models [47–49].

2.3. Model experiment(construction)

ML models, including RF, XGB, and LGB, which are commonly used in the application of ML models in various fields, were used in this study [20,26–29,50–52]. In our preliminary tests, relatively poor prediction accuracies were obtained using Linear, Ridge and Lasso regression models. All the programs for running the ML models were implemented using Python (ver. 3.10.13) and the Scikit_learn (ver. 1.3.0), xgboost (ver. 1.7.3), and LightGBM (ver. 4.0.0) libraries were used. The hyperparameters of the applied ML models were optimized using tree-structured Parzen Estimator algorithm, RMSE, and cross validation results. The optimized hyperparameter values for XGB, LGB, and RF were shown in Table 1S [53,54].

2.4. Performance evaluation metrics

The accuracies of the ML models were evaluated using MAE, RMSE, and R^2, also known as the coefficient of determination (Table 1) [55,56]. MAE represents the average absolute error between the predicted and actual values, and RMSE is the square root of the mean squared error (MSE), which is the average of the squared errors. The MAE and RMSE values closer to 0 indicate higher prediction performance. The R^2 value reflects how well the generated model explains the data for both the actual and predicted values, with a value closer to 1 indicating a higher model performance.

Furthermore, the SHAP value method was employed to analyze the contribution of each input variable to the predicted target variable. SHAP assigns a 'SHAPley value' to each input variable (or feature), precisely measuring the contribution of individual variable to the model output in comparison to a baseline prediction—the mean prediction in the data. These values play a pivotal role in understanding the impact of specific features on the predictive decisions of a model [57].

3. Result and discussion

3.1. Statistical analysis of collected data

Tables 2 and 3 show the results of a simple statistical analysis of the collected data, including the outliers [eight input variables: frequency, power density, pH, temperature, initial concentration, solubility, vapor pressure, and octanol–water partition coefficient (K_{ow}); one target variable: pseudo-1st-order reaction rate constant] with P_{ele} and P_{cal}, respectively. Table 2S and 3S depict the statistical analysis results without outliers. Fig. 1S–4S also show the histogram plots of data with P_{ele} including the outliers (374 sets), data with P_{ele} excluding the outliers (326 sets), data with P_{cal} including the outliers (330 sets), and data with P_{cal} excluding the outliers (275 sets). Although many more datasets for the ultrasonic degradation of organic pollutants under a wide range of experimental conditions were collected compared to previous research [22,31,58], the data were not distributed evenly, which could lead to low accuracies in the predicted target values using the developed ML algorithms for the input variables in the untrained or poorly trained ranges.

Before the ML analysis, the linear relationship between the variables was investigated using Pearson correlation with the data including the outliers as shown in Fig. 2. The Pearson correlation coefficient ranges from −1.00 to + 1.00 and represents the strength and direction of the linear relationship between two variables [19,25,59]. Most variables in the collected data did not demonstrate a linear relationship. For the data with P_{cal}, a relatively strong linear relationship between vapor pressure and rate constant was obtained, which could be due to the presence of large outliers [the maximum vapor pressure was 10^{15} times higher than the minimum pressure (malachite green: 2.4×10^{-13} mmHg; carbon

Table 2
Descriptive statistics of the collected data with P_{ele} (374 sets) (the outliers were included).

	Frequency (kHz)	Power density (kW/L)	pH	Temp.	Initial conc. (mM)	Solubility (g/L)	Log K_{ow}	Vapor pressure (mmHg)	Rate constants ($\times 10^{-3}$/sec)
mean	358	0.82	5.92	22.8	2.25	12.88	2.64	0.25	0.72
std	283	2.92	1.87	5.59	28.34	47.9	1.80	1.31	1.23
min	20	8×10^{-5}	2.0	10	3.67×10^{-5}	1.30×10^{-6}	−1.20	2.40×10^{-13}	7.56×10^{-5}
25 %	200	0.16	5.5	20	0.01	0.04	1.46	4.74×10^{-6}	0.16
50 %	300	0.24	6.0	23.5	0.03	0.2	2.9	9.01×10^{-5}	0.39
75 %	497	0.48	6.5	25	0.10	4.3	3.83	3.08×10^{-3}	0.90
max	1,700	50	13.0	80	545.9	818	6.91	12.0	19.19

Table 3
Descriptive statistics of the collected data with P_{cal} (330 sets) (the outliers were included).

	Frequency (kHz)	Power density (kW/L)	pH	Temp.	Initial conc. (mM)	Solubility (g/L)	Log K_{ow}	Vapor pressure (mmHg)	Rate constants ($\times 10^{-3}$/sec)
mean	432	0.47	5.84	24.6	1.11	15.67	2.74	16.66	1.24
std	396	1.79	1.96	9.59	6.42	53.64	2.03	69.08	4.42
min	20	0.02	2.0	1	4.76×10^{-5}	1.30×10^{-6}	−9.7	2.57×10^{-12}	7.56×10^{-5}
25 %	35	0.08	5.5	20	26.7×10^{-3}	0.02	1.94	6.68×10^{-6}	0.08
50 %	404	0.15	6.0	25	0.086	0.50	2.9	9.01×10^{-5}	0.23
75 %	600	0.25	7.0	25	0.5	10	3.83	0.087	0.78
max	1,700	12.6	12.7	70	100	717	6.91	297	44.33

