Energy Proceedings
Vol 41, 2024
ISSN 2004-2965

Predicting the Degradation Kinetic Constants of Organic Pollutants in Sonoelectrochemical System using Machine Learning Methods

Yongyue Zhou[1], Yangmin Ren[1], Shiyu Sun[1], Fengshi Guo[1], Mingcan Cui[1], Jeehyeong Khim[1*]

1 Environmental, and Architectural Engineering Department, Korea University, Republic of Korea

(*Jeehyeong Khim: hyeong@korea.ac.kr)

ABSTRACT

The petroleum industry is one of the fastest-growing sectors and has made a significant contribution to the economic growth of developing countries. Wastewater generated by the petroleum industry contains a variety of organic pollutants. These organic compounds exist in highly complex forms in discharged water and cause environmental hazards. Sono-electrochemical system is emerging as a future trend due to its clean and non-secondary pollution characteristics. This process combines ultrasound and electrochemical methods to enhance the reaction rate constants for pollutant degradation. However, in the system design process, the complex interactions between EC, US, pollutants, and environmental parameters significantly impact the outcomes. Therefore, predicting the kinetic constants of organic compound degradation in US-EC systems within complex reaction systems is challenging. In this study, Machine learning models such as Artificial Neural Networks (ANN), Support Vector Machines (SVM), and Extreme Gradient Boosting (XGBoost) were employed to predict the degradation rates of organic compounds in US-EC systems. Comparative analysis of the prediction results from different models showed that XGBoost performed exceptionally well, with R^2 and RMSE values of 0.97 and 0.0006, respectively. SHAP analysis was conducted to evaluate the impact of design parameters on the model's predictive performance, and the results indicated that ultrasonic frequency, ultrasonic power, and the distance 'r' from the ultrasonic transducer to the electrode had the most significant influence on the model's predictive performance. This method effectively guides the parameter design of US-EC systems and enables accurate predictions of the degradation rates of organic compounds.

Keywords: Sonoelectrochemical, Design parameter, Machine learning, SHAP model, Organics degradation, Prediction method

NONMENCLATURE

Abbreviations	
ANN	Artificial Neural Networks
SVM	Support Vector Machines
XGBoost	Extreme Gradient Boosting
AOP	Advanced oxidation processes
US-EC	Sonoelectrochemical
US	Ultrasound
EC	Electrochemical
ML	Machine learning
SHAP	SHapley's Additive Interpretation
CV	5-fold cross-validation
RMSE	Root mean squared error
BDD	Boron-doped diamond
Symbols	
R^2	Coefficient of determination
$y_{i,exp}$	Actual value
$y_{i,pred}$	Forecasted value
$y_{i,average}$	Mean of the value
n	Number of the datasets

1. INTRODUCTION

The petroleum industry has become one of the fastest growing industries and has made an important contribution to the economic growth of developing countries with the continuous development of energy. However, wastewater produced through the petroleum industry contains a variety of pollutants such as petroleum hydrocarbons, phenol, ammonia, sulfides and other organic compounds. Easily discharged into water,

This is a paper for 15th International Conference on Applied Energy (ICAE2023), Dec. 3-7, 2023, Doha, Qatar.

causing serious harm to the environment and health[1]. Due to the complex structure of the compound, it is difficult to degrade and is usually not degraded by biological treatment. In order to minimize the impact of pollutants on the environment, advanced oxidation processes (AOP) have received more and more attention as an advanced pollutant treatment technology. AOP technology primarily generates and utilizes free radicals to oxidize and react with many types of organic compounds, producing shorter and simpler organic compounds, or complete mineralization[2]. In AOP technology, sonoelectrochemical (US-EC) oxidation process combines two powerful technologies of ultrasound (US) and electrochemical (EC), and has become a rapidly developing field in recent years because of its clean, safe, and efficient treatment technology[3]. However, challenges remain with this technology.

