Carbon Research

ORIGINAL ARTICLE

Open Access

Predictive capability of rough set machine learning in tetracycline adsorption using biochar

Paramasivan Balasubramanian[1], Muhil Raj Prabhakar[1], Chong Liu[2*], Pengyan Zhang[3] and Fayong Li[3]

Abstract

Machine learning algorithms investigate relationships in data to deliver useful outputs. However, past models required complete datasets as a prerequisite. In this study, rough set-based machine learning was applied using real-world incomplete datasets to generate a prediction model of biochar's adsorption capacity based on key attributes. The predictive model consists of *if–then* rules classifying properties by fulfilling certain conditions. The rules generated from both complete and incomplete datasets exhibit high certainty and coverage, along with scientific coherence. Based on the complete dataset model, optimal pyrolysis conditions, biomass characteristics and adsorption conditions were identified to maximize tetracycline adsorption capacity (> 200 mg/g) by biochar. This study demonstrates the capabilities of rough set-based machine learning using incomplete practical real-world data without compromising key features. The approach can generate valid predictive models even with missing values in datasets. Overall, the preliminary results show promise for applying rough set machine learning to real-world, incomplete data for generating biomass and biochar predictive models. However, further refinement and testing are warranted before practical implementation.

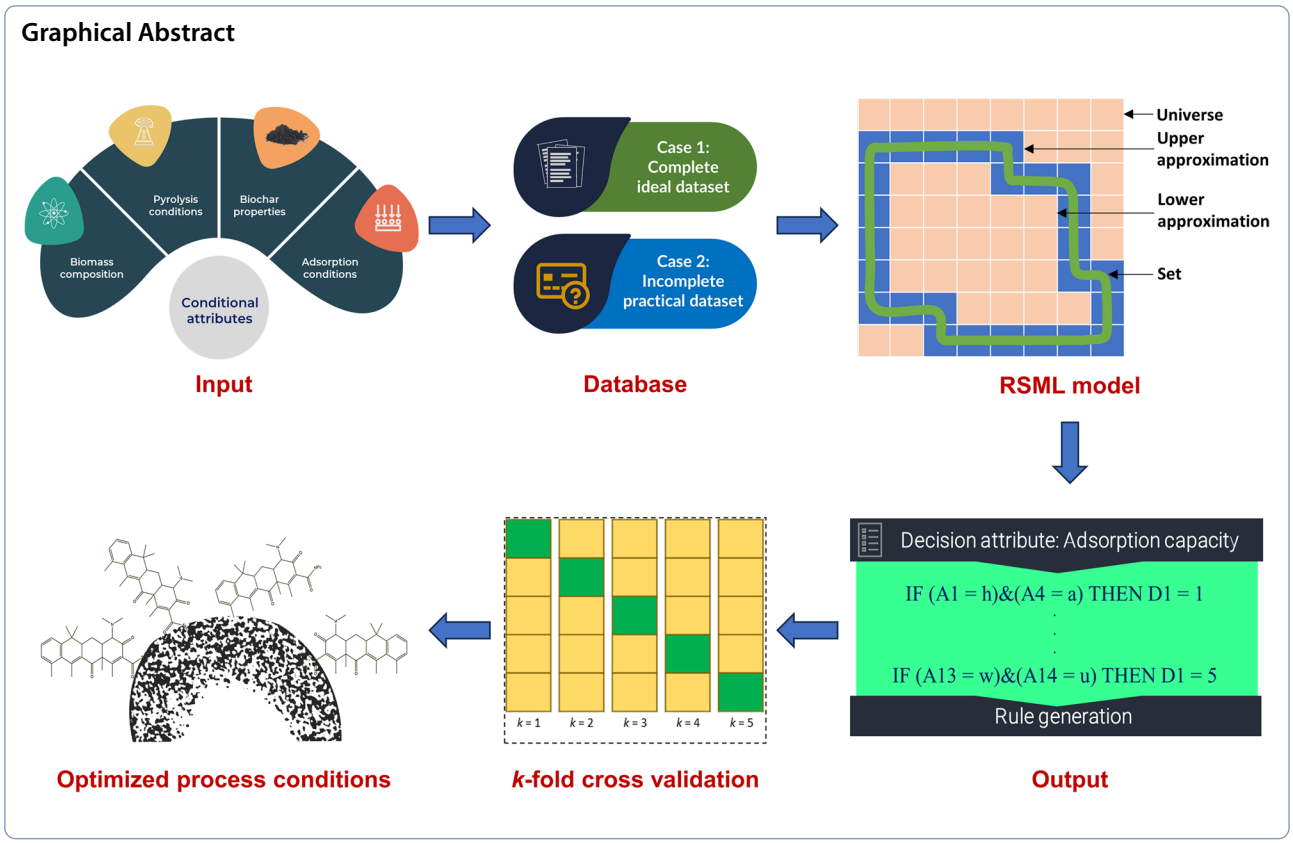
Highlights

- It is the first explainable AI-based rough set model to study the tetracycline adsorption capacity of biochar.
- Usage of an incomplete Practical dataset through RSML evaded the biasness due to imputations.
- Higher accuracy and precision of incomplete Practical datasets revealed the uniqueness of the model.

Keywords Tetracycline, Adsorption, Rough set, Machine learning, Biochar

Handling Editor: Fengchang Wu.

*Correspondence:
Chong Liu
17609858895@163.com
Full list of author information is available at the end of the article



1 Introduction

Tetracycline (TC) is a persistent organic pollutant found in surface, ground, and drinking water, which can cause endocrine disruption and transmit antibiotic-resistance genes, offering serious human health and environmental dangers (Bilal et al. 2020; Zhang et al. 2020). Researchers have paid much attention to the problems of incomplete metabolism and TC emissions recently (Zeng et al. 2021), as they are frequently used as an antimicrobial agent and feed additive in agriculture and livestock production (Gopal et al. 2020). Chemical oxidation, biological treatment, and physical removal are the main TC wastewater treatments (Phoon et al. 2020). The utilization of biological methods for the elimination of TC from wastewater poses significant challenges due to its antimicrobial properties (Zhu et al. 2021a). Because of its intrinsic benefits, such as simplicity, cheap cost, and great efficiency, adsorption is regarded as an excellent technology for treating TC (Cheng et al. 2021). Due to its distinctive qualities, including a large specific surface area, homogeneous pore distribution, and a high abundance of surface functional groups, biochar (BC) has received extensive research as an adsorbent for removing contaminants from wastewater (Akhil et al. 2021).

Van der Waals forces and hydrogen bonds, along with covalent and ionic bonding, are primarily responsible for the absorption of pollutants into BC (Thangaraj & Solomon 2019). As a result, the BC's characteristics, the adsorption environment, and the ratio of adsorbate to absorbent all play major roles in the adsorption process. Previous research has extensively assessed the traditional kinetic and isothermal adsorption models (Chen et al. 2018; Jang and Kan 2019; Liu et al. 2021). Results indicated that electrostatic interactions and chemisorption are among the potential adsorption mechanisms. Within the same framework, the relationship between each influencing element and the amount of sorption can be determined using a normal controlled-variable experimental approach. However, traditional batch sorption studies are time-consuming and inefficient for choosing the opt BC (Li et al. 2022). Predicting adsorption efficiency, improving process parameters, and understanding the adsorption mechanism require realistic tools, which further urged to explore the advancements in machine learning algorithms (Luo et al. 2023; Cao et al. 2023).

A subset of artificial intelligence known as machine learning (ML) relies on automated, data-driven model construction. The creation of ML models involves the use of various training techniques. Recent works applied ML

algorithms to carbon-based materials for TC adsorption (Taoufik et al. 2022; Zhu et al. 2021b). However, model predictions should be improved. The study by Zhu et al. (2021b) used carbon-based materials, including activated carbon and BC, with different compositions, so prediction models for both would result in large variance; secondly, their study had a small database, and the best correlation coefficient (R^2) was only 0.8944, necessitating optimization of the ML model (Leng et al. 2022; Yang et al. 2022). Integrated learning models must be utilized to predict TC adsorption on a single BC to test the prediction effect.

For pattern identification based on ambiguous information, many ML techniques have been developed, including rough set theory (RST) (Pawlak 1982), fuzzy set theory (Goguen 1974), and evidence theory (Dempster 1967). These theories' approximation-based methodology enables them to identify structural links in noisy and erratic data. In order to categorize things depending on the supplied data, Pawlak (1982) originally suggested RST, which is applied in classification, prediction, and decision analysis tasks, through which rough set-based machine learning (RSML) was developed. The input data for RSML's information tables are the object characteristics that are further divided into attributes for conditions and decisions (Pawlak 1997). The factors that place the object into a certain judgment class are known as condition attributes. A subset known as reduct can be created by removing duplicate condition attributes. The rough set algorithm will provide a list of categorization rules using the reduct as its base. Multiple reducts are frequently generated in a case study, and the condition attributes that appear in all the reducts are regarded as the core attributes. If the information system's essential properties were deleted, the classification rules would be rendered useless (Pawlak 1997). The use of RSML has many benefits, including the ability to produce rules based on various decision classes and validate the rules through their scientific coherency.

Though traditional ML models like artificial neural networks (ANN), support vector machines (SVM) and linear regression have been successful in many applications. They are criticized for being oblique (Rudin and Radin 2019), hard to interpret (Rudin 2019) and mandate of interpreting tools like SHAP (Shapley additive explanations) (Merrick and Taly 2020). Explainable artificial intelligence (XAI) has grown in popularity for applications that require more scientific outcomes apart from the statistical performance of the developed models as a result of these restrictions (Calegari et al. 2020). An XAI method called RSML creates *if–then* rules for categorizing data based on identified influential parameters. RSML has the advantage of locating hidden patterns in data sets and is based on Pawla's (1982) RST to categorize items based on already existing knowledge. The *if–then* rules produced by RSML characterize and generalize data to predict the future outcome qualities using these guidelines. Though recent RSML research has been undertaken in a variety of domains. To date, only very few limited papers have reported on RSML for biomass pyrolysis, such as its BC energy potential (Tang et al. 2023), BC surface properties (Ang et al. 2023), bio-oil properties (Chong et al. 2022). The other studies using RSML include forecasting CO_2 storage integrity (Aviso et al. 2019), building energy usage (Lei et al. 2021), detection of city energy usage and greenhouse gaseous (GHG) emissions (Aviso et al. 2021), estimating water quality (Albuquerque et al. 2021), forecasting impeller service life (Zhao et al. 2019), and evaluating variable effectiveness and success criteria for construction projects (Akbari et al. 2018).