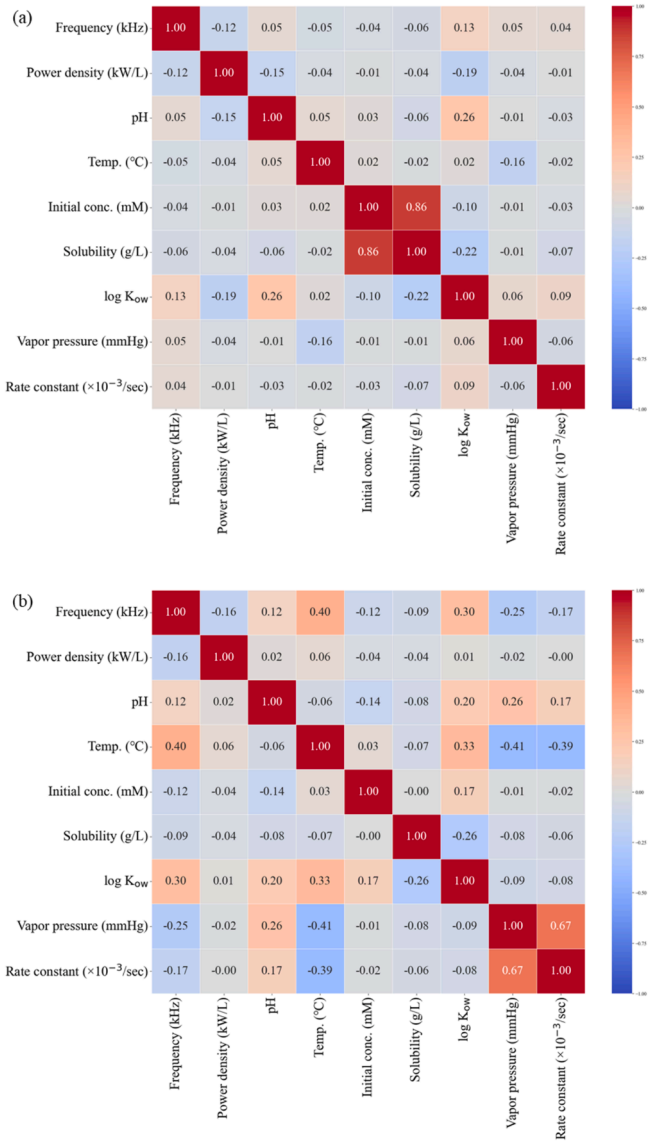


Fig. 2. Pearson correlation matrix of eight input variables and one target variable: a) data with P_{ele} including the outliers, b) data with P_{cal} including the outliers.

disulfide: 297 mmHg)]. As shown in Fig. 5S(b), a weak linear relationship between the vapor pressure and rate constant was obtained using data that excluded outliers.

ML algorithms are widely recognized for their ability to handle nonlinear relationships and offer more flexible solutions when the data distribution is unknown or nonnormal. Therefore, given the limitations of simple statistical methods in providing valuable conclusions, we opted for ML methods to analyze our data further [47]. The feature engineering techniques were not used in this study because of relatively small number of features [60–62].

Table 4
Evaluation of the performance of ML models for the rate constant prediction using the RMSE, MAE, and R^2 ["Train" represents the accuracy evaluation result for the test using training data (randomly selected 90% of the data) and "Test" represents the accuracy evaluation result for the test using untraining data (remaining 10% of the data) in K-fold cross validation.].

Evaluation of ML models		Data with P_{ele} including the outliers	Data with P_{ele} excluding the outliers	Data with P_{cal} including the outliers	Data with P_{cal} excluding the outliers
RF	RMSE (Train)	0.4265	0.0368	0.2185	0.0020
	MAE (Train)	0.1220	0.0742	0.0988	0.0187
	R^2 (Train)	0.5937	0.9134	0.7791	0.9334
	RMSE (Test)	0.9679	0.2286	0.5358	0.0160
	MAE (Test)	0.3883	0.2740	0.2357	0.0744
	R^2 (Test)	−0.2514	0.4041	−0.1461	0.4792
XGB	RMSE (Train)	0.4588	0.0505	0.0791	0.0019
	MAE (Train)	0.1683	0.1109	0.0686	0.0243
	R^2 (Train)	0.5614	0.8811	0.9198	0.9360
	RMSE (Test)	0.9586	0.2144	0.5457	0.0160
	MAE (Test)	0.3487	0.2616	0.1813	0.0678
	R^2 (Test)	−0.1954	0.4490	−2.4308	0.4709
LGB	RMSE (Train)	0.4231	0.0318	0.0020	0.0001
	MAE (Train)	0.1103	0.0559	0.0197	0.0052
	R^2 (Train)	0.5975	0.9251	0.9979	0.9972
	RMSE (Test)	0.9688	0.2380	0.3436	0.0140
	MAE (Test)	0.3702	0.2892	0.1775	0.0665
	R^2 (Test)	−0.2378	0.3414	0.2291	0.5357

3.2. Comparison of generated models

The performance of the trained ML models was evaluated for the prediction of rate constants based on their accuracy for four types of data using the evaluation metrics of RMSE, MAE, and R^2: 1) data with P_{ele} including the outliers (374 sets), 2) data with P_{ele} excluding the outliers (326 sets), 3) data with P_{cal} including the outliers (330 sets), and 4) data with P_{cal} excluding the outliers (275 sets). Table 4 shows the averaged

