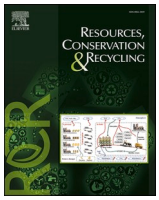
Resources, Conservation & Recycling 190 (2023) 106847

Contents lists available at ScienceDirect

Resources, Conservation & Recycling

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



Review

Machine learning for sustainable development and applications of biomass and biomass-derived carbonaceous materials in water and agricultural systems: A review

Hannah Szu-Han Wang, Yuan Yao *

Center for Industrial Ecology, Yale School of the Environment, Yale University, 380 Edwards Street, New Haven, CT, 06511, United States

ARTICLE INFO

Keywords:
Biomass-derived material
Machine learning
Sustainability
Water
Agriculture
Biochar

ABSTRACT

Biomass-derived materials (BDM) have broad applications in water and agricultural systems. As an emerging tool, Machine learning (ML) has been applied to BDM systems to address material, process, and supply chain design challenges. This paper reviewed 53 papers published since 2008 to understand the capabilities, current limitations, and future potentials of ML in supporting sustainable development and applications of BDM. Previous ML applications were classified into three categories based on their objectives – material and process design, end-use performance prediction, and sustainability assessment. These ML applications focus on identifying critical factors for optimizing BDM systems, predicting material features and performances, reverse engineering, and addressing data challenges for sustainability assessments. BDM datasets show large variations, and ~75% of them possess < 600 data points. Ensemble models and state-of-the-art neural networks (NNs) perform and generalize well on such datasets. Limitations for scaling up ML for BDM systems lie in the low interpretability of the ensemble and NN models and the lack of studies in sustainability assessment that consider geo-temporal dynamics. A workflow is recommended for future ML studies for BDM systems. More research is needed to explore ML applications for sustainable development, assessment, and optimization of BDM systems.

1. Introduction

Biomass is widely considered a renewable alternative to fossil fuels and is expected to play an essential role in combating climate change (Stegmann et al., 2020). The concept of bioeconomy has been mentioned in the national policies of more than 40 countries (El-Chichakli et al., 2016). According to the European Commission, a bioeconomy is the "production of renewable biological resources and the conversion of these resources and waste streams into value-added products, such as food, feed, bio-based products, and bioenergy" (Commission and Innovation, 2012). In addition to food, feed, and bioenergy that have been intensively explored in the literature (Lan et al., 2020b), biomass-derived materials (BDM) have obtained increasing interest. Various biomass, such as vegetation, wood, aquatic biomass, or animal wastes, have been considered renewable feedstock for material production. Researchers have explored different biomass precursors to produce biosorbent, biochar, and biomass-derived activated carbon that have broad applications in agricultural and water systems. Biosorbents throughout this article refer to dried biomass without further manufacture; biochar is derived from biomass through various carbonization processes; biomass-derived activated carbon is usually upgraded from biochar with activation, which consists of a series of reactions between activation agents and reactive carbon components within the biochar (Cha et al., 2016).

Typical applications of BDM include soil amendment and wastewater treatment. Activated carbon is one of the most effective adsorbents. In addition to conventional usages such as removing pollutants from aqueous solution, soil, and gas, it has gained popularity in high-value applications, for example, energy storage, catalyst support, and medical applications (D. P. Yang et al., 2019). BDMs are essential due to their capability to combat climate change. For instance, biochar is considered a carbon-negative technology to deliver 3.4–6.3 $PgCO_2e$/year Greenhouse Gas (GHG) emission reduction globally (Lehmann et al., 2021).

The technical, economic, and environmental performance of BDM depends on the combinations of biomass species, conversion technologies, and BDM applications. For example, the effectiveness of biochar application in soil amendment or water treatment highly depends on biochar's physical and chemical properties (Mohan et al., 2014;

* Corresponding author.
E-mail address: y.yao@yale.edu (Y. Yao).

https://doi.org/10.1016/j.resconrec.2022.106847
Received 2 July 2022; Received in revised form 19 October 2022; Accepted 19 December 2022
Available online 5 January 2023

Pignatello et al., 2015). These material properties are governed by conversion pathways, operational conditions, and feedstocks (Suliman et al., 2017; Sun et al., 2014), which also determine the economic viability and environmental impacts (Liao et al., 2020). Large-scale production of BDM is limited due to the complex supply chain, large feedstock quantity and quality variability, challenges in controlling and optimizing biomass conversion, and economic constraints (Liao and Yao, 2021).

Researchers have leveraged Machine Learning (ML) to address the challenges in BDM development and applications. Previous studies have reviewed ML applications in different industrial sectors, including the chemical industry (Liao et al., 2022), bioenergy (Liao and Yao, 2021), power generation (Donti and Kolter, 2021), transportation (Veres and Moussa, 2020), and buildings (Hong et al., 2020). Several review papers discussed ML applications in agriculture (Liakos et al., 2018) and water treatment (Huang et al., 2021; Li et al., 2021; Sundui et al., 2021). However, none of the previous studies have (1) reviewed ML applications of BDM and their applications in agriculture and water treatment systems across the entire life cycle – specifically, from biomass cultivation to BDM production and end-use applications; (2) reviewed ML applications in the sustainability assessment of BDM from diverse biomass feedstock and conversion technologies; (3) discussed interpretability of ML for large-scale BDM system deployment; (4) recommended a workflow to assist future ML applications to BDM systems. A holistic review of ML applications across the BDM life cycle is needed to reveal the unique capacities, potentials, and challenges of ML in supporting systems-wide design and optimization of BDM for their sustainable applications in agriculture and water systems.

This review addresses this need. The literature search and screening methods are discussed in Section 2, with brief overviews of ML and BDM systems. Fifty-three papers were reviewed and categorized based on their objectives, ML methods, and input and output variables (Section 3). The benefits and limitations of existing ML applications are discussed. For each category, this review focuses on answering three questions, including why ML is helpful, how ML has been used from past advances and current developments, and what limitations of ML applications need to be addressed in future research. Future research directions and a recommended workflow are discussed in Section 4.

2. Material and methods

Fifty-three papers were collected through a three-stage process. In the first stage, a search in the Web of Science database was performed using keywords: "machine learning" AND "biochar" and "machine learning" AND "activated carbon". The search resulted in 75 articles published from 2008 to 2021. In the second stage, the introduction sections of the 75 articles were screened to identify additional relevant literature, and Google Scholar was used to identify additional literature, leading to 85 papers. Finally, review papers were excluded, and all articles were filtered based on their relevance to ML and three BDM explored in this study, including biosorbent, biochar and its byproducts, and biomass-derived activated carbon and their applications. This results in 53 papers. The three types of BDM were selected because of their broad applications in water treatment and soil amendment in the agriculture sector. As an emerging field, most papers reviewed are published after 2015. The following sections introduce BDM and ML techniques covered in this review.

2.1. Biomass-derived materials

The BDM supply chain is similar to other biomass-based systems; it involves biomass cultivation, biomass production and harvest, pretreatment, BDM production, distribution, and final application (De Meyer et al., 2014), and sometimes recycling. BDM discussed in this article includes biosorbent, biochar, and biomass-derived activated carbon.

As defined in the Introduction section, biosorbents usually do not undergo intensive thermo-chemical conversions needed for biochar or activated carbon. Biosorbents discussed in this review are biomass dried in an air oven at a temperature $\leq$ 105 °C; their innate porous and chemical structures allow them to act as adsorbents, such as dried sawdust (Prakash et al., 2008) and agricultural waste (Parveen et al., 2017). They remove pollutants through biosorption.

Biochar is the product of various carbonization processes, including pyrolysis, gasification, and hydrothermal carbonization (Cha et al., 2016). These processes yield different mass fractions (wt%) of solids (biochar and ash), liquids (tar and bio-oil), and syngas (a mixture of H_2, CO, CO_2, CH_4, etc.) (Cha et al., 2016; Inayat et al., 2022; Jalalifar et al., 2020). Pyrolysis is a heating procedure operated from 300 – 900 °C without oxygen. Depending on heating rates and temperatures, there are three types of pyrolysis – slow, fast, and flash. Slow pyrolysis favors biochar production; fast and flash pyrolysis majorly produce bio-oil (Inayat et al., 2022; Jalalifar et al., 2020). Gasification is a thermo-chemical partial oxidation process that converts biomass to syngas, and it has liquids and solids as byproducts (Cha et al., 2016; Wu et al., 2023). Pyrolysis and gasification generally require a separate drying step to obtain high product yields and reduce the process energy consumption (Cha et al., 2016); hydrothermal carbonization allows the direct conversion of wet biomass into hydrochar under self-generated pressure and low temperature (180−350 °C) (Liu et al., 2021).

Biochar can be upgraded to activated carbon by activation processes. Different activation agents have been explored. Physical activation uses gas agents (e.g., CO_2, H_2O); chemical activation uses chemical agents (e.g., HNO_3, NaOH) (Cha et al., 2016). Activation can graft functional groups on material surface and generate more pores within materials, which may benefit the desired end-use, such as water treatment (X. Yang et al., 2019).

2.2. Machine learning

Machine learning (ML) is: "a computer program is said to learn from experience E concerning some class of tasks T, and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E" (Mitchell, 1997). Current paradigms of ML include supervised learning, unsupervised learning, and reinforcement learning (Jordan and Mitchell, 2015).

In supervised learning, there are inputs $x \in X$ and outputs $y \in Y$. The inputs x are called features, covariates, or predictors; x is often a fixed-dimensional vector of numbers, such as chemical elemental compositions (C%, H%, O%) of a biomass feedstock. The output y is known as the label, target, or response. The experience E is given as a training dataset D with sample size N illustrated in Eq.(1); the performance measure P is measured by the empirical risk $L(\theta)$ defined in Eq. (2) (Murphy, 2022), where $l(y_n, f(x_n;\theta))$ is the gap between observed value and predicted value; θ is the parameter that determines f. The task T would be learning θ from D such that it minimizes the empirical risk, i.e., the learned $\theta^* = \textit{argmin}\, L(\theta)$, where *argmin* represents "the argument that minimizes $L(\theta)$".

$$Data\ (D) = \{(x_n, y_n)\}_{n=1}^{N} \tag{1}$$

$$L(\theta) =^{\triangle} \frac{1}{N}\sum_{n=1}^{N} l(y_n, f(x_n;\theta)) \tag{2}$$

Many different f exist, including decision trees, decision forests, logistic regression, support vector machines, neural networks, kernel machines, and Bayesian classifiers (Friedman et al., 2001; Jordan and Mitchell, 2015). Various learning algorithms have been proposed to estimate disparate mapping types, such as backpropagation, gradient descent, expectation-maximization (EM) algorithm, boosting, and multiple kernel learning that combine the outputs of learning algorithms (Jordan and Mitchell, 2015; Murphy, 2022).

Unsupervised learning involves the analysis of unlabeled data (i.e., $D = \{x_n : n = 1 : N\}$) under assumptions about the structural properties of the data (e.g., algebraic, combinatorial, or probabilistic) (Jordan and Mitchell, 2015). Two common types of unsupervised learning tasks are dimension reduction and clustering. The dimension reduction method assumes high-dimensional data lie on a low-dimensional manifold and aims to identify that manifold explicitly from data (Jordan and Mitchell, 2015). Popular dimension reduction methods include principal components analysis, manifold learning, factor analysis, random projections, and autoencoders (Friedman et al., 2001). Clustering involves finding a partition of the observed data without explicit labels indicating the desired partition (Jordan and Mitchell, 2015). Often, leveraging dimension reduction methods can assist the clustering procedure.

Reinforcement learning is a class of problems where the system or agent must learn how to interact with its environment. This can be encoded using a policy $a = \pi(x)$, specifying which action to take in response to each possible input x (Murphy, 2022). An example would be a robot learning the biochar application for soil amendment according to regional environmental conditions data. In this case, the environmental conditions data is the input x, and the output a can be whether to apply the biochar or not. That is, x is a set of joint positions and angles for all the robot limbs, and the a is a set of actuation or motor control signals (Murphy, 2022). Although reinforcement learning has not been employed in BDM systems, as the biochar example here, it has the potential to empower real-time decision-making. This paper reviewed different ML models as listed in Table 1. Model types were assigned according to (Murphy, 2022).

ML applications reviewed in this study were categorized into three groups based on their objectives: material and process design optimization (M&P design, Section 3.1), end-use performance prediction (Section 3.2), and sustainability assessment (Section 3.3). For each group, this paper aims to answer three questions, including why ML is helpful, how ML has been used from past advances and current developments, and what limitations of ML applications need to be addressed in future research.

3. Results

Depending on the objectives, different ML applications have diverse dataset sizes. Fig. 1 shows the distribution of dataset sizes of three application groups. Six outliers were excluded (Hough et al., 2017; Karri and Sahu, 2018; Prakash et al., 2008; Shen et al., 2019; Wehrle et al., 2021; Zhu et al., 2020) because they contained model-extrapolated data, porous carbon materials that were not derived from biomass, or spectra data.

Fig. 1 shows that ML applications for end-use performance predictions have the largest dataset size range due to their various application scenarios, e.g., gas adsorption and soil amendment studies compiled $>$ 1000 data points for model training, while 75% of the studies for other applications contained a dataset of $<$ 600. Studies using literature data usually report datasets with a size $>$ 100; studies using first-hand experimental data commonly have smaller datasets (mainly observed in M&P design: size $<$ 100). This review includes studies that used small datasets and different methods to prevent overfitting, such as feature selection (Pathy et al., 2020) and early stopping methods (Selvarajoo et al., 2020). Others integrated the NN framework with spatial interpolation methods such as Kriging (Ismail et al., 2019) or optimization techniques (Ewees and Elaziz, 2018) to enhance the performance.

3.1. ML applications for material design and process optimization of BDM

BDM development commonly relies on trial-and-error experiments with different combinations of biomass feedstocks and conversion processes. Due to laboriousness, past experimental studies have focused on a single or a few material design combinations (Varma, 2019). Furthermore, it is difficult to tailor biomass conversion and feedstock

Table 1
Machine learning model types.

Model type	Models included	Supervised/ Unsupervised
Linear (LM)	(1) Linear regression (LR): Multiple linear regression (MLR)	Supervised
	(2) Generalized linear models (GLM)	
	(3) Linear Discriminant Analysis: Naïve Bayes Classifier (NBC)	
Neural Network (NN)	(1) * Feedforward Neural Network (FFNN)	Supervised
	(2) Adaptive Neuro-Fuzzy Inference System (ANFIS)	
	(3) Radial basis neural network: generalized regression neural network (GRNN)	
	(4) Cascade forward backpropagation	
	(5) Recurrent network (RNN): Jordan network, Elman backpropagation	
Exemplar-based methods	K-nearest neighbor (KNN)	Supervised
Kernel methods	(1) Gaussian Processes (GP)	Supervised
	(2) Support-vectorMachine-based algorithms (SVM): for regression – Support-vector Machine Regression (SVR) or Least-Square SVM (LS-SVM)	
Trees, Forests, Bagging, Boosting (TFBB)	(1) Decision Trees (DT): for regression – Regression Trees (RT)	Supervised
	(2) Random Forests (RF)	
	(3) Bagging	
	(4) Boosting: gradient boosting (GB), extreme gradient boosting (XGB)	
Bayesian Network	Bayesian Network	Supervised
Clustering	Clustering	Unsupervised

* FFNN is also called Multilayer perceptron (MLP).

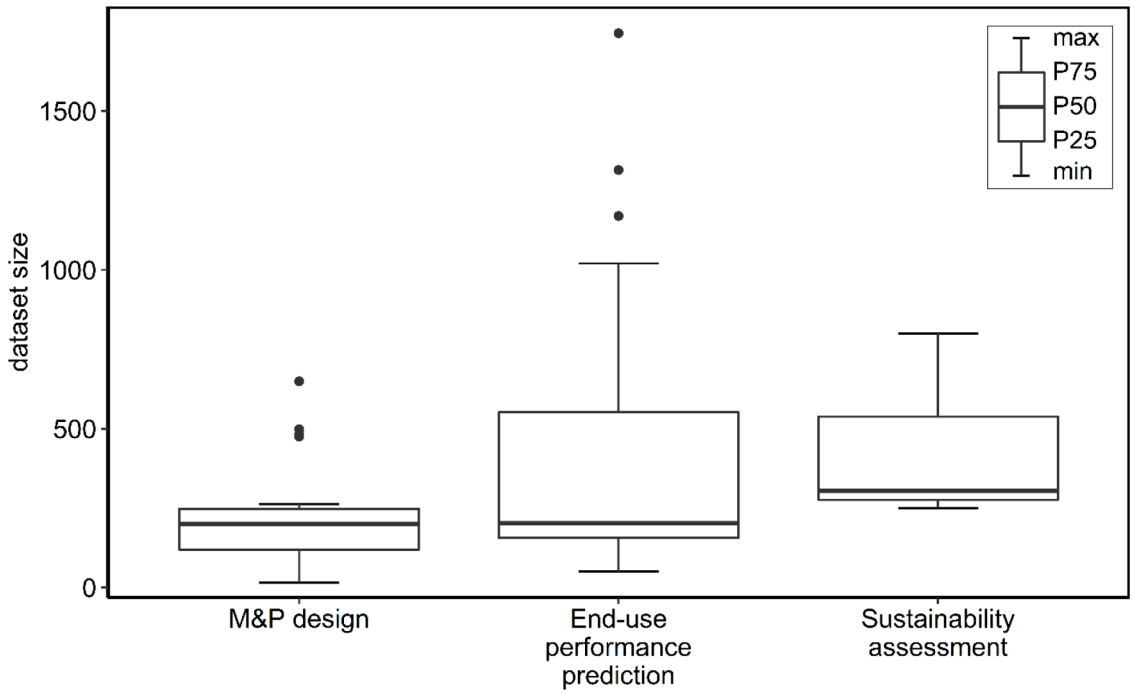


Fig. 1. Dataset size distribution of ML applications reviewed in this study (P = percentile, black dots = outliers).

selection for desired material properties, given the challenges in understanding feedstock-process-property relationships with limited experimental data. To overcome the difficulties of traditional experimental approaches, previous studies applied ML in two ways: (1) predicting product yields and properties and (2) predicting thermochemical conditions required to achieve desired material/process performance, enabling reverse engineering to identify/optimize production pathways given material properties. As biosorbents require few processing steps, there is barely any optimization need; thus, literature has explored chiefly biochar and biomass-derived activated carbon production.

Utilizing ML to predict product yields and features based on various feedstock and thermal chemical treatments has received the most attention. In total, there are 19 papers describing 31 models. Inputs of these models include feedstock and process features (Fig. 2). Feedstock features include elemental compositions (C, H, O, N, S wt%), structural components (lignin, cellulose, and hemicellulose wt%), particle size, proximate analysis data (ash, fixed carbon, and volatile compound wt %), and higher and lower heating values (LHV, HHV). Process conditions include chemical and heat pretreatment, pyrolysis conditions (e.g., temperatures, rate, and residence time), and activation conditions (e.g., impregnation ratio of material, activation agent, and reaction time). Common model outputs are product yields. Recent studies also include BDM characteristics, such as chemical compositions, fuel properties (e.g., HHV, energy recovery efficiency), sorbent capacities, and specific capacitance (see Table S2 for detailed inputs and outputs of each application).

With respect to algorithms, TFBB and NNs were mainly used (Fig. 2). Notably, deep neural networks-based (DNN, i.e., ANN with more than two hidden layers), DT, SVM-based, LR, RF, and XGB have been employed. SVM-based regression includes SVR and LS-SVM. DNN-based includes FFNN, ANFIS, and integration of DNN with optimization techniques such as Kriging and gray wolf optimization (GWO) (Table S2). According to the "No free lunch" theory (Wolpert and Macready, 1997) – all algorithms, on average, have similar performances under specific constraints. That is, one algorithm can perform better in some instances but worse in others. We summarized the basics, strengths, and weakness of common algorithms applied to BDM (Table S1) and provided rationales for why they work/does not work well on datasets used for M&P design.

M&P design encompasses highly non-linear processes, uncertain measurements, various correlated and uncorrelated features with a wide range of values and units, a combination of different data types (e.g., feedstock type is categorical data; process/product parameters is numeric data), and relatively small dataset size. The prediction task can

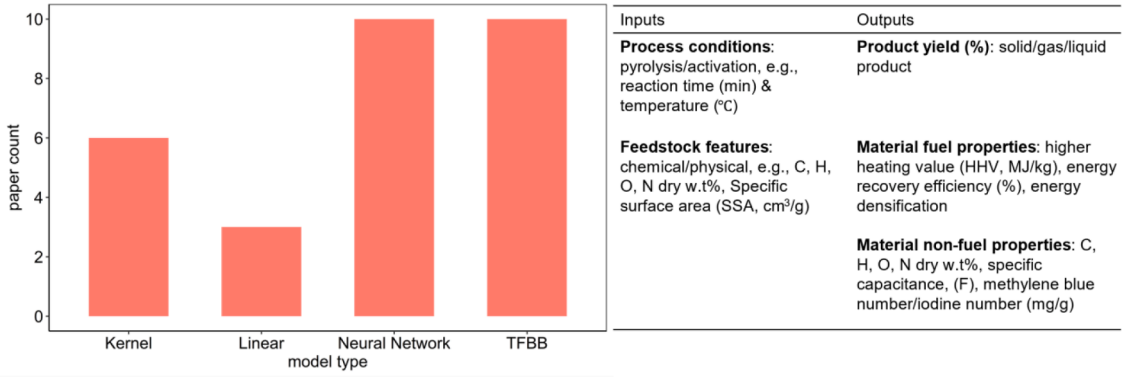


Fig. 2. Summary of ML applications for material design and process optimization of BDM.

be either multi-input-single-output (MISO) or multi-input-multi-output (MIMO) – MIMO is feasible because outputs of material features are correlated, and multi-task learning can exploit the relatedness (Ben-David and Schuller, 2003). It is recommended to apply several candidate algorithms and choose the one that suits one's specific situation. Here, we present some observations from past ML applications.