BC feedstocks include numerous forms of biomass, such as crop wastes, agricultural residues, and algae. Biomass qualities (i.e., volatile matter, ash content, and carbon content) and pyrolysis conditions (i.e., temperature, heating rate and retention time) affect BC's adsorption efficiency. However, the desired BC characteristics aimed for TC removal were dispersed, with no clear conclusion. It is essential to study the conditions that produce BC for TC adsorption. RSML can categorize characteristics to study conditional and decision attributes' hidden relationships. This study examines the impact of biomass qualities and operation conditions on BC surface properties for TC removal. To the best of the authors' knowledge, no prior literature exists that applies RSML algorithms to build performance prediction models for equilibrium sorption capacity for TC based on the combination of fifteen factors. However, few studies are available for predicting TC adsorption using BC through traditional ML methods (Zhang et al. 2023; Zhou et al. 2023). Thus, this study attempts to discover the applicability of rough sets in developing the prediction model that assists in choosing the optimized conditions for accomplishing maximum equilibrium sorption capacity.

The objective of this study was to create a general RSML model to forecast the sorption capacity of TC on BC based on the adsorbent characteristics and sorption circumstances. This study utilized four key steps in RST to develop a prediction model that included discretization, identification of core attributes and generation of reducts, generation of decision rules, and evaluation. Thus, the novelty of the current study is that it explored the comparative performance evaluation of RSML with incomplete and complete datasets. Since RSML produces a set of decision rules that express conditional and decision qualities, further validating and analyzing these

rules results in the selection of decisive, interpretable rules. These guidelines should determine the biomass parameters and operating conditions that produce BC with the appropriate characteristics for maximizing TC adsorption.

2 Materials and methods

2.1 Collection, pre-processing of input BC data

Twenty-two types of BC and 295 sets of experimental adsorption data for the TC adsorption by BC were gathered (Tables S1 and S2) from recently published literature that was pertinent (Chen et al. 2018, 2021; Choi et al. 2020; Fan et al. 2020; Jang and Kan 2019; Kim et al. 2020; Shen et al. 2020; Wang et al. 2018; Xu et al. 2020; Zhang et al. 2019; Zheng et al. 2021). Data were taken from published studies using Plot Digitizer v3 (https://plotdigitizer.com/.) without author bias (Wilschut et al. 2022).

Fifteen key parameters such as pyrolysis temperature (T_{py}, °C), pH of the BC in water (*pH Char*), total carbon in the BC (*C*, w%), molar ratio of oxygen and nitrogen to carbon [*(O+N)/C*], molar ratio of oxygen to carbon (*O/C*), molar ratio of hydrogen to carbon (*H/C*), ash content (*Ash*, w%), surface area (*BET*, m^2/g^{-1}), pore volume (*PV*, cm^3/g^{-1}), and BC pore size (*PS*, nm), adsorption temperature (T_a, °C) and adsorption solution pH (*pH Solution*) were compiled from the published literature and dataset was developed. Further, the initial concentration of TC (C_{To}, mg/L), BC dosage (C_{char}, g/L) and the ratio of TC to BC (C_o, $mmol/g^{-1}$) were computed using Eq. (1) as outlined in Zhu et al. (2019),

$$C_o = \frac{C_{To}}{(C_{char} \times 444.4)} \quad (1)$$

where, C_{To} and C_{char} represents the initial concentration of TC (mg/L^{-1}) and BC (g/L^{-1}) considered in the corresponding study.

2.2 Translation and classification of BC's conditional attributes

Fifteen crucial criteria were considered and grouped into four categories to frame the general rule for defining BC characteristics aimed at TC adsorption. The adsorption efficiency was expressed by equilibrium sorption capacity Q_e (mg/g^{-1}), and the four categories of input data considered were pyrolysis conditions, BC characteristics, adsorption conditions and the initial concentration ratio of TC to BC. Firstly, the pyrolysis temperature (T_{py}, °C) were taken into consideration. Secondly, BC properties such as pH of the BC in water (*pH Char*), total carbon in the BC (*C*, w%), molar ratio of oxygen and nitrogen to carbon [*(O+N)/C*], molar ratio of oxygen to carbon (*O/C*), molar ratio of hydrogen to carbon (*H/C*), ash content (*Ash*, w%), surface area (*BET*, m^2/g^{-1}), pore volume (*PV*, cm^3/g^{-1}), and BC pore size (*PS*, nm) were considered. Thirdly, the adsorption conditions such as adsorption temperature (T_a, °C) and aqueous solution pH (*pH Solution*) were taken into consideration. Finally, the initial concentration of TC (C_{To}, mg/L), the initial concentration of BC (C_{char}, g/L), and its ratio of TC to BC (C_o, $mmol/g^{-1}$) were also considered. These conditional attributes were identified and translated into measurable properties.

2.3 Development of RSML model

The sample dataset and the information table for the present study of TC adsorption on biochar using case 1 (Ideal dataset) are shown in Table S3a and b. Likewise, case 2 containing Practical datasets is shown in Table S4a and b. The RSML model was developed using a tabular decision table representation of the database. Each row in the table represents an object, and the columns represent attributes corresponding to each object, which can be categorized as condition attributes or decision attributes. The information table is represented as *S=(U, A)*, where *U* is a universal set of non-empty finite objects and A is a non-empty finite set of attributes. Large datasets may contain indiscernible objects, which are objects that perform similarly in their attributes or features. Figure 1 depicts the framework of the RSML algorithm and how it works for the current study. The reduction of condition attributes was done by removing existing data that did not affect the final decision, resulting in fewer attributes considered, reducing the redundancy of data while retaining its basic features. The reduct refers to a subset of indispensable attributes that can partition the database with the same level of discrimination as the original set of attributes, while the core refers to the intersection of all reducts and represents the essential attributes set that cannot be excluded from the decision system without losing the equivalence class structure. The RSML algorithm predicts decision rules based on training data in the information, which explicates the class of output provided the set of conditional attributes is satisfied. The reduction of the set is performed based on the indiscernibility of objects, which depicts the relation between two or more objects with dissimilar target conditional attributes. The main purpose of this step is to negate the attributes that have no effects on the decision and to keep the influencing conditional attributes. The final information table is then used for approximation, reduction, and rule generation using open-source ROSE2 software, as outlined in Tang et al. (2023). More information on RST and its software implementation can be found in Prędki et al. (1998) and Prędki and Wilk (1999).

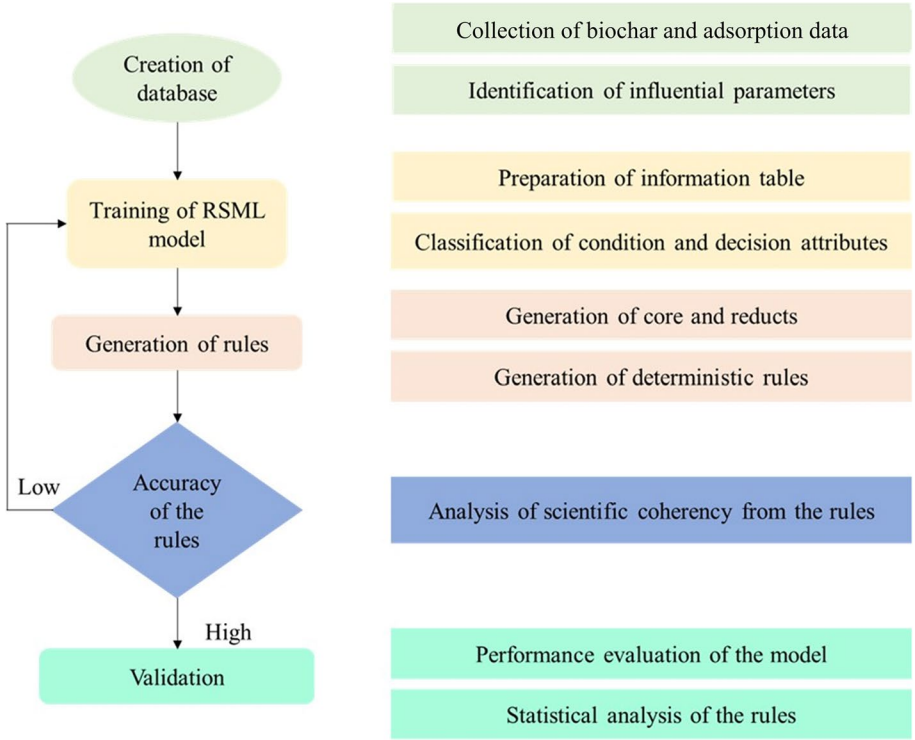


Fig. 1 Overview of the RSML algorithm used in TC adsorption on BC

2.4 Performance evaluation of the developed RSML model

From the standpoint of Bayesian probability, the effectiveness of the created rules can be evaluated quantitatively based on the coverage, certainty, and strength of the rules (Pawlak 2002). As the rules produced by RSML may not always be deterministic, it is mandatory to evaluate the model inconsistencies through these three key criteria for understanding its impact on model performances. A rule's generalization capacity is measured by its strength and coverage, while its predictive accuracy is assessed with certainty. It is noteworthy to mention that higher strength, certainty, and coverage features are preferred to be considered as the well-trained RSML model. A decision rule's strength (Str_x) is defined as the percentage of data points ($supp_x$ *[C, D]*) in a dataset (card [*U*]) that support it by adhering to the rule (Eq. [2]). The likelihood of an object being assigned to a decision class (*Dx*) if it demonstrates a particular set of conditional characteristics (*Cx*) is used to quantify a rule's certainty (Cer_x) (Eq. [3]). During the phase of selecting the decision rules, a rule with a higher certainty value is considered. The coverage (Cov_x) of a rule is the percentage of objects that it successfully categorizes in each decision class (Eq. [4]).