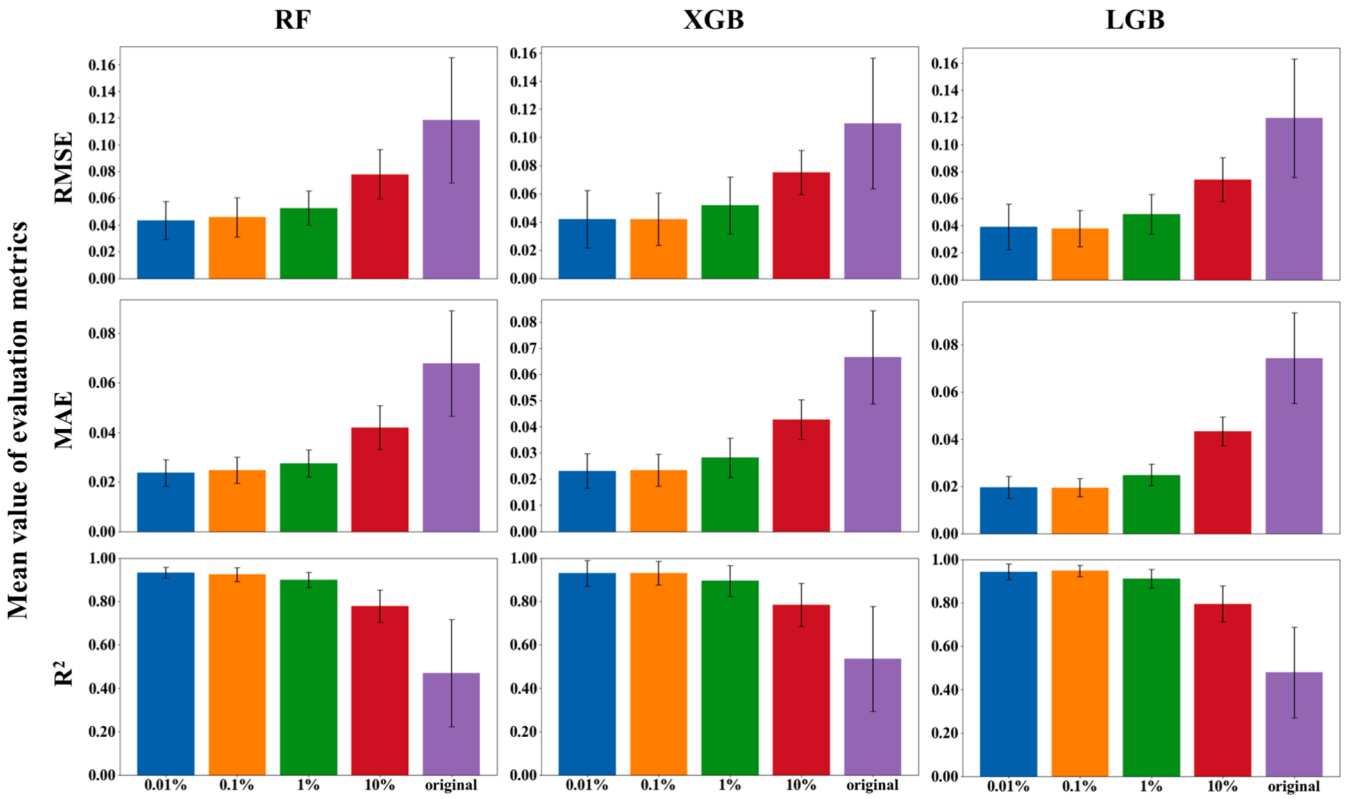


Fig. 3. Comparison of performance of ML models using the augmented data.

results from 10 iterations for the RMSE, MAE, and R^2 for the cases of "Train" and "Test": "Train" represents the accuracy evaluation result for the test using training data (randomly selected 90 % of the data) and "Test" represents the accuracy evaluation result for the test using untraining data (remaining 10 % of the data) in K-fold cross validation. The MAE and RMSE values closer to 0 and a value of R^2 closer to 1 indicate a higher prediction performance [55,56].

Higher accuracy was obtained for the data with P_{cal} than those with P_{ele}, which indicated that the P_{cal} data might be more appropriate for quantitatively analyzing sonochemical activity and applying the ML algorithms. It should be noted that the P_{ele} data included the measured P_{ele} using a power meter, and the displayed (or provided by the manufacturer) P_{ele} on a device. Moreover, the results using data that excluded outliers showed higher accuracy because of the more focused ML effect for smaller scopes [42–45]. However, this could significantly reduce diversity in the collected datasets and narrow the variable scope of the ML application. More datasets should be collected to reduce the outliers and appropriate outlier handling methods should be developed to ensure diversity in datasets [63,64].

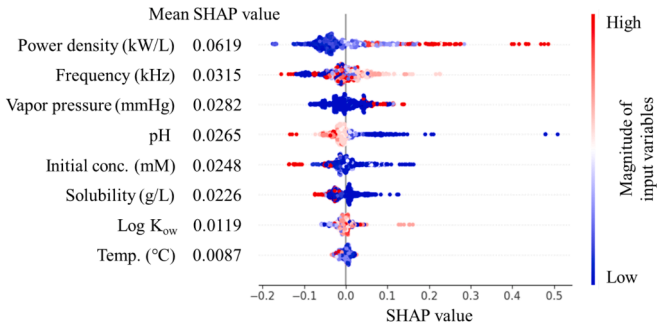


Fig. 4. Input variable contribution analysis on the target value (rate constants) using the SHAP method.