Table 2 displays a summary of studies that compared more than 2 models. Overall, SVM outperforms RF when the kernel was optimized to radial basis function (RBF), and models that do not assume linear relationships between inputs and outputs are more suitable (Table 2). The dataset from (Cao et al., 2016) has been examined with LS-SVM, FFNN, ANFIS, and extended ANFIS (Ewees and Elaziz, 2018). The performance ranking: extended ANFIS (with GWO) > LS-SVM > ANFIS > NN. The reason is that the extended ANFIS models the BDM most comprehensively – it incorporates uncertain fuzzy rules presented in BDM and adjusts for small datasets. LS-SVM lacks stochasticity but predicts with high accuracy on small datasets, and it finds global optimal, ANFIS and FFNN may converge to local optimum for extremely non-linear processes, while ANFIS performs slightly better than FFNN due to the fuzzy consolidation.

For heterogeneous tabular (also called structured) datasets (e.g., data used in M&P design studies), ensemble gradient-boosted trees (GBDT, e.g., XGB) have dominated over NN (Shwartz-Ziv and Armon, 2022). The reason may be that Tree-based methods can directly process input variables while NNs require data preprocessing (e.g., data standardization or normalization). Additionally, GBDT can achieve high accuracy with a small dataset, while conventional NN requires larger datasets. The phenomenon has also been observed in previous BDM datasets (Thiruvengadam et al., 2021). Despite the superior performance of GBDT, its interpretability is inferior to DT, and flexibility is poorer than NN or Kernel methods. In particular, DT is a white box where one can see how the model is trained; NN and Kernel methods allow incorporation with mechanistic processing models. The scientific interpretability and performance trade-offs should be considered for model selection (more ML interpretability discussions in Section 3.4).

In terms of functionality, ML is useful in capturing hidden patterns in complex datasets such as those from biomass conversion mechanisms. (Alaba et al., 2020) utilized multiple Artificial Neural Networks (ANNs) with varying architectures to predict thermogravimetric curves, which showed the degradation mechanism of rice husk pyrolysis. Not all studies rely on experimental data. For example, (Thiruvengadam et al., 2021) employed extreme gradient boosting (XGB) to build generalizable predictive models for material properties, and the data were obtained from computational expensive pyrolytic polygeneration kinetic modeling (e.g., detailed gaseous and liquid product types and corresponding yields).

Aside from output predictions, some studies have used ML to identify critical input features. For example, (Zhu et al., 2019a) used pyrolysis conditions and the properties of lignocellulose biomass as inputs to train an RF model. They determined that pyrolysis temperatures were more significant in influencing the yields and carbon contents than biomass properties. (Li et al., 2020) leveraged SVM and RF to predict biochar yield and fuel properties (e.g., HHV, energy recovery efficiency, and energy densification). They concluded that elemental compositions (C, N, H wt%) are critical for determining fuel properties.

Few studies have investigated the reverse engineering perspective of ML applications, which identifies process conditions required to meet desired material or process properties. For example, (Jalalifar et al., 2020) developed a computational fluid dynamic model and an SVR model using particle swarm optimization algorithms to identify optimum pyrolysis conditions for maximum yield of bio-oil. Similarly, (Mathew et al., 2020) used multi-response optimization techniques to determine production conditions for producing activated carbon with optimal super capacitance and lowest resistance.

Overall, ML employment in this category has guided material design and process optimization. ML can help identify the most influential factors for material development and process optimization; furthermore, ML can be used in a reverse engineering fashion to develop tailored biomass conversion processes for desired BDM properties. One promising direction in this paradigm is using ML for rapid screen and exploration of diverse biomass species and conversion processes for BDM development to reduce experimental efforts. For example, ML models based on features of biomass feedstock (e.g., elemental compositions) may be used to predict the material and process performance of BDM derived from new biomass feedstock (as long as their composition data are available).

The main challenge of large-scale ML applications is data availability. Many studies reviewed in this section have used small datasets. The use of physics-informed machine learning to address small datasets and allow the incorporation of the laws of physics has been discussed in the literature (Eichelsdörfer et al., 2021; Karniadakis et al., 2021). For example, future ML applications can incorporate conversion reaction rules (e.g., pyrolysis or activation mechanisms) as constraints into ML algorithms such that the models can learn based on the known relationship and adapt to smaller dataset sizes and have enhanced interpretability (Ji and Deng, 2021).

Table 2
Summary of ML applications for material design and process optimization of BDM.

ML type	ref	Winner *	Competitor algorithms	SVM kernel	Objective category**
TFBB	(Li et al., 2015)	DT	MLR	–	A(a, b)
	(Jiang et al., 2019a)	RF	MLR; SVM	polynomial	A(b)
	(Jiang et al., 2019b)	RF	MLR; SVM	polynomial	A(a, b)
	(Thiruvengadam et al., 2021)	XGB	FFNN	–	A(b), B
Kernel	(Cao et al., 2016)	LS-SVM	FFNN	RBF	A(b)
	(J. Li et al., 2020)	SVM	RF	RBF	A(a,b)
	(Li et al., 2021)	SVM; FFNN	RF	RBF	A(a), B
NN	(Hough et al., 2017)	FFNN	DT	–	A(a), B
	(Ewees and Elaziz, 2018)***	ANFIS-GWO	ANFIS; FFNN; LS-SVM	RBF	A(b), B
	(Ismail et al., 2019)	FFNN-Kriging	FFNN	–	A(b)

* Winner is the model with the lowest test RMSE (if RMSE is not available, R^2 or other metrics were used).

** Objective category:
A. Material property prediction: (a) energy related; (b) non-energy related
B. Reverse engineering (estimate optimal input combination for desired output).

*** (Ewees and Elaziz, 2018) used the data from (Cao et al., 2016).

3.2. End-use performance prediction

The most common end-use applications for BDM included in this review are pollutant treatments and soil amendments, which are essential environmental management practices for safe water and healthy soil. Conventionally, laborious trial experiments are necessary for selecting treatments for specific goals and sites. Thirty papers using 73 models were investigated (Table S3). Input/output variables and the number of applications by different ML algorithms are shown in Fig. 3: similar to M& P design, the most popular ML algorithms are TFBB and NNs. The difference is the inclusion of Clustering and Exemplar methods and a greater variety of NN and TFBB models (Table 3).

The input and output variables are application-dependent; therefore,

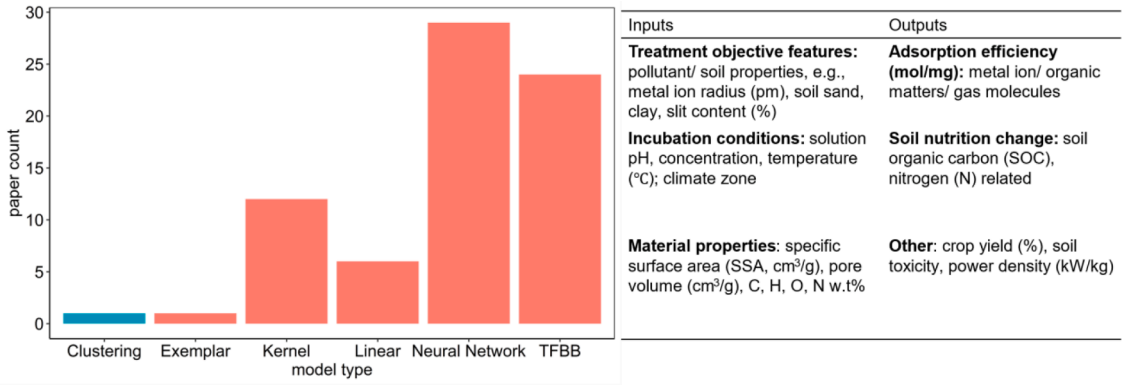


Fig. 3. Summary of ML applications for predicting the end-use performance of BDM.

ML applications are discussed in the following sections by their applications.

3.2.1. Wastewater treatment

BDMs remediate wastewater through adsorption mechanisms. Wastewater is a multi-component system consisting of organic compounds (e.g., dye) and inorganic components (e.g., metal ions and nutrients). The sorption of these chemicals onto carbonaceous sorbents is concentration-dependent (non-linear). The complex interactions between adsorbents and adsorbates are challenging to be captured by traditional modeling methods (Sigmund et al., 2020). Essentially, adsorption capacity (Q_e) under an equilibrium concentration of the chemical (C_e) is a function of chemical properties, adsorbent properties, and C_e: $logK_d = logQ_e/C_e = f(chemical,\ adsorbent,\ C_e)$, where $logK_d$ quantifies the extent of adsorption (Zhang et al., 2020). Ideally, combining the chemical and adsorbent properties can better model the adsorption mechanisms, which is almost impossible to achieve by simple regression models (Zhang et al., 2020).

Different ML models have been applied to BDM applications for wastewater treatment. The earliest ML application was found for a sawdust sorbent, which leveraged ANN to predict Cu(II) adsorption efficiency onto the sawdust (Prakash et al., 2008). Over the years, more ML techniques have been adopted to investigate pollutant removal efficiencies from biomass-derived activated carbon and biochar. ML models, including ANN (e.g., DNN, RNN, and ANFIS), SVM (e.g., SVR), GP, LM (e.g., NBC), KNN, DT, boosting/bagging, boosting and bagging of DT (e.g., XGB, boosted regression trees), and RF have been proved to be powerful for predicting adsorption efficiencies and identifying influential factors for the adsorption performance (de Miranda Ramos Soares et al., 2020; Zhang et al., 2020).

ML model development in this category has advanced through the following stages; each stage moves closer toward authentically reflecting real-world adsorption systems, and they share the same output – the adsorption capacity (Zhang et al., 2020):

(a) Building predictive models under a particular isotherm model assumption for a single/multi-component system, based on inputs from adsorbent dosage and solution features (de Miranda Ramos Soares et al., 2020; Dolatabadi et al., 2018; el Hanandeh et al., 2021; Karri and Sahu, 2018; Li et al., 2019; Mazaheri et al., 2017; Mojiri et al., 2020, 2019; Nguyen et al., 2021; Parveen et al., 2017; Prakash et al., 2008; Talebkeikhah et al., 2020);
(b) Constructing predictive models for a single/multi-component system based on inputs from adsorbent (e.g., surface areas and elemental components) and solution features (Afolabi et al., 2020; Ke et al., 2021a; Sigmund et al., 2020; Zhang et al., 2020; Zhao et al., 2021; Zhu et al., 2021, 2019b);
(c) Leveraging unsupervised learning and supervised learning to improve adsorption efficiency predictions according to different metal ion and adsorption environment combinations (Ke et al., 2021b).

At stage (a), inputs are often solution pH, initial chemical concentration (from one type of dye molecule or metal ion), chemical solution temperature, and the contacting time of sorbents and sorbates. For example, (Parveen et al., 2017) developed a support vector regression model for predicting the sorption capacity of Cr(VI) onto a biosorbent agricultural waste 'maize bran.' The input features included the contact time of Cr(VI) and maize bran, initial Cr(VI) concentration, pH of the Cr (VI) solution, and the adsorption temperature. (Dolatabadi et al., 2018) built ANN and ANFIS models to predict the simultaneous adsorption capacity of dye and Cu(II) onto the sawdust. The input variables included initial dye concentration, initial Cu(II) concentration, contact time of the sawdust, and the mixture solution (dye and Cu(II)).

At stage (b), input factors are extended to adsorbent and adsorbate properties. For instance, (Zhu et al., 2019b) developed RF and ANN models to predict heavy metal ions' adsorption capacity onto biochar. They trained ML models by adsorbent features, including pH of biochar in water, cation exchange capacity, ash content, biochar particle size, the carbon content in biochar, biochar stability $(O + N)$/C, biochar polarity H/C), and adsorbate properties from 6 types of heavy metal ion solutions (Pb(II), Cd(II), Ni(II) As(II), Cu(II), Zn(II). Their results showed that RF was more robust than ANN, because ANN predicted negative adsorption capacity values when the actual adsorption capacity is extremely low. This phenomenon that ANN failed for predictions at boundary cases was again found in their later work (Zhu et al., 2021). Additionally, they found that RF models could be generalized to adsorption prediction for other heavy metal ions. Based on the dataset collected from (Zhu et al., 2019b) and with the same input features, (Zhao et al., 2021) further employed Kernel Extreme Learning Machine (KELM, a variation of SVM) and Kriging to model the adsorption behavior of biochar in the multi-component heavy metal ion system.

Algorithms before stage (c) belonged to supervised learning. At stage (c), research has begun to use unsupervised learning. Most recently, leveraging the dataset from (Zhu et al., 2019b), another study (Ke et al., 2021b) divided the data into clusters using an unsupervised learning technique, the fuzzy C-means clustering (FCM) method. Eventually, 4 clusters were uncovered, and each cluster represented a kind of treatment-adsorption environment combination, characterized by biochar characteristics and adsorption conditions. After clustering, a backpropagation neural network model (BPNN) was deployed to predict adsorption efficiency under each cluster. This integrated FCM-BPNNs (test RMSE = 0.036) showed accuracy improvements compared to BPNN alone (test data RMSE = 0.050). For any heavy-metal-ion-contained wastewater site to be treated by some BDMs, if one can first classify the environment-BDM combination into one of the 4 clusters, the heavy metal ion removal prediction can be significantly improved.

Table 3
ML applications for predicting the end-use performance of BDM.

ML type	ref	Winner*	Competitor algorithms	SVM kernel	Objective category***
TFBB	(Mazaheri et al., 2017)	BRT	FFNN	–	A(a, b)
	(Cipullo et al., 2019)	RF	FFNN	–	C
	(Zhu et al., 2019b)	RF	FFNN	–	A(a)
	(De Miranda Ramos Soares et al., 2020)	RF	FFNN	–	A(b)
	(Ke et al., 2021a)	RF; Bagging (SVM-FFNN)	FFNN; GP; M5Tree**; SVM; Bagging**	RBF	A(a)
	(Maulana Kusdhany and Lyth, 2021)	RF; XGB	MLR; SVM	RBF	B
	(Nguyen et al., 2021)	RF	CUBIST**; GLM; KNN; MLR; SVM	RBF	A(c): NH_4—N
	(Zhu et al., 2021)	RF	FFNN; GBT	–	A(b)
	(Palansooriya et al., 2022)	RF	FFNN; SVM	RBF	C, 1.A (b)****
Kernel	(Parveen et al., 2017)	SVM	FFNN; MLR	RBF	A(a)
	(Talebkeikhah et al., 2020)	SVM	ANFIS; DT; FFNN; GMDH**; RBFNN; RF	RBF	A(a)
	(Nguyen et al., 2021)	SVM	CUBIST**; GLM; KNN; MLR; RF	RBF	A(b): BOD_5**
	(Zhao et al., 2021)	GP (Kriging)	KELM	–	A(a)
NN	(Dolatabadi et al., 2018)	ANFIS	FFNN	–	A(a, b)
	(Zhang et al., 2020)	FFNN	Bagging; SVM	RBF	A(b)
	(Zhou et al., 2020)	FFNN	GLM; RF; SVM	RBF	D
	(El Hanandeh et al., 2021)	GRNN**	Elman NN; FFNN; GB	–	A(a)
	(Ke et al., 2021b)	FCM-FFNN**	FFNN	–	A(a)

* Model with the lowest test RMSE is designated as the winner (if RMSE is not available, R^2 or other metrics were used).

** Bagging in (Ke et al., 2021a) built bagged models with combinations of the four models – FFNN, GP, M5Tree, SVM;

M5Tree: a Decision Tree learner; CUBIST: an extension of M5Tree; GMDH: grouped method of data handling; FCM-FFNN is unsupervised-supervised framework; KNN is K-nearest-neighbor, which is an Exemplar framework; GRNN: General regression neural network

BOD_5: Biological oxygen demand during 5 days.

*** A. Pollutant removal: (a) Metal ion, (b) Organic matter, (c) Non-organic matter; B. Gas molecule adsorption; C. Soil amendment; D. Electrode.

**** 1.A(b) inherits from M&P design, which is reverse engineering.

Besides algorithms, the abundance of training data is critical for improving prediction accuracy. Data from adsorption experiments are often limited; therefore, efforts have been spent on harnessing values from limited data beyond algorithms selection or hyperparameter optimization. For example, (Zhang et al., 2020) improved prediction accuracy by employing a cosine similarity approach that mined the available data before building models. The mining approach identified the most relevant adsorption isotherm data concerning the prediction target and then utilized mined data to build models – if one tries to predict the adsorption of phenol on a granular activated carbon (GAC), the cosine similarity approach suggested training models based on the adsorption data of phenol or phenol-like chemicals on GACs.

Although ML models can help decision-making for wastewater treatment, (Mendoza-Castillo et al., 2018) mentioned a few pitfalls of ML applications if inappropriate output variables were chosen. Specifically, they conducted a multi-metallic adsorption test on BDM and built separate ANN models with different output variables – removal percentage, adsorption capacities, and solute concentrations after adsorption (i.e., adsorption equilibrium concentration). Their results showed that ML models failed to predict adsorption efficiencies when models were trained using adsorption equilibrium concentrations or removal percentages alone as the output variables.

In general, for a fixed sorbent, its adsorption efficiency (q_e in Eq. (3)) increases with the target sorbate's initial concentration ($[M_i]_0$). During a Langmuir-type adsorption process, at equilibrium ($[M_i]_e$), ($[M_i]_e - [M_i]_0$) stays constant. For a multi-component system, because a target ion's adsorption may be inhibited by other ions, resulting in $[M_i]_e$ approaches $[M_i]_0$, making $[M_i]_e - [M_i]_0$ no longer to be constant. When ML training uses $[M_i]_e$ as output and $[M_i]_0$ as input, q_e estimated using ML results show a decreasing trend with increasing $[M_i]_0$, which contradicts physical observation. This contradiction will not exist if ML training uses q_e as output and $[M_i]_0$ as input. Thus, it is critical to consider these dynamic adsorption phenomena and choose appropriate output variables when training the model. Mendoza-Catillo et al. also pointed out that no single ML model fits all. They suggested testing with different ML models until finding the optimal one.

$$q_e = \left([M_i]_e - [M_i]_0\right) \times V/m \tag{3}$$

Where V is the solution volume (L), and m is the sorbent mass (g) (Mendoza-Castillo et al., 2018).

3.2.2. Soil amendment

Among BDMs, biochar has gained the most attention for soil amendment. Numerous literature has discussed the benefits of biochar in storing carbon and combating climate change (Lehmann et al., 2021), improving soil water retention (Razzaghi et al., 2020) and fertility (Vijay et al., 2021), and remediating problem soils (Yu et al., 2019). Biochar can impact soil conditions through various mechanisms affecting microbial activities, including adsorption processes and soil pH adjustments. The interactions between biochar and soil are complex and challenging to be modeled by traditional regression methods. Previous studies have used ML as a powerful tool to understand the underlying relationships between biochar and the soil environment and predict the effectivities of biochar for soil amendment

ML in soil amendment has different goals. Over the years, three subfields based on ML purposes have developed: organic matter preservation (C sequestration (Ding et al., 2018; Shen et al., 2019; Wehrle et al., 2021) and N conservation (Liu et al., 2019; Wehrle et al., 2021)), pollutant removal (Cipullo et al., 2019; Palansooriya et al., 2022), and crop production improvement (Dokoohaki et al., 2019; Dumortier et al., 2020).

For organic matter (C and N) preservation, previous studies applied ML to identify critical factors and optimal strategies for biochar applications. (Ding et al., 2018) adopted boosted regression trees (BRTs) algorithms to identify factors determining the impact of biochar on soil carbon priming. The key factors are incubation conditions (incubation time and soil moisture) and biochar properties (biochar C/N ratio, nitrogen content, pyrolysis time, and biochar pH), while soil properties (N, slit content, C/N ratio, pH, land-use type) are less critical factors. (Liu et al., 2019) built RF models to understand how soil properties, biochar type, biochar addition level/rate, and climate zone impact soil N preservation after biochar amendment. Additionally, they utilized the predictive model to identify optimal biochar application strategies according to global soil conditions.

Due to organic matter's uneven and dynamic distribution, measuring SOC across large geospatial and temporal scales is challenging. ML has been used to facilitate SOC measurement using fast screening spectral methods such as ground penetrating radar (GPR) and portable mid-

infrared spectroscopy (MIRS). These studies can support further investigation of biochar application performance. (Shen et al., 2019) used GPR signal attributes as inputs and built a Naïve Bayes model to predict soil organic carbon (SOC) in biochar-amended soil. ML can also provide predictions under noisy spectra data. (Wehrle et al., 2021) utilized SVM and kernel methods to calibrate the large-variation-portable MIRS spectra and build predictive models to evaluate organic C and N components following soil amendments. Those studies have focused on SOC measurement instead of BDM's impact on soil; therefore, this review does not provide further discussions. Instead, readers are referred to literature in the SOC fast screening field (Heuvelink et al., 2021; Sothe et al., 2022; Zhou et al., 2022).

Pollutant removal is vital for amending problematic soil. Two studies have investigated ML applications for biochar applied to soil contaminated by heavy metals. (Cipullo et al., 2019) built RF and ANN models with experimental data to predict heavy metal bioavailability concentration and toxicity of biochar-treated soil and identify critical factors determining the remediation performance. (Palansooriya et al., 2022) collected past literature data addressing heavy metal immobilization, and leveraged RF, SVR, and NN techniques to predict heavy metal immobilization efficiency in biochar-amended soils. They concluded that ML models performed well in prediction and ML methods have different strengths. RF model provided insights on critical features that drive bioavailability and toxicity of the soil, while ANN models offered accurate predictions of the toxicity change after biochar or traditional compost amendment.

Biochar is expected to improve soil health and productivity, which are crucial for sustainable crop production to meet increasing food demands (Vijay et al., 2021). Previous studies have applied ML to predict crop yield following soil amendment. (Dokoohaki et al., 2019) studied the Bayesian network (BN) model and generalized additive model (GM) to predict crop yield response according to different soil conditions. Their results showed BN model outperformed. In addition, they discovered that regions with poor soil quality displayed a higher probability of yield increase after biochar addition. Based on the BN model in (Dokoohaki et al., 2019), another study (Dumortier et al., 2020) predicted location-specific yield responses across the U.S. for six types of crops. They also evaluated the financial return for farmers and the indirect environmental impacts following the biochar amendment, which are discussed in the sustainability assessment section.