$$\text{Strength}, Str_x = \frac{\text{supp}_x(\text{C}, \text{D})}{\text{card}(\text{U})} \tag{2}$$

$$\text{Certainty}, Cer_x = \frac{\text{supp}_x(\text{C}, \text{D})}{\text{card}(\text{C}\{\text{x}\})} \tag{3}$$

$$\text{Coverage}, Cov_x = \frac{\text{supp}_x(\text{C}, \text{D})}{\text{card}(\text{D}\{\text{x}\})} \tag{4}$$

2.5 Validation of the developed RSML model

The RSML model was validated using the *k*-fold cross-validation of the dataset to assess the performance and forecast accuracy of the reduct sets and the resulting decision rules. There were no duplicates between the validation set and the training set. The validation step is intended to evaluate the model's performance on new data, which has never been used before to train the model. Positive performance on the validation set implies that the model has mastered the applicable general principles correctly. As demonstrated by the quantum of examples covered in the training dataset, underlying patterns in the data should be turned into rules with an appropriate balance of prediction accuracy and generalization strength. However, if the created rules exhibit low accuracy or coverage, it may be due to the presence of coverage clusters with distinct behavior in the dataset. Therefore, to improve the accuracy and coverage of the predictions, the RSML

will be revised again by consolidating the best rules or minimizing the conflicting rules. These revisions should be stopped while the rules with the desired certainty and coverage are attained.

The *k*-fold cross-validation generated an N x $N+1$ confusion matrix, where N is the number of desired classes in the output attribute. The diagonal region with higher values in the confusion matrix indicated the true positive values, which the model accurately predicted. The values in the off-diagonal region were the mispredicted values. The lower values in the off-diagonal region represented the higher accuracy of the model. Specific validation metrics like accuracy, precision, recall and F1 score were calculated using (Eqs. 5, 6, 7 and 8) from the confusion matrix to validate the model efficiency. Accuracy and precision were used to evaluate the model performance based on prediction effectiveness. On the other hand, recall quantifies the efficacy of a classification model in accurately identifying all pertinent instances within a given dataset. The F1-score metric was employed to assess the comprehensive performance of a classification model. It represents the harmonic mean of recall and precision. All the four metrics possessed values ranging from 0–100%.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

$$\text{F1 score} = 2 * \frac{\text{Precision * Recall}}{\text{Precision + Recall}} \tag{8}$$

2.6 Comparative evaluation of the developed RSML model with the other classifiers

The developed RSML model was comparatively evaluated with existing classifier models through the Pycaret tool (Ali 2020), which is an open-source low-code machine learning library in Python that consists of several machine learning libraries and frameworks, such as scikit-learn, XGBoost, LightGBM, CatBoost, spaCy, Optuna, Hyperopt, Ray, and a few more. The data has been passed through multiclass classifiers and validated using the *k*-fold cross-validation of the dataset to assess the performance and forecast accuracy of the reduct sets and the resulting decision rules.

3 Results and discussion

This study is carried out with 295 datasets collected from various literature. The datasets without any missing values were segregated as a separate database with a total of 94 datasets, and analysis was carried out as case 1. In case 2, the collected 295 datasets, of which 201 datasets had missing values, were designated as Practical datasets and considered for RSML analysis. Since the RSML algorithm is known to handle incomplete datasets, this study compared the impact of using Ideal datasets and Practical datasets in machine learning.

3.1 Exploratory data analysis

In this study, fifteen input parameters under four broad characteristics of pyrolysis conditions (T_{py}), feedstock's characteristics (ratio of ultimate analyses such as *[(O+N)/C]*, *(O/C)*, *(H/C)*, and *ash*) and BC characteristics (such as *pH Char*, pore size (*PS*), *BET* surface area, and total pore volume (*PV*)) and adsorption experimental conditions (such as adsorption temperature [T_a], pH of the aqueous solution [*pH Solution*], initial TC concentration [C_{To}], BC dose [C_{char}], and initial concentration ratio of TC to BC [C_o]) were considered as the condition attributes. Meanwhile, the adsorption capacity of TC on BC was selected as the decision attribute under the RSML study. The rationale behind selecting all these fifteen parameters for the RSML study was as follows: as the pyrolytic temperature (T_{py}) plays a major role in dictating the BC characteristics, it has been considered in the present study. In case the pyrolytic temperature is low, the resulting BC can show weak acidity due to the incomplete release of alkali salts. However, the average pH of the BC (*pH Char*) samples considered in the current study was around 9.2 and 9.1 in case 1 and case 2, respectively, and the BC was loaded into the aqueous solution containing TC for its removal using adsorption studies. Yet, both the *pH Char* and the pH of the aqueous solution (*pH Solution*) were taken into account for consideration as the pH of adsorption solutions was adjusted with either acid or base in all batch adsorption studies. Apart from the *pH Solution*, the initial TC concentration (C_{To} in g/L^{-1}), initial BC concentration (C_{Char} in g/L^{-1}) and the ratio of TC to BC (C_o in mmol/g^{-1}) were also considered in the present study.

The exploratory data analysis (EDA) was performed by adapting the univariate graphical method to illustrate the utmost fundamental statistical explanations of data distribution. A violin plot was employed for data visualization as it is highly effective in displaying data distribution in a clear and informative way (as shown in Figs. 2 and 3 for both cases, respectively). The median of the data is represented as the white dot, and the bar represents the interquartile region where the majority

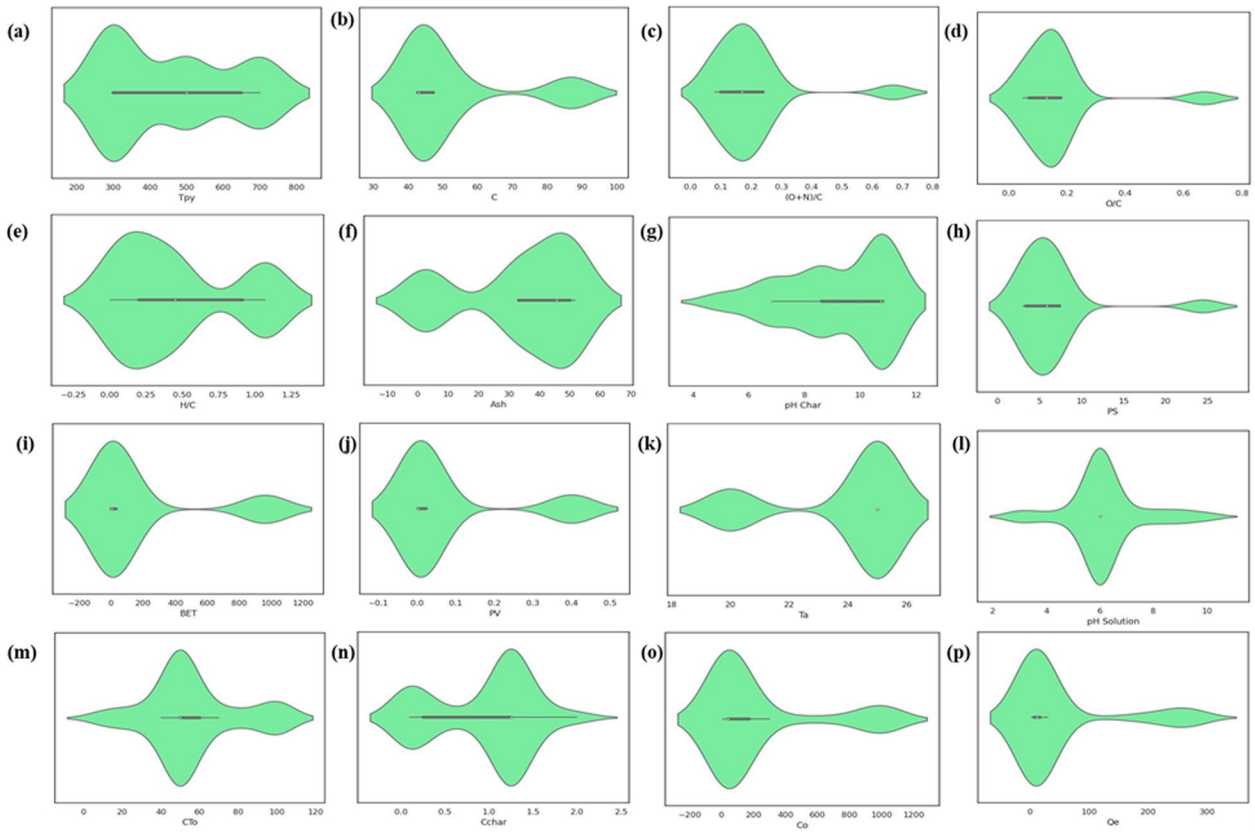


Fig. 2 Violin plot depicts the distribution and density of Ideal dataset (case 1): (**a**) pyrolysis temperature (T, °C), (**b**) total carbon in the BC (C, w%), (**c**) molar ratio of oxygen and nitrogen to carbon [(O + N)/C], (**d**) molar ratio of oxygen to carbon (O/C), (**e**) molar ratio of hydrogen to carbon (H/C), (**f**) ash content (Ash, w%), (**g**) pH of the BC in water (pH_H_2O), (**h**) BC pore size (PS, nm), (**i**) surface area (BET, $m^2.g^{-1}$), (**j**) pore volume (PV, cm^3/g^{-1}), (**k**) adsorption temperature (T, °C), (**l**) solution pH (pH_sol), (**m**) initial concentration of TC (mg/L^{-1}), (**n**) initial concentration of BC (g/L^{-1}), (**o**) the initial concentration ratio of TC to BC (C_0, $mmol/g^{-1}$) and (**p**) equilibrium sorption capacity Q_e ($mg/g.^{-1}$). (The value of y-axis ranges from -1 to + 1)

of the data are present in the distribution, and the four quartiles (25%, 50%, 75%, and 100%) provide insights into the data distribution along with its range.