It was found that the evaluation metrics for the test using "Train" data resulted in higher accuracy than the test using "Test" data for RMSE, MAE, and R^2. This was because the trained ML model was tested on the data sets that it had already experienced for the cases of "Train". On the other hand, the evaluation metrics for "Test" were obtained from the new data sets that the ML model had never experienced [65]. This indicates that ML methods might be somewhat vulnerable for less-trained or unseen data, which is why more related data should be accumulated over wider ranges for the potential use of ML. The results in Table 4 also indicated that relatively low accuracy for "Test" might represent the data in wide ranges for the variables collected in this study [66]. Among the three ML models, the XGB model showed the most reliable results, especially for the data with P_{cal} excluding the outliers. The difficulty of determining which ML model is more appropriate for handling collected data has previously been reported [19,25,30,67].

3.3. Data augmentation

To improve the accuracy of ML performance, data augmentation has been widely utilized to address data scarcity in the image and natural language domains [68–70]. Data augmentation has been also used to increase the number of chemical structure datasets in chemistry [71–76]. However, in the chemical and environmental engineering fields, its application in ML/DL has rarely been reported. Numerical data can be augmented using a Gaussian distribution, or by randomly generating new datasets of input variables from the original data [77]. However, data that are randomly generated or significantly different from the original data may result in low accuracy of the ML model, despite the use of more datasets. In this study, we suggested a novel noise injection method to increase the size of the training data. New datasets were generated by adding values ranging from ± 0.01 % to 10 % of the original dataset (data with P_{cal} excluding the outliers), which were used for the prediction of the rate constants in the ML models.

As shown in Fig. 3, the augmented data resulted in significantly

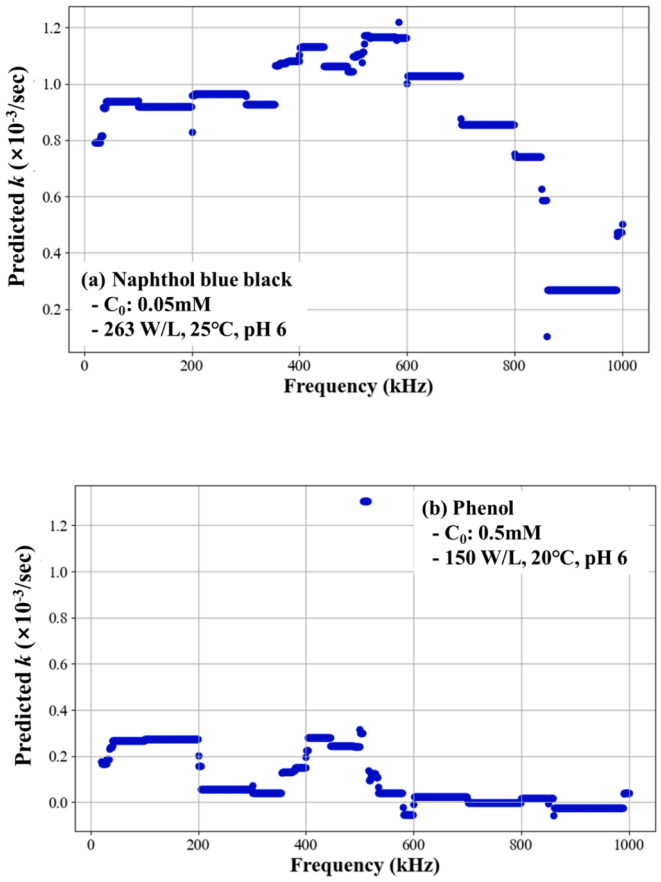


Fig. 5. Predicted pseudo-1st-order reaction kinetic constants for the sonochemical degradation of naphthol blue black or phenol under various frequency conditions (20 to 1,000 kHz) using the XGB model with the augmented data.

higher accuracy for all applied models, and the smallest change in the original datasets (adding a value of 0.01 %) induced the most reliable ML performance in this study. This indicates that an increase in the number of datasets with a slight change in the variables could noticeably enhance the ML performance. However, it is not clear whether augmentation methods such as noise injection are statistically reliable because few ML results using data augmentation have been reported in the context of chemical and environmental engineering data.

It has been known that ML models are barely understood intuitively and they are called as black-box models [57,78]. To understand the contribution of each input variable to the predicted target values (rate constants in this study), the outcome of the trained ML method, the SHAP method, was applied to the aforementioned augmented data. In Fig. 4, the input variables are listed from top to bottom in the order in which they predominantly affect the ML predicted target value, and the most influential variables were the power density (Mean SHAP value: 0.0619) and frequency (Mean SHAP value: 0.0315) in this study. Vapor pressure, pH, initial conc., and solubility also made considerable contributions to the ML performance. In addition, more positive SHAP values of the input variables have a more positive impact on the predicted target values, whereas negative SHAP values have the opposite effect. Thus, the application of a high power density (positive SHAP values/red color and negative SHAP values/blue color), moderate frequency (positive SHAP values/white color and negative SHAP value/red and blue color), and low pH (positive SHAP values/blue color and negative SHAP values/red color) can result in high rate constants for sonochemical degradation. This result appeared somewhat consistent with previous sonochemistry results [79,80] and further data accumulation was required to better understand the contribution of the input variables in the ML analysis.