One major limitation of current ML models is the incapability of predicting the long-term effects of biochar in the soil due to the lack of experimental data. The long-term carbon permanence of biochar has been studied in some literature (e.g., 63−82% of the initial carbon in biochar remains in the soil after 100 years (Woolf et al., 2021)). However, the long-term impact of biochar on crop yields has huge uncertainty (Ye et al., 2020) due to the lack of long-term data. Another critical factor not included in previous ML literature is the co-application of fertilizers. A meta-analysis (Ye et al., 2020) of field studies reported the vital role of nutrient addition in determining the crop yield response to biochar application and called for prioritizing nutrient selection for future biochar research. This may also be a promising future direction that ML can support.

3.2.3. Miscellaneous applications

In addition to wastewater treatment and soil amendment, there are other emerging end-use applications of BDM recently, including gas molecular uptake (Maulana Kusdhany and Lyth, 2021; Zhang et al., 2019; Zhu et al., 2020) and electric double-layer capacitors (Su et al., 2019; Zhou et al., 2020). Kusdhany and Lyth predicted the capability of BDM in adsorbing H_2 for clean energy storage; Zhang et al. and Zhu et al. predicted the CO_2 adsorption capability of BDM. The input variables for gas molecule and energy storage predictions include detailed material textural properties, such as ultra-micropore volume, micropore volume, mesopore volume, and specific surface area. NN, SVM-based, RF, and XGB methods have been adopted. Su et al., 2019 and Zhou et al., 2020 investigated electric double-layer capacitors derived from various feedstocks, among which biomass-derived data points account for a small portion. It was observed that RF and NN had high performance, with NN succeeding in predicting extremely low capacitance (Zhou et al., 2020) and RF capable of exporting results for researchers to explain the impact of each input variable (Su et al., 2019).

ML models can capture complicated interactions between materials and wastewater/soil systems. Previous studies show that given a site condition, ML can enable the selection of effective biomass-derived materials for specific pollutant removal or desired crop yield improvements without costly experiments.

In terms of model performance, we summarized several observations from previous studies (Table 3). Performances are application-dependent; in some cases, RF wins over SVM, FFNN; in other cases, SVM wins over RF, NN; in other cases, NN is optimal. Ensemble methods (e.g., RF and BRT) and modified NN models (e.g., ANFIS, Kriging, FCM-ANN, GRNN) possessed superior performances on end-use datasets. Different from applications in M&P design (Table 2), RF can win over SVM with RBF kernel for particular objectives. The reason can be that end-use datasets share a similar data structure with M&P datasets and may include considerably more noise. The noises came from measurements for various objectives in the system: adsorbents, adsorbates, and incubation conditions. Ensemble methods are more robust to noises for small-size datasets (Olson et al., 2018; Sagi and Rokach, 2018); thus, they are more generalizable. The performances were improved for the modified NNs in the previous studies; they were integrated with clustering methods, optimization techniques, and fuzzy rules to accommodate the need for modeling stochastic and small-size datasets. In addition, ML techniques can identify influential variables for treatment performances, enable what-if scenarios investigation for different combinations of input changes, and allow decision-makers to adjust operational parameters for better output performances.

3.3. Sustainability assessment

Given the broad coverage of different environmental, economic, and social themes in the concept of "sustainability," sustainability assessment is considered the most complex appraisal method (Sala et al., 2015). Different approaches have been explored previously, including environmental life cycle assessment (LCA), life cycle cost analysis (LCC), social LCA, and life cycle sustainability assessment (Costa et al., 2019; Onat et al., 2017; Sala et al., 2015). Understanding the potential sustainability implications of new materials, such as BDMs, is critical to further design and optimization of those technologies towards sustainability (van Schoubroeck et al., 2021; Yao and Huang, 2019).

In this section, four articles and six models were identified, containing the fewest articles (Fig. 4), and none is related to biosorbents (Table S4). The input and output variables were similar to those in M&P design and end-use performance prediction; the main difference is that studies in this section leveraged the predicted values from ML to conduct sustainability assessment. BDM production can be energy-intensive and have high environmental footprint (Lan et al., 2020; Liao et al., 2020). Assessing the environmental footprint of BDM is often challenging due to the lack of life cycle inventory (LCI) data for various feedstocks and process conditions. LCI data commonly include mass and energy balances and emissions to land, water, and air. Many LCAs rely on static LCI data for fixed biomass feedstock and process conditions, making their conclusions and results challenging to be used for varied biomass species and process operations.

ML has been used to link LCI data with critical feedstock and process parameters, allowing for estimating LCI data needed for further sustainability assessment. These ML applications can be classified into two groups depending on the specific impacts. The first group of ML applications estimated the environmental or economic consequences directly associated with activities in the life cycle of BDM (e.g., biomass cultivation or conversion). The second group explored the use of ML in

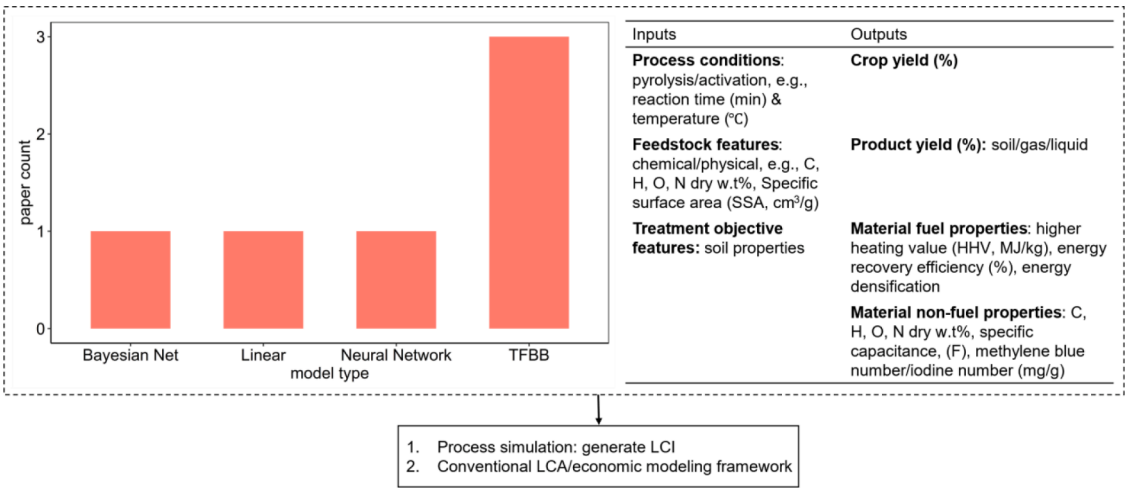


Fig. 4. Summary of ML applications for sustainability assessment of BDM.

understanding the impacts caused by adopting BDM (e.g., crop yield increase or land-use change). The following sections discuss these two groups of applications.

3.3.1. *Direct impact assessment*

Previous studies mainly focused on assessing the potential environmental impacts of biomass conversion. ML models were trained on inputs such as feedstock elemental compositions and thermochemical conditions (pyrolysis conditions: pyrolysis temperatures, rate, and time; activation conditions: activation agent and activation time). Outputs include the yields and energy contents (e.g., HHV) of BDMs and byproducts. The trained ML models estimated the yields and properties of biochar derived from diverse biomass types, which are input to other process-based models to estimate the LCI data for LCA and economic analysis.

For example, (Liao et al., 2020) used a combination of kinetic and ANN models to predict the yields and compositions of activated carbon that are input to the process simulation models developed in AspenPlus, a software for whole chemical plant simulation, to generate gate-to-gate LCI data such as energy consumption and air emissions. (Cheng et al., 2020b; Cheng et al., 2020a) used RF models to predict the yield, energy content, and C and N contents of biochar. Further, these predicted data were used as inputs to estimate energy return on investment (EROI) and global warming potential (GWP). In addition, they conducted an economic analysis to examine the trade-offs between economic and environmental performance for various combinations of feedstock characteristics and pyrolysis temperatures. The results showed superior climate benefits but inferior economic feasibility of lignocellulosic biochar with lower pyrolysis temperature than sludge and crop residual-based products. Unlike traditional BDM LCAs that only include a limited number of biomass feedstock, ML-enhanced LCA includes a variety of biomass feedstocks and can directly assess the environmental implications of different operational conditions.

One potential future direction is adopting ML to optimize the thermochemical conversion processes of BDM production to reduce environmental impacts while maintaining desired material properties. Previous LCA studies have found that thermochemical conversion processes for manufacturing biochar and activated carbon make large contributions to the life cycle environmental impacts (Osman et al., 2022; Smebye et al., 2017). Therefore, ML applications regarding this aspect may empower sustainability-informed material/process design and optimization.

3.3.2. *Indirect impact assessment*

Only one study used ML to evaluate the indirect impact of BDM. (Dumortier et al., 2020) estimated the land use and GHG implications of crop yield changes induced by biochar application. The increased crop yields in the United States lead to lower commodity prices globally, resulting in reduced agricultural land use (Kauffman et al., 2014) and associated GHG emissions. The authors used the Bayesian Network model to estimate the location-specific crop yield changes in response to management options, the properties of biochar and soil, and biomass conversion parameters. The global carbon implications were estimated using an agricultural commodity model based on the crop yield changes.

As discussed in previous sections, some studies have used ML to predict the conversion yields and soil effects of biochar. Those ML applications can be combined with different land-use change models to understand the induced GHG emission implications at various locations and times using a similar approach presented in (Dumortier et al., 2020). In addition to land-use change, other indirect impacts of biochar applications have been investigated in LCA literature (Tisserant et al., 2022), such as reduced fertilizer application and decreased N_2O emissions due to weakened nitrogen leaching. Similar to SOC changes, those indirect impacts are location-dependent, and the LCA study relies on generic data ranges (Tisserant et al., 2022). ML applications may allow location-specific estimation of these indirect impacts and support dynamic, regionalized LCA for biochar applications.

3.4. *Interpretability of ML models for BDM systems*

Interpretability of ML models has obtained increasing attention in the field of Artificial Intelligence, although they have not been broadly discussed in the field of BDM (Han et al., 2022; Marcinkevičs and Vogt, 2020; Pearl, 2022; Rudin, 2019). Being able to interpret the interactions between parameters and causal relationships between inputs and outputs are keys to develop robust BDM systems (Marcinkevičs and Vogt, 2020). There are two aspects of interpretability: explainable ML and interpretable ML.

Explainable ML refers to a collection of post hoc methods used to explain complicated models (Rudin, 2019). Most of the ML interpretations reviewed in this study are explainable ML. For example, the RF variable importance analysis was conducted in several studies (Cipullo et al., 2019; Nguyen et al., 2021; Zhu et al., 2021, 2020, 2019b), which quantifies the importance of variables according to the prediction error they reduce when being adopted to construct the model. SHapley Additive exPlanations (SHAP) Dependence plots have been utilized to help diagnose the positive or negative impacts of factors on outcomes (Lundberg et al., 2017). Example implementations include (Li et al., 2020; Maulana Kusdhany and Lyth, 2021; Pathy et al., 2020). However, these post-hoc explanations do not reflect physical constraints of a

system, causality, and transferability (the ability that the model can transfer learned information to unfamiliar situations) (Lipton, 2016).

Interpretable ML are models that are trained transparently with human-understandable steps and the weights learned by the model have physical meanings (Lipton, 2016). For example, a DT with a reasonable depth allows for understanding the decision process at each tree split. In the literature reviewed, DT has not been used in many studies because of its moderate performance (Hough et al., 2017; Li et al., 2015). (Li et al., 2015) visualized the M&P design factors identified at each step during the DT training process; thus, given a new data point, users can follow the decision criteria to make predictions. Another frequently mentioned interpretable ML are linear models because weights in simple linear models can be interpreted as strengths of associations between features and predictions (Lipton, 2016).

Either DT or Linear models may not be ideal for BDM systems due to their complexity and nonlinear relationships between inputs and outputs (Hough et al., 2017; Li et al., 2015). Another interpretable ML example for such a complicated system would be physics-informed ML that incorporates physical principles into data-driven models and as a result allows for learning with less data. For example, (Ji and Deng, 2021) proposed a chemical reaction neural network, where they encode parameters in a chemical reaction that follows Arrhenius law (dependent on temperature) into nodes: ln[A], ln[B], ln[C], ln[D], −1/RT, lnT for elementary reactions involving four species of [A, B, C, D] with corresponding stoichiometric coefficients: [v_A, v_B, v_C, v_D]:

$$v_A A + v_B B \rightarrow v_C C + V_D D$$

They encoded the number of reactions as the number of hidden neurons; then, they trained the neural network with stochastic gradient descent. As physical constraints are encoded in the framework, the resulting learned weights from stochastic gradient descent are interpretable, i.e., they are the corresponding v_A, v_B, v_C, v_D and coefficients for −1/RT and lnT. Furthermore, the predictions fall within the system constraints.

For high-stake decisions, which are decisions that involve the existence of large financial and/or emotional prospective, loss outcomes, and the presence of high costs to reverse a decision once it is made (e.g., whether to purchase a flood insurance policy for one's house (Kunreuther et al., 2002), whether deploy BDM for a large-scale water treatment plant), interpretable ML is preferred (Rudin, 2019). A detailed comparison of ML models, physics-based models, and physics-informed ML models is provided in Table S12 for further demonstration.

4. Discussions

Material and process design, end-use operation optimization, and sustainability assessment problems raised in BDM studies are related to computational sustainability, an interdisciplinary research field that aims to develop computational models, methods, and tools to empower sustainable development (Gomes, 2009). Addressing those problems can potentially advance both ML and BDM communities to achieve a more sustainable society. This section discusses the main limitations of current ML applications for BDM and highlights future research directions. Current ML algorithms are not designed to solve problems in the BDM system. Almost all ML applications reviewed in this paper directly apply off-shelf ML packages and tune hyper-parameters accordingly for better predictions. Based on the review, this approach has achieved desirable accuracies in imputing data and predicting material properties and end-use performance. Nevertheless, to enable large-scale BDM deployment, customized interpretable ML may be more desirable.

Large-scale BDM deployment is related to two tasks that interpretable ML may resolve – causal inference optimization and prediction for real-world physics-constrained systems. Therefore, interpretable ML for BDM can be one potential future direction. As BDM deployment is a type of high-stake decision, it is critical to build an interpretable model that reveals the cause-effect relationships between different process/material/logistic parameters and the techno-environmental-social impacts. A holistic understanding of the supply chain can enable sustainable supply-chain-wide optimization for various BDM systems, deployment sites, and different stakeholders involved.

Many ML applications for BDM systems are trained on lab-scale experimental data, therefore the predictions and insights from these applications are more likely to be applicable to the lab-scale results. Previous studies show the potential discrepancy between lab-scale and industrial-scale data when assessing early-stage technologies (Tsoy et al., 2020; van Schoubroeck et al., 2021; Yao and Masanet, 2018). Constructing physics-informed ML models may remediate this. Unlike conventional ML, which learns everything from scratch and from patterns in the data, MLs with embedded physical principles allow the models to learn based on existing knowledge. These physics-informed models require less data, and the predictions generalize well to unseen datasets governed by physical laws (Chen et al., 2021; Eivazi and Vinuesa, 2022), which is applicable in the case of experiment BDM v.s. real-world BDM data (Karniadakis et al., 2021). In addition, the learned weights of parameters in physics-informed ML are physically meaningful. They can also resolve the prediction failures that pure-data-driven ANNs encountered, as mentioned previously by (Zhu et al., 2021, 2019b). As a result, physics-informed ML can enable broader and more practical applications of ML for BDM research and development.

Most BDM studies have focused on the application of supervised learning. However, there are many opportunities for applying other ML paradigms. For example, some studies show the economic benefits and environmental variations of using blended biomass for bio-products given the regional and seasonal variations of biomass availability and quality (Lan et al., 2021, 2020a). Previous ML applications only focused on single biomass feedstock using supervised learning. Unsupervised learning techniques can be introduced to explore the mixture of different biomass feedstocks by classifying a large variety of biomass into different groups based on their characteristics. Additionally, a combination of unsupervised-supervised learning framework has the potential to support sustainability assessment and sustainability-informed design/optimization, given the possibility of clustering similar feedstock types, processing conditions, and application scenarios. Furthermore, BDM supply chains often involve different stakeholders, such as landowners, material producers, biorefineries, and end-users. Disparate stakeholders can provide feedbacks that would benefit the efficient and sustainable design and operation of the entire BDM supply chain. ML approaches such as RL can take feedback into the learning process, supporting and enabling real-time optimization for sustainable BDM systems.

Previous studies combined ML with LCA and economic analysis for location- or spatial-specific assessment of direct and indirect environmental and economic impacts. However, no studies have explored the social implications. Social LCA is an emerging tool to assess the social impacts of individual products or corporations. One of the main challenges in applying social LCA is the difficulties in developing and obtaining sufficient data for region-specific social indicators (Macombe et al., 2013; Siebert et al., 2018). ML has been used to generate regional socioeconomic indicators; thus, it could offer a new means of addressing the data gaps in social LCA.

Based on the findings discussed above, we provide a workflow recommendation (*Fig. 5*) that includes the database knowledge discovery process (Fayyad et al., 1996). In the data exploration stage, frequently used parameters for feature engineering may be included, e.g., reaction parameters (temperature, time), feedstock properties (C, H, N, lignin wt%), BDM texture properties, incubation environment conditions, and others listed in Table S5–6. The data quality and quantity need to be assessed before building a model. Although it is difficult to determine a certain data set size due to the disparate nature of data and the complex variations of ML algorithms, a widely used rule-of-thumb is that the sample size needs to be at least a factor 10–100 times the number of the features (Alwosheel et al., 2018; Jain and

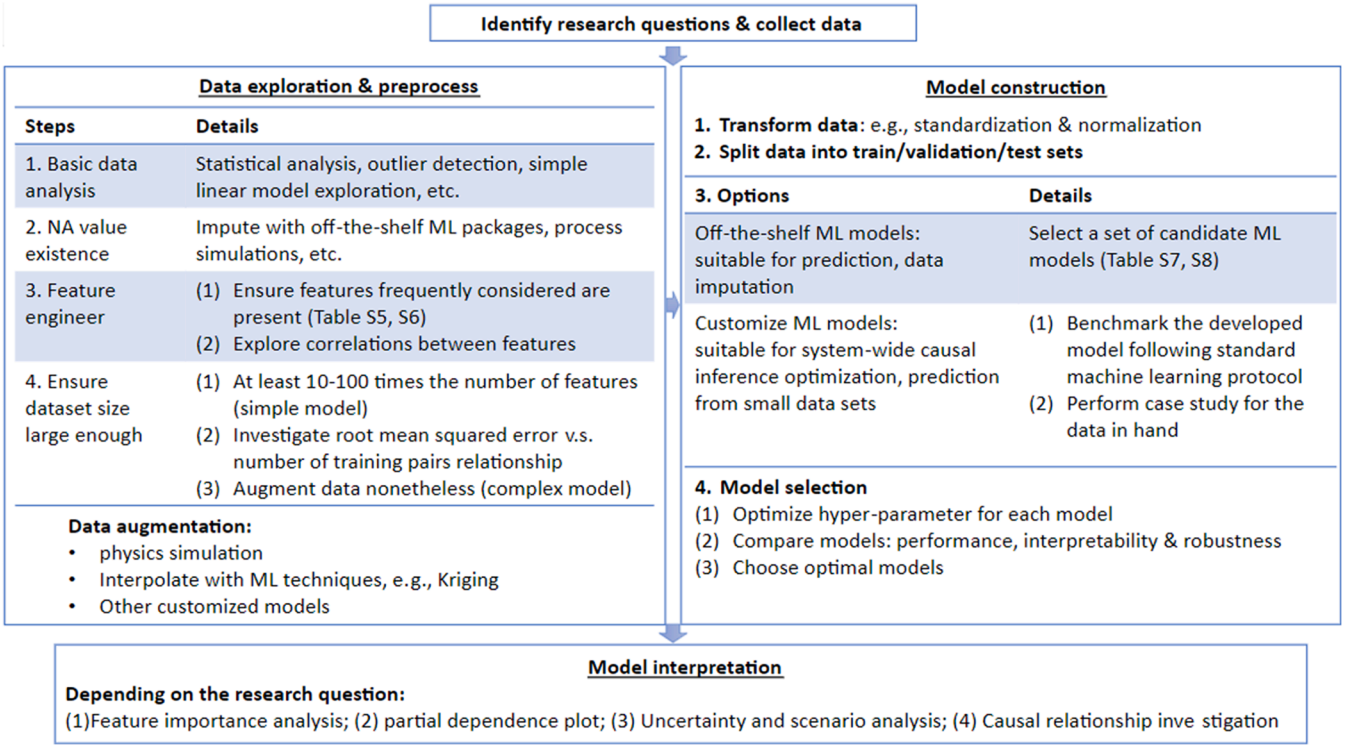


Fig. 5. Machine learning workflow recommendation.

Chandrasekaran, 1982; Raudys and Jain, 1991). Interpolation techniques such as simulation, Kriging, or other customized ML can be deployed to enlarge datasets if needed. For a complicated deep neural network model, as the number of parameters increases, more data are needed. For example, (Hough et al., 2017) demonstrated the relationship between the number of training pairs and mean squared error, showing that > 15,000 training pairs improved the neural network model.

At the model construction stage, using off-the-shelf ML packages may be sufficient for ML projects focusing on the prediction of certain outputs, and it is recommended to compare the performance of various models (simple to complex) for algorithm selection. To support system-wide decision making, it is recommended to develop customized physics-informed ML models to ensure their robustness and applicability for real-world issues. We recommend that for new algorithms, researchers should benchmark the performance following standard ML protocols, e.g., cross-validation and comparing outputs/computation time with the results from popular ML packages. For ML model selection, a candidate set of models, e.g., XGB, SVM (kernel: RBF), FCM-FFNN, RF, and others in Table S7–8, need to be optimized and compared.