As efforts are made to optimize the key important factors to achieve maximum Q_e, the factors that influence the Q_e tend to vary in synchrony with each other. Therefore, it is necessary to emphasize the relationship between various feedstock and pyrolysis variables, as well as the ratio of TC to BC used. Figure 4 depicts the Pearson correlation matrix on the relationship among the fifteen input parameters with the desired output features of BC for attaining the maximum equilibrium sorption capacity (Q_e) of TC. The sign of the Pearson correlation coefficient determines the type of correlation between parameters. The magnitude of the coefficient indicates the degree to which one parameter influences the others. As observed from Fig. 4a, factors such as *C*, *BET*, *PV*, C_{To}, and C_o had significant positive correlations with the Q_e in case 1. However, in the case of the Practical dataset (case 2) (Fig. 4b), T_{py} and *C* showed a weak positive correlation Q_e whereas *BET*, *PV*, C_{To} and C_o exhibited significant positive correlations with Q_e.

The BCs considered in the study were rich in carbon content from approximately 40% to the maximum value of 90% in case 1, while the carbon content ranged from approximately 30% to 92%. It is well evident that the carbon content increases with increasing pyrolytic temperature and accumulates during the thermochemical processing of biomass feedstocks. However, the feedstock ratios of the ultimate analyses, such as [*(N + O)/C*], *(H/C)* and *(O/C),* indicate the polarity indices, aromaticity, and hydrophilicity of BC, respectively. While the higher [*(N + O)/C*] ratio represents higher polarity, the higher *(H/C)* and *(O/C)* ratio signifies lower aromaticity and higher hydrophilicity, respectively. Further, the BC characteristics such as BET surface area, pore structure and pore volume were also considered in the RSML study.

However, our earlier studies revealed that the initial TC concentration, BET surface area and pore volume had a significant positive correlation with the adsorption capacity of TC on BC (Zhang et al. 2023). This is because

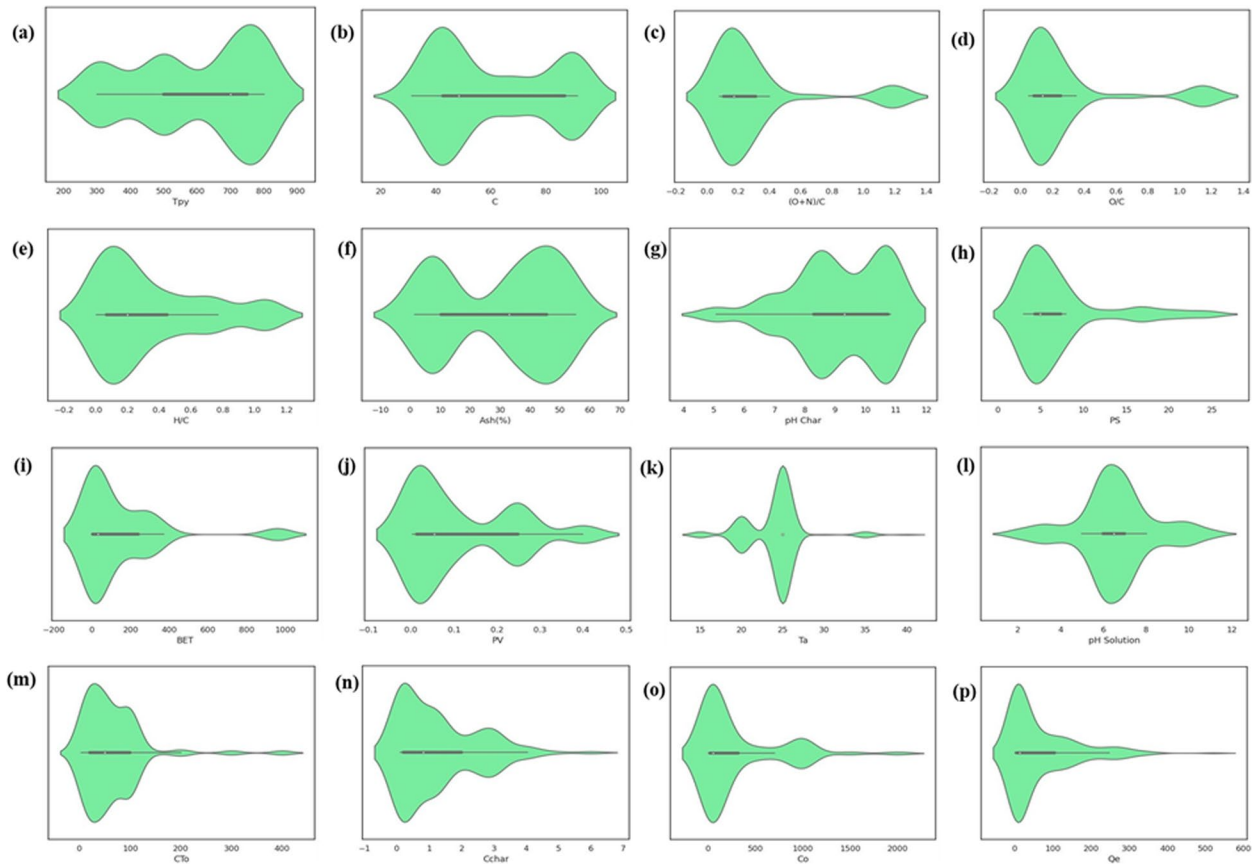


Fig. 3 Violin plot depicts the distribution and density of Practical dataset (case 2): (**a**) pyrolysis temperature (T, °C), (**b**) total carbon in the BC (C, w%), (**c**) molar ratio of oxygen and nitrogen to carbon [(O + N)/C], (**d**) molar ratio of oxygen to carbon (O/C), (**e**) molar ratio of hydrogen to carbon (H/C), (**f**) ash content (Ash, w%), (**g**) pH of the BC in water (pH_H_2O), (**h**) BC pore size (PS, nm), (**i**) surface area (BET, m^2/g^{-1}), (**j**) pore volume (PV, cm^3/g^{-1}), (**k**) adsorption temperature (T, °C), (**l**) solution pH (pH_sol), (**m**) initial concentration of TC (mg/L^{-1}), (**n**) initial concentration of BC (g/L^{-1}), (**o**) the initial concentration ratio of TC to BC (C_0, $mmol/g^{-1}$) and (**p**) equilibrium sorption capacity Q_e (mg/g^{-1}). (The value of y-axis ranges from -1 to + 1)

the adsorption capacity per unit of adsorbed mass will increase while the adsorption dosage rises, though the adsorption value is constant. Likewise, in the case of constant adsorbent (BC), the increase in BET surface area could also lead to the higher adsorption of adsorbates (TC) per unit mass of adsorbent. More pore volume signifies higher pores per unit of adsorbent, which results in higher adsorption capacity.

3.2 Assessment of cores and reducts

The fifteen conditional attributes related to the highest TC adsorption capacity on BC (decision attribute) were derived from cores and reducts observed through an RSML study. Two cases, an Ideal dataset ($n = 94$) and a Practical dataset ($n = 295$), were considered for the current RSML study. The sample dataset and the information table for studying TC adsorption on biochar are shown in Tables S3 and S4 for case 1 and Tables S5 and S6 for case 2. With the established decision system for TC adsorption on BC with the highest capacity, 4 cores and 15 reducts were identified and generated by RST using ROSE2 software for case 1. However, for case 2, only 6 cores and 7 reducts were generated. As the core factors, pH of the aqueous solution, initial TC concentration, BC dosage, and initial concentration ratio of TC to BC were known as the most decisive subset of attributes in the decision table for both cases. In other words, the *pH Solution*, C_{To}, C_{char} and C_o cannot be excluded from the decision system without influencing the classification power of the adsorption capacity. It is noteworthy to mention that when the datasets were incomplete, as in case 2, the decision system demanded the inclusion of two additional attributes, such as pyrolysis temperature and adsorption temperature, as the cores in addition to the existing four cores. Table 1 represents the number of cores and reducts generated along with their respective number of generated rules for TC adsorption on

