3P4-4

Predicting the rate reaction constant of sonochemical process using machine learning

Iseul Na[1,2†], Yeji Lee[1,2], Suwan An[1], and Younggyu Son[1,2*]
([1] Dept. Environ. Eng., Kumoh Nat'l Inst. Technol.; [2] Dept. Energy Eng. Converg., Kumoh Nat'l Inst. Technol.)

1. Introduction

Recently, researchers in a variety of fields have been applying machine learning (ML) techniques to data analysis[1-4]. Recent advancements in ML have been leveraged in Advanced Oxidation Processes (AOPs) to analyze the experimental data or collecting data from existing research to analyze the variables and predict the degradation efficiency[2-4]. Recently, researchers have reported the application of ML techniques to ultrasound-based AOPs[3,4]. Most studies used only a small amount of their own experimental data for ML training and performance evaluation. This study aims to investigate the applicability of ML to sonochemical single process data from previous studies. A total of 618 datasets from 89 previous research papers were used for the training of ML.

2. Experiment and analytical methodology

A flow chart for ML using data from previous research on sonochemical processes in this study is shown in Fig. 1. A total of 618 data was collated from 89 research papers covering pollutant degradation via sonochemical processes. The collected data was imputed for missing values. Nine features related to ultrasound, solution, and pollutants were selected based on three distinct characteristics and utilized as input variables for ML. These features included: - Ultrasonic: frequency, power density, sonicator type (bath, probe) - Solution: initial pH, Temperature - Pollutants: initial concentration, solubility, vapor pressure, log K_{ow}.

The collected data was divided according to the power condition. By statistically analyzing the datasets, outliers, whose values were higher or lower than three standard deviations(std) of each variable were determined and removed. ML models including Random forest, XGBoost, LightGBM were used for prediction of pseudo-1st-order reaction rate constants. The performance of models was evaluated using root mean squared error (RMSE), mean absolute error (MAE) and R squared (R^2). RMSE and MAE values closer to 0, R^2 values closer to 1 indicating a higher model performance. The SHAP value method was employed to analyze the contribution of each input variable to the predicted target variable.

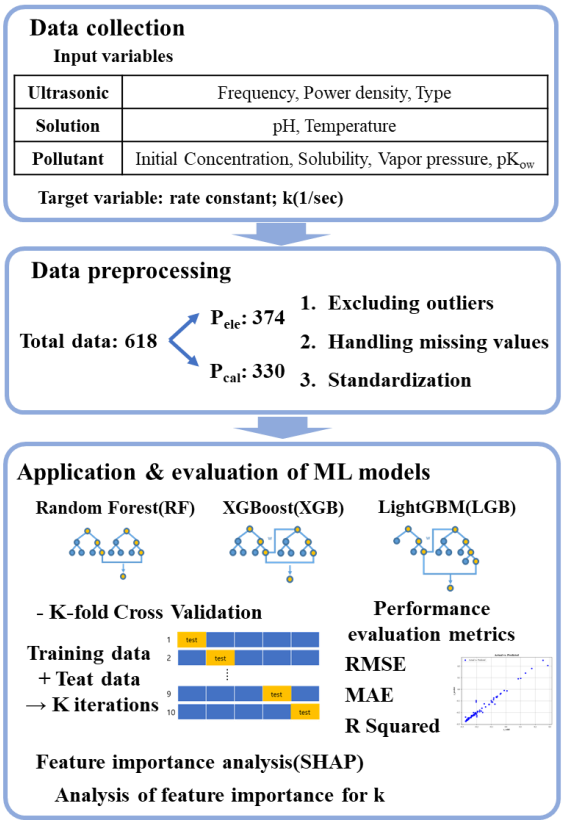

Fig. 1 Flow chart of ML models for predicting rate constants.

3. Result and discussion

The distributions of the numerical data of P_{ele} and P_{cal} datasets were shown in Fig. 2. Even though the outlier data has been removed, it is observed that the boxplot image includes points that are considered to be outliers. A much larger number of datasets for the sonochemical degradation of organic pollutants under a wide range of experimental conditions were collected than in previous research, the data were not distributed evenly.

The performance of the trained ML models was evaluated for the predication of rate constants. 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.