Depending on the research questions, different analyses can be conducted to use and interpret ML models. Previous ML applications for BDM systems have used feature importance analysis, partial dependence plot (i.e., input-output relationship analysis), and SHAP. Uncertainty and scenario analysis have been widely used in supporting decision-making related to sustainability (van Schoubroeck et al., 2021) and bio-based material optimization and applications (Lan et al., 2022). ML models can assist in simulations and prediction of what-if scenarios for decision making. Investigating causal relationships is another capability of ML that can support not only process/material optimization but also enhance fundamental knowledge of bio-based materials.

5. Conclusion

Fifty-three papers were reviewed to understand ML applications of BDM in water and agricultural systems. We categorized the applications into three categories – M&P design, end-use performance predictions, and sustainability assessment. In M&P design, ML has been used to identify critical factors for optimizing BDM characteristics, predict BDM features, and reverse engineer; in the end-use class, ML has been mainly employed to identify essential factors that optimize BDM performances for wastewater treatment and soil amendment; in the sustainability assessment category, ML has been adopted to address the data challenge – researchers leveraged the prediction results from M&P design and end-use to generate life cycle inventory data, and further conduct LCA and estimate other economic matrices to assess the sustainability aspect of BDM in water and agricultural systems.

BDM datasets are heterogeneous tabular data with small sizes (75% of the datasets are composed of < 600 data points) and may contain considerable noise. Although the optimal model differs case by case, integrated NN and ensemble models such as RF and XGB usually perform well. One major limitation for adopting ML to assist BDM development and optimization is the limited interpretability of ensemble and NN models. Physics-informed ML can be explored in future research to incorporate mechanistic principles to improve interpretability and model predictions against physical constraints. Limited studies have focused on ML applications for BDM sustainability assessment. As an emerging computational tool, ML may support faster assessment for biomass systems that are highly dynamic at both temporal and spatial scales. More research is needed to explore practical ML applications for sustainable BDM development and optimization considering economic, environmental, social aspects, and geo-temporal dynamics.

Declaration of Competing Interest

The authors declare no competing financial interests to affect the work reported in this paper.

Data availability

This is a review paper and the data collected from literature were documented in the supporting information.

Acknowledgments

The authors thank the funding support from the U.S. National Science Foundation and Yale University. This material is based upon work supported by the National Science Foundation under Grant No. 2038439. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.resconrec.2022.106847.

References

Afolabi, I.C., Popoola, S.I., Bello, O.S., 2020. Machine learning approach for prediction of paracetamol adsorption efficiency on chemically modified orange peel. Spectrochim. Acta A Mol. Biomol. Spectrosc. 243, 118769 https://doi.org/10.1016/j.saa.2020.118769.

Alaba, P.A., Popoola, S.I., Abnisal, F., Lee, C.S., Ohunakin, O.S., Adetiba, E., Akanle, M. B., Abdul Patah, M.F., Atayero, A.A.A., Wan Daud, W.M.A., 2020. Thermal decomposition of rice husk: a comprehensive artificial intelligence predictive model. J. Therm. Anal. Calorim 140, 1811–1823. https://doi.org/10.1007/s10973-019-08915-0.

Alwosheel, A., van Cranenburgh, S., Chorus, C.G., 2018. Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis. J. Choice Modell. 28, 167–182. https://doi.org/10.1016/J.JOCM.2018.07.002.

Ben-David, S., Schuller, R., 2003. Exploiting task relatedness for multiple task learning. Lect. Notes Artif. Intell. 2777, 567–580. https://doi.org/10.1007/978-3-540-45167-9_41/COVER/ (Subseries of Lecture Notes in Computer Science).

Cao, H., Xin, Y., Yuan, Q., 2016. Prediction of biochar yield from cattle manure pyrolysis via least squares support vector machine intelligent approach. Bioresour. Technol. 202, 158–164. https://doi.org/10.1016/j.biortech.2015.12.024.

Cha, J.S., Park, S.H., Jung, S.C., Ryu, C., Jeon, J.K., Shin, M.C., Park, Y.K., 2016. Production and utilization of biochar: a review. J. Ind. Eng. Chem. 40, 1–15. https://doi.org/10.1016/j.jiec.2016.06.002.

Chen, D., Bai, Y., Ament, S., Zhao, W., Guevarra, D., Zhou, L., Selman, B., van Dover, R. B., Gregoire, J.M., Gomes, C.P., 2021. Automating crystal-structure phase mapping by combining deep learning with constraint reasoning. Natu. Mach. Intell. 3, 812–822. https://doi.org/10.1038/s42256-021-00384-1, 20219 3.

Cheng, F., Luo, H., Colosi, L.M., 2020a. Slow pyrolysis as a platform for negative emissions technology : an integration of machine learning models, life cycle assessment, and economic analysis. Energy Convers. Manag. 223, 113258 https://doi.org/10.1016/j.enconman.2020.113258.

Cheng, F., Porter, M.D., Colosi, L.M., 2020b. Is hydrothermal treatment coupled with carbon capture and storage an energy-producing negative emissions technology? Energy Convers. Manag. 203, 112252 https://doi.org/10.1016/j.enconman.2019.112252.

Cipullo, S., Snapir, B., Prpich, G., Campo, P., Coulon, F., 2019. Prediction of bioavailability and toxicity of complex chemical mixtures through machine learning models. Chemosphere 215, 388–395. https://doi.org/10.1016/j.chemosphere.2018.10.056.

Commission, E., Innovation, D.-.G. for R. and, 2012. Innovating for sustainable growth: a bioeconomy for Europe. Publications Office. https://doi.org/10.2777/6462.

Costa, D., Quinteiro, P., Dias, A.C., 2019. A systematic review of life cycle sustainability assessment: current state, methodological challenges, and implementation issues. Sci. Total Environ. 686, 774–787. https://doi.org/10.1016/J.SCITOTENV.2019.05.435.

de Meyer, A., Cattrysse, D., Rasinmäki, J., van Orshoven, J., 2014. Methods to optimise the design and management of biomass-for-bioenergy supply chains: a review. Renew. Sustain. Energy Rev. 31, 657–670. https://doi.org/10.1016/J.RSER.2013.12.036.

de Miranda Ramos Soares, A.P., de Oliveira Carvalho, F., de Farias Silva, C.E., da Silva Gonçalves, A.H., de Souza Abud, A.K., 2020. Random forest as a promising application to predict basic-dye biosorption process using orange waste. J. Environ. Chem. Eng. 8, 103952 https://doi.org/10.1016/j.jece.2020.103952.

Ding, F., Zwieten, L.van, Zhang, W., Weng, Z.H., Shi, S., Wang, J., 2018. A meta-analysis and critical evaluation of influencing factors on soil carbon priming following biochar amendment 1507–1517.

Dokoohaki, H., Miguez, F.E., Laird, D., Dumortier, J., 2019. Where should we apply biochar ?.

Dolatabadi, M., Mehrabpour, M., Esfandyari, M., Alidadi, H., 2018. Modeling of simultaneous adsorption of dye and metal ion by sawdust from aqueous solution using of ANN and ANFIS. Chemom. Intell. Lab. Syst. 181, 72–78. https://doi.org/10.1016/j.chemolab.2018.07.012.

Donti, P.L., Kolter, J.Z., 2021. Machine learning for sustainable energy systems. 10.1146/annurev-environ-020220-061831 46, 719–747. 10.1146/ANNUREV-ENVIRON-020220-061831.

Dumortier, J., Dokoohaki, H., Elobeid, A., Hayes, D.J., Laird, D., Miguez, F.E., 2020. Global land-use and carbon emission implications from biochar application to cropland in the United States 258. https://doi.org/10.1016/j.jclepro.2020.120684.

Eichelsdörfer, J., Kaltenbach, S., Koutsourelakis, P.-.S., 2021. Physics-enhanced neural networks in the small data regime. In: Workshop at the 35th Conference on Neural Information Processing Systems.

Eivazi, H., Vinuesa, R., 2022. Physics-informed deep-learning applications to experimental fluid mechanics. https://doi.org/10.48550/arxiv.2203.15402.

el Hanandeh, A., Mahdi, Z., Imtiaz, M.S., 2021. Modelling of the adsorption of Pb, Cu and Ni ions from single and multi-component aqueous solutions by date seed derived biochar: comparison of six machine learning approaches. Environ. Res. 192, 110338 https://doi.org/10.1016/j.envres.2020.110338.

El-Chichakli, B., von Braun, J., Lang, C., Barben, D., Philp, J., 2016. Policy: five cornerstones of a global bioeconomy. Nature 535, 221–223. https://doi.org/10.1038/535221a, 20167611 535.

Ewees, A.A., Elaziz, M.A., 2018. Improved adaptive neuro-fuzzy inference system using gray wolf optimization: a case study in predicting biochar yield. J. Intellig. Syst. 29, 924–940. https://doi.org/10.1515/jisys-2017-0641.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., 1996. From data mining to knowledge discovery in databases. AI Mag. 17 https://doi.org/10.1609/AIMAG.V17I3.1230, 37–37.

Friedman, J., Hastie, T., Tibshirani, R., 2001. The Elements of Statistical Learning. Springer series in statistics, New York.

Gomes, C., 2009. Computational sustainability: computational methods for a sustainable environment, economy, and society. Bridge, Natl. Acad. Eng. 39, 5–13.

Han, T., Srinivas, S., Lakkaraju, H., 2022. Which explanation should I choose? A function approximation perspective to characterizing post hoc explanations. doi:10.48550/arXiv.2206.01254.

Heuvelink, G.B.M., Angelini, M.E., Poggio, L., Bai, Z., Batjes, N.H., van den Bosch, R., Bossio, D., Estella, S., Lehmann, J., Olmedo, G.F., Sanderman, J., 2021. Machine learning in space and time for modelling soil organic carbon change. Eur. J. Soil Sci. 72, 1607–1623. https://doi.org/10.1111/EJSS.12998.

Hong, T., Wang, Z., Luo, X., Zhang, W., 2020. State-of-the-art on research and applications of machine learning in the building life cycle. Energy Build 212, 109831. https://doi.org/10.1016/J.ENBUILD.2020.109831.

Hough, B.R., Beck, D.A.C., Schwartz, D.T., Pfaendtner, J., 2017. Application of machine learning to pyrolysis reaction networks: reducing model solution time to enable process optimization. Comput. Chem. Eng. 104, 56–63. https://doi.org/10.1016/J.COMPCHEMENG.2017.04.012.

Huang, R., Ma, C., Ma, J., Huangfu, X., He, Q., 2021. Machine learning in natural and engineered water systems. Water Res. 205, 117666 https://doi.org/10.1016/J.WATRES.2021.117666.

Inayat, A., Ahmed, A., Tariq, R., Waris, A., Jamil, F., Ahmed, S.F., Ghenai, C., Park, Y.K., 2022. Techno-economical evaluation of bio-oil production via biomass fast pyrolysis process: a review. Front. Energy Res. 9, 993. https://doi.org/10.3389/FENRG.2021.770355/XML/NLM.

Ismail, H.Y., Shirazian, S., Skoretska, I., Mynko, O., Ghanim, B., Leahy, J.J., Walker, G. M., Kwapinski, W., 2019. ANN-Kriging hybrid model for predicting carbon and inorganic phosphorus recovery in hydrothermal carbonization. Waste Manag. 85, 242–252. https://doi.org/10.1016/j.wasman.2018.12.044.

Jain, A.K., Chandrasekaran, B., 1982. 39 Dimensionality and sample size considerations in pattern recognition practice. Handbook of Statistics 2, 835–855. https://doi.org/10.1016/S0169-7161(82)02042-2.

Jalalifar, S., Masoudi, M., Abbassi, R., Garaniya, V., Ghiji, M., Salehi, F., 2020. A hybrid SVR-PSO model to predict a CFD-based optimised bubbling fluidised bed pyrolysis reactor. Energy 191, 116414. https://doi.org/10.1016/j.energy.2019.116414.

Ji, W., Deng, S., 2021. Autonomous discovery of unknown reaction pathways from data by chemical reaction neural network. J. Phys. Chem. A 125, 1082–1092. https://doi.org/10.1021/ACS.JPCA.0C09316/ASSET/IMAGES/LARGE/JP0C09316_0014.JPEG.

Jiang, W., Xing, X., Li, S., Zhang, X., Wang, W., 2019a. Synthesis, characterization and machine learning based performance prediction of straw activated carbon. J Clean Prod 212, 1210–1223. https://doi.org/10.1016/J.JCLEPRO.2018.12.093.

Jiang, W., Xing, X., Zhang, X., Mi, M., 2019b. Prediction of combustion activation energy of NaOH/KOH catalyzed straw pyrolytic carbon based on machine learning. Renew Energy 130, 1216–1225. https://doi.org/10.1016/j.renene.2018.08.089.

Jordan, M.I., Mitchell, T.M., 2015. Machine learning: trends, perspectives, and prospects. Science 349 (1979), 255–260. https://doi.org/10.1126/SCIENCE.AAA8415/ASSET/AB2EF18A-576D-464D-B1B6-1301159EE29A/ASSETS/GRAPHIC/349_255_F5.JPEG.

Karniadakis, G.E., Kevrekidis, I.G., Lu, L., Perdikaris, P., Wang, S., Yang, L., 2021. Physics-informed machine learning. Nat. Rev. Phys. 3, 422–440. https://doi.org/10.1038/s42254-021-00314-5, 20216 3.

Karri, R.R., Sahu, J.N., 2018. Modeling and optimization by particle swarm embedded neural network for adsorption of zinc (II) by palm kernel shell based activated carbon from aqueous environment. J. Environ. Manage. 206, 178–191. https://doi.org/10.1016/j.jenvman.2017.10.026.

Kauffman, N., Dumortier, J., Hayes, D.J., Brown, R.C., Laird, D.A., 2014. Producing energy while sequestering carbon? The relationship between biochar and agricultural productivity. Biomass Bioenergy 63, 167–176. https://doi.org/10.1016/J.BIOMBIOE.2014.01.049.

Ke, B., Nguyen, H., Bui, X., Bui, H., Choi, Y., 2021a. Predicting the sorption efficiency of heavy metal based on the biochar characteristics, metal sources, and environmental conditions using various novel hybrid machine learning models. Chemosphere 276, 130204. https://doi.org/10.1016/j.chemosphere.2021.130204.

Ke, B., Nguyen, H., Bui, X.N., Bui, H.B., Nguyen-Thoi, T., 2021b. Prediction of the sorption efficiency of heavy metal onto biochar using a robust combination of fuzzy

C-means clustering and back-propagation neural network. J. Environ. Manage. 293 https://doi.org/10.1016/j.jenvman.2021.112808.
Kunreuther, H., Meyer, R., Zeckhauser, R., Slovic, P., Schwartz, B., Schade, C., Luce, M. F., Lippman, S., Krantz, D., Kahn, B., Hogarth, R., 2002. High stakes decision making: normative, descriptive and prescriptive considerations. Mark. Lett. 13, 259–268. https://doi.org/10.1023/A:1020287225409, 20023 13.
Lan, K., Kelley, S.S., Nepal, P., Yao, Y., 2020. Dynamic life cycle carbon and energy analysis for cross-laminated timber in the Southeastern United States. Environmental Research Letters 15, 124036. https://doi.org/10.1088/1748-9326/abc5e6.
Lan, K., Ou, L., Park, S., Kelley, S.S., English, B.C., Yu, T.E., Larson, J., Yao, Y., 2021. Techno-Economic Analysis of decentralized preprocessing systems for fast pyrolysis biorefineries with blended feedstocks in the southeastern United States. Renew. Sustain. Energy Rev. 143, 110881 https://doi.org/10.1016/J.RSER.2021.110881.
Lan, K., Ou, L., Park, S., Kelley, S.S., Yao, Y., 2020a. Life cycle analysis of decentralized preprocessing systems for fast pyrolysis biorefineries with blended feedstocks in the southeastern United States. Energy Technol. 8, 1900850 https://doi.org/10.1002/ENTE.201900850.
Lan, K., Park, S., Yao, Y., 2020b. Key issue, challenges, and status quo of models for biofuel supply chain design. Biofuels for a More Sustainable Future: Life Cycle Sustainability Assessment and Multi-Criteria Decision Making 273–315. https://doi.org/10.1016/B978-0-12-815581-3.00010-5.
Lan, K., Zhang, B., Yao, Y., 2022. Circular utilization of urban tree waste contributes to the mitigation of climate change and eutrophication. One Earth 5, 944–957. https://doi.org/10.1016/J.ONEEAR.2022.07.001.
Lehmann, J., Cowie, A., Masiello, C.A., Kammann, C., Woolf, D., Amonette, J.E., Cayuela, M.L., Camps-Arbestain, M., Whitman, T., 2021. Biochar in climate change mitigation. Nat. Geosci. 14, 883–892. https://doi.org/10.1038/s41561-021-00852-8, 202112 14.
Li, J., Pan, L., Suvarna, M., Tong, Y.W., Wang, X., 2020. Fuel properties of hydrochar and pyrochar: prediction and exploration with machine learning. Appl. Energy 269, 115166. https://doi.org/10.1016/j.apenergy.2020.115166.
Li, L., Flora, J.R.V., Caicedo, J.M., Berge, N.D., 2015. Investigating the role of feedstock properties and process conditions on products formed during the hydrothermal carbonization of organics using regression techniques. Bioresour. Technol. 187, 263–274. https://doi.org/10.1016/J.BIORTECH.2015.03.054.
Li, L., Rong, S., Wang, R., Yu, S., 2021. Recent advances in artificial intelligence and machine learning for nonlinear relationship analysis and process control in drinking water treatment: a review. Chem. Eng. J. 405, 126673 https://doi.org/10.1016/J.CEJ.2020.126673.
Li, M., Wei, D., Liu, T., Liu, Y., Yan, L., Wei, Q., Du, B., Xu, W., 2019. EDTA functionalized magnetic biochar for Pb(II) removal: adsorption performance, mechanism and SVM model prediction. Sep. Purif. Technol. 227, 115696 https://doi.org/10.1016/j.seppur.2019.115696.
Liakos, K.G., Busato, P., Moshou, D., Pearson, S., Bochtis, D., 2018. Machine learning in Agriculture: A review. Sensors 18, 2674. doi:10.3390/s18082674.
Liao, M., Kelley, S., Yao, Y., 2020. Generating energy and greenhouse gas inventory data of activated carbon production using machine learning and kinetic based process simulation. ACS Sustain. Chem. Eng. 8, 1252–1261. https://doi.org/10.1021/acssuschemeng.9b06522.
Liao, M., Lan, K., Yao, Y., 2022. Sustainability implications of artificial intelligence in the chemical industry: a conceptual framework. J. Ind. Ecol. 26, 164–182. https://doi.org/10.1111/JIEC.13214.
Liao, M., Yao, Y., 2021. Applications of artificial intelligence-based modeling for bioenergy systems: a review. GCB Bioenergy 13, 774–802. https://doi.org/10.1111/gcbb.12816.
Lipton, Z.C., 2016. The mythos of model interpretability. Commun. ACM 61, 35–43. https://doi.org/10.48550/arxiv.1606.03490.
Liu, Q., Liu, B., Zhang, Y., Hu, T., Lin, Z., Liu, G., Wang, X., Ma, J., Wang, H., Jin, H., Ambus, P., Amonette, J.E., Xie, Z., 2019. Biochar application as a tool to decrease soil nitrogen losses (NH 3 volatilization, N 2 O emissions, and N leaching) from croplands: options and mitigation strength in a global perspective. Glob. Chang. Biol. 25, 2077–2093. https://doi.org/10.1111/gcb.14613.
Liu, Ziyun, Wang, Z., Chen, H., Cai, T., Liu, Zhidan, 2021. Hydrochar and pyrochar for sorption of pollutants in wastewater and exhaust gas: a critical review. Environ. Pollut. 268, 115910 https://doi.org/10.1016/j.envpol.2020.115910.
Lundberg, S.M., Allen, P.G., Lee, S.-I., 2017. A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. 30.
Macombe, C., Leskinen, P., Feschet, P., Antikainen, R., 2013. Social life cycle assessment of biodiesel production at three levels: a literature review and development needs. J. Clean. Prod. 52, 205–216. https://doi.org/10.1016/J.JCLEPRO.2013.03.026.
Marcinkevičs, R., Vogt, J.E., 2020. Interpretability and explainability: a machine learning zoo mini-tour. https://doi.org/10.48550/arxiv.2012.01805.
Mathew, S., Karandikar, P.B., Kulkarni, N.R., 2020. Modeling and optimization of a jackfruit seed-based supercapacitor electrode using machine learning. Chem. Eng. Technol. 43, 1765–1773. https://doi.org/10.1002/ceat.201900616.
Maulana Kusdhany, M.I., Lyth, S.M., 2021. New insights into hydrogen uptake on porous carbon materials via explainable machine learning. Carbon N Y 179, 190–201. https://doi.org/10.1016/j.carbon.2021.04.036.
Mazaheri, H., Ghaedi, M., Ahmadi Azqhandi, M.H., Asfaram, A., 2017. Application of machine/statistical learning, artificial intelligence and statistical experimental design for the modeling and optimization of methylene blue and Cd(ii) removal from a binary aqueous solution by natural walnut carbon. Phys. Chem. Chem. Phys. 19, 11299–11317. https://doi.org/10.1039/c6cp08437k.
Mendoza-Castillo, D.I., Reynel-Ávila, H.E., Sánchez-Ruiz, F.J., Trejo-Valencia, R., Jaime-Leal, J.E., Bonilla-Petriciolet, A., 2018. Insights and pitfalls of artificial neural network modeling of competitive multi-metallic adsorption data. J. Mol. Liq. 251, 15–27. https://doi.org/10.1016/j.molliq.2017.12.030.
Mitchell, T.M., 1997. Machine Learning, 1st ed. McGraw-Hill, Inc., USA.
Mohan, D., Sarswat, A., Ok, Y.S., Pittman, C.U., 2014. Organic and inorganic contaminants removal from water with biochar, a renewable, low cost and sustainable adsorbent – A critical review. Bioresour Technol 160, 191–202. https://doi.org/10.1016/J.BIORTECH.2014.01.120.
Mojiri, A., Kazeroon, R.A., Gholami, A., 2019. Cross-linked magnetic chitosan/activated biochar for removal of emerging micropollutants from water: optimization by the artificial neural network. Water (Switzerland) 11, 1–18. https://doi.org/10.3390/w11030551.
Mojiri, A., Ohashi, A., Ozaki, N., Aoi, Y., Kindaichi, T., 2020. Integrated anammox-biochar in synthetic wastewater treatment: performance and optimization by artificial neural network. J. Clean. Prod. 243, 118638 https://doi.org/10.1016/j.jclepro.2019.118638.
Murphy, K.P., 2022. Probabilistic Machine Learning: An Introduction. MIT Press.
Nguyen, X.C., Ly, Q.V., Peng, W., Nguyen, V.H., Nguyen, D.D., Tran, Q.B., Huyen Nguyen, T.T., Sonne, C., Lam, S.S., Ngo, H.H., Goethals, P., Le, Q.van, 2021. Vertical flow constructed wetlands using expanded clay and biochar for wastewater remediation: a comparative study and prediction of effluents using machine learning. J. Hazard. Mater. 413 https://doi.org/10.1016/j.jhazmat.2021.125426.
Olson, M., Wyner, A.J., Berk, R., 2018. Modern neural networks generalize on small data sets. Adv. Neural Inf. Process Syst. 3619–3628, 2018-Decem.
Onat, N.C., Kucukvar, M., Halog, A., Cloutier, S., 2017. Systems thinking for life cycle sustainability assessment: a review of recent developments, applications, and future perspectives. Sustainability 9, 706. https://doi.org/10.3390/SU9050706, 20179, 706.
Osman, A.I., Elgarahy, A.M., Mehta, N., Al-Muhtaseb, A.H., Al-Fatesh, A.S., Rooney, D. W., 2022. Facile synthesis and life cycle assessment of highly active magnetic sorbent composite derived from mixed plastic and biomass waste for water remediation. ACS Sustain. Chem. Eng. 10, 12433–12447. https://doi.org/10.1021/ACSSUSCHEMENG.2C04095/ASSET/IMAGES/MEDIUM/SC2C04095_M015.GIF.
Palansooriya, K.N., Li, J., Dissanayake, P.D., Suvarna, M., Li, L., Yuan, X., Sarkar, B., Tsang, D.C.W., Rinklebe, J., Wang, X., Ok, Y.S., 2022. Prediction of soil heavy metal immobilization by biochar using machine learning. Environ. Sci. Technol. 56, 4187–4198. https://doi.org/10.1021/acs.est.1c08302.
Parveen, N., Zaidi, S., Danish, M., 2017. Development of SVR-based model and comparative analysis with MLR and ANN models for predicting the sorption capacity of Cr(VI). Process Saf. Environ. Protect. 107, 428–437. https://doi.org/10.1016/j.psep.2017.03.007.
Pathy, A., Meher, S., P, B., 2020. Predicting algal biochar yield using eXtreme gradient boosting (XGB) algorithm of machine learning methods. Algal Res. 50, 102006 https://doi.org/10.1016/j.algal.2020.102006.
Pearl, J., 2022. Causal Diagrams for Empirical Research (With Discussions), in: Probabilistic and Causal Inference. ACM, New York, NY, USA, pp. 255–316. https://doi.org/10.1145/3501714.3501734.
Prakash, N., Manikandan, S.A., Govindarajan, L., Vijayagopal, V., 2008. Prediction of biosorption efficiency for the removal of copper (II) using artificial neural networks 152, 1268–1275. https://doi.org/10.1016/j.jhazmat.2007.08.015.
Pignatello, J.J., Uchimiya, M., Abiven, S., Schmidt, M.W.I., 2015. Evolution of biochar properties in soil. In: Biochar for Environmental Management—Science, Technology and Implementation, pp. 195–233.
Raudys, S.J., Jain, A.K., 1991. Small sample size effects in statistical pattern recognition: recommendations for practitioners. IEEE Trans. Pattern Anal. Mach. Intell. 13, 252–264. https://doi.org/10.1109/34.75512.
Razzaghi, F., Obour, P.B., Arthur, E., 2020. Does biochar improve soil water retention? A systematic review and meta-analysis. Geoderma 361, 114055. https://doi.org/10.1016/J.GEODERMA.2019.114055.
Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intellig. 206–215. https://doi.org/10.1038/s42256-019-0048-x, 2019 1:5 1.
Sagi, O., Rokach, L., 2018. Ensemble learning: a survey. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 8, e1249. https://doi.org/10.1002/WIDM.1249.
Sala, S., Ciuffo, B., Nijkamp, P., 2015. A systemic framework for sustainability assessment. Ecol. Econ. 119, 314–325. https://doi.org/10.1016/J.ECOLECON.2015.09.015.
Selvarajoo, A., Muhammad, D., Arumugasamy, S.K., 2020. An experimental and modelling approach to produce biochar from banana peels through pyrolysis as potential renewable energy resources. Model Earth Syst. Environ. 6, 115–128. https://doi.org/10.1007/s40808-019-00663-2.
Shen, X., Foster, T., Baldi, H., Dobreva, I., Burson, B., Hays, D., Tabien, R., Jessup, R., 2019. Quantification of soil organic carbon in biochar-amended soil using ground penetrating radar (GPR). Remote Sens. (Basel) 11, 1–12. https://doi.org/10.3390/rs11232874.
Shwartz-Ziv, R., Armon, A., 2022. Tabular data: deep learning is not all you need. Inf. Fusion 81, 84–90. https://doi.org/10.1016/J.INFFUS.2021.11.011.
Siebert, A., Bezama, A., O'Keeffe, S., Thrän, D., 2018. Social life cycle assessment: in pursuit of a framework for assessing wood-based products from bioeconomy regions in Germany. Int. J. Life Cycle Assess. 23, 651–662. https://doi.org/10.1007/S11367-016-1066-0/FIGURES/5.
Sigmund, G., Gharasoo, M., Hüffer, T., Hofmann, T., 2020. Deep learning neural network approach for predicting the sorption of ionizable and polar organic pollutants to a wide range of carbonaceous materials. Environ. Sci. Technol. 54, 4583–4591. https://doi.org/10.1021/acs.est.9b06287.
Smebye, A.B., Sparrevik, M., Schmidt, H.P., Cornelissen, G., 2017. Life-cycle assessment of biochar production systems in tropical rural areas: comparing flame curtain kilns