Previous studies on the influence of different design parameters on the results of US-EC systems generally used the response surface design method[4,5]. However, this approach has certain limitations. First, the experimental design may require a lot of time and resources for problems with multiple factors and large dimensions. Second, the experimental design must take into account the interaction between the factors, otherwise the response surface model may not accurately describe the relationship between the response variable and the factors. In addition, response surface models can only predict responses within the range of factor levels used in the experiment, but not responses outside this range. Finally, the response surface model is an empirical model whose accuracy is influenced by experimental design and data quality. It is difficult for researchers to estimate the relative contributions of all factors and focus on the key factors that influence contaminant degradation rates and reactor design. To overcome the limitations of existing models, a more comprehensive approach is urgently needed to summarize the effects of various conditions on contaminant degradation in US-EC systems and make predictions for future work in this area.

Recently, advances in computer science have led to a growing interest in the use of artificial intelligence for prediction. Machine learning (ML) methods have proven to be powerful prediction tools as they can solve linear and complex nonlinear problems[6,7]. This method is more flexible than the traditional method of multifactorial impact analysis. Machine learning can process various types of data, including structured and unstructured data, and can automatically recognize and learn complex nonlinear relationships between input and output variables, resulting in more accurate prediction models. At the same time, machine learning methods can train models on limited data sets, saving time and resources.

This study uses machine learning methods to predict the degradation rate of organic pollutants, and chooses the most general ANN, SVM and XGBoost as models. Among them, ANN is used in the form of supervised learning to identify complex relationships between input and output and evaluate the results of input data, and SVM is implemented by applying various nonlinear mapping functions to analyze and present small samples, nonlinear, models with multidimensional and local minima. As an ensemble model, XGBoost is estimated using the average of many simple models, which can capture complex information in the data. It has been used for pollutant degradation prediction[8]. However, studies of this ML approach have not yet been applied to predict the degradation of pollutants in US-EC systems. In order to understand how variables affect the prediction results. This study performed SHapley's Additive Interpretation (SHAP) analysis. SHAP can generate feature influences for each instance and provide guaranteed explanations for them[9].

The overall goal of this study is to predict the reaction rate of organic pollutants degradation using three ML models in a US-EC system and compare the accuracy of the models. Finally, SHAP analysis is used to evaluate the impact of design parameters on the model's predictive performance.

2. METHODS

2.1 Data collection and preprocessing

In order to predict the degradation rate represented by the k-value, various influencing parameters were taken into account in this study. Fifteen variables were selected as input factors, including six EC indicators, two US indicators, four pollutant indicators and two environmental indicators. The dataset included 90 sets of experimental data and another 14 sets derived from literature sources. The datasets were compiled by searching for the keywords 'sonoelectrochemical' and 'pollutant degradation' on Google Scholar[10] and the relevant information was extracted for use as input and output data. To understand the statistical relationships between each parameter and the 'k' value, boxplots were used for analysis. To enable machine learning algorithms to process the data effectively, two

categorical input features, namely, electrode material and chemical formula, were transformed into numerical values using the one-hot encoding technique. This transformation step is crucial as machine learning algorithms work exclusively with numerical data. In addition, all numerical input features were standardized to fall within a range of 0 to 1 before being fed into the neural network. This standardization step ensures that variables measured at different scales contribute equally to the training process of the model[11].

2.2 Model training and testing

In this study, using machine learning methods to predict the degradation rate of drugs including IBP, and chooses the most general ANN, SVM and XGBoost as models. In order to train and test the model, the entire data set is divided into two parts, of which 90% are used as the training data set, and the 10% are used as the test data set. Verify the generalization ability of the trained model. In the training procedure, the grid search method is used to find the optimal hyperparameters and a 5-fold cross-validation (CV) method is used to reduce the bias arising from the random sampling of the training set. In ANN, use Relu as activation function and 3-layers of ANN was constructed. The SVM was built using a RBF kernel. Set to the value range of the cost constant, epsilon, and gamma were optimized. The entire construction process is shown in Fig.1.To evaluate the performance of the proposed method, the most widely used evaluation metrics were adopted: the Coefficient of determination (R^2) and root mean squared error (RMSE) were utilized to compare the accuracy of predictions. The equations were given in eq (1) and (2).