3.4. Application of trained ML model

The applicability of the trained ML models using augmented data was investigated to predict the pseudo-1st-order reaction kinetic constant for sonochemical degradation in the frequency range of 20–1,000 kHz. Fig. 5 shows the predicted results of the ML models for two pollutants, naphthol blue black and phenol, which were the most reported pollutants in our collected data. Other input variables, such as the power density, initial concentration, temperature, and pH, were determined as the median values of the collected datasets of the target pollutant (naphthol blue black or phenol), as shown in Fig. 5.

Higher rate constants were obtained in the frequency ranges of 100–600 and 400–500 kHz for naphthol blue black and phenol, respectively. As no significant trends in the rate constants were observed for the two pollutants, it was difficult to determine whether the predicted results were reliable. It is notable that the datasets of naphthol blue black (21 sets) were obtained from two papers from a single research group, indicating that limited experimental conditions were applied using frequencies of 585, 860, and 1140 kHz. Thus, the predicted results for the low-frequency range had relatively low accuracy. For phenol, datasets (23 sets) were obtained from eight papers by various research groups, indicating that a variety of experimental conditions and their results were trained in the ML models. The frequencies of the phenol datasets collected were 20, 30, 35, 200, 300, 500, 520, and 534 kHz. However, less training of the ML models under certain conditions may be conducted for phenol, despite the more varied frequency conditions. Consequently, it is recommended that more data over wider experimental conditions including ultrasound-based AOPs from different research groups be accumulated and analyzed to obtain more reliable predicted results using ML models.

4. Conclusion

In this study, novel ML models, including RF, XGB, and LGB, were used to predict pseudo-1st-order reaction rate constants for the sonochemical degradation of aqueous organic pollutants. A total of 618 sets of data, including ultrasonic, solution, and pollutant characteristics, were used for training the ML models. The ML models were well trained with high accuracy for the prediction of rate constants using data with P_{cal} excluding the outliers; the XGB model exhibited the highest performance among the applied models. Data augmentation can significantly enhance the accuracy of an ML model by increasing the size of the training data. However, the application of the trained ML model for the two selected pollutants in the frequency range of 20–1,000 kHz resulted in uncertain predictions. Additional data on a wider range of experimental conditions from different research groups are required for the use of ML models in sonochemistry.

CRediT authorship contribution statement

Iseul Na: Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Taeho Kim:** Validation, Methodology, Investigation. **Pengpeng Qiu:** Methodology, Validation, Writing – review & editing. **Younggyu Son:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, 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.

Acknowledgments

This work was supported by the National Research Foundation of Korea [RS-2024-00350023] and the Korea Ministry of Environment (MOE) as part of the "Subsurface Environment Management (SEM)" Program [RS-2021-KE001466].

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ultsonch.2024.107032.