to other production methods. Biomass Bioenergy 101, 35–43. https://doi.org/10.1016/J.BIOMBIOE.2017.04.001.
Sothe, C., Gonsamo, A., Arabian, J., Snider, J., 2022. Large scale mapping of soil organic carbon concentration with 3D machine learning and satellite observations. Geoderma 405, 115402. https://doi.org/10.1016/J.GEODERMA.2021.115402.
Stegmann, P., Londo, M., Junginger, M., 2020. The circular bioeconomy: its elements and role in European bioeconomy clusters. Resour., Conserv. Recycl.: X 6, 100029. https://doi.org/10.1016/J.RCRX.2019.100029.
Su, H., Lin, S., Deng, S., Lian, C., Shang, Y., Liu, H., 2019. Predicting the capacitance of carbon-based electric double layer capacitors by machine learning. Nanoscale Adv. 1, 2162–2166. https://doi.org/10.1039/C9NA00105K.
Suliman, W., Harsh, J.B., Fortuna, A.M., Garcia-Pérez, M., Abu-Lail, N.I., 2017. Quantitative effects of biochar oxidation and pyrolysis temperature on the transport of pathogenic and nonpathogenic escherichia coli in biochar-amended sand columns. Environ. Sci. Technol. 51, 5071–5081. https://doi.org/10.1021/ACS.EST.6B04535/ASSET/IMAGES/LARGE/ES-2016-04535P_0006.JPEG.
Sun, Y., Gao, B., Yao, Y., Fang, J., Zhang, M., Zhou, Y., Chen, H., Yang, L., 2014. Effects of feedstock type, production method, and pyrolysis temperature on biochar and hydrochar properties. Chem. Eng. J. 240, 574–578. https://doi.org/10.1016/J.CEJ.2013.10.081.
Sundui, B., Ramirez Calderon, O.A., Abdeldayem, O.M., Lázaro-Gil, J., Rene, E.R., Sambuu, U., 2021. Applications of machine learning algorithms for biological wastewater treatment: updates and perspectives. Clean. Technol. Environ. Policy 23, 127–143. https://doi.org/10.1007/S10098-020-01993-X/TABLES/4.
Talebkeikhah, F., Rasam, S., Talebkeikhah, M., Torkashvand, M., Salimi, A., Moraveji, M. K., 2020. Investigation of effective processes parameters on lead (II) adsorption from wastewater by biochar in mild air oxidation pyrolysis process. Int. J. Environ. Anal. Chem. https://doi.org/10.1080/03067319.2020.1777291.
Thiruvengadam, S., Edmund Murphy, M., Tan, J.S., 2021. Mathematically modelling pyrolytic polygeneration processes using artificial intelligence. Fuel 295, 120488. https://doi.org/10.1016/j.fuel.2021.120488.
Tisserant, A., Morales, M., Cavalett, O., O'Toole, A., Weldon, S., Rasse, D.P., Cherubini, F., 2022. Life-cycle assessment to unravel co-benefits and trade-offs of large-scale biochar deployment in Norwegian agriculture. Resour. Conserv. Recycl. 179, 106030 https://doi.org/10.1016/J.RESCONREC.2021.106030.
Tsoy, N., Steubing, B., van der Giesen, C., Guinée, J., 2020. Upscaling methods used in ex ante life cycle assessment of emerging technologies: a review. Int. J. Life Cycle Assess. 25, 1680–1692. https://doi.org/10.1007/S11367-020-01796-8/FIGURES/3.
van Schoubroeck, S., Thomassen, G., van Passel, S., Malina, R., Springael, J., Lizin, S., Venditti, R.A., Yao, Y., van Dael, M., 2021. An integrated techno-sustainability assessment (TSA) framework for emerging technologies. Green Chem. 23, 1700–1715. https://doi.org/10.1039/D1GC00036E.
Varma, R.S., 2019. Biomass-derived renewable carbonaceous materials for sustainable chemical and environmental applications. ACS Sustain. Chem. Eng. 7, 6458–6470. https://doi.org/10.1021/ACSSUSCHEMENG.8B06550/ASSET/IMAGES/LARGE/SC-2018-06550Y_0014.JPEG.
Veres, M., Moussa, M., 2020. Deep Learning for intelligent transportation systems: a survey of emerging trends. IEEE Trans. Intellig. Trans. Syst. 21, 3152–3168. https://doi.org/10.1109/TITS.2019.2929020.
Vijay, V., Shreedhar, S., Adlak, K., Payyanad, S., Sreedharan, V., Gopi, G., Sophia van der Voort, T., Malarvizhi, P., Yi, S., Gebert, J., Aravind, P.v., 2021. Review of large-scale biochar field-trials for soil amendment and the observed influences on crop yield variations. Front. Energy Res. 9, 499. https://doi.org/10.3389/FENRG.2021.710766/BIBTEX.
Wehrle, R., Welp, G., Pätzold, S., 2021. Total and hot-water extractable organic carbon and nitrogen in organic soil amendments: their prediction using portable mid-infrared spectroscopy with support vector machines. Agronomy 11, 659. https://doi.org/10.3390/agronomy11040659.
Wolpert, D.H., Macready, W.G., 1997. No free lunch theorems for optimization. IEEE Trans. Evol. Comput. 1, 67–82. https://doi.org/10.1109/4235.585893.
Woolf, D., Lehmann, J., Ogle, S., Kishimoto-Mo, A.W., McConkey, B., Baldock, J., 2021. Greenhouse gas inventory model for biochar additions to soil. Environ. Sci. Technol. 55, 14795–14805. https://doi.org/10.1021/ACS.EST.1C02425/ASSET/IMAGES/MEDIUM/ES1C02425_M006.GIF.
Wu, N., Lan, K., Yao, Y., 2023. An integrated techno-economic and environmental assessment for carbon capture in hydrogen production by biomass gasification. Resour. Conserv. Recycl. 188, 106693 https://doi.org/10.1016/J.RESCONREC.2022.106693.
Yang, D.P., Li, Z., Liu, M., Zhang, X., Chen, Y., Xue, H., Ye, E., Luque, R., 2019a. Biomass-derived carbonaceous materials: recent progress in synthetic approaches, advantages, and applications. ACS Sustain. Chem. Eng. 7, 4564–4585. https://doi.org/10.1021/ACSSUSCHEMENG.8B06030/ASSET/IMAGES/LARGE/SC-2018-06030K_0014.JPEG.
Yang, X., Wan, Y., Zheng, Y., He, F., Yu, Z., Huang, J., Wang, H., Ok, Y.S., Jiang, Y., Gao, B., 2019b. Surface functional groups of carbon-based adsorbents and their roles in the removal of heavy metals from aqueous solutions: a critical review. Chem. Eng. J. 366, 608–621. https://doi.org/10.1016/J.CEJ.2019.02.119.
Yao, Y., Huang, R., 2019. A parametric life cycle modeling framework for identifying research development priorities of emerging technologies: a case study of additive manufacturing. Procedia CIRP 80, 370–375. https://doi.org/10.1016/J.PROCIR.2019.01.037.
Yao, Y., Masanet, E., 2018. Life-cycle modeling framework for generating energy and greenhouse gas emissions inventory of emerging technologies in the chemical industry. J. Clean. Prod. 172, 768–777. https://doi.org/10.1016/J.JCLEPRO.2017.10.125.
Ye, L., Camps-Arbestain, M., Shen, Q., Lehmann, J., Singh, B., Sabir, M., 2020. Biochar effects on crop yields with and without fertilizer: a meta-analysis of field studies using separate controls. Soil Use Manag. 36, 2–18. https://doi.org/10.1111/SUM.12546.
Yu, H., Zou, W., Chen, J., Chen, H., Yu, Z., Huang, J., Tang, H., Wei, X., Gao, B., 2019. Biochar amendment improves crop production in problem soils: a review. J. Environ. Manage. 232, 8–21. https://doi.org/10.1016/J.JENVMAN.2018.10.117.
Zhang, K., Zhong, S., Zhang, H., 2020. Predicting aqueous adsorption of organic compounds onto biochars, carbon nanotubes, granular activated carbons, and resins with machine learning. Environ. Sci. Technol. 54, 7008–7018. https://doi.org/10.1021/acs.est.0c02526.
Zhang, Z., Schott, J.A., Liu, M., Chen, H., Lu, X., Sumpter, B.G., Fu, J., Dai, S., 2019. Prediction of carbon dioxide adsorption via deep learning. Angew. Chem. Int. Ed. 58, 259–263. https://doi.org/10.1002/anie.201812363.
Zhao, Y., Li, Y., Fan, D., Song, J., Yang, F., 2021. Application of kernel extreme learning machine and Kriging model in prediction of heavy metals removal by biochar. Bioresour. Technol. 329, 124876 https://doi.org/10.1016/j.biortech.2021.124876.
Zhou, M., Gallegos, A., Liu, K., Dai, S., Wu, J., 2020. Insights from machine learning of carbon electrodes for electric double layer capacitors. Carbon N Y 157, 147–152. https://doi.org/10.1016/j.carbon.2019.08.090.
Zhou, Y., Zhao, X., Guo, X., Li, Y., 2022. Mapping of soil organic carbon using machine learning models: combination of optical and radar remote sensing data. Soil Sci. Soc. Am. J. 86, 293–310. https://doi.org/10.1002/SAJ2.20371.
Zhu, X., Li, Y., Wang, X., 2019a. Machine learning prediction of biochar yield and carbon contents in biochar based on biomass characteristics and pyrolysis conditions. Bioresour. Technol. 288, 121527 https://doi.org/10.1016/J.BIORTECH.2019.121527.
Zhu, X., Tsang, D.C.W., Wang, L., Su, Z., Hou, D., Li, L., Shang, J., 2020. Machine learning exploration of the critical factors for CO2 adsorption capacity on porous carbon materials at different pressures. J. Clean. Prod. 273, 122915 https://doi.org/10.1016/j.jclepro.2020.122915.
Zhu, X., Wan, Z., Tsang, D.C.W., He, M., Hou, D., Su, Z., Shang, J., 2021. 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, Y.S., Sik, Y., 2019b. 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.

Machine Learning for Sustainable Development and Applications of Biomass and Biomass-Derived Materials in Water and Agricultural Systems: A Review

Supporting Information

Hannah Szu-Han Wang[1], Yuan Yao[1,*]

[1]Center for Industrial Ecology, Yale School of the Environment, Yale University, 380 Edwards Street, New Haven, Connecticut, 06511, United States

*Email: y.yao@yale.edu

Number of pages: 60

Number of tables: 12

Table of Contents

Table S1 is a summary for properties of common ML algorithms in BDM, including basic introduction, strengths, weaknesses, and some remediations to address the weaknesses.

Table S1 Common ML algorithms in BDM: basics, strengths, and weaknesses (Friedman et al., 2001; Murphy, 2022)

Type	Algorithm	Basics	Strengths	Weaknesses	*Remediation
Linear	Linear regression (LR)	The expected value of the output y $\in \mathbb{R}$ is assumed to be a linear function of input $x \in \mathbb{R}^D$: $E[y\|x] = w^T x$, where w is the parameter that will be learned	(a) Highly interpretable (b) Easy to fit data	Lower generalizability on complex non-linear dataset: generalized linear models (GLM) make the strong assumption that input-output mapping is linear	Increase model flexibility: perform feature transformation by replacing x with $\phi(x)$. E.g., polynomial transform: $\phi(x) = [1, x, x^2, \dots]$
TFBB	Decision trees (DT)	Consists of a set of nested decision rules (nodes). At each node i, the feature dimension d_i of x is compared to a threshold t_i. If the $input > t_i$, it passes down to the left branch; to the right otherwise. At the leaves of the tree are the predicted output.	(a) Highly interpretable (b) Fast to fit (c) relatively robust to outliers (d) automatic variable selection (e) insensitive to input transformation → can handle various data types & no need to standardize the data (f) handle missing data	(a) lower prediction accuracy (because of the greedy nature) (b) unstable and predictions highly vary if the training data is perturbed: small changes to the input data can have large effects on DT (because the change at the top of the tree will affect the rest)	Reduce variance: ensemble learning (e.g., bagging: re-running the same learning algorithm on different subsets of the data to result in sufficiently diverse base models.)
	Random Forest (RF)	An ensemble of DT that leverages bagging and bootstrap samples. At each decision node i, RF tries to	(a) Offer mechanisms for assessing the importance of an input variable	(a) Limit performance for low-dimensional data: because of the	Prune the tree depth or reduce the number of selected features

		decorrelate the base model learners further by learning tress based on a randomly chosen subset of input variables and a randomly chosen subset of data cases.	(b) Offer proximity measure to measure the similarity of two samples and detect outliers (c) inherit advantages (b)~(f) from DT (d) reduces prediction variance	reduced randomization effect (b) tradeoff between the computation complexity and number of trees: prediction can be slow for large forests	
	Boosting	$F_m :=$ the mth tree or any general function approximator (e.g., NN), and $\beta_m :=$ the corresponding weight. Boosting sequentially fits the additive model $f = \sum_{m=1}^{M} \beta_m F_m$. First, fit F_1 on the original data; then weight the data samples by the errors made by F_1. Next, fit F_2 to the weighted data set. Keep iterating until fitting for M components. If $F_m$has an accuracy higher than 0.5; the final ensemble model will have higher accuracy than any F_m (i.e. boosted accuracy)	(a) fast and easy to program (b) able to flexibly combine with any base learner F_m	(a) limit performance for insufficient data and base learners that are too complex or weak (b) susceptible to noise	(a) for insufficient data, use a modification of boosting that combines human expertise. E.g., (Schapire et al., 2002) (b) add a regularization term to prevent overfitting (this method also works for other algorithms). E.g., extreme gradient boosting (XGB)

	Bagging	Bagging means bootstrap aggregating – a form of ensemble learning. One would fit M different base models to different randomly sampled subsets of data (sampled with replacement, i.e., bootstrap sampling); this encourages the different models to make diverse predictions. We would sample until we have a total of N examples per model, where N is the number of original data points, and an example may appear multiple times.	(a) enhance robustness and generalization: bagging prevents the ensemble model from relying too much on any individual training example	(a) does not always improve performance: each base model only sees 63% of the unique input examples on average. For deep networks, fewer training data may affect performance; thus, bagged DNNs do not usually work well	(a) decorrelate the base learners further by learning based on a randomly chosen subset of input variables and randomly chosen subset of data cases. E.g., RF
NN	Deep neural network (DNN)	DNN consists of network of nodes and layers – input, output, and ≥ 2 hidden layers. Each layer l is composed of combinations of feature transformation functions ϕ, defined by a vector of parameters θ_l. i.e., $f(x;\theta) = f_L(f_{L-1}(\dots(f_1(x))\dots))$, where $f_l(x) = f(x;\theta_l)$. Gradient descent is most commonly used to train the model.	(a) easy to handle multi-task learning (b) highly flexible: adaptive to layer modification during the training process (online learning) (c) superior performance on homogeneous datasets: images, text, video, audio	(a) suffers from overfitting for highly nonlinear processes (b) may converge to local minimum due to gradient descent (c) black box: lack of interpretability of the relationship between inputs and outputs (d) needs larger dataset and feature engineering details to achieve high accuracy (e) computational expensive for training	(a) Incorporate fuzzy logic (an uncertain logic rule occurs in biochar pyrolysis system) to improve prediction accuracy, e.g., ANFIS (b) Use meta-heuristics like gray wolf optimization (GWO) to address the challenge of stucking at local optima and training on small dataset

				(f) performance is affected by initial point (g) typically inferior to tabular data that lack spatial structure like homogeneous datasets	
	Recurrent neural network (RNN)	Let t be time, and the prediction of output y_t depends on both input x_t and a hidden state of the system h_t. RNN maps the input space of sequences to an output space of sequences as the sequence is processed. That is:	(a) useful for generating sequences of real-valued feature vectors (e.g., pen strokes for hand-written characters) and time series real-value sequences	(a) expensive to train, as they need to maintain long term hidden state	(a) To make training easier, use convolutional neural networks (CNN) that compute a function of some local neighborhood, and return an output.
Kernel	Support Vector Machine (SVM)	Finds the decision boundary that maximizes the margin of support vectors to the boundary. It consists of kernel function and supporting hyperplane. A kernel function maps input variables to a higher dimensional place such that they can be separated into different classes by a hyperplane. The vectors on the boundaries are called support vectors. SVM can be extended to regression tasks through SVR and LS-SVM.	(a) superior performance (than RF) on clean and outlier free data (b) work with a variety of data: handle nonlinearly separable data sets; kernels can be defined on non-vector inputs; kernels can combine different types of data (Noble, 2006) (c) kernels allow SVM to incorporate	(a) prediction performance is affected by chosen kernel function, and needs trial and errors to find the optimal kernel function (Noble, 2006): e.g., linear, polynomial, spline, gaussian radial basis (GRB). GRB generally performs well. (b) data needs to be on a similar scale (to calculate the "distance" and	(a) construct a set of kernel functions and leverage cross-validation to test the optimal one (Noble, 2006) (b) normalize the input dataset before applying the algorithm (c) leverage sampling and probabilistic kernel function to incorporate stochastic properties of the input features (relevance vector machine, RVM)

			prior knowledge (Noble, 2006) (d) Superior performance on small dataset (e) fast on data sets of thousands of examples (Noble, 2006) (f) generalize well on sparse features (g) SVR is a non-parametric technique. Hence, the model output does not rely on distributions of the underlying dependent and independent variables (Jalalifar et al., 2020a)	maximize the "margin') (c) lack of uncertainty	
*Remediations are some methods that are developed to address the weakness					

Table S2 listed the detailed summary of each paper in the category material and process (M&P) design. Papers within M&P usually have two main objectives: material property prediction or reverse engineering. For material property prediction, it can be energy related and non-energy related. We assigned letter numbers to the objectives for the summary:

A. Material property prediction
    a. Energy related
    b. Non-energy related
B. Reverse engineering (estimate optimal input combination for desired output)

Due to the space limit, we chose root mean squared error (RMSE) as the model evaluation parameter. If RMSE is not available, the coefficient of determination (R^2) was displayed. We preferred RMSE over R^2, because R^2 may not necessarily indicate goodness-of-fit. In practice, it is recommended to examine several evaluation matrices (e.g., mean absolute error MAE) to determine model goodness-of-fit – RMSE penalizes outliers more as it squares the error, while MAE is less affected by outliers. In addition, we make a note of whether the models belong to multi-input single-output (MISO) or multi-input multi-output (MIMO), and whether it came from first-hand experiments. MISO refers to models that were trained on multiple input variables and produced one output variable; MIMO refers to models that produced multiple output variables at once.