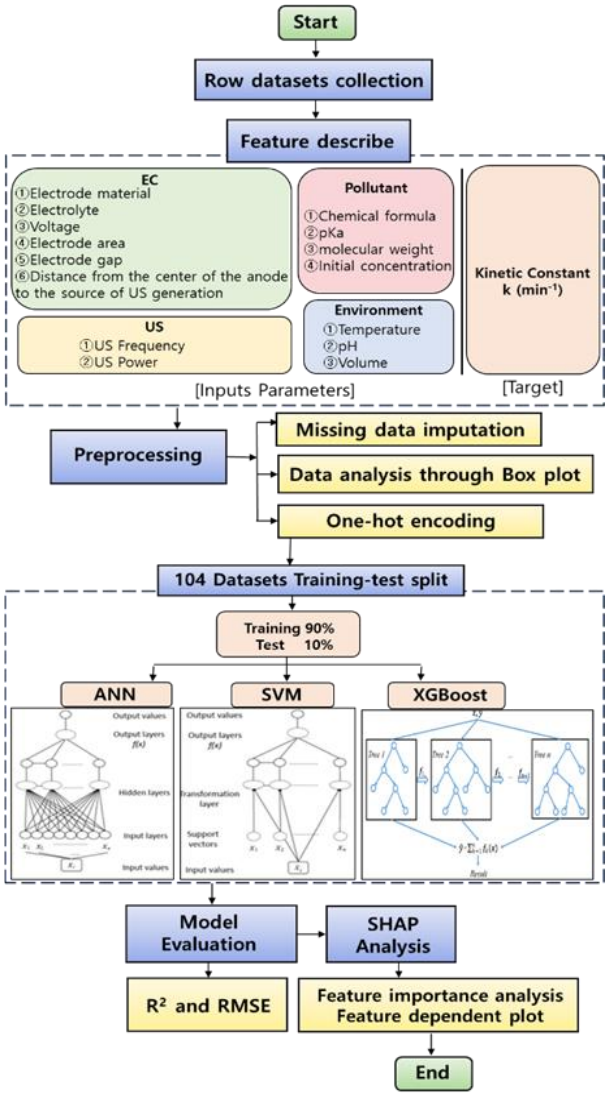

Fig. 1. Flowchart of the strategies of the modeling framework to predict the kinetic constant of pollutants degradation process in US-EC system.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_{i,exp} - y_{i,pred})^2}{\sum_{i=1}^{n}(y_{i,exp} - \overline{y_{average}^{target}})^2} \quad (1)$$

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(y_{i,exp} - y_{i,pred})^2}{n}} \quad (2)$$

2.3 Model performance evaluation

SHAP analysis was used to evaluate variables' importance to models. SHAP analysis is a locally accurate and consistent feature attribution method that provides more stable rankings than previous importance measures[12]. SHAP values attribute the marginal contribution from each predictor variable to each prediction (measured relative to the average prediction).

3. RESULTS AND DISCUSSION

3.1 Section of material and methods

According to the flowchart, the first step involves the collection of raw data sets. A total of 104 datasets were collected, containing 15 input features and 1 target variable, with consistent units. Next, during the feature engineering phase, we categorized the database based on the target pollutant and response system conditions. To ensure that no data were overlooked, we were guided by relevant literature studies when selecting descriptors for this study. The primary characteristics of the target pollutant included drug-related information such as molecular formula, molecular weight, initial pollutant concentration, and pKa coefficient. Reaction system conditions included parameters such as ultrasonic frequency, ultrasonic input power, electrode material,