References

[1] X. Zhou, W. Liang, K.I.K. Wang, H. Wang, L.T. Yang, Q. Jin, Deep-Learning-Enhanced Human Activity Recognition for Internet of Healthcare Things, IEEE Internet Things 7 (2020) 6429–6438.
[2] S. Shamshirband, M. Fathi, A. Dehzangi, A.T. Chronopoulos, H. Alinejad-Rokny, A review on deep learning approaches in healthcare systems: Taxonomies, challenges, and open issues, J. Biomed. Inform. 113 (2021) 103627.
[3] A.A. Abdullah, M.M. Hassan, Y.T. Mustafa, A Review on Bayesian Deep Learning in Healthcare: Applications and Challenges, IEEE Access 10 (2022) 36538–36562.
[4] M.J. Bianco, P. Gerstoft, J. Traer, E. Ozanich, M.A. Roch, S. Gannot, C.-A. Deledalle, Machine learning in acoustics: Theory and applications, J. Acoust. Soc. Am. 146 (2019) 3590–3628.
[5] T.P. Nagarhalli, V. Vaze, N.K. Rana, Impact of Machine Learning in Natural Language Processing: A Review, in, Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV) 2021 (2021) 1529–1534.
[6] M.R. Bachute, J.M. Subhedar, Autonomous Driving Architectures: Insights of Machine Learning and Deep Learning Algorithms, Mach. Learn. Appl. 6 (2021) 100164.
[7] D. Hoang, K. Wiegratz, Machine learning methods in finance: Recent applications and prospects, Eur. Financ. Manag. 29 (2023) 1657–1701.
[8] H. Chen, C. Zhang, H. Yu, Z. Wang, I. Duncan, X. Zhou, X. Liu, Y. Wang, S. Yang, Application of machine learning to evaluating and remediating models for energy and environmental engineering, Appl. Energ. 320 (2022) 119286.
[9] V.G. Sharmila, S.P. Shanmugavel, J.R. Banu, A review on emerging technologies and machine learning approaches for sustainable production of biofuel from biomass waste, Biomass. Bioenergy 180 (2024) 106997.
[10] A. Konya, P. Nematzadeh, Recent applications of AI to environmental disciplines: A review, Sci. Total Environ. 906 (2024) 167705.
[11] S.B.H.S. Asadollah, A. Sharafati, D. Motta, Z.M. Yaseen, River water quality index prediction and uncertainty analysis: A comparative study of machine learning models, J. Environ. Chem. Eng. 9 (2021) 104599.
[12] S. Zhong, K. Zhang, M. Bagheri, J.G. Burken, A. Gu, B. Li, X. Ma, B.L. Marrone, Z. J. Ren, J. Schrier, W. Shi, H. Tan, T. Wang, X. Wang, B.M. Wong, X. Xiao, X. Yu, J.-J. Zhu, H. Zhang, Machine Learning: New Ideas and Tools in Environmental Science and Engineering, Environ. Sci. Technol. 55 (2021) 12741–12754.
[13] J. Kang, J. Choi, D. Lee, Y. Son, UV/persulfate processes for the removal of total organic carbon from coagulation-treated industrial wastewaters, Chemosphere 346 (2024) 140609.
[14] D. Lee, J. Kang, Y. Son, Effect of violent mixing on sonochemical oxidation activity under various geometric conditions in 28-kHz sonoreactor, Ultrason. Sonochem. 101 (2023) 106659.
[15] Y. Son, Advanced Oxidation Processes Using Ultrasound Technology for Water and Wastewater Treatment, in: M. Ashokkumar (Ed.), Handbook of Ultrasonics and Sonochemistry, Springer Singapore, Singapore, 2016, pp. 1–22.
[16] Y. Son, M. Lim, J. Khim, M. Ashokkumar, Attenuation of UV Light in Large-Scale Sonophotocatalytic Reactors: The Effects of Ultrasound Irradiation and TiO_2 Concentration, Ind. Eng. Chem. Res. 51 (2012) 232–239.
[17] Z. Fang, R. Huang, P. Chelme-Ayala, Q. Shi, C. Xu, M. Gamal El-Din, Comparison of UV/Persulfate and UV/H_2O_2 for the removal of naphthenic acids and acute toxicity towards Vibrio fischeri from petroleum production process water, Sci. Total Environ. 694 (2019) 133686.
[18] Y. Lee, S. Lee, M. Cui, J. Kim, J. Ma, Z. Han, J. Khim, Improving sono-activated persulfate oxidation using mechanical mixing in a 35-kHz ultrasonic reactor: Persulfate activation mechanism and its application, Ultrason. Sonochem. 72 (2021) 105412.
[19] C.-M. Kim, Z.H. Jaffari, A. Abbas, M.F. Chowdhury, K.H. Cho, Machine learning analysis to interpret the effect of the photocatalytic reaction rate constant (k) of semiconductor-based photocatalysts on dye removal, J. Hazard. Mater. 465 (2024) 132995.
[20] S. Jiang, Y. Zhou, W. Xu, Q. Xia, M. Yi, X. Cheng, Machine learning-driven optimization and application of bimetallic catalysts in peroxymonosulfate activation for degradation of fluoroquinolone antibiotics, J. Chem. Eng. 486 (2024) 150297.
[21] Y. Zhou, Y. Ren, M. Cui, F. Guo, S. Sun, J. Ma, Z. Han, J. Khim, Sonoelectrochemical system mechanisms, design, and machine learning for predicting degradation kinetic constants of pharmaceutical pollutants, J. Chem. Eng. 478 (2023) 147266.
[22] A. Esmaeili, S. Pourranjabar Hasan Kiadeh, A. Ebrahimian Pirbazari, F. Esmaeili Khalil Saraei, A. Ebrahimian Pirbazari, A. Derakhshesh, F.-S. Tabatabai-Yazdi, CdS nanocrystallites sensitized ZnO nanosheets for visible light induced sonophotocatalytic/photocatalytic degradation of tetracycline: From experimental results to a generalized model based on machine learning methods, Chemosphere 332 (2023) 138852.