E-mail: †rosee1330@kumoh.ac.kr, *yson@kumoh.ac.kr

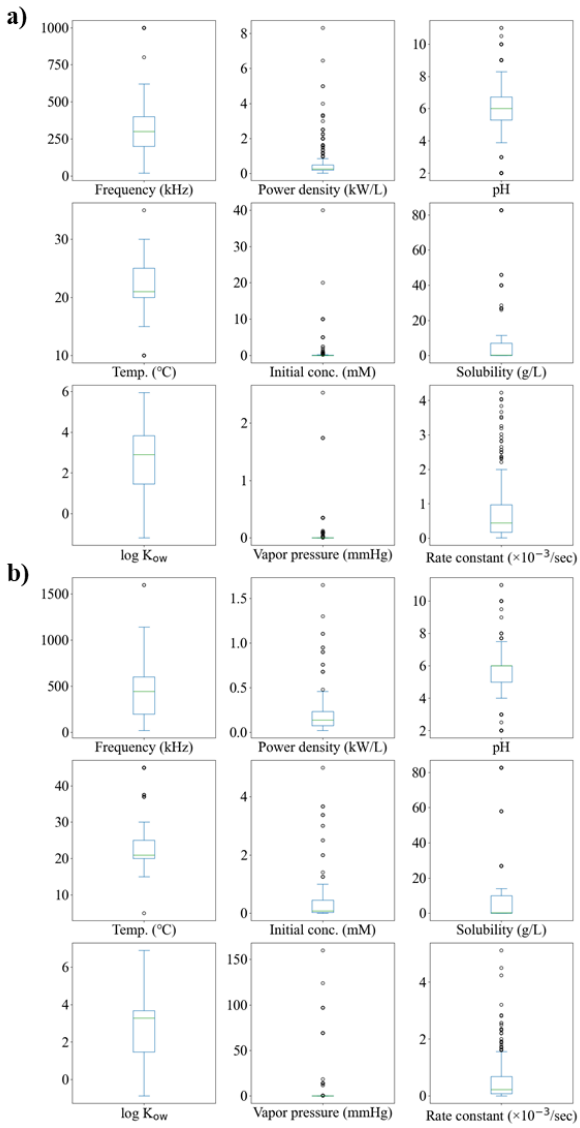

Fig. 2 Descriptive statistics of the collected data: a) P_{ele} data, b) P_{cal} data.

Table I. Evaluation of the performance of ML models for the rate constant prediction.

ML	Evaluation metrics	P_{ele}	P_{cal}
RF	MAE	0.2702	0.0759
	RMSE	0.4363	0.1238
	R^2	0.4951	0.4325
XGB	MAE	0.2722	0.0679
	RMSE	0.4787	0.1181
	R^2	0.3288	0.4706
LGB	MAE	0.2649	0.0673
	RMSE	0.4430	0.1185
	R^2	0.4505	0.4700

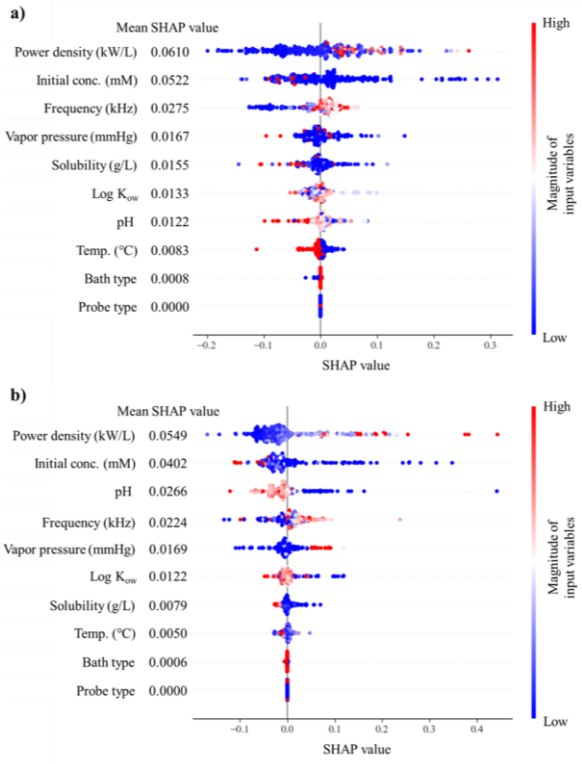

Fig. 3 Input variable contribution analysis on the ta rget variable (rate constants) using the SHAP method: a) P_{ele} data, b) P_{cal} data.

To understand the contribution of each input variables to the predicted target values, SHAP value method applied to the datasets. In Fig. 3., the input variables are listed from top to bottom in order to the most influential variables were power density, initial concentration in this study. In addition, more positive SHAP values of the input variables have a more positive impact on the predicted target, whereas negative SHAP values have the opposite effect. The application of high power density, low pH, moderate frequency can result in high rate constants for sonochemical degradation.

Acknowledgment
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].

References

1) N. Tanaka, K. Watanabe, K. Matsuoka, K. Azumagawa, T. Kozawa, T. Ikeda, Y. Komuro, D. Kawana, Jpn. J. Appl. Phys., 60, 066503 (2021).
2) Y. Zhou, Y. Ren, M. Cui, F. Guo, S. Sun, J. Ma, Z. Han, J. Khim, J. Chem. Eng. 478, 147266 (2023).
3) J. Glienke, W. Schillberg, M. Stelter, P. Braeutigam, Ultrason. Sonochem., 82 105867 (2022).
4) T. Zhu, Y. Yu, M. Chen, Z. Zong, C. Tao, J. Environ. Chem. Eng., 12, 112473 (2024).