electrolyte concentration, voltage, electrode area, distance and electrode position. Environmental conditions such as temperature, volume and pH were also taken into account. All of these factors were taken into account when evaluating their effects on the reaction kinetics. Prior to applying machine learning, the data were subjected to manual categorization, where the headings for each group of influencing factors (e.g., EC_material, EC_Electrolyte, EC_Voltage, etc.) were changed in the preprocessing steps. Although it was expected that all data could be taken directly from literature sources, there was insufficient information on the electrode positions in ultrasonic reactors. The missing data affected the size of the database and the prediction accuracy of the model. Therefore, the missing data was supplemented by calculating the positions based on the given reactor dimensions, electrode spacing and immersion depths. Based on the distribution in the boxplots, outliers were identified, which mainly originated from data sets from other literature with different experimental conditions than in our study. To ensure the accuracy of variable significance, these outliers were removed, reducing the database from 104 to 102 data sets.

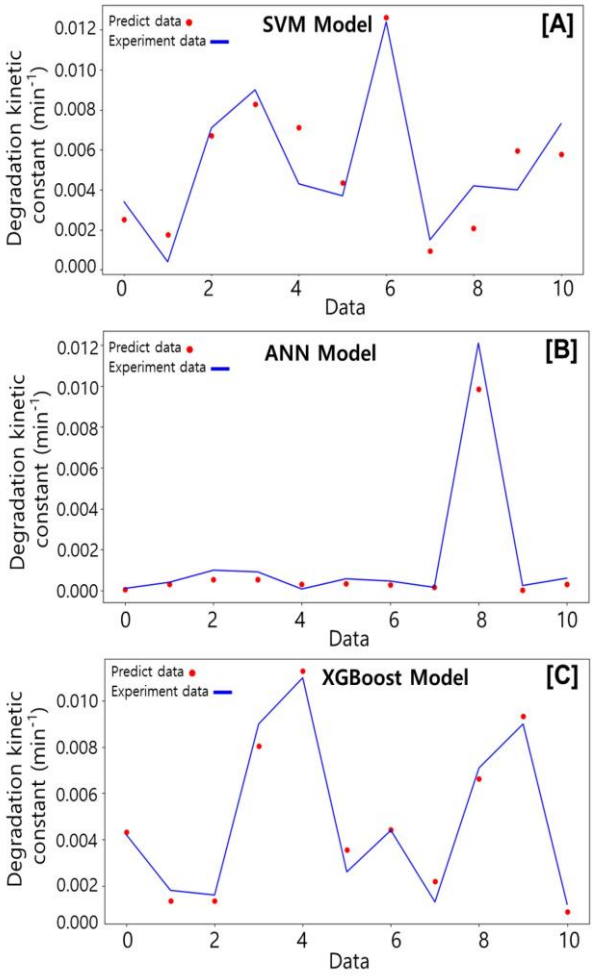

Fig. 2. Results of original and predicted values under [A] SVR, [B] ANN, and [C] XGBoost models.

Two categorical variables within the input data (reaction material and chemical formula) were transformed using one-hot encoding. Following preprocessing, all datasets were randomly divided into two groups, with 90% of the data used for model training and the remaining 10% for validation. Fig 2 illustrates the results of comparing all tested k values with the original k values.

The R^2 for the SVR, ANN, and XGBoost models were 0.81, 0.95, and 0.97, respectively, the RMSE results were as follows:0.006, 0.007, 0.0006. These results indicate that XGBoost outperforms ANN and SVR in terms of predictive performance. Furthermore, the model Train Score is 0.99 for XGBoost which means model is not overfitting. This may be due to the regularization term added to the XGBoost model to control the complexity of the model and reduce the risk of overfitting.

3.2 SHAP analysis

This study not only evaluated how well the k value prediction model can make accurate predictions, but also analyzed the insights provided by the trained model on the importance of different variables using SHAP.