Table S1 Detailed summary of each paper (M&P design)

Biochar/hydrochar							
Ref	# Data	ML method	Input variables	Output variables	RMSE (or R^2)	Objectives	Experiment

(Li et al., 2015) (1st attempt to aggregate literature data to predict product features based on various feedstocks, reaction conditions)	340, each output: (i) Yield: 263 (ii) Carbon content: 248 (iii) Energy content: 220 (iv) Normalized $C_{(s)}$: 244 (v) Normalized $C_{(l)}$: 203 (vi) Normalized $C_{(g)}$: 188	(i) Multiple linear regression (MLR) (ii) Regression tree (RT) *MISO	(i) process related: reaction time (t), reaction temperature, initial feedstock concentration, heating rate, heating time, heating time /reaction time (HT/t), reactor volume, volume ratio (% of reactor volume filled with liquid and feedstock) (ii) feedstock composition (%, dry wt.): C, H, O, ash, volatile matter, fixed carbon contents	(i) product yield (ii) carbon content (iii) energy content (iv) normalized C in solid (v) normalized C in liquid (vi) normalized C in gas	For each output variable (MLR/RT): (i) 9.41/7.47 (ii) 4.66/3.94 (iii) 1.92/1.83 (iv) 0.047/0.027 (v) 0.049/0.028 (vi) 0.004/0.004 RT wins	A(a, b)	
(Cao et al., 2016)	33	(i) ANN (ii) Least-square support vector	(i) process related: heating rate,	Yield	ANN/LS-SVM: 0.8347/0.3652	A(b)	Yes

		machine (LS-SVM), with radial basis function (RBF) kernel *MISO	pyrolysis temperature, holding time (ii) feedstock physical feature: moisture content, sample mass		LS-SVM wins		
(Hough et al., 2017)	250,000 (rich generated data, exclude from Figure 1)	(i) ANN (FFNN) (ii) Decision tree (DT) *MISO (single net) *MIMO (full net)	(i) process related: maximum pyrolysis temperature (Tmax), heating rate (ii) feedstock composition (% dry wt.): C, H content	(i) chemical compositions in solid (ii) chemical compositions in gas	R^2>0.982 for all outputs and for both ANN and DT Accuracy: Single net the most accurate; Full net ≈ DT	A(a)	
(Ewees and Elaziz, 2018)	33	Adaptive neuro-fuzzy inference system and gray wolf optimization algorithm (ANFIS-GWO): a hybrid between the GWO and ANFIS, in which the	(i) process related: heating rate, pyrolysis temperature, holding time (ii) feedstock physical feature: moisture	Yield	ANFIS-GWO/ANFIS-PSO/ANFIS-GA/ ANFIS-GOA/ANFIS-SCA/ANFIS-WOA/ ANFIS-flower/ANFIS/ LS-SVM/NN:	A(b), B	Yes same as (Cao et al., 2016)

		parameters of the ANFIS are determined by using the GWO algorithm. Compared with: - original ANFIS - seven optimized ANFIS with different meta-heuristic algos: particle swarm optimization (PSO), genetic algorithm (GA), grasshopper optimization algorithm (GOA), sine-cosine algorithm (SCA), whale optimization algorithm (WOA), flower pollination algorithm, LS-SVM, regression NN	content, sample mass		0.259/0.263/0.263/ 0.388/0.294/0.311/ 0.307/0.720/0.365/0.835 ANFIS-GWO wins		

		*MISO					
(Jiang et al., 2019b)	130	(i) Linear Regression (LR) (ii) Support vector regression (SVR), kernel: polynomial (iii) Random Forest Regression (RFR) *MISO	(i) catalyst related condition: amount of NaOH, KOH, wt% of NaOH-KOH (ii) feedstock physical feature: straw used to prepare the material	corresponding combustion AE of each interval from the conversion rate 0.1 to 0.94	LR/SVR/RFR: 5.66/4.56/1.90 RFR wins	A(a, b)	
(Zhu et al., 2019a)	(i) Yield: 245 (ii) Carbon content in biochar (C-char): 128	Random Forest (RF) *MISO	(i) feedstock composition (% dry wt.): lignin, cellulose hemicellulose, ash, C, H, O, N content (ii) feedstock physical	(i) Biochar Yield (ii) C-char	For each output variable: (i) 3.4028 (ii) 5.8123	A(b)	

			feature: particle size (PS) (iii) process related: heating rate (HR), highest treatment temperature (HTT), residence time (RT)				
(Ismail et al., 2019)	21	ANN-Kriging hybrid Compared with original ANN *MISO	process related: hydrothermal time, hydrothermal temperature	(i) inorganic phosphorous carbon (ii) carbon content	For each output variable (ANN/ANN-Kriging): (i) 7.0731/3.3124 (ii) 6.7333/3.1340	A(b)	Yes
(Jalalifar et al., 2020)	82	SVR-PSO Compared Kernels: Linear, Polynomial, Gaussian (radial basis function, RBF) *MISO	Process related: pyrolysis conditions	Product (bio-oil) yields	SVR-PSO with different Kernels (linear/ polynomial/RBF): 18.58/30.51/2.21 Sequentially applied SVR	B	Yes

					and then leveraged PSO to find corresponding optimal values		
(Alaba et al., 2020)	-	ANN *curve prediction	(i) feedstock composition (% dry wt.): C, H, N, S, O, volatile matter, ash, fixed carbon, ash composition, heating values (e.g., higher heating value (HHV), lower heating value (LHV)) (ii) feedstock physical feature: water content	(i) thermal gravimetric curve (TG) (ii) differential TG (DTG) curve	0.02 ~ 0.03, depending on thermal decomposition temperature	A(b)	Yes
(Pathy et al., 2020)	91	eXtreme Gradient Boosting (XGB) *MISO	(i) feedstock composition % dry wt): C, H, O, N; H/C, O/C, N/C, ash, fixed carbon,	(i) Biochar Yield (ii) Biochar composition (C, H, O, N)	For each output variable R^2: (i) 0.844 (II) 0.66	A(b)	

			and volatile compound (ii) process related: pyrolysis condition (pyrolysis temperature, heating rate and residence time)				
(L. Li et al., 2020)	(i) Yield: 649; (ii) Energy: 475	RF *MISO	(i) feedstock composition (% dry wt.): ash content, volatile matter, fixed carbon, C, H, O, cellulose, hemicellulose, lignin (ii) process related: reaction time, temperature (iii) initial solids concentration	(i) Hydrochar yield (ii) energy content	Detailed models were described in (Li et al., 2018). For each output variable (R^2): (i) 0.946 (ii) 0.952	A(a)	
(J. Li et al., 2020)	(i) Hydrochar: 248	(i) SVR (ii) RF	(i) feedstock composition	(i) Yield	For SVR, RBF kernel has the	A(a, b)	

	(ii) Pyrochar: 165	*1st MIMO respectively for hydrochar and pyrochar	(% dry wt.): C, H, N, O, fixed carbon, ash, and volatile matter (ii) process related: operational conditions of hydrothermal carbonization (HTC) (temperature HT, reaction time Ht, and water content in reactor WC); pyrolysis (temperature PT, heating rate PHR and reaction time Pt)	(ii) Higher Heating value (HHV) (iii) energy recovery efficiency (ER) (iiii) energy densification (ED)	lowest RMSE (compared with linear and poly kernel), so SVR with RBF kernel was summarized here (RF/SVR): (1) hydrochar: 6.2/3.88 (b) pyrochar: 4.23/4.18 SVR wins		
(Selvarajoo et al., 2020)	196	ANN (FFNN) *MISO	(i) process related: heating temperature, heating rate, residence time	Yield	0.5954	A(b)	Yes

(Thiruvengadam et al., 2021)	498	(i) XGB (ii) ANN *MISO	(i) feedstock type: cotton, rice husk, soybean,… (ii) process-related: chemical pretreatment, heat pretreatment, pyrolysis conditions, chemical post-treatment conditions	(i) % biochar, liquid, gas yields (ii) Yields of gaseous products (detailed gaseous product type) (iii) Yields of liquid products (detailed liquid product type) (iv) Biochar physical properties (v) Biochar chemical properties (vi) Biochar sorbent capacities	Average % error was presented: In all outputs, XGB outperformed ANN	A(b), B	
(Li et al., 2021)	248	(i) RF (ii) SVM, kernel: RBF (iii) Deep neural network (DNN)	(i) feedstock composition: C, H, N, O, fix carbon (Fc), ash (A), volatile matter (V)	(i) Yield (ii) Fuel properties (FP, including HHV and energy recovery ER)	For each output variable (RF/SVR/DNN): (i) 10.83/7.50/7.05	A(a), B	

		(integrated ML with multi-objective optimization) *MIMO	(ii) feedstock physical feature: water content (iii) process related: HTC conditions, e.g., reaction time (t), temperature (T)	(iii) carbon capture (C char and carbon recovery CR) (iv) carbon stability (represented by atomic ratios: N/C, H/C, O/C)	(ii) HHV: 2.82/1.27/1.53; ER: 13.18/8.05/7.59 (iii) C char: 3.91/2.52/2.91; CR: 12.59/7.72/7.15 (iv) H/C: 0.15/0.08/0.08; O/C: 0.14/0.06/0.06; N/C: 0.01/0.01/0.01 For some output SVR wins; for some DNN wins. Overall, SVR and DNN are comparable.		
(Tsekos et al., 2021)	482	ANN *MISO	(i) feedstock component (% dry wt.): cellulose, hemicellulose, lignin, ash,	(i) Biochar yield (ii) Liquid yield (iii) Gas yield	For each output variable (reduced/full model): (i) 5.1/5.9	A(b)	

			moisture content (ii) process related: pyrolysis temperature, heating rate, holding time, gas residence time (iii) feedstock physical feature: average particle size, sample size	(i.e., pyrolysis product composition)	(ii) 9.3/6.9 (iii) 5.6/6		
Biomass-derived AC							
(Jiang et al., 2019a)	60 sets of experiments With experimental design (DoE)	(i) LR (ii) SVR, kernel: polynomial (iii) RFR *MISO	(i) process related: impregnation ratio in grams of activation chemical to biomass, heating rate, pyrolytic temperature	(i) Methylene blue number (MBN) (characterizes the number of mesopores) (ii) Iodine number (IN) (characterizes the number of micropores)	RFR was more generally suitable for MBN and IN prediction	A(b)	Yes

(Mathew et al., 2020)	15 sets of experiments With DoE	ANN-PSO *MISO	(i) process related: impregnation ratio in grams of activation chemical to biomass, temperature of activation (ii) feedstock component: C(%) of biomass to other material used to make the electrode	(i) Specific capacitance (ii) Equivalent series resistance (ESR)	For each output variable R^2: (i) 0.9975 (ii) 0.9788	A(b), B	Yes
(Liao et al., 2019)	168	ANN *MISO	(i) feedstock composition: (% dry wt.): C, H, O fixed carbon, volatile matter, ash (ii) process related: carbonization conditions (carbonization temperature	(i) yield (ii) Brunauer–Emmett–Teller (BET) specific surface area	< 0.1 for all outputs		

			and time); activation conditions: (activation temperature, time), and steam to biochar ratio				

Table S3 summarizes the papers in the category end-use performance prediction. Papers within this category have five main objectives: pollutant removal efficiency prediction, gas molecule adsorption capacity prediction, soil amendment efficiency prediction, electrode capacitance prediction, and spectra measurement prediction. For pollutant removal, depending on the pollutant type, we further divided the category into metal ion, organic matter, and non-organic matter. We assigned letter numbers to the objectives as the following:

A. Pollutant removal
    a. Metal ion
    b. Organic matter
    c. Non-organic matter
B. Gas molecule adsorption
C. Soil amendment
D. Electrode
E. Spectra measurement

Table S2 Detailed summary of each paper (end-use)

Biochar/hydrochar							
Ref	# Data	ML method	Input variables	Output variables	RMSE (or R^2)	Objectives	Experiment
(Ding et al., 2018)	1170	Boosted regression trees (BRT) *MISO	(i) soil properties (ii) biochar properties (iii) incubation conditions	Decomposition of native soil organic carbon	R^2: 0.724	C	

(Liu et al., 2019)	(i) Crop production: 1314 (ii) Soil NH_3 volatilization: 163 (iii) Soil N_2O emissions: 552 (iv) Soil N leaching: 181	RF (random forest regression) *MISO	(i) soil properties (ii) biochar properties (iii) incubation conditions (iv) climate zone (v) scale of the experiment (field or lab)	(i) Crop production (ii) Soil NH_3 volatilization (iii) N_2O emissions (iv) N leaching	For each output variable: (i) 19 (ii) 31 (iii) 40 (iiii) 18	C	
(Cipullo et al., 2019)	6-month experimental data	(i) ANN (FFNN) (ii) RF *MISO	1st stage: (i) Soil type (ii) Amendment (iii) Total concentration at t = 0 (iv) Time 2nd stage: (i) Soil type (ii) Amendment (iii) Bioavailbility conc. At time t (prediction form 1st stage)	1st stage: Bioavailable concentration of various pollutants at time t 2nd stage: Toxicity at time t	(R^2) 1st stage: For all pollutants, RF ≥ FFNN 2nd stage: RF is slightly better than FFNN in most toxicity indices	C	Yes

			(iv) Time				
(Shen et al., 2019)	3868 (exclude from Figure 1 since it's outside of scope)	Naïve Bayes Classifier	(i) GPR signal maximum amplitude (ii) GPR signal intensity (iii) GPR signal area (iv) GPR signal energy	(i) soil C content (in %) (ii) soil C structure (iii) soil moisture levels		E	Yes
(Zhu et al., 2019b)	353	(i) ANN (FFNN) (ii) RF *MISO	(i) biochar properties: pH of biochar in water (pHH_2O), surface area of biochar, cation exchange capacity (CEC), ash content, biochar particle size (PS), mass percentage of total carbon in the biochar (C), molar ratio of oxygen and nitrogen to carbon [(O+N)/ C], molar ratio of oxygen to carbon (O/C), and molar ratio of hydrogen to carbon (H/C) (ii) incubation conditions, including solution pH,	Adsorption capacity	Prediction performance: RF is slightly better than ANN Generalizability: RF is better than ANN	A(a)	

			adsorption temperature (T, ^{o}C) (iii) initial concentration ratio of heavy metals to biochars (iv) adsorbate properties: heavy metal charge number, ion radius (r, nm), and electronegativity (χ).				
(Li et al., 2019)	156	SVM (directed acyclic graph SVM) *MISO	(i) adsorbate properties: contaminant (heavy metal) type (ii) incubation condition: temperature, pH, adsorbent dosage, contact time, contaminant concentration (iii) biochar properties: BET surface area, adsorption capacity (iv) sorption classes	Adsorption capacity levels (level 1: adsorption capacity $<$50; level 2: 50$\leq$adsorption capacity$<$100; class 3: adsorption capacity $\geq$100)	Classification accuracy: 99.4%	A(a)	Test cases were experimental data

(Mojiri et al., 2020)	data from 119-day experiment	ANN (FFNN) *MISO	incubation condition: treatment time, nitrogen loading rate (NLR), ammonia concentration, nitrite concentration	Nitrogen removal (ammonia and nitrite), namely Total Nitrogen (TN) removal	1.14	A(c)	Yes
(Sigmund et al., 2020) (1st to include sorbent properties)	467, including different sorbents biochar and activated carbon	ANN (FFNN) *MIMO	(i) adsorbent properties: content of carbon (C, %), hydrogen (H, %), H/C, oxygen (O, %), O/C, SSA (m2/g), pH (C is a proxy for homogeneity, SSA is a proxy for porosity and accessible sorption sites, H/C is a proxy for aromaticity, and O/C is a proxy for polarity, and the experimental pH is linked to the material's surface charge (negative charge increasing with pH). (ii) adsorbate properties: A^- (ionized negatively charged species);	Freundlich isotherm constants: log K_F and n	(R^2): (i) log K_F: 0.98 (ii) n: 0.91	A(b)	

			$logD_{ow}$ (pH-dependent hydrophobicity parameter); five Abraham solute parameters (E, S, A, B, and V): E (excess molar refraction), S (dipolarity/polarizability), A (H-bond acidity), B (H-bond basicity), V (molar volume)				
(De Miranda Ramos Soares et al., 2020)	202	(i) ANN (ii) RF *MISO	(i) incubation condition: Salinity (g/L), rotation (rpm), temperature, contact time (min), adsorbent dosage (%), pH (ii) Initial Dye Concentration (mg/L)	(i) Final Dye Concentration (mg/L) (ii) Adsorption capacity (mg/g) (iii) removal rate (%)	(RF/ANN): (i) 0.034/0.04 (ii) 0.022/0.026 (iii) 0.039/0.044 RF wins (because it's better at capture data variation)	A(b)	Yes
(Zhang et al., 2020) (1^{st} to include BET and V_t of sorbent)	586 isotherms (four carbon materials: biochar, CNTs, GAC and polymeric	(i) ANN (ii) SVM (iii) Bagging (used cosine similarity)	(i) the equilibrium concentration log Ce (μM) (ii) adsorbate properties: 5 Abraham descriptors	The adsorption coefficient log Kd (L/g).	For each adsorbent: RMSE of Bagging and NN win over SVM; MAE of Bagging ≥NN.	A(b)	

	resin; total 586*7 data points, since each isotherm has 7 data points)	*MISO	(E, S, A, B, and V) for the chemicals (iii) adsorbent properties: BET in m^2/g and V_t in cm^3/g		Considering RMSE & MAE, NN is preferred		
(Wehrle et al., 2021)	162 (exclude from Figure 1, because it's outside of scope)	SVM (kernel: RBF)	portable MIRS spectra of soil sample treated with different types of organic amendment	(i) total organic carbon (TOC) (ii) total nitrogen (TN) (iii) ratio of TOC to TN (CN-ratio) (iv) hot water extractable carbon (hwC) (v) hot water extractable nitrogen (hwN) (vi) hwC/hwN (hwCN-ratio) (vii) proportion of hwC to TOC (hwCprop) (viii) proportion of hwN to TN (hwNprop)		E, C	Yes
(Zhu et al., 2021)	110 different carbon materials,	(i) RF (ii) Gradient boosting	(i) adsorbent properties: total carbon content (C,	adsorption capacity (Q, mg/g) of CBMs	For either TC or SMX prediction: ANN's R^2 was	A(b)	

	including biochar and activated carbon, but did not specify how many sets of isotherm data were considered	trees (GBDT) (iii) ANN *MISO	wt.%), molar ratio of hydrogen to carbon (H/C), molar ratio of oxygen to carbon (O/C), molar ratio of oxygen and nitrogen to carbon [(O + N)/C] (representing the polarity of adsorbents), ash content (ash, wt.%), Brunauer-Emmett-Teller surface area (BET, m2/g), and point of zero charge (pHpzc) (ii) incubation condition: adsorption temperature (T, °C) and solution pH (pHsol) (iii) initial concentration of TC or SMX in comparison to CBMs dosage (C_0, mg/g)	for antibiotics (tetracycline, TC; sulfamethoxazole, SMX)	higher (ANN>RF>GBDT), but some predicted values from ANN at a low adsorption capacity were negative possibly due to the activation function. RF is preferred		
(El Hanandeh et al., 2021)	476	(i) FFNN, cascade forward network,	(i) initial concentration of the metal ions (CiPb, CiCu, CiNi in	heavy metal (Pb(II), Cu(II), Ni(II)) sorption efficiency	Several backpropagation algorithms were tested, and those	A(a)	Yes