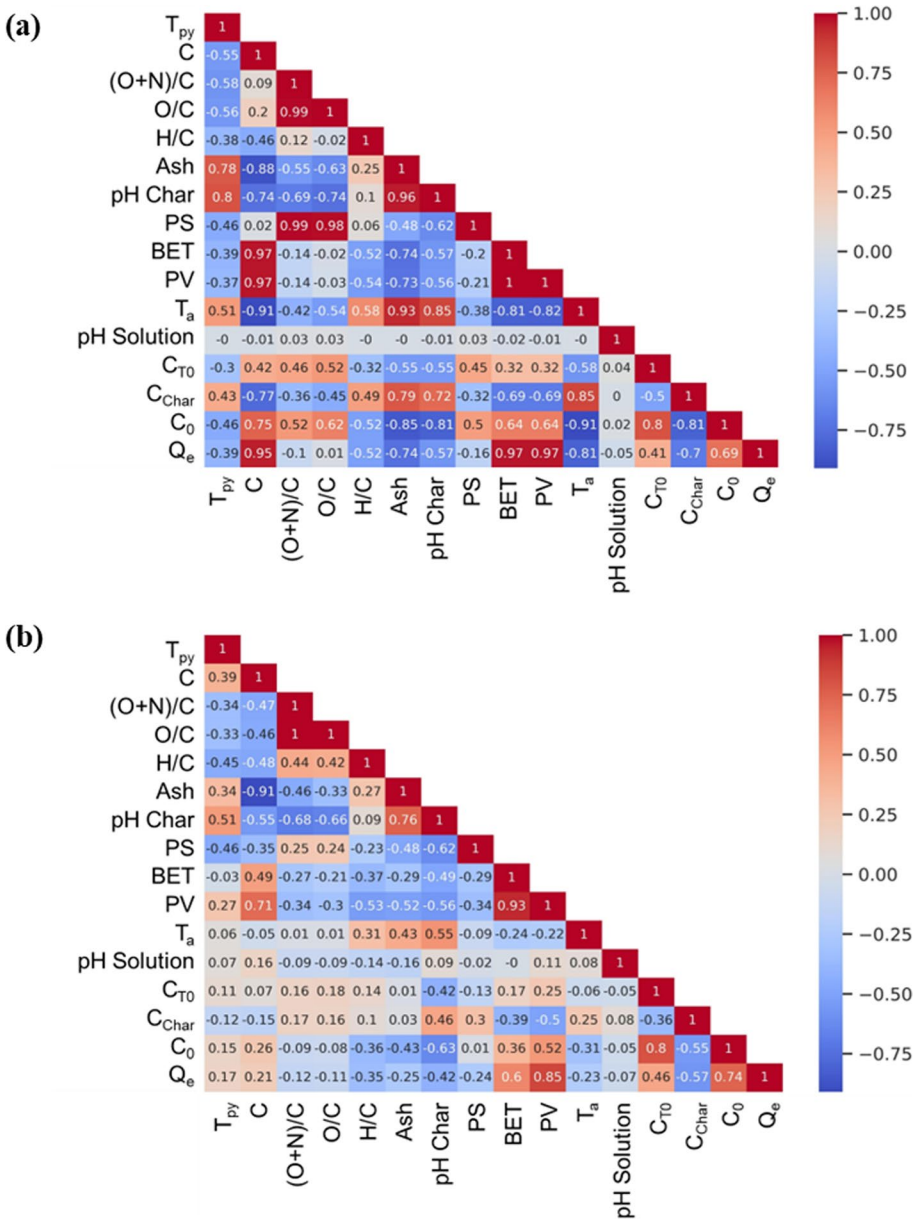


Fig. 4 Pearson correlation matrix (**a**) Ideal dataset (case 1) and (**b**) Practical dataset (case 2), depicting relationship between the input and output features of TC adsorption using BC for achieving maximum equilibrium sorption capacity

BC. It is noteworthy to highlight that the total number of generated rules for a case cannot be summed up under each reduct, as the rules may overlap across various reducts. For instance, in case 1 of the Ideal dataset, 15 reducts were induced, and approximately 36 rules were generated. Therefore, summing up those rules under 15 reducts would result in a higher number than the total number of generated rules.

3.3 Rules generated for TC adsorption by BC

RST induced 36 rules from all 15 reducts for case 1 of the Ideal dataset, whereas 6 rules were classified into class 1 decision of Q_e greater than 200 mg/g (Table 2). Likewise, 2, 13 and 11 rules were generated for class 2, 4 and 5 decisions of Q_e, respectively. No rules could be generated for class 3 decisions as it had only one dataset. Further, two approximate rules were generated for

Table 1 Cores, reducts and the number of rules generated through RSML for the two different datasets of TC adsorption on BC

Case	Core	Reducts	No of rules generated
Case 1: Ideal dataset	pH Solution, C_{To}, C_{char}, C_o	1: {T_{py}, C, pH Solution, C_{To}, C_{char}, C_o}	17
		2: {T_{py}, O/C, pH Solution, C_{To}, C_{char}, C_o}	14
		3: {C, O/C, pH Solution, C_{To}, C_{char}, C_o}	12
		4: {T_{py}, BET, pH Solution, C_{To}, C_{char}, C_o}	13
		5: {C, BET, pH Solution, C_{To}, C_{char}, C_o}	15
		6: {T_{py}, PV, pH Solution, C_{To}, C_{char}, C_o}	16
		7: {C, PV, pH Solution, C_{To}, C_{char}, C_o}	13
		8: {T_{py}, pH Char, pH Solution, C_{To}, C_{char}, C_o}	17
		9: {O/C, pH Char, pH Solution, C_{To}, C_{char}, C_o}	9
		10: {(O + N)/C, pH Solution, C_{To}, C_{char}, C_o}	7
		11: {H/C, pH Solution, C_{To}, C_{char}, C_o}	3
		12: {pH Char, BET, pH Solution, C_{To}, C_{char}, C_o}	12
		13: {ASH, pH Solution, C_{To}, C_{char}, C_o}	3
		14: {pH Char, PV, pH Solution, C_{To}, C_{char}, C_o}	13
		15: {PS, pH Solution, C_{To}, C_{char}, C_o}	6
Case 2: Practical dataset	T_{py}, T_a, pH Solution, C_{To}, C_{char}, C_o	1: {T_{py}, (O + N)/C, PS, T_a, pH Solution, C_{To}, C_{char}, C_o}	45
		2: {T_{py}, O/C, PS, T_a, pH Solution, C_{To}, C_{char}, C_o}	53
		3: {T_{py}, C, T_a, pH Solution, C_{To}, C_{char}, C_o}	45
		4: {T_{py}, (O + N)/C, PV, T_a, pH Solution, C_{To}, C_{char}, C_o}	47
		5: {T_{py}, O/C, PV, T_a, pH Solution, C_{To}, C_{char}, C_o}	53
		6: {T_{py}, ASH, PS, T_a, pH Solution, C_{To}, C_{char}, C_o}	35
		7: {T_{py}, BET, T_a, pH Solution, C_{To}, C_{char}, C_o}	35

the decision attributes of Q_e, except class 1. RSML was able to provide classification between classes 1 to 5 with relatively high certainty and coverage. All the generated rules, along with their respective relative strength, coverage factor, and certainty, are listed in Supplementary Table S5.

As a rule of thumb, the rules with high certainty and coverage should be selected for further consideration. It is noteworthy to mention that all the rules generated in both cases had 100% certainty, which revealed the well-trained model. Figure 5 presents the strength and coverage for all the generated rules in both cases. Further, the low dataset during model development might also cause the generated rules to have the weakest classification power.

For instance, in the case of Class 1 under the Ideal dataset, rules (2, 3, 5, 6) reveal the pH range of 3–7 was much more suitable for enhanced TC adsorption, which can be obviously corroborated from the studies of Zhang et al. (2019) over the different pyrolytic conditions and pH range. Further, their experiments revealed that biochar produced at higher temperatures exhibited maximal adsorption capacity of TC at acidic pH due to the deprotonation of BC surface leading to electrostatic interactions between BC and TC.

In the RSML for case 2 of the Practical dataset, the number of rules generated was 92, including six approximate rules. Out of which, 15 deterministic rules for class 1 were generated. 15, 9, 26 and 21 deterministic rules for classes 2, 3, 4 and 5, respectively, were created by RSML as shown in Table S6. The higher number of rules generated might be due to the fact that data was three-fold higher when compared to case 1, although the reducts were half as high as those of case 1.

For instance, considering rule 1 of case 2 (Practical dataset) from Table 2, the predicted process condition for achieving $Q_e > 200$ mg/g involves biochar properties [$O/C >= 1$], pH of the solution for adsorption reaction at 6 with TC to BC ratio (C_0) greater than 2. This rule provides one of the possible conditions for achieving $Q_e > 200$ mg/g with the aid of 3 main parameters rather than focusing on all the 15 parameters. Out of these 3 parameters, the pH of the solution is a crucial parameter for the adsorption phenomenon to occur at a higher rate, and the TC to BC ratio is required to ensure the presence of enough adsorbent (BC) for capturing TC.

Table 2 List of rules generated for TC adsorption on BC using RSML algorithms for class 1 decision (Q_e > 200 mg/g) using ideal and practical dataset

Rule No	Rule	Strength	Relative strength	Coverage (%)	Accuracy (%)
CASE 1: Ideal dataset for Class 1 Decision (Q_e > 200 mg/g)					
1	(1 < = C_0 < 2)	4	4	33.33	100
2	(0.1 < = (O/C) < 0.2) & (pH Solution = 6) & (C_0 > = 2)	3	3	25	100
3	(0.1 < = (O/C) < 0.2) & (pH Solution = 7) & (C_0 > = 2)	1	1	8.33	100
4	(0.5 < = C_0 < 1)	2	2	16.67	100
5	(T_{py} = 300) & (0.1 < = (O + N/C) < 0.2) & (pH Solution = 5)	1	1	8.33	100
6	(C = 87) & (pH Solution = 3)	1	1	8.33	100
CASE 2: Practical dataset for Class 1 Decision (Q_e > 200 mg/g)					
1	((O/C) > = 1) & (pH Solution = 6) & (C_0 > = 2)	3	3	11.11	100
2	(T_{py} = 300) & (1 < = C_0 < 2)	4	4	14.81	100
3	(PV > = 0.1) & (pH Solution = 6) & (100 < = C_{To} < 200)	2	2	7.41	100
4	((O/C) > = 1) & (0.2 < = (H/C) < 0.3) & (pH Solution = 5)	1	1	3.7	100
5	((H/C) < 0.1) & (PV > = 0.1) & (pH Solution = 3)	1	1	3.7	100
6	(0.2 < = (O + N/C) < 0.3) & (T_a = 20) & (pH Solution = 9)	1	1	3.7	100
7	(T_a = 40)	1	1	3.7	100
8	(0.1 < = (O/C) < 0.2) & ((H/C) < 0.1) & (pH Solution = 5)	1	1	3.7	100
9	((O/C) > = 1) & (0.2 < = (H/C) < 0.3) & (pH Solution = 4)	1	1	3.7	100
10	(T_{py} = 300) & (0.5 < = C_0 < 1)	2	2	7.41	100
11	(T_a = 30)	1	1	3.7	100
12	((O/C) > = 1) & (0.2 < = (H/C) < 0.3) & (pH Solution = 7)	1	1	3.7	100
13	(T_{py} = 750) & (PV > = 0.1)	1	1	3.7	100
14	(0.1 < = (O + N/C) < 0.2) & (PS = 4) & (pH Solution = 7)	1	1	3.7	100
15	(BET = 3) & (pH Solution = 8) = > (Qe = 1)	1	1	3.7	100

3.4 Cross-validation of the generated rules for TC adsorption by BC

The confusion matrix depicts the classification efficiency of datasets under each class for both cases and the highest values in the diagonal of the matrix revealed the appropriate classification. However, for case 1, class 2 and class 3 had zero values, which had not been appropriately classified. It might be due to the very low dataset in the respective classes. On the contrary, the classification of the datasets based on the trained model in case 2 for all the classes was found to be relatively satisfactory (as seen in Table 3) than that of case 1, which might be due to the higher number of datasets.