[23] S. Khademakbari, A. Ebrahimian Pirbazari, F. Esmaeili Khalil Saraei, A. Esmaeili, A. Ebrahimian Pirbazari, A. Akbari Kohnehsari, A. Derakhshesh, Designing of plasmonic 2D/1D heterostructures for ultrasound assisted photocatalytic removal of tetracycline: Experimental results and modeling, J. Alloys Compd. 975 (2024) 172994.
[24] Z. Xiao, B. Yang, X. Feng, Z. Liao, H. Shi, W. Jiang, C. Wang, N. Ren, Density Functional Theory and Machine Learning-Based Quantitative Structure-Activity Relationship Models Enabling Prediction of Contaminant Degradation Performance with Heterogeneous Peroxymonosulfate Treatments, Environ. Sci. Technol. 57 (2023) 3951–3961.
[25] S.-Z. Zhang, S. Chen, H. 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.
[26] N. Taoufik, W. Boumya, M. Achak, H. Chennouk, R. Dewil, N. 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.
[27] T. Zhu, Y. Yu, M. Chen, Z. Zong, C. Tao, An innovative method for predicting oxidation reaction rate constants by extracting vital information of organic contaminants (OCs) based on diverse molecular representations, J. Environ. Chem. Eng. 12 (2024) 112473.
[28] Y. Sun, Z. Zhao, H. Tong, B. Sun, Y. Liu, N. Ren, S. You, Machine Learning Models for Inverse Design of the Electrochemical Oxidation Process for Water Purification, Environ. Sci. Technol. 57 (2023) 17990–18000.
[29] Y. Ren, M. Cui, Y. Zhou, Y. Lee, J. Ma, Z. Han, J. Khim, Zero-valent iron based materials selection for permeable reactive barrier using machine learning, J. Hazard. Mater. 453 (2023) 131349.
[30] A. Ajami Yazdi, A. Ebrahimian Pirbazari, F. Esmaeili Khalil Saraei, A. Esmaeili, A. Ebrahimian Pirbazari, A. Akbari Kohnehsari, A. Derakhshesh, Design of 2D/2D β-$Ni(OH)_2$/ZnO heterostructures via photocatalytic deposition of nickel for sonophotocatalytic degradation of tetracycline and modeling with three supervised machine learning algorithms, Chemosphere 352 (2024) 141328.
[31] J. Glienke, W. Schillberg, M. Stelter, P. Braeutigam, Prediction of degradability of micropollutants by sonolysis in water with QSPR - a case study on phenol derivates, Ultrason. Sonochem. 82 (2022) 105867.
[32] Y. Son, J. Choi, Effects of gas saturation and sparging on sonochemical oxidation activity in open and closed systems, part II: NO_2^-/NO_3^- generation and a brief critical review, Ultrason. Sonochem. 92 (2023) 106250.
[33] D. Ghosh, A. Sarkar, A.G. Basu, S. Roy, Effect of plastic pollution on freshwater flora: A meta-analysis approach to elucidate the factors influencing plant growth and biochemical markers, Water Res. 225 (2022) 119114.
[34] I. Na, Y. Son, Sonochemical oxidation activity in 20-kHz probe-type sonicator systems: The effects of probe positions and vessel sizes, Ultrason. Sonochem. 108 (2024) 106959.
[35] Y. Son, Y. No, J. Kim, Geometric and operational optimization of 20-kHz probe-type sonoreactor for enhancing sonochemical activity, Ultrason. Sonochem. 65 (2020) 105065.
[36] Y. Son, Simple design strategy for bath-type high-frequency sonoreactors, J. Chem. Eng. 328 (2017) 654–664.
[37] Y. Son, M. Lim, M. Ashokkumar, J. Khim, Geometric Optimization of Sonoreactors for the Enhancement of Sonochemical Activity, J. Phys. Chem. C 115 (2011) 4096–4103.
[38] T. Khuyen Viet Bao, A. Yoshiyuki, K. Shinobu, Influence of Liquid Height on Mechanical and Chemical Effects in 20 kHz Sonication, Jpn. J. Appl. Phys. 52 (2013) 07HE07.
[39] Y. Asakura, T. Nishida, T. Matsuoka, S. Koda, Effects of ultrasonic frequency and liquid height on sonochemical efficiency of large-scale sonochemical reactors, Ultrason. Sonochem. 15 (2008) 244–250.
[40] P. Isnard, S. Lambert, Aqueous solubility and n-octanol/water partition coefficient correlations, Chemosphere 18 (1989) 1837–1853.
[41] J. Dai, L. Jin, L. Wang, Z. Zhang, Determination and estimation of water solubilities and octanol/water partition coefficients for derivates of benzanilides, Chemosphere 37 (1998) 1419–1426.
[42] T.V. Pollet, L. van der Meij, To Remove or not to Remove: the Impact of Outlier Handling on Significance Testing in Testosterone Data, Adap. Human Behav. Physiol. 3 (2017) 43–60.
[43] D. Dave, T.H.R. Varma, A.M. Méan, A Review of various statestical methods for Outlier Detection, International Journal of Computer Science & Engineering Technology (IJCSET) 5 (2) (2014) 137–140.
[44] M.B. Johansen, P.A. Christensen, A simple transformation independent method for outlier definition, Clinical Chemistry and Laboratory Medicine (CCLM) 56 (2018) 1524–1532.
[45] H. Liu, S. Shah, W. Jiang, On-line outlier detection and data cleaning, Comput. Chem. Eng. 28 (2004) 1635–1647.
[46] S.K. Kwak, J.H. Kim, Statistical data preparation: management of missing values and outliers, Korean J Anesthesiol 70 (2017) 407–411.
[47] L. Li, S. Rong, R. Wang, S. Yu, Recent advances in artificial intelligence and machine learning for nonlinear relationship analysis and process control in drinking water treatment: A review, J. Chem. Eng. 405 (2021) 126673.