(1st model to address the mutual interactions of key process parameters on the adsorption capacity in multi-solute systems)		partial recurrent network (Elman NN), radial basis network (generalized regression NN, called GRNN) (ii) Gradient boosting *MIMO	mg/L): binary and ternary solutions of Pb^{2+}, Cu^{2+}, and Ni^{2+} (ii) incubation condition: the pH of the solution, contact time (t in minutes), temperature (T in ℃)	several cases: single, binary, ternary multi-component solutions	with Bayesian regularization performed the best because Bayesian regularization back propagation is more suitable for smaller datasets with considerable noise. GRNN provided the best predictions and was able to capture the physical constraints of the system		
(Zhao et al., 2021)	353	(i) Kernel extreme learning machine (KELM) (ii) Kriging (also called guassian process regression)	(i) biochar properties: pH of biochar in water (pH_{H2O}), Specific surface area (SA, m^2/g), cation exchange capacity (CEC, cmol(+)/kg), Ash (%), particle size (PS, mm), C	heavy metal sorption efficiency three cases: Pb(II), Cd(II), Zn(II), Cu(II), Ni(II), As(III) separately, and six ions altogether	KELM wins in single-ion prediction; Kriging wins in multi-ion prediction Reason: Kriging is characterized by interpolation,	A(a)	

		*MISO	(dry w.t.%), (O + N)/C, O/C, H/C (ii) incubation conditions: pH_{solute}, adsorption environment temperature (T, °C) (iii) initial concentration ratio of heavy metal to biochar: C_0 (mmol/g) (iv) adsorbate properties: the number of charges (N_{charge}), ionic radius (r), and electronegativity (χ)		and less data capacity does not allow it to perform better training		
(Ke et al., 2021a)	353	(i) SVM (ii) RF (iii) ANN (iv) M5Tree (v) Gaussian process (GP) (vi) Bagging: each individual	(i) biochar properties: biochar surface area (BSA), percentage of ash (A), cation exchange capacity (CEC), particle size of biochar (PSB), pH of biochar in wastewater (pHww), percentage of carbon in biochar (C), the ratio of oxygen and carbon (O/C), the	heavy metal sorption efficiency	Single model: RF is superior; GP has the lowest performance (implying that Gaussian distribution is not strong enough to explain the relationship between input	A(a)	

		model was bagged with each other *MISO	ratio of hydrogen and carbon (H/C), ratio of O and N with C [(O + N)/C], (ii) incubation conditions: solution pH (pHsol), heavy metal concentration in wastewater (CO), pyrolysis temperature (TP), and environmental temperature (Tenvi)		and output variables) Bagged models: SVM-ANN is the best. Ensemble model is not always better, e.g., SVM alone performs better than SVM-GP. Ensemble models based on SVM, RF, M5Tree are suitable for predictions; those based on GP showed higher error as sorption efficiency increases		
(Ke et al., 2021b)	353	(i) ANN (Backpropagation neural network, BPNN) (ii) Fuzzy C-means	(i) biochar production conditions and biochar properties: pyrolysis temperature (TP), the ratio of hydrogen and carbon (H/C),	sorption efficiency	ANN/FCM-ANN: 0.050/0.036 FCM-ANN performs better	A(a)	

		clustering + BPNN *MISO	percentage of carbon in biochar (C), ratio of oxygen and nitrate with carbon [(O + N)/C], the ratio of oxygen and carbon (O/C), percentage of ash (A), particle size of biochar (PSB), biochar surface area (BSA), cation exchange capacity (CEC) (ii) incubation condition: pH of biochar in wastewater (pHw), environmental temperature (Tenvi), solution pH (pHs), heavy metal concentration in wastewater (CO)				
(Nguyen et al., 2021)	Data from pilot-scale 21-week treatment system	(i) RF (ii) SVM (iii) K-nearest neighbor (iv) GLM	influents concentration's: (i) pH (ii) Suspended solids (TSS) (iii) NH_4-N	effluent concentration's; (i) NH_4-N (ii) BOD_5	KNN has the worst performance (i) RF wins (SVM & CUBIST are comparable)	A(b, c)	Yes

		(v) LR (vi) CUBIST (an extension of M5 model tree) *MISO	(iv) Biological oxygen demand during 5 days (BOD_5) (v) Chemical oxygen demand (COD) (vi) NO_3-N (vii) Hydraulic loading rate (HLR)		(ii) SVM wins (except KNN, all others are comparable)		
(Palansoori ya et al., 2022)	162	(i) RF (ii) SVR (iii) NN *MISO	(i) Biochar production conditions and biochar properties: pyrolysis temperature, biochar pH (pH_{BC}), C, H, O, N contents (dry wt.%), H/C, O/C, (O+N)/C, ash content, surface area (SA) (ii) incubation conditions: biochar application rate in soil, experimental duration (time), available heavy metal content in soil (Avail. HM), (iii) soil properties: soil pH, soil	(i) biochar surface area (goal: impute missing surface area data points) (ii) heavy metal immobilization efficiency	(RF/ SVR/ NN/ updated-RF): (i) All performed well (ii) Updated RF is RF built with reduced input 14 features (originally 20 features) 11.99/ 15.73/ 10.54/9.92 Updated RF wins	C, 1.A(b)	

			electrical conductivity (EC) (iv) adsorbate properties: heavy metal properties, e.g., molecular weight, electronegativity, ionic radius, valency				
Biosorbent							
(Prakash et al., 2008)	256 experimental data + 4864 generated with interpolation (they generated data to produce sufficient data to train the network effectively) (rich extrapolated data, exclude from Figure 1)	RNN (Elman) *MISO	(i) initial Cu ion concentration (ii) incubation condition: pH, temperature (iii) biosorbent property: particle size	adsorption efficiency	0.046	A(a)	

(Parveen et al., 2017)	124	(i) SVM (kernel: rbf) (ii) MLR (iii) ANN	(i) initial Cr(VI) concentration (ii) incubation condition: temperature, contact time, pH	adsorption capacity (mg/g)	(SVR/MLR/ANN): 0.0159/ 0.1549/ 0.1540 SVR wins	A(a)	
(Dolatabadi et al., 2018)	50 sets of experimental data	(i) ANN (FFNN) (iii) ANFIS *MISO	(i) initial dye concentration (ii) initial Cu concentration (iii) incubation condition: contact time adsorbent, dosage	removal efficiency (%): (i) dye (ii) Cu(II)	(ANN/ANFIS): (i) 0.676/0.426 (ii) 1.248/0.353 ANFIS wins	A(a, b)	Yes
Biomass-derived AC							
(Mazaheri et al., 2017)	52 with experimental design	(i) Response surface methodology (RSM) (ii) BRT (iii) ANN *MISO	(i) incubation condition: stirring time (min), pH, concentrations of methylene blue (MB), concentrations of Cd(II) (ii) adsorbent mass (mg)	percentage removal (%) (i) methylene blue dye (ii) Cd(II)	(RSM/BRT/ANN): (i) 0.0180/ 0.00292/ 0.00475 (ii) 0.01125/ 0.00426/ 0.00477 BRT and ANN win over RSM;	A(a, b)	Yes

					BRT performs the best		
(Karri and Sahu, 2018)	50 with experimental design. This experimental data was used to generate 270 datasets (rich extrapolated data, exclude from Figure 1)	(i) RSM (ii) ANN (ANN-PSO) *MISO	(i) incubation condition: pH, residence time, reaction temperature (ii) initial concentration (iii) activated carbon dosage	Zn (II) removal (%)	(RSM/ANN-PSO): 2.632/0.983 ANN-PSO is preferred	A(a)	Yes
(Zhou et al., 2020)	70	(i) Generalized Linear Regression (GLR) (ii) SVM (iii) RF (iv) ANN	(i) activated carbon properties: specific surface area (micro), specific surface area (meso) (ii) Scan Rate	(i) Specific Capacitance (ii) Power Density	(GLR/ SVM/ RF/ ANN): (i) 54.91/ 40.16/ 38.13/ 36.40 SVM and RF fail to predict when the capacitance approaches zero (although the predicted value is not negative, it is far from the observed value).	D	

					Overall, ANN is preferred (ii) only ANN is used to further predict power density		
(Mojiri et al., 2019)	50	ANN *MISO	incubation condition: micropollutant concentration (mg/L), pH	Micropollutant removal	1.14	A(b)	Yes
(Zhang et al., 2019)	1020	DNN *MISO	(i) activated carbon textural properties: micropore volume (Vmicro), mesopore volume (Vmeso), total pore volume (Vtotal), Specific Surface Area (BET)	CO_2 adsorption	-	B	Yes
(Talebkeikhah et al., 2020)		(i) SVM (kernel: RBF) (ii) group method of data handling (GMDH) (iii) DT (iv) RF	(i) incubation condition: pH, contact time (ii) adsorbent dosage (iii) initial Pb (II) concentration	Pb (II) adsorption capacity	SVM wins In addition, coupling of MLP and ANFIS with grasshopper optimization algorithm (GOA) increases accuracy.	A(a)	Yes

		(v) Radial basis function (RBF) (vi) ANFIS (vii) Multilayer perceptron (MLP) *MISO					
(Zhu et al., 2020)	6244 (wide-range of porous carbon materials, exclude from Figure 1)	RF *MISO	(i) activated carbon properties: chemical compositions (CC), e.g., wr% of C, H, O, N (ii) activated carbon textural properties: BET surface area, micropore volume, mesopore volume, ultramicropore volume (iii) incubation conditions: temperature (T), pressure (P)	CO_2 adsorption capacity (Q, mmol/g)	0.148~0.266 depending on adsorption conditions	B	
(Afolabi et al., 2020)	495	ANN	(i) adsorbate initial concentration	Adsorption efficiency	0.0243	A(b)	Yes

		*MISO	(ii) incubation condition: adsorbate temperature, adsorbent and adsorbate contact time				
(Maulana Kusdhany and Lyth, 2021)	1745	(i) LR (ii) SVR (kernel: rbf, linear) (iii) XGB (iv) RF	(i) activated carbon properties: wt% of C, H, O, and N (ii) activated carbon textural properties: micropore volume, ultramicropore volume, total pore volume, BET specific surface area (iii) incubation condition: pressure	excess hydrogen uptake (wt%)	LR/ SVR(L)/ SVR (rbf)/ XGBT/ RF: 1.166/ 1.180/ 0.863/ 0.547/ 0.542 RF performs the best (XGBT is comparable)	B	

Table S4 describes the papers in the category of sustainability assessment. Papers within this category use ML to predict material properties, and further plug the predictions into traditional LCA framework to assess sustainability impact. For model performance evaluation, if RMSE is not available, we listed other available evaluation matrices. Other matrices that were used in this category: R^2 and mean absolute deviation (MAD).

Objectives of the ML models in this category contains the objectives from M&P design and end-use performance prediction categories. Therefore, we denoted the objectives as the following:

1. Objectives from M&P design: 1.X, where X consists of the letter numbers listed in M&P design. For example, 1.A(a) indicates energy related material property prediction.
2. Objectives from end-use performance prediction: 2.X, where X consists of the letter numbers listed in end-use performance prediction. For example, 2.C indicates soil amendment performance prediction.

Table S3 Detailed summary of each paper (sustainability assessment)

Biochar/hydrochar							
Ref	# Data	ML method	Input variables	Output variables	RMSE (or others)	Objectives	Experiment
(Dokoohaki et al., 2019)	1260	(i) Generalized additive model (GAM) (ii) Bayesian network (BN)	(i) soil properties: Soil organic carbon (SOC), sand, silt, clay content, CEC and soil pH (ii) biochar production conditions and biochar properties: carbon, nitrogen, ash content, pH, carbon-to-nitrogen (C:N) ratio, highest pyrolysis temperature (HPT), feedstock, and	the effects of biochar application on the crop yield response ratio	Mean absolute difference (MAD) of (GAM/BN): 0.10/0.18 BN wins	2.C and further quantified economic aspects and indirect GHG emissions	

		*MISO	thermochemical process. Biochar feedstock was classified into woody, non-woody, and manure, while pyrolysis type was characterized as fast and slow (iii) latitude (iv) N fertilizer and biochar application rates				
(Cheng et al., 2020b)	800	(i) MLR (ii) RT (iii) RF *MISO	(i) feedstock properties (wt% of C, H, N, O, and ash) (ii) process related conditions (reaction temperature, heating rate, and residence time)	(i) Yield: biocrude yield, hydrochar yield, gas, aqueous co-product (ACP) yield, gas yield (ii) Product properties: HHV and C% for hydrochar and biocrude	For each output variable: RF has the lowest RMSE RF wins	1.A(a) and further calculated LCA and economic w/ the predicted data	
(Cheng et al., 2020a)	(i) Yields: 538 (ii) Energy: 276 (iii) C-char: 305 (iv) N-char: 276	RF *MISO	(i) feedstock properties (wt% of C, H, N, O, and ash) (ii) process related conditions (reaction temperature, heating	(i) Biochar yields (ii) Biochar properties: HHV, C%, N%	For each output variable: (i) 4 (ii) HHV: 1; C%: 0.02; N%: 0.002	1.A(a) and further calculated LCA w/ the predicted data	

			rate, and residence time)				
Biomass-derived AC							
(Liao et al., 2020)	250	Pyrolysis kinetic model + ANN *MISO	(i) process related conditions: pyrolysis time, pyrolysis temperature, activation time, activation temperature, steam to biochar ratio (ii) feedstock properties: wt% of C, H, O in the biomass	(i) Kinetic generated gas/solid products (ii) ANN predicted properties (biochar yield)	R^2: 0.971	1.A(a) and generated LCI with Aspen and predicted output variables	

Table S5 and Table S6 showed the importance of input parameters for the M&P and end-use categories. Different colors encode different input parameter types. The summary was not done for the sustainability assessment category because the number of studies are not abundant enough to make consensus observation. Even within M&P and end-use categories, not every study conducted importance analysis for their models. Therefore, here we only summarized for the studies that conducted importance analysis; each study may include multiple models built for same or different output objectives. For every output objective, the authors usually chose the model with best performance to conduct feature importance analysis within one study. That is, the features are counted every time they are identified as important for predicting a kind of output variable.

The two tables displayed the input parameters that have been considered in more than 5 models (column name: total) among all studies that conducted importance analysis. Furthermore, we summarized the number of times that the input parameters have been identified as one of the top 3 influential factors (column name: n_top). In addition, we calculated the ratio for n_top/total (column name: ratio) to weigh the occurrence for the importance of the factors, and eventually ranked by ratio.

It was observed that for M&P, material production process factors such as HT/t, Tfinal, Solids amount; feedstock properties such as PS, C dry wt% are frequently considered to be influential when they were included in the model. For the end-use category, BDM texture properties such as meso-pore volume, ultra-micro pore volume, and specific surface area; incubation condition such as gas pressure and BDM dosage are usually detected to be influential to end-use performance when they are included in the model.

Table S4 M&P input parameters considered more than 5 times (ranked by ratio)

Input type	Input param	total	n_top	ratio
material production process	Heating time/reaction time ratio (HT/t)	13	7	0.54
material production process	Heating (pyrolysis) temperature (T_{final})	45	24	0.53
feedstock properties	feedstock particle size (PS)	6	3	0.50
feedstock properties	C dry wt% (C_{feed})	38	18	0.47
material production process	Solid amount ($Solids_{initial}$)	13	6	0.46
feedstock properties	N dry wt% (N_{feed})	21	8	0.38
feedstock properties	Lignin content	6	2	0.33

feedstock properties	H dry wt% (H_{feed})	38	10	0.26
feedstock properties	Ash dry wt% (Ash_{feed})	41	9	0.22
feedstock properties	Moisture content (MC)	14	3	0.21
material production process	Heating time (HT)	12	2	0.17
feedstock properties	Hemicellulose content	6	1	0.17
feedstock properties	Cellulose content	6	1	0.17
material production process	Holding time (reaction time, residence time) (t)	44	5	0.11
feedstock properties	Fixed carbon content (FC_{feed})	36	4	0.11
material production process	Volume ratio (VR)	12	1	0.08
feedstock properties	O dry wt% (O_{feed})	38	3	0.08
material production process	Heating rate (HR)	29	2	0.07
feedstock properties	Volatile matter (Vm_{feed})	36	2	0.06
material production process	Volume (V)	12	0	0.00

Table S5 end-use input parameters considered more than 5 times (ranked by ratio)

input type	input param	total	n_top	ratio
BDM properties (texture)	BDM meso pore volume	9	9	1.00
BDM properties (texture)	BDM ultramicro pore volume	10	8	0.80
incubation condition (environment conc.)	gas pressure	9	7	0.78
Incubation condition (BDM dosage)	BDM dosage	6	4	0.67

BDM properties (texture)	BDM Specific Surface Area	19	9	0.47
incubation condition (environment pH)	solution/soil pH	16	7	0.44
adsorbate properties	soil loam content	7	3	0.43
BDM properties (texture)	BDM micro pore volume	10	4	0.40
adsorbate properties	soil sand content	8	3	0.38
adsorbate properties	soil clay content	8	3	0.38
adsorbate properties	soil slit content	8	3	0.38
incubation condition (environment temperature)	temperature	9	3	0.33
Incubation condition (environment conc.)	equilibrium conc.	6	2	0.33
BDM properties (chemical component)	BDM H content	10	3	0.30
Incubation condition (time)	incubation time	11	3	0.27
BDM properties (pH)	BDM pH	12	2	0.17
BDM properties (chemical component)	BDM H/C (polarity)	7	1	0.14
BDM properties (chemical component)	BDM O/C (polarity)	7	1	0.14
BDM properties (chemical component)	BDM N content	14	1	0.07
BDM properties (chemical component)	BDM C content (homogenity)	21	1	0.05
BDM properties (chemical component)	BDM O content	10	0	0.00
BDM properties (chemical component)	BDM total chemical component	8	0	0.00

We show the winner algorithms for studies that performed multiple ML on the same dataset for M&P (Table S7) and end-use (Table S8) categories. The objective category inherits from Table S2.

Table S6 M&P winner algorithms for studies compare more than 2 ML algorithms

ML Type	ref	Winner*	Competitor algorithms	SVM kernel	Objective category**
TFBB	(Li et al., 2015)	DT	MLR	-	A(a, b)
	(Jiang et al., 2019a)	RF	MLR; SVM	polynomial	A(b)
	(Jiang et al., 2019b)	RF	LR; SVM	polynomial	A(a, b)
	(Thiruvengadam et al., 2021)	XGB	FFNN	-	A(b), B
Kernel	(Cao et al., 2016)	LS-SVM	FFNN	RBF	A(b)
	(J. Li et al., 2020)	SVM	RF	RBF	A(a,b)
	(Li et al., 2021)	SVM;FFNN	RF	RBF	A(a), B
NN	(Hough et al., 2017)	FFNN	DT	-	A(a), B
	(Ewees and Elaziz, 2018)	ANFIS-GWO	ANFIS; FFNN; LS-SVM	RBF	A(b), B
	(Ismail et al., 2019)	FFNN-Kriging	FFNN	-	A(b)

* Winner: Model with the lowest test RMSE was designated as winner (if RMSE is not available, R^2 or other metrics were used)

** Objective category:

A. Material property prediction: (a) energy related; (b) non-energy related
B. Reverse engineering (estimate optimal input combination for desired output)

*** (Ewees and Elaziz, 2018) used the data from (Cao et al., 2016)

Table S7 end-use winner algorithms for studies compare more than 2 ML algorithms

ML type	ref	Winner*	Competitor algorithms	SVM kernel	Objective category***
TFBB	(Mazaheri et al., 2017)	BRT	FFNN	-	A(a, b)
	(Cipullo et al., 2019)	RF	FFNN	-	C
	(Zhu et al., 2019b)	RF	FFNN	-	A(a)
	(De Miranda Ramos Soares et al., 2020)	RF	FFNN	-	A(b)
	(Ke et al., 2021a)	RF; Bagging (SVM-FFNN)	FFNN; GP; M5Tree**; SVM; Bagging**	RBF	A(a)
	(Maulana Kusdhany and Lyth, 2021)	RF; XGB	MLR; SVM	RBF	B
	(Nguyen et al., 2021)	RF	CUBIST**; GLM; KNN; MLR; SVM	RBF	A(c): NH_4-N
	(Zhu et al., 2021)	RF	FFNN; GBT	-	A(b)
	(Palansooriya et al., 2022)	RF	FFNN; SVM	RBF	C, 1.A(b)****
Kernel	(Parveen et al., 2017)	SVM	FFNN; MLR	RBF	A(a)
	(Talebkeikhah et al., 2020)	SVM	ANFIS; DT; FFNN; GMDH**; RBFNN; RF	RBF	A(a)
	(Nguyen et al., 2021)	SVM	CUBIST**; GLM; KNN; MLR; RF	RBF	A(b): BOD_5**
	(Zhao et al., 2021)	GP (Kriging)	KELM	-	A(a)
NN	(Dolatabadi et al., 2018)	ANFIS	FFNN	-	A(a, b)
	(Zhang et al., 2020)	FFNN	Bagging; SVM	RBF	A(b)
	(Zhou et al., 2020)	FFNN	GLM; RF; SVM	RBF	D

	(El Hanandeh et al., 2021)	GRNN	Elman NN; FFNN; GB	-	A(a)
	(Ke et al., 2021b)	FCM-FFNN**	FFNN	-	A(a)

*: Model with the lowest test RMSE was designated as winner (if RMSE is not available, R^2 or other metrics were used)

**: Bagging in (Ke et al., 2021a) built bagged models with combinations of the four models – FFNN, GP, M5Tree, SVM;

M5Tree: a Decision Tree learner; CUBIST: an extension of M5Tree; GMDH: grouped method of data handling; FCM-FFNN is unsupervised-supervised framework; KNN is K-nearest-neighbor, which is an Exemplar framework;

BOD_5: Biological oxygen demand during 5 days

***: A. Pollutant removal: (a) Metal ion, (b) Organic matter, (c) Non-organic matter; B. Gas molecule adsorption; C. Soil amendment; D. Electrode

****: 1.A(b) inherits from M&P design, which is reverse engineering

Table S9, Table S10, and Table S11 grouped references by algorithms. One study may be seen in several algorithm groups because it compares multiple models.