It is evident that evaluating the performance of machine learning models is challenging with limited data. Thus, *k*-fold cross-validation is well-suited for research using sparse datasets. The current study employed *k*-fold (10-fold) cross-validation to evaluate model performance after randomly splitting the datasets into training (*k*-1) sets and the rest as test sets. This process was repeated 10 times to serve as the test set once for each subset. The classification model was built using the training set and validated on the test set. Model efficiency was quantified based on true positives, true negatives, false positives and false negatives. Precision and recall are statistical measures used for validation. Precision refers to the proportion of correctly predicted positive observations among all predicted positive observations, while recall refers to the proportion of actual positives that are correctly identified. The F1-score is another statistical measure used to evaluate the developed RSML. Table 4 presents the cross-validation attributes of RSML algorithms for both cases. In the case of class 1 (Q_e > 200 mg/g), the accuracy noticed was 89.25% for case 1, while it was 93.22% for case 2. Likewise, precision was shown to be slightly higher in case 2 than in case 1, which signifies that even in the practical dataset, the trained model could exhibit the most correctly predicted positive observations without any imputations. However, recall and F1-score have shown a declining trend in case 2, which might be due to the large amount of missing data in the Practical dataset. Similarly, the interpretations of the remaining classes between the two cases shall be drawn from Table 4. As mentioned earlier, the number of rules generated in class 2 and class 3 of case 1 were two and

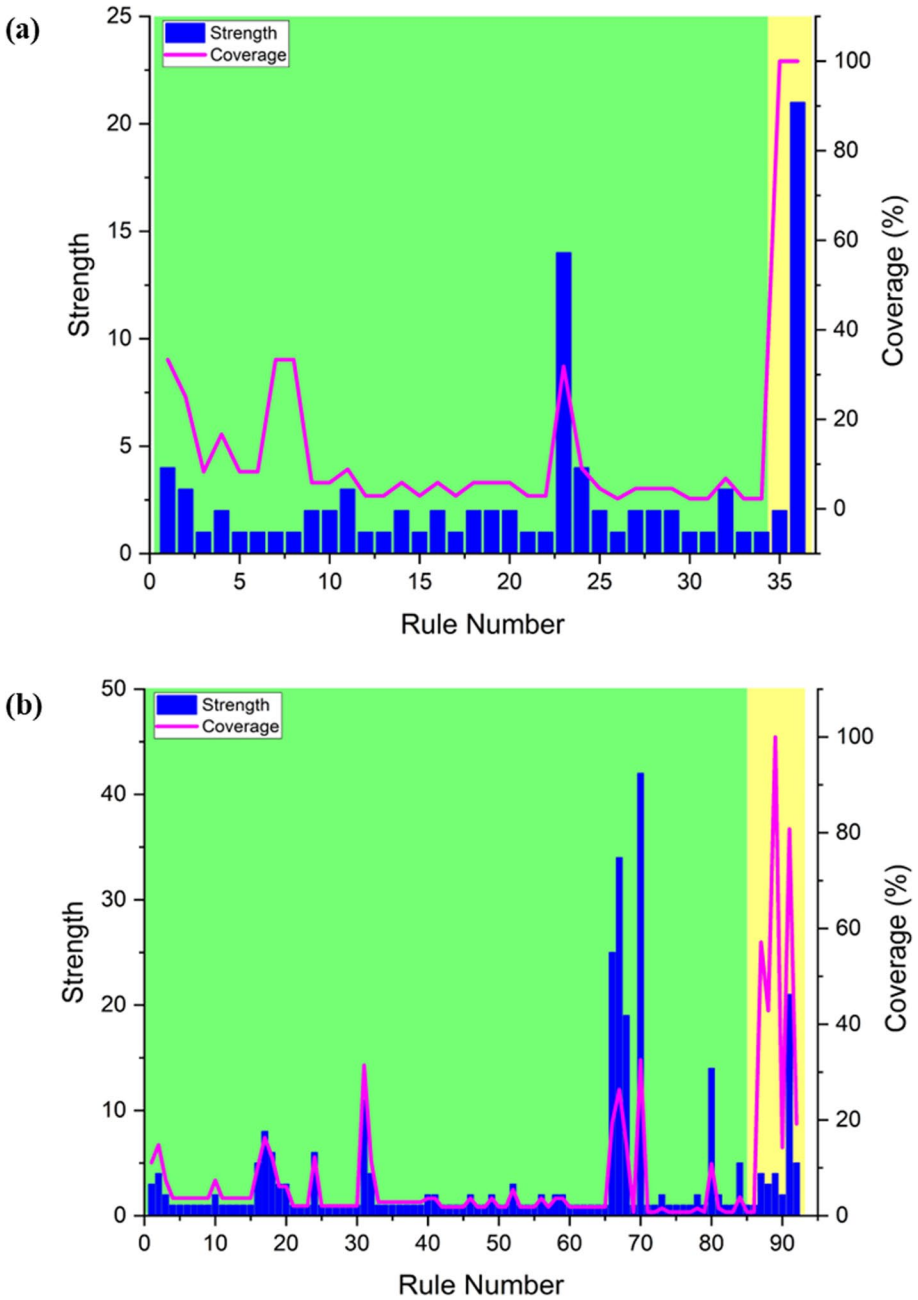


Fig. 5 Plot of (**a**) Ideal dataset (case 1) and (**b**) Practical dataset (case 2) between strength (number of data) of the rules and its coverage for the rules generated for TC adsorption on BC

zero due to the very low availability of the dataset, and the precision, recall, and F1 scores were zero for all, respectively.

3.5 Comparative evaluation of RSML algorithm with the other classifier models for TC adsorption by BC

A wide array of ML algorithms has been performed to identify the best model that yields the greatest performance in the given classification task. Several commonly used classification algorithms from pycaret, such as Extra Trees Classifier, Gradient Boosting Classifier, Random Forest Classifier, Extreme Gradient Boosting, Decision Tree Classifier, Light Gradient Boosting Machine, K Neighbors Classifier, Logistic Regression, Linear Discriminant Analysis, Ridge Classifier, Naive Bayes, Quadratic Discriminant Analysis, AdaBoost Classifier, and SVM—Linear Kernel were utilized to compare

Table 3 Confusion matrix for the two different datasets of TC adsorption on BC using RSML algorithm

Class	Case 1: Ideal dataset							Case 2: Practical dataset						
	No of data ($n=94$)	1	2	3	4	5	None	No of data ($n=295$)	1	2	3	4	5	None
Class 1: > 200	12	9	0	0	3	0	0	27	13	12	1	1	0	0
Class 2: 200 > x > 100	3	1	0	1	1	0	0	49	1	39	6	0	3	0
Class 3: 100 > x > 50	1	0	1	0	0	0	0	35	0	6	24	2	3	0
Class 4: 50 > x > 10	34	5	2	2	19	6	0	55	4	8	8	27	8	0
Class 5: 10 > x > 0	44	1	2	4	13	23	1	129	1	6	1	15	106	0

Table 4 Performance evaluation of RSML algorithms for the two different datasets of TC adsorption on BC

Class	Case 1: Ideal dataset				Case 2: Practical dataset			
	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
Class 1: > 200	89.25	56.25	75.00	64.29	93.22	68.42	48.15	56.52
Class 2: 200 > x > 100	91.40	NaN	NaN	NaN	85.76	54.93	79.59	65.00
Class 3: 100 > x > 50	91.40	NaN	NaN	NaN	90.85	60.00	68.57	64.00
Class 4: 50 > x > 10	65.59	52.78	55.88	54.29	84.41	60.00	49.09	54.00
Class 5: 10 > x > 0	72.04	79.31	53.49	63.89	87.46	88.33	82.17	85.14

Table 5 Comparative evaluation of RSML algorithms with other classifier models for practical datasets of TC adsorption on BC

No	Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
1	Rough-Set Machine Learning	88.34	66.36	65.51	64.93
2	Extra Trees Classifier	81.05	82.99	80.20	80.38
3	Gradient Boosting Classifier	80.60	82.08	81.26	80.55
4	Random Forest Classifier	79.64	81.88	79.68	79.19
5	Extreme Gradient Boosting	79.64	79.95	80.23	78.70
6	Decision Tree Classifier	79.14	82.13	79.15	79.16
7	Light Gradient Boosting Machine	78.19	78.88	79.21	77.98
8	K Neighbors Classifier	73.83	78.66	74.36	73.31
9	Logistic Regression	72.88	76.22	73.33	73.16
10	Linear Discriminant Analysis	68.93	68.55	68.45	66.59
11	Ridge Classifier	66.48	61.90	66.84	62.31
12	Naive Bayes	66.00	64.55	63.60	62.57
13	Quadratic Discriminant Analysis	54.31	47.34	56.61	49.36
14	Ada Boost Classifier	38.69	44.46	40.47	37.49
15	SVM—Linear Kernel	37.45	35.21	39.65	30.93

the performance of TC adsorption on biochar with RSML. The model comparison has been done with the case 2 practical dataset as it performed better than the case 1 Ideal dataset. Table 5 summarizes the performance metrics of the RSML algorithm along with the other classifier algorithms. Since the Practical dataset was segregated based on users' prejudiced Q_e values, it resulted in an imbalanced dataset. This explains the reason behind low precision, recall and F1 score values rather than accuracy. However, the developed RSML model exhibited better accuracy when compared to the other classifier models, which may be due to the fact that RSML works based on rough set theory for deducing better approximations. Herein, the RSML model performed well with an accuracy of 88.34%, followed by the extra trees classifier (81.05%) and gradient boosting classifier (80.60%).