The analysis of feature importance was divided into two groups: categorical variables and numerical variables. Among the categorical variables, shown in Fig. 3A, there was a significant effect on the reaction rate when BDD (boron-doped diamond) was used as the electrode material. This effect is due to the higher oxidation potential of BDD compared to other electrode materials, which allows it to effectively oxidize more pollutants and increase the degradation rate[13]. In contrast, other categorical variables had no significant effect on the results. Therefore, the main focus of this study was to investigate the effects of numerical variables. The importance of numerical variables, as shown in Fig.3B, was assessed in descending order of importance. It was found that US power and frequency played a central role in predicting reaction time, as their values significantly influenced the outcome. These variables were followed in importance by the distance from the origin to the coordinate position (r), the applied voltage, the electrode area, the electrode spacing and the electrolyte. The other input parameters had a negligible influence on the results. It is noteworthy that the two most important predictors in terms of weighting

contained key information for the prediction of the k value, indicating their crucial role in predicting the reaction rate constant. Consequently, it is advisable to prioritize the consideration of US power and frequency when developing a US-EC system.

To gain deeper insights into the positive or negative relationships between different indicators and the outcome variable, a special variant of SHAP, called Tree SHAP, was used for model interpretation in this study. Fig. 3C shows a summary SHAP diagram that combines the attribute importance and attribute effects. Each point in this diagram represents a feature and its corresponding Shapley value for a particular instance. The vertical position on the y-axis is determined by the feature, while the horizontal position on the x-axis is determined by the Shapley value. The color represents the value of the feature and ranges from low to high.

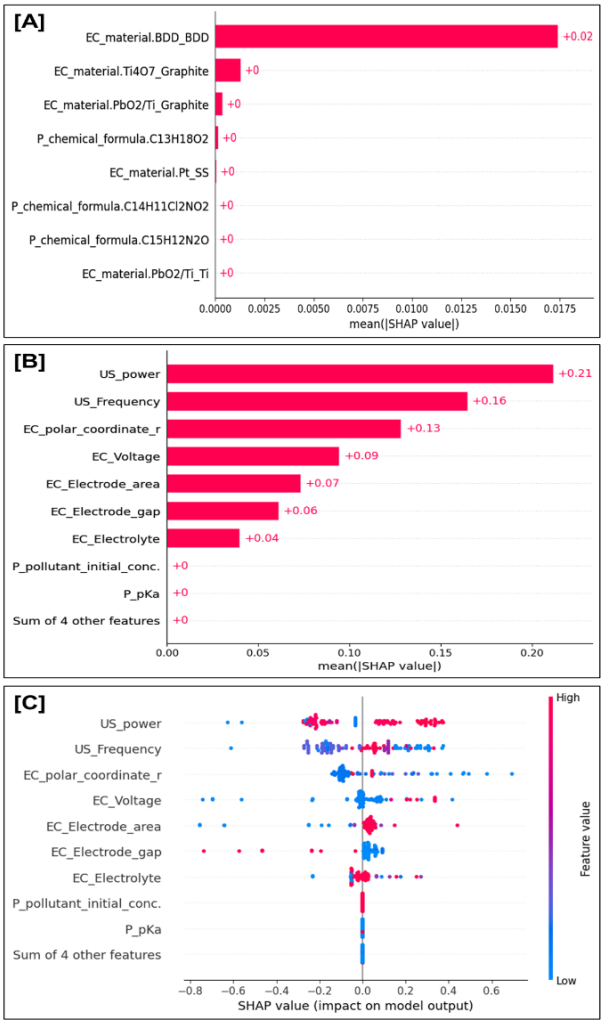

Fig. 3. SHAP feature dependency diagram for important indicators for [A] categorical and [B] numerical variables. [C] Distribution of SHAP values for all instances of each feature.

When overlapping points align along the y-axis, this provides a clear view of the distribution of Shapley values for each trait. Among these traits, US power had the greatest impact on the model, with higher trait values associated with higher Shapley values, implying a greater response rate constant. This was closely followed by US frequency and distance from origin to coordinate position (r), both of which had a significant influence, with lower feature values leading to higher reaction rate constants. In addition, the applied voltage also played a decisive role, with a higher voltage having a more positive effect on the result. Although larger electrode areas and smaller electrode spacing contributed to higher pollutant degradation rates, their significance was comparatively small. The SHAP analysis provides insight into specific areas and shows the expected trends in result changes under certain conditions. Consequently, combining this information with the actual experimental conditions is essential to determine whether a particular feature can be selected for intervention.