[48] T.T. Wong, P.Y. Yeh, Reliable Accuracy Estimates from k-Fold Cross Validation, IEEE t. Knowl. Data En. 32 (2020) 1586–1594.
[49] X. Zhang, C.-A. Liu, Model averaging prediction by K-fold cross-validation, J. Econometrics 235 (2023) 280–301.
[50] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, in: LightGBM: a Highly Efficient Gradient Boosting Decision Tree, in, Curran Associates Inc., Long Beach, California, USA, 2017, pp. 3149–3157.
[51] L. Breiman, Random Forests, Mach. Learn. 45 (2001) 5–32.
[52] T. Chen, C. Guestrin, in: Xgboost: A Scalable Tree Boosting System, in, Association for Computing Machinery, San Francisco, California, USA, 2016, pp. 785–794.
[53] H.-P. Nguyen, J. Liu, E. Zio, A long-term prediction approach based on long short-term memory neural networks with automatic parameter optimization by Tree-structured Parzen Estimator and applied to time-series data of NPP steam generators, Appl. Soft Comput. 89 (2020) 106116.
[54] T.T. Khoei, S. Ismail, N. Kaabouch, Boosting-based Models with Tree-structured Parzen Estimator Optimization to Detect Intrusion Attacks on Smart Grid, in: *2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, 2021, pp. 0165-0170.
[55] D. Chicco, M.J. Warrens, G. Jurman, The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation, PeerJ Comput. Sci. 7 (2021) e623.
[56] T. Chai, R.R. Draxler, Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature, Geosci. Model Dev. 7 (2014) 1247–1250.
[57] E. Štrumbelj, I. Kononenko, Explaining prediction models and individual predictions with feature contributions, Knowl. Inf. Syst. 41 (2014) 647–665.
[58] Q. Ren, C. Yin, Z. Chen, M. Cheng, Y. Ren, X. Xie, Y. Li, X. Zhao, L. Xu, H. Yang, W. Li, Efficient sonoelectrochemical decomposition of chlorpyrifos in aqueous solution, Microchem. J. 145 (2019) 146–153.
[59] B. Deng, P. Chen, P. Xie, Z. Wei, S. Zhao, Iterative machine learning method for screening high-performance catalysts for H_2O_2 production, Chem. Eng. Sci. 267 (2023) 118368.
[60] S.A. Bini, Artificial Intelligence, Machine Learning, Deep Learning, and Cognitive Computing: What Do These Terms Mean and How Will They Impact Health Care? J. Arthroplasty 33 (2018) 2358–2361.
[61] M. Di Mauro, G. Galatro, G. Fortino, A. Liotta, Supervised feature selection techniques in network intrusion detection: A critical review, Eng. Appl. Artif. Intel. 101 (2021) 104216.
[62] T. Verdonck, B. Baesens, M. Óskarsdóttir, S. vanden Broucke, Special issue on feature engineering editorial, Mach. Learn., 113 (2024) 3917-3928.
[63] A. Smiti, A critical overview of outlier detection methods, Comput. Sci. Rev. 38 (2020) 100306.
[64] H. Wang, M.J. Bah, M. Hammad, Progress in Outlier Detection Techniques: A Survey, IEEE Access 7 (2019) 107964–108000.
[65] M.K. Uçar, M. Nour, H. Sindi, K. Polat, The Effect of Training and Testing Process on Machine Learning in Biomedical Datasets, Math. Probl. Eng. 2020 (2020) 2836236.
[66] X. Ying, An Overview of Overfitting and its Solutions, J. Phys. Conf. Ser. 1168 (2019) 022022.
[67] R. Mohammadzadeh Kakhki, Y. Jafarian shahri, M. Mohammadpoor, KCl mediated Ag/Co@Fe_2O_3/C_3N_4 heterojunction as a highly efficient visible photocatalyst for tetracycline degradation: Application of machine learning, J. Mol. Struct. 1299 (2024) 137139.
[68] R.M. Zur, Y. Jiang, L.L. Pesce, K. Drukker, Noise injection for training artificial neural networks: A comparison with weight decay and early stopping, Med. Phys. 36 (2009) 4810–4818.
[69] U. Khan, S. Zahid, M.A. Ali, A. Ul-Hasan, F. Shafait, TabAug: Data Driven Augmentation for Enhanced Table Structure Recognition, in: J. Lladós, D. Lopresti, S. Uchida (Eds.), Document Analysis and Recognition – ICDAR 2021, Springer International Publishing, Cham, 2021, pp. 585–601.
[70] K. Maharana, S. Mondal, B. Nemade, A review: Data pre-processing and data augmentation techniques, Glob. Transit. Proc. 3 (2022) 91–99.
[71] Y. Zhang, L. Wang, X. Wang, C. Zhang, J. Ge, J. Tang, A. Su, H. Duan, Data augmentation and transfer learning strategies for reaction prediction in low chemical data regimes, Org. Chem. Front. 8 (2021) 1415–1423.
[72] R. Magar, Y. Wang, C. Lorsung, C. Liang, H. Ramasubramanian, P. Li, A. Barati Farimani, AugLiChem: data augmentation library of chemical structures for machine learning, Mach. Learn.: Sci. Technol. 3 (2022) 045015.
[73] N. Ulrich, K.-U. Goss, A. Ebert, Exploring the octanol–water partition coefficient dataset using deep learning techniques and data augmentation, Commun. Chem. 4 (2021) 90.
[74] S. Zhong, J. Hu, X. Yu, H. Zhang, Molecular image-convolutional neural network (CNN) assisted QSAR models for predicting contaminant reactivity toward OH radicals: Transfer learning, data augmentation and model interpretation, J. Chem. Eng. 408 (2021) 127998.
[75] X. Wu, Y. Zhang, J. Yu, C. Zhang, H. Qiao, Y. Wu, X. Wang, Z. Wu, H. Duan, Virtual data augmentation method for reaction prediction, Sci. Rep. 12 (2022) 17098.
[76] F. Fricke, M. Brandalero, S. Liehr, S. Kern, K. Meyer, S. Kowarik, R. Hierzegger, S. Westerdick, M. Maiwald, M. Hübner, Artificial Intelligence for Mass Spectrometry and Nuclear Magnetic Resonance Spectroscopy Using a Novel Data Augmentation Method, IEEE Trans. Emerg. Top. Comput. 10 (2022) 87–98.
[77] M. Arslan, M. Guzel, M. Demirci, S. Ozdemir, SMOTE and Gaussian Noise Based Sensor Data Augmentation, in: *2019 4th International Conference on Computer Science and Engineering (UBMK)*, 2019, pp. 1-5.
[78] Z. Li, Extracting spatial effects from machine learning model using local interpretation method: An example of SHAP and XGBoost, Comput. Environ. Urban Syst. 96 (2022) 101845.
[79] Z. Wei, R. Spinney, R. Ke, Z. Yang, R. Xiao, Effect of pH on the sonochemical degradation of organic pollutants, Environ. Chem. Lett. 14 (2016) 163–182.
[80] R.J. Wood, J. Lee, M.J. Bussemaker, A parametric review of sonochemistry: Control and augmentation of sonochemical activity in aqueous solutions, Ultrason. Sonochem. 38 (2017) 351–370.