Table 6 Optimal range of condition attributes for Class 1 TC adsorption capacity of $Q_e > 200$ mg/g on BC

Decision attributes	Condition attributes			
	Ratio of ultimate analysis	Pyrolysis conditions	Biochar properties	Adsorption conditions
Case 1: Ideal dataset	0.1 < (O + N)/C < 0.2, 0.1 < O/C < 0.19 C = 87%	T_{py} = 300 °C		3 < pH Solution < 7
Case 2: Practical dataset	0.1 < (O + N)/C < 0.3, O/C > 1 0.1 < H/C < 0.3,	300 °C < T_{py} < 700 °C	PV > 0.1 cm^3/g PS = 4 nm BET = 3 $m^2.g^{-1}$	3 < pH Solution < 9, 20 °C < T_a < 40 °C, 100 mg/L < C_{To} < 200 mg/L, C_o > 0.5 mmol/g

3.6 Comparative analysis of the optimized range of conditions needed for maximizing Q_e of TC on BC

Table 6 summarizes the selected rules to obtain the maximum TC adsorption capacity (say more than 200 mg/g) on BC. Based on the RSML model trained on Ideal dataset, the pyrolysis temperature conditions of 300 °C with the biomass characteristics based on the ratio of ultimate analysis such as carbon content in BC greater than 87%, *O/C* between 0.1–0.19 and *(O + N)/C* between 0.1–0.2 in the adsorption conditions of pH of solution between 3 to 7 would result in TC adsorption capacity (Q_e) > 200 mg/g on BC. While in case 2 of the model that was trained on the Practical dataset, the RSML demanded additional parameters on BC properties (such as pore volume, pore size and surface area) to predict the $Q_e > 200$ mg/g of TC on BC.

It is noteworthy to mention that the researchers are trying to produce BC with the desired attributes for the maximized adsorption of target pollutants. That is why the prime focus of the discussion was confined to the interpretations for class 1 in both cases. Apart from the accurate rules, RSML offers approximate rules that might classify the output in either of the classes. Herein, in case 1, two approximate rules were generated, and both exhibited 100% coverage and certainty. While in case 2, seven approximate rules with coverage ranged from 14.2 to 100%, and the certainty of 100% was achieved. The results indicate that RST can effectively select relevant attributes that improve predictive performance. According to RST, its feature selection method has the ability to eliminate irrelevant data, simplifying the decision-making process. This likely explains why rough set modeling yielded satisfactory prediction results in this study.

4 Conclusion

Using an Ideal and Practical dataset, two rule-based RSML models were developed to estimate TC adsorption capacity on BC. Both models produced scientifically coherent decision rules. However, the Practical dataset model performed better. The model trained with the Ideal dataset suggested T_{py}, *C*, *O/C*, *(O + N)/C*, and *pH Solution* were essential for $Q_e > 200$ mg/g. Yet, the model developed with a Practical dataset demanded BC properties in addition to the aforementioned attributes to achieve the same purpose. The model trained with the Practical dataset provided that the pyrolysis temperature at 300 ℃ with TC to BC ratio between 1 and 2 is required to achieve an adsorption capacity greater than 200 mg/g. This study demonstrated an interpretable RSML tool to estimate BC adsorption capacity using a Practical dataset without imputation, thus minimizing bias and variances during decision-making.

Abbreviations

TC	Tetracycline
BC	Biochar
ML	Machine learning
RSML	Rough set-based machine learning
ANN	Artificial neural networks
SVM	Support vector machines
XAI	Explainable artificial intelligence
GHG	Greenhouse gases
pH_{Char}	pH of the BC in water
C	Total carbon in the BC
[(O + N)/C]	Molar ratio of oxygen and nitrogen to carbon
O/C	Molar ratio of oxygen to carbon
H/C	Molar ratio of hydrogen to carbon
Ash	Ash content
BET	Surface area
PV	Pore volume
PS	Pore size
T_a	Adsorption temperature
pH Solution	Adsorption solution pH
C_{To}	Initial concentration of TC
C_{char}	BC dosage
C_o	Ratio of TC to BC
T_{py}	Pyrolysis temperature

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1007/s44246-024-00129-w.

Supplementary Material 1.

Acknowledgements

The authors thank their institutions for providing the necessary research facilities to carry out this work.

Authors' contributions
Paramasivan Balasubramanian wrote the original draft, designed methodology and conducted formal analysis. Muhil Raj Prabhakar contributed to methodology, formal analysis, Software, and reviewing and editing the manuscript. Chong Liu obtained resources and performed project administration, supervision, and validation. Pengyan Zhang wrote the original draft and obtained resources. Fayong Li contributed to the review and editing process and formal analysis.

Funding
No funding was available.

Availability of data and materials
Datasets generated during the current study are available from the corresponding author upon reasonable request.

Declarations

Competing interests
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.

Author details
[1]Department of Biotechnology & Medical Engineering, National Institute of Technology Rourkela, Rourkela 769008, India. [2]Department of Chemical & Materials Engineering, University of Auckland, Auckland 0926, New Zealand. [3]College of Water Resources and Architectural Engineering, Tarim University, Xinjiang 843300, China.

Received: 25 December 2023 Revised: 18 April 2024 Accepted: 21 April 2024
Published online: 07 May 2024

References

Akbari S, Khanzadi M, Gholamian MR (2018) Building a rough sets-based prediction model for classifying large-scale construction projects based on sustainable success index. Eng Constr Archit Manag 25(4):534–558. https://doi.org/10.1108/ECAM-05-2016-0110

Akhil D, Lakshmi D, Kartik A, Vo DVN, Arun J, Gopinath KP (2021) Production, characterization, activation and environmental applications of engineered biochar: a review. Environ Chem Lett 19(3):2261–2297. https://doi.org/10.1007/S10311-020-01167-7

Albuquerque LG, de Oliveira Roque F, Valente-Neto F, Koroiva R, Buss DF, Baptista DF, Hepp LU, Kuhlmann ML, Sundar S, Covich AP, Pinto JOP (2021) Large-scale prediction of tropical stream water quality using Rough Sets Theory. Eco Inform 61:101226. https://doi.org/10.1016/J.ECOINF.2021.101226

Ali M (2020) PyCaret: an open source, low-code machine learning library in Python

Ang JC, Tang JY, Chung BYH, Chong JW, Tan RR, Aviso KB, Chemmangattuvalappil NG, Thangalazhy-Gopakumar S (2023) Development of predictive model for biochar surface properties based on biomass attributes and pyrolysis conditions using rough set machine learning. Biomass Bioenerg 174:106820. https://doi.org/10.1016/J.BIOMBIOE.2023.106820

Aviso KB, Janairo JIB, Promentilla MAB, Tan RR (2019) Prediction of CO2 storage site integrity with rough set-based machine learning. Clean Technol Environ Policy 21(8):1655–1664. https://doi.org/10.1007/S10098-019-01732-X

Aviso KB, Capili MJ, Chin HH, van Fan Y, Klemeš JJ, Tan RR (2021) Detecting patterns in energy use and greenhouse gas emissions of cities using machine learning. Chem Eng Trans 88:403–408. https://doi.org/10.3303/CET2188067

Bilal M, Mehmood S, Rasheed T, Iqbal HMN (2020) Antibiotics traces in the aquatic environment: persistence and adverse environmental impact. Curr Opin Environ Sci Health 13:68–74. https://doi.org/10.1016/J.COESH.2019.11.005

Calegari R, Ciatto G, Omicini A (2020) On the integration of symbolic and sub-symbolic techniques for XAI: A survey. Intell Artific 14(1):7–32. https://doi.org/10.3233/IA-190036

Cao S, Luo Y, Li T, Li J, Wu L, Liu G (2023) Machine learning assisted screening of doped metals phosphides electrocatalyst towards efficient hydrogen evolution reaction. Mol Catal 551:113625. https://doi.org/10.1016/j.mcat.2023.113625

Chen T, Luo L, Deng S, Shi G, Zhang S, Zhang Y, Deng O, Wang L, Zhang J, Wei L (2018) Sorption of tetracycline on H3PO4 modified biochar derived from rice straw and swine manure. Biores Technol 267:431–437. https://doi.org/10.1016/J.BIORTECH.2018.07.074

Chen Y, Liu J, Zeng Q, Liang Z, Ye X, Lv Y, Liu M (2021) Preparation of Eucommia ulmoides lignin-based high-performance biochar containing sulfonic group: synergistic pyrolysis mechanism and tetracycline hydrochloride adsorption. Biores Technol 329:124856. https://doi.org/10.1016/J.BIORTECH.2021.124856

Cheng N, Wang B, Wu P, Lee X, Xing Y, Chen M, Gao B (2021) Adsorption of emerging contaminants from water and wastewater by modified biochar: a review. Environ Pollut 273:116448. https://doi.org/10.1016/J.ENVPOL.2021.116448

Choi YK, Choi TR, Gurav R, Bhatia SK, Park YL, Kim HJ, Kan E, Yang YH (2020) Adsorption behavior of tetracycline onto Spirulina sp. (microalgae)-derived biochars produced at different temperatures. Sci Total Environ 710:136282. https://doi.org/10.1016/J.SCITOTENV.2019.136282

Chong JW, Thangalazhy-Gopakumar S, Tan RR, Aviso KB, Chemmangattuvalappil NG (2022) Estimation of fast pyrolysis bio-oil properties from feedstock characteristics using rough-set-based machine learning. Int J Energy Res 46(13):19159–19176. https://doi.org/10.1002/ER.8201

Dempster AP (1967) Upper and lower probability inferences based on a sample from a finite univariate population. Biometrika 54(3–4):515–528. https://doi.org/10.1093/BIOMET/54.3-4.515

Fan SS, Liu WP, Wang JT, Hu HM, Yang YN, Zhou N (2020) Preparation of tea waste biochar and its application in tetracycline removal from aqueous solution. Huan Jing Ke Xue 41(3):1308–1318. https://doi.org/10.13227/J.HJKX.201908179