4. CONCLUSIONS

This study aimed to use machine learning to predict the degradation of organic compounds in the US-EC system under different influencing parameters. The results can be summarized as follows:

1. Summarized 104 datasets of in the US-EC system, and used SVR, ANN, and XGBoost models to predict the reaction rate constants for organic pollutants degradation. The best prediction model is XGBoost, with R^2 and RMSE of 0.97 and 0.0006 respectively. The model has excellent prediction performance.

2. The SHAP evaluation results show that the ultrasonic frequency, ultrasonic power and the distance r between the electrode and the ultrasonic emission source have the most significant impact on the model prediction performance.

ACKNOWLEDGEMENT

Funding: This work was supported by the Korean Ministry of the Environment as a Subsurface Environment Management (SEM) project (No.2021002170003 and 202300231376) and the Provincial Nanjing City, Department of Science project (No.202201002). Dr. Y.G. Ahn (Korea Basic Science Institute, Western Seoul Center) is gratefully acknowledged for the help with data analysis.

DECLARATION OF INTEREST STATEMENT

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. All authors read and approved the final manuscript.

REFERENCE

[1] W.F. Elmobarak, B.H. Hameed, F. Almomani, A.Z. Abdullah, A Review on the Treatment of Petroleum Refinery Wastewater Using Advanced Oxidation Processes, Catalysts. 11 (2021) 782.

[2] Giwa, A., Yusuf, A., Balogun, H. A., Sambudi, N. S., Bilad, M. R., Adeyemi, I., ... & Curcio, S. (2021). Recent advances in advanced oxidation processes for removal of contaminants from water: A comprehensive review. Process Safety and Environmental Protection, 146, 220-256.

[3] B. Thokchom, A.B. Pandit, P. Qiu, B. Park, J. Choi, J. Khim, A review on sonoelectrochemical technology as an upcoming alternative for pollutant degradation, Ultrason Sonochem. 27 (2015)210–234.

[4] P. Finkbeiner, M. Franke, F. Anschuetz, A. Ignaszak, M. Stelter, P. Braeutigam, Sonoelectrochemical degradation of the anti-inflammatory drug diclofenac in water, Chemical Engineering Journal. 273 (2015) 214–222.

[5] G. Donoso, J.R. Dominguez, T. González, S. Correia, E.M. Cuerda-Correa, Electrochemical and sonochemical advanced oxidation processes applied to tartrazine removal. Influence of operational conditions and aqueous matrix, Environ Res. 202 (2021).

[6] A. Mosavi, P. Ozturk, K.W. Chau, Flood prediction using machine learning models: Literature review, Water (Switzerland). 10 (2018).

[7] S. Chibani, F.X. Coudert, Machine learning approaches for the prediction of materials properties, APL Mater. 8 (2020).

[8] 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, Science of the Total Environment. 807 (2022).

[9] S. Wang, Y. Zhou, X. You, B. Wang, L. Du, Quantification of the antagonistic and synergistic effects of Pb2+, Cu2+, and Zn2+ bioaccumulation by living Bacillus subtilis biomass using XGBoost and SHAP, J Hazard Mater. 446 (2023).

[10] Retrieved from https://scholar.google.com, (n.d.).

[11] P. Rodríguez, M.A. Bautista, J. Gonzàlez, S. Escalera, Beyond one-hot encoding: Lower dimensional target embedding, Image Vis Comput. 75 (2018) 21–31.

[12] S.M. Lundberg, P.G. Allen, S.-I. Lee, A Unified Approach to Interpreting Model Predictions, n.d.

[13] Y. Zhou, M. Cui, Y. Ren, Y. Lee, J. Ma, Z. Han, J. Khim, Evaluation of anode materials in sonoelectrochemistry processes: Kinetic, mechanism, and cost estimation, Chemosphere. 306 (2022) 135547.