Goguen JA (1974) Concept representation in natural and artificial languages: Axioms, extensions and applications for fuzzy sets. Int J Man Mach Stud 6(5):513–561. https://doi.org/10.1016/S0020-7373(74)80017-9

Gopal G, Alex SA, Chandrasekaran N, Mukherjee A (2020) A review on tetracycline removal from aqueous systems by advanced treatment techniques. RSC Adv 10(45):27081–27095. https://doi.org/10.1039/D0RA04264A

Jang HM, Kan E (2019) Engineered biochar from agricultural waste for removal of tetracycline in water. Biores Technol 284:437–447. https://doi.org/10.1016/J.BIORTECH.2019.03.131

Kim JE, Bhatia SK, Song HJ, Yoo E, Jeon HJ, Yoon JY, Yang Y, Gurav R, Yang YH, Kim HJ, Choi YK (2020) Adsorptive removal of tetracycline from aqueous solution by maple leaf-derived biochar. Biores Technol 306:123092. https://doi.org/10.1016/J.BIORTECH.2020.123092

Lei L, Chen W, Wu B, Chen C, Liu W (2021) A building energy consumption prediction model based on rough set theory and deep learning algorithms. Energy Build 240:110886. https://doi.org/10.1016/J.ENBUILD.2021.110886

Leng L, Zhang W, Liu T, Zhan H, Li J, Yang L, Li J, Peng H, Li H (2022) Machine learning predicting wastewater properties of the aqueous phase derived from hydrothermal treatment of biomass. Biores Technol 358:127348. https://doi.org/10.1016/J.BIORTECH.2022.127348

Li X, Huang Y, Liang X, Huang L, Wei L, Zheng X, Albert HA, Huang Q, Liu Z, Li Z (2022) Characterization of biochars from woody agricultural wastes and sorption behavior comparison of cadmium and atrazine. Biochar 4(1):1–12. https://doi.org/10.1007/S42773-022-00132-7

Liu C, Hu X, Xu Q, Zhang S, Zhang P, Guo H, You Y, Liu Z (2021) Response surface methodology for the optimization of the ultrasonic-assisted rhamnolipid treatment of oily sludge. Arab J Chem 14(3):102971. https://doi.org/10.1016/J.ARABJC.2020.102971

Luo Y, Du X, Wu L, Wang Y, Li J, Ricardez-Sandoval L (2023) Machine-learning-accelerated screening of double-atom/cluster electrocatalysts for the oxygen reduction reaction. J Phys Chem C 127(41):20372–20384. https://doi.org/10.1021/acs.jpcc.3c05753

Merrick L, Taly A (2020) The explanation game: explaining machine learning models using shapley values. Lecture notes in computer science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture

Notes in Bioinformatics), 12279 LNCS, p 17–38. https://doi.org/10.1007/978-3-030-57321-8_2
Pawlak Z (1982) Rough sets. Int J Comput Inform Sci 11(5):341–356. https://doi.org/10.1007/BF01001956
Pawlak Z (1997) Rough set approach to knowledge-based decision support. Eur J Oper Res 99(1):48–57. https://doi.org/10.1016/S0377-2217(96)00382-7
Pawlak Z (2002) Rough sets, decision algorithms and Bayes' theorem. Eur J Oper Res 136(1):181–189. https://doi.org/10.1016/S0377-2217(01)00029-7
Phoon BL, Ong CC, Mohamed Saheed MS, Show PL, Chang JS, Ling TC, Lam SS, Juan JC (2020) Conventional and emerging technologies for removal of antibiotics from wastewater. J Hazard Mater 400:122961. https://doi.org/10.1016/J.JHAZMAT.2020.122961
Prędki B, Wilk S (1999) Rough set based data exploration using ROSE system. Lect Notes Comput Sci 1609:172–180. https://doi.org/10.1007/BFB0095102
Prędki B, Słowiński R, Stefanowski J, Susmaga R, Wilk S (1998) ROSE - software implementation of the rough set theory. Lect Notes Comput Sci 1424:605–608. https://doi.org/10.1007/3-540-69115-4_85
Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1(5):206–215. https://doi.org/10.1038/s42256-019-0048-x
Rudin C, Radin J (2019) Why are we using black box models in AI when we don't need to? A lesson from an explainable AI Competition. Harvard Data Sci Rev 1(2):2019. https://doi.org/10.1162/99608F92.5A8A3A3D
Shen Q, Wang Z, Yu Q, Cheng Y, Liu Z, Zhang T, Zhou S (2020) Removal of tetracycline from an aqueous solution using manganese dioxide modified biochar derived from Chinese herbal medicine residues. Environ Res 183:109195. https://doi.org/10.1016/J.ENVRES.2020.109195
Tang JY, Chung BYH, Ang JC, Chong JW, Tan RR, Aviso KB, Chemmangattuvalappil NG, Thangalazhy-Gopakumar S (2023) Prediction model for biochar energy potential based on biomass properties and pyrolysis conditions derived from rough set machine learning. https://doi.org/10.1080/09593330.2023.2192877
Taoufik N, Boumya W, Achak M, Chennouk H, Dewil R, Barka N (2022) The state of art on the prediction of efficiency and modeling of the processes of pollutants removal based on machine learning. Sci Total Environ 807:150554. https://doi.org/10.1016/J.SCITOTENV.2021.150554
Thangaraj B, Solomon PR (2019) Immobilization of lipases – a review. Part I: enzyme immobilization. ChemBioEng Rev 6(5):157–166. https://doi.org/10.1002/CBEN.201900016
Wang H, Fang C, Wang Q, Chu Y, Song Y, Chen Y, Xue X (2018) Sorption of tetracycline on biochar derived from rice straw and swine manure. RSC Adv 8(29):16260–16268. https://doi.org/10.1039/C8RA01454J
Wilschut RA, De Long JR, Geisen S, Hannula SE, Quist CW, Snoek B, Steinauer K, Wubs ERJ, Yang Q, Thakur MP (2022) Combined effects of warming and drought on plant biomass depend on plant woodiness and community type: a meta-analysis. Proc Royal Soc B 289(1984):20221178. https://doi.org/10.1098/RSPB.2022.1178
Xu D, Gao Y, Lin Z, Gao W, Zhang H, Karnowo K, Hu X, Sun H, Syed-Hassan SSA, Zhang S (2020) Application of biochar derived from pyrolysis of waste fiberboard on tetracycline adsorption in aqueous solution. Front Chem 7:510935. https://doi.org/10.3389/FCHEM.2019.00943
Yang Y, Yuan Y, Zhang G, Wang H, Chen YC, Liu Y, Tarolli CG, Crepeau D, Bukartyk J, Junna MR, Videnovic A, Ellis TD, Lipford MC, Dorsey R, Katabi D (2022) Artificial intelligence-enabled detection and assessment of Parkinson's disease using nocturnal breathing signals. Nat Med 28(10):2207–2215. https://doi.org/10.1038/s41591-022-01932-x
Zeng G, Liu Y, Ma X, Fan Y (2021) Fabrication of magnetic multi-template molecularly imprinted polymer composite for the selective and efficient removal of tetracyclines from water. Front Environ Sci Eng 15(5):1–12. https://doi.org/10.1007/S11783-021-1395-5
Zhang P, Li Y, Cao Y, Han L (2019) Characteristics of tetracycline adsorption by cow manure biochar prepared at different pyrolysis temperatures. Biores Technol 285:121348. https://doi.org/10.1016/J.BIORTECH.2019.121348
Zhang P, Liu C, Lao D, Nguyen XC, Paramasivan B, Qian X, Inyinbor AA, Hu X, You Y, Li F (2023) Unveiling the drives behind tetracycline adsorption capacity with biochar through machine learning. Sci Rep 13(1):1–12. https://doi.org/10.1038/s41598-023-38579-8
Zhang X, Li J, Yan S, Tyagi RD, Chen J (2020) Physical, chemical, and biological impact (hazard) of hospital wastewater on environment: presence of pharmaceuticals, pathogens, and antibiotic-resistance genes. In: Current developments in biotechnology and bioengineering: environmental and health impact of hospital wastewater. Elsevier, p 79–102. https://doi.org/10.1016/B978-0-12-819722-6.00003-1
Zhao B, Ren Y, Gao D, Xu L (2019) Prediction of service life of large centrifugal compressor remanufactured impeller based on clustering rough set and fuzzy Bandelet neural network. Appl Soft Comput 78:132–140. https://doi.org/10.1016/J.ASOC.2019.02.018
Zheng Z, Zhao B, Guo Y, Guo Y, Pak T, Li G (2021) Preparation of mesoporous batatas biochar via soft-template method for high efficiency removal of tetracycline. Sci Total Environ 787:147397. https://doi.org/10.1016/J.SCITOTENV.2021.147397
Zhou BQ, Yang RC, Li HP, Wang YJ, Zhang CY, Xiao ZJ, He ZQ, Pang WH (2023) Numeric and nonnumeric information input to predict adsorption amount, capacity and kinetics of tetracyclines by biochar via machine learning. Chem Eng J 471:144636. https://doi.org/10.1016/J.CEJ.2023.144636
Zhu TT, Su ZX, Lai WX, Zhang YB, Liu YW (2021a) Insights into the fate and removal of antibiotics and antibiotic resistance genes using biological wastewater treatment technology. Sci the Total Environ 776:145906. https://doi.org/10.1016/J.SCITOTENV.2021.145906
Zhu X, Wan Z, Tsang DCW, He M, Hou D, Su Z, Shang J (2021b) Machine learning for the selection of carbon-based materials for tetracycline and sulfamethoxazole adsorption. Chem Eng J 406:126782. https://doi.org/10.1016/J.CEJ.2020.126782
Zhu X, Wang X, Ok YS (2019) The application of machine learning methods for prediction of metal sorption onto biochars. J Hazard Mater 378:120727. https://doi.org/10.1016/j.jhazmat.2019.06.004

Publisher's Note