Table S8 Algorithm occurrence – M&P

Category	Algorithm	Ref
M&P design	ANFIS	(Ewees and Elaziz, 2018)
M&P design	ANFIS-GWO	(Ewees and Elaziz, 2018)
M&P design	DT	(Hough et al., 2017), (Li et al., 2015)
M&P design	FFNN	(Alaba et al., 2020), (Cao et al., 2016), (Ewees and Elaziz, 2018), (Hough et al., 2017), (Ismail et al., 2019), (Li et al., 2021), (Liao et al., 2019), (Mathew et al., 2020), (Selvarajoo et al., 2020), (Thiruvengadam et al., 2021), (Tsekos et al., 2021), (Ismail et al., 2019)
M&P design	FFNN-Kriging	(Ismail et al., 2019)
M&P design	LS-SVM	(Cao et al., 2016), (Ewees and Elaziz, 2018)
M&P design	MLR	(Jiang et al., 2019a), (Jiang et al., 2019b), (Li et al., 2015)
M&P design	RF	(J. Li et al., 2020), (Jiang et al., 2019a), (Jiang et al., 2019b), (L. Li et al., 2020), (Li et al., 2021), (Zhu et al., 2019a)
M&P design	SVM	(J. Li et al., 2020), (Jiang et al., 2019a), (Jiang et al., 2019b), (Li et al., 2021)
M&P design	SVM-PSO	(Jalalifar et al., 2020)
M&P design	XGB	(Pathy et al., 2020), (Thiruvengadam et al., 2021)

Table S9 Algorithm occurrence – end-use

Category	Algorithm	Ref
end-use	ANFIS	(Dolatabadi et al., 2018), (Talebkeikhah et al., 2020)
end-use	Bagging	(Ke et al., 2021a), (Zhang et al., 2020)
end-use	BRT	(Ding et al., 2018), (Mazaheri et al., 2017)
end-use	CFNN	(El Hanandeh et al., 2021
end-use	CUBIST	(Nguyen et al., 2021)
end-use	DT	(Talebkeikhah et al., 2020)
end-use	Elman NN	(El Hanandeh et al., 2021)
end-use	FCM-FFNN	(Ke et al., 2021b)
end-use	FFNN	(Afolabi et al., 2020), (Cipullo et al., 2019), (De Miranda Ramos Soares et al., 2020), (Dolatabadi et al., 2018), (El Hanandeh et al., 2021), (Ke et al., 2021a), (Ke et al., 2021b), (Mazaheri et al., 2017), (Mojiri et al., 2019), (Mojiri et al., 2020), (Palansooriya et al., 2022), (Parveen et al., 2017), (Sigmund et al., 2020), (Talebkeikhah et al., 2020), (Zhang et al., 2019), (Zhang et al., 2020), (Zhou et al., 2020) , (Zhu et al., 2019b), (Zhu et al., 2021)
end-use	FFNN-PSO	(Karri and Sahu, 2018)
end-use	GB	(El Hanandeh et al., 2021)
end-use	GBT	(Zhu et al., 2021)
end-use	GLM	(Nguyen et al., 2021), (Zhou et al., 2020)
end-use	GMDH	(Talebkeikhah et al., 2020)
end-use	GP	(Ke et al., 2021a), (Zhao et al., 2021)
end-use	GRNN	(El Hanandeh et al., 2021)
end-use	KELM	(Zhao et al., 2021)
end-use	KNN	(Nguyen et al., 2021)
end-use	M5Tree	(Ke et al., 2021a)
end-use	MLR	(Maulana Kusdhany and Lyth, 2021), (Nguyen et al., 2021), (Parveen et al., 2017)
end-use	Naïve Bayes Classifier	(Shen et al., 2019)
end-use	RBF	(Talebkeikhah et al., 2020)
end-use	RF	(Cipullo et al., 2019), (De Miranda Ramos Soares et al., 2020), (Ke et al., 2021a), (Liu et al., 2019), (Maulana Kusdhany and Lyth, 2021), (Nguyen et al., 2021), (Palansooriya et al., 2022),

		(Talebkeikhah et al., 2020), (Zhou et al., 2020) , (Zhu et al., 2019b), (Zhu et al., 2020), (Zhu et al., 2021)
end-use	RNN	(Prakash et al., 2008)
end-use	SVM	(Ke et al., 2021a), (Li et al., 2019), (Maulana Kusdhany and Lyth, 2021), (Nguyen et al., 2021), (Palansooriya et al., 2022), (Parveen et al., 2017), (Talebkeikhah et al., 2020), (Wehrle et al., 2021), (Zhang et al., 2020), (Zhou et al., 2020)
end-use	XGB	(Maulana Kusdhany and Lyth, 2021)

Table S10 Algorithm occurrence – sustainability

Category	Algorithm	ref
sustainability	BN	(Dokoohaki et al., 2019)
sustainability	DT	(Cheng et al., 2020b)
sustainability	FFNN	(Liao et al., 2020)
sustainability	GAM	(Dokoohaki et al., 2019)
sustainability	MLR	(Cheng et al., 2020b)
sustainability	RF	(Cheng et al., 2020a), (Cheng et al., 2020b)

Table S11 compares physics-based, pure ML, and physics-informed ML model. Let us use the example described in (Ji and Deng, 2021). Let an elementary reaction involving four species of [A, B, C, D] with corresponding stoichiometric coefficients: [v_A, v_B, v_C, v_D]:

$$v_A A + v_B B \rightarrow v_C C + V_D D$$

Suppose this is an adsorption reaction, and we would like to predict the adsorption capacity for future, a comparison of pros and cons of using physics-based models, pure ML models, and physics-informed ML models are presented in Table S12:

Table S12 Comparison of physics-based, pure ML, physics-informed ML model

	Physics-based models	Pure ML models	Physics-informed ML models
About	choose several candidate kinetic models to fit the experimental data, e.g., First-order, second-order kinetic; then, the kinetic model that fits better can be used to interpret the kinetic processes underlying the system – whether the rate-determining step is diffusion or binding with functional groups, and the magnitude of the rate constant k can provide physical insights of how fast the reaction happens	fit a deep neural network model with input parameters: ln[A], ln[B], ln[C], ln[D], lnT, t, -1/RT, where [X] represents element X's concentration, T: temperature, t: incubation time, R: gas constant. Then the number of neurons in the hidden layer and the number of layers is chosen such that the model fits the data best. As a result, the corresponding weight do are difficult to interpret.	there are multiple ways to incorporate physical principles into the Machine Learning, and we refer interested readers to (Karniadakis et al., 2021). One way to incorporate physics principle is encode the parameters in the law as input neuron; for the hidden layer, design each node as number of reactions; output nodes as targets for predictions. This way, the ML model is designed as a digital twin to the chemical reaction; therefore, the learned weights will have physical meanings for interpretation (Ji and Deng, 2021).
Pros	highly interpretable	(1) computationally efficient (2) not limited to a specific type of kinetic model. That is, the prediction will perform	the best of both world

		well if instead the data follow kinetic process other than first-order or second-order.	
Cons	(1) if the true data fall outside the candidate models, the prediction is poor (2) computation expensive		encoding representation is challenging for complex systems

Reference:

Afolabi, I.C., Popoola, S.I., Bello, O.S., 2020. Machine learning approach for prediction of paracetamol adsorption efficiency on chemically modified orange peel. Spectrochim Acta A Mol Biomol Spectrosc 243, 118769. https://doi.org/10.1016/j.saa.2020.118769

Alaba, P.A., Popoola, S.I., Abnisal, F., Lee, C.S., Ohunakin, O.S., Adetiba, E., Akanle, M.B., Abdul Patah, M.F., Atayero, A.A.A., Wan Daud, W.M.A., 2020. Thermal decomposition of rice husk: a comprehensive artificial intelligence predictive model. J Therm Anal Calorim 140, 1811–1823. https://doi.org/10.1007/s10973-019-08915-0

Cao, H., Xin, Y., Yuan, Q., 2016. Prediction of biochar yield from cattle manure pyrolysis via least squares support vector machine intelligent approach. Bioresour Technol 202, 158–164. https://doi.org/10.1016/j.biortech.2015.12.024

Cheng, F., Luo, H., Colosi, L.M., 2020a. Slow pyrolysis as a platform for negative emissions technology : An integration of machine learning models, life cycle assessment, and economic analysis. Energy Convers Manag 223, 113258. https://doi.org/10.1016/j.enconman.2020.113258

Cheng, F., Porter, M.D., Colosi, L.M., 2020b. Is hydrothermal treatment coupled with carbon capture and storage an energy-producing negative emissions technology? Energy Convers Manag 203, 112252. https://doi.org/10.1016/j.enconman.2019.112252

Cipullo, S., Snapir, B., Prpich, G., Campo, P., Coulon, F., 2019. Prediction of bioavailability and toxicity of complex chemical mixtures through machine learning models. Chemosphere 215, 388–395. https://doi.org/10.1016/j.chemosphere.2018.10.056

de Miranda Ramos Soares, A.P., de Oliveira Carvalho, F., de Farias Silva, C.E., da Silva Gonçalves, A.H., de Souza Abud, A.K., 2020. Random Forest as a promising application to predict basic-dye biosorption process using orange waste. J Environ Chem Eng 8, 103952. https://doi.org/10.1016/j.jece.2020.103952

Ding, F., Zwieten, L. van, Zhang, W., Weng, Z.H., Shi, S., Wang, J., 2018. A meta-analysis and critical evaluation of influencing factors on soil carbon priming following biochar amendment 1507–1517.

Dokoohaki, H., Miguez, F.E., Laird, D., Dumortier, J., 2019. Where should we apply biochar ?

Dolatabadi, M., Mehrabpour, M., Esfandyari, M., Alidadi, H., 2018. Modeling of simultaneous adsorption of dye and metal ion by sawdust from aqueous solution using of ANN and ANFIS. Chemometrics and Intelligent Laboratory Systems 181, 72–78. https://doi.org/10.1016/j.chemolab.2018.07.012

el Hanandeh, A., Mahdi, Z., Imtiaz, M.S., 2021. Modelling of the adsorption of Pb, Cu and Ni ions from single and multi-component aqueous solutions by date seed derived biochar: Comparison of six machine learning approaches. Environ Res 192. https://doi.org/10.1016/j.envres.2020.110338

Ewees, A.A., Elaziz, M.A., 2018. Improved Adaptive Neuro-Fuzzy Inference System Using Gray Wolf Optimization: A Case Study in Predicting Biochar Yield. Journal of Intelligent Systems 29, 924–940. https://doi.org/10.1515/jisys-2017-0641

Friedman, J., Hastie, T., Tibshirani, R., others, 2001. The elements of statistical learning. Springer series in statistics New York.

Hough, B.R., Beck, D.A.C., Schwartz, D.T., Pfaendtner, J., 2017. Application of machine learning to pyrolysis reaction networks: Reducing model solution time to enable process optimization. Comput Chem Eng 104, 56–63. https://doi.org/10.1016/J.COMPCHEMENG.2017.04.012

Ismail, H.Y., Shirazian, S., Skoretska, I., Mynko, O., Ghanim, B., Leahy, J.J., Walker, G.M., Kwapinski, W., 2019. ANN-Kriging hybrid model for predicting carbon and inorganic phosphorus recovery in hydrothermal carbonization. Waste Management 85, 242–252. https://doi.org/10.1016/j.wasman.2018.12.044

Jalalifar, S., Masoudi, M., Abbassi, R., Garaniya, V., Ghiji, M., Salehi, F., 2020a. A hybrid SVR-PSO model to predict a CFD-based optimised bubbling fluidised bed pyrolysis reactor. Energy 191, 116414. https://doi.org/10.1016/j.energy.2019.116414

Jalalifar, S., Masoudi, M., Abbassi, R., Garaniya, V., Ghiji, M., Salehi, F., 2020b. A hybrid SVR-PSO model to predict a CFD-based optimised bubbling fluidised bed pyrolysis reactor. Energy 191, 116414. https://doi.org/10.1016/J.ENERGY.2019.116414

Ji, W., Deng, S., 2021. Autonomous Discovery of Unknown Reaction Pathways from Data by Chemical Reaction Neural Network. Journal of Physical Chemistry A 125, 1082–1092. https://doi.org/10.1021/ACS.JPCA.0C09316/ASSET/IMAGES/LARGE/JP0C09316_0014.JPEG

Jiang, W., Xing, X., Li, S., Zhang, X., Wang, W., 2019a. Synthesis, characterization and machine learning based performance prediction of straw activated carbon. J Clean Prod 212, 1210–1223. https://doi.org/10.1016/J.JCLEPRO.2018.12.093

Jiang, W., Xing, X., Zhang, X., Mi, M., 2019b. Prediction of combustion activation energy of NaOH/KOH catalyzed straw pyrolytic carbon based on machine learning. Renew Energy 130, 1216–1225. https://doi.org/10.1016/j.renene.2018.08.089

Karniadakis, G.E., Kevrekidis, I.G., Lu, L., Perdikaris, P., Wang, S., Yang, L., 2021. Physics-informed machine learning. Nature Reviews Physics 2021 3:6 3, 422–440. https://doi.org/10.1038/s42254-021-00314-5

Karri, R.R., Sahu, J.N., 2018. Modeling and optimization by particle swarm embedded neural network for adsorption of zinc (II) by palm kernel shell based activated carbon from aqueous environment. J Environ Manage 206, 178–191. https://doi.org/10.1016/j.jenvman.2017.10.026

Ke, B., Nguyen, H., Bui, X., Bui, H., Choi, Y., 2021a. Predicting the sorption efficiency of heavy metal based on the biochar characteristics, metal sources, and environmental conditions using various novel hybrid machine learning models. Chemosphere 276, 130204. https://doi.org/10.1016/j.chemosphere.2021.130204

Ke, B., Nguyen, H., Bui, X.N., Bui, H.B., Nguyen-Thoi, T., 2021b. Prediction of the sorption efficiency of heavy metal onto biochar using a robust combination of fuzzy C-means clustering and back-propagation neural network. J Environ Manage 293. https://doi.org/10.1016/j.jenvman.2021.112808

Li, J., Pan, L., Suvarna, M., Tong, Y.W., Wang, X., 2020. Fuel properties of hydrochar and pyrochar: Prediction and exploration with machine learning. Appl Energy 269, 115166. https://doi.org/10.1016/j.apenergy.2020.115166

Li, J., Zhu, X., Li, Y., Tong, Y.W., Ok, Y.S., Wang, X., 2021. Multi-task prediction and optimization of hydrochar properties from high-moisture municipal solid waste: Application of machine learning on waste-to-resource. J Clean Prod 278, 123928. https://doi.org/10.1016/j.jclepro.2020.123928

Li, L., Flora, J.R.V., Berge, N.D., 2020. Predictions of energy recovery from hydrochar generated from the hydrothermal carbonization of organic wastes. Renew Energy 145, 1883–1889. https://doi.org/10.1016/j.renene.2019.07.103

Li, L., Flora, J.R.V., Caicedo, J.M., Berge, N.D., 2015. Investigating the role of feedstock properties and process conditions on products formed during the hydrothermal carbonization of organics using regression techniques. Bioresour Technol 187, 263–274. https://doi.org/10.1016/J.BIORTECH.2015.03.054

Li, L., Wang, Y., Xu, J., Flora, J.R.V., Hoque, S., Berge, N.D., 2018. Quantifying the sensitivity of feedstock properties and process conditions on hydrochar yield, carbon content, and energy content. Bioresour Technol 262, 284–293. https://doi.org/10.1016/J.BIORTECH.2018.04.066

Li, M., Wei, D., Liu, T., Liu, Y., Yan, L., Wei, Q., Du, B., Xu, W., 2019. EDTA functionalized magnetic biochar for Pb(II) removal: Adsorption performance, mechanism and SVM model prediction. Sep Purif Technol 227, 115696. https://doi.org/10.1016/j.seppur.2019.115696

Liao, M., Kelley, S., Yao, Y., 2020. Generating Energy and Greenhouse Gas Inventory Data of Activated Carbon Production Using Machine Learning and Kinetic Based Process Simulation. ACS Sustain Chem Eng 8, 1252–1261. https://doi.org/10.1021/acssuschemeng.9b06522

Liao, M., Kelley, S.S., Yao, Y., 2019. Artificial neural network based modeling for the prediction of yield and surface area of activated carbon from biomass. Biofuels, Bioproducts and Biorefining 13, 1015–1027. https://doi.org/10.1002/BBB.1991

Liu, Q., Liu, B., Zhang, Y., Hu, T., Lin, Z., Liu, G., Wang, X., Ma, J., Wang, H., Jin, H., Ambus, P., Amonette, J.E., Xie, Z., 2019. Biochar application as a tool to decrease soil nitrogen losses (NH 3 volatilization, N 2 O emissions, and N leaching) from croplands: Options and mitigation strength in a global perspective. Glob Chang Biol 25, 2077–2093. https://doi.org/10.1111/gcb.14613

Mathew, S., Karandikar, P.B., Kulkarni, N.R., 2020. Modeling and Optimization of a Jackfruit Seed-Based Supercapacitor Electrode Using Machine Learning. Chem Eng Technol 43, 1765–1773. https://doi.org/10.1002/ceat.201900616

Maulana Kusdhany, M.I., Lyth, S.M., 2021. New insights into hydrogen uptake on porous carbon materials via explainable machine learning. Carbon N Y 179, 190–201. https://doi.org/10.1016/j.carbon.2021.04.036

Mazaheri, H., Ghaedi, M., Ahmadi Azqhandi, M.H., Asfaram, A., 2017. Application of machine/statistical learning, artificial intelligence and statistical experimental design for the modeling and optimization of methylene blue and Cd(ii) removal from a binary aqueous solution by natural walnut carbon. Physical Chemistry Chemical Physics 19, 11299–11317. https://doi.org/10.1039/c6cp08437k

Mojiri, A., Kazeroon, R.A., Gholami, A., 2019. Cross-linked magnetic chitosan/activated biochar for removal of emerging micropollutants from water: Optimization by the artificial neural network. Water (Switzerland) 11, 1–18. https://doi.org/10.3390/w11030551

Mojiri, A., Ohashi, A., Ozaki, N., Aoi, Y., Kindaichi, T., 2020. Integrated anammox-biochar in synthetic wastewater treatment: Performance and optimization by artificial neural network. J Clean Prod 243, 118638. https://doi.org/10.1016/j.jclepro.2019.118638

Murphy, K.P., 2022. Probabilistic Machine Learning: An introduction. MIT Press.

Nguyen, X.C., Ly, Q.V., Peng, W., Nguyen, V.H., Nguyen, D.D., Tran, Q.B., Huyen Nguyen, T.T., Sonne, C., Lam, S.S., Ngo, H.H., Goethals, P., Le, Q. van, 2021. Vertical flow constructed wetlands using expanded clay and biochar for wastewater remediation: A comparative study and prediction of effluents using machine learning. J Hazard Mater 413. https://doi.org/10.1016/j.jhazmat.2021.125426

Noble, W.S., 2006. What is a support vector machine? Nature Biotechnology 2006 24:12 24, 1565–1567. https://doi.org/10.1038/nbt1206-1565

Palansooriya, K.N., Li, J., Dissanayake, P.D., Suvarna, M., Li, L., Yuan, X., Sarkar, B., Tsang, D.C.W., Rinklebe, J., Wang, X., Ok, Y.S., 2022. Prediction of Soil Heavy Metal Immobilization by Biochar Using Machine Learning. Environ Sci Technol 56, 4187–4198. https://doi.org/10.1021/acs.est.1c08302

Parveen, N., Zaidi, S., Danish, M., 2017. Development of SVR-based model and comparative analysis with MLR and ANN models for predicting the sorption capacity of Cr(VI). Process Safety and Environmental Protection 107, 428–437. https://doi.org/10.1016/j.psep.2017.03.007

Pathy, A., Meher, S., P, B., 2020. Predicting algal biochar yield using eXtreme Gradient Boosting (XGB) algorithm of machine learning methods. Algal Res 50, 102006. https://doi.org/10.1016/j.algal.2020.102006

Prakash, N., Manikandan, S.A., Govindarajan, L., Vijayagopal, V., 2008. Prediction of biosorption efficiency for the removal of copper (II) using artificial neural networks 152, 1268–1275. https://doi.org/10.1016/j.jhazmat.2007.08.015

Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence 2019 1:5 1, 206–215. https://doi.org/10.1038/s42256-019-0048-x

Schapire, R.E., Rochery, M., Rahim, M., Gupta, N., 2002. Incorporating Prior Knowledge into Boosting.

Selvarajoo, A., Muhammad, D., Arumugasamy, S.K., 2020. An experimental and modelling approach to produce biochar from banana peels through pyrolysis as potential renewable energy resources. Model Earth Syst Environ 6, 115–128. https://doi.org/10.1007/s40808-019-00663-2

Shen, X., Foster, T., Baldi, H., Dobreva, I., Burson, B., Hays, D., Tabien, R., Jessup, R., 2019. Quantification of soil organic carbon in biochar-amended soil using ground penetrating radar (GPR). Remote Sens (Basel) 11, 1–12. https://doi.org/10.3390/rs11232874

Sigmund, G., Gharasoo, M., Hüffer, T., Hofmann, T., 2020. Deep Learning Neural Network Approach for Predicting the Sorption of Ionizable and Polar Organic Pollutants to a Wide Range of Carbonaceous Materials. Environ Sci Technol 54, 4583–4591. https://doi.org/10.1021/acs.est.9b06287

Talebkeikhah, F., Rasam, S., Talebkeikhah, M., Torkashvand, M., Salimi, A., Moraveji, M.K., 2020. Investigation of effective processes parameters on lead (II) adsorption from wastewater by biochar in mild air oxidation pyrolysis process. Int J Environ Anal Chem. https://doi.org/10.1080/03067319.2020.1777291

Thiruvengadam, S., Edmund Murphy, M., Tan, J.S., 2021. Mathematically modelling pyrolytic polygeneration processes using artificial intelligence. Fuel 295, 120488. https://doi.org/10.1016/j.fuel.2021.120488

Tsekos, C., Tandurella, S., de Jong, W., 2021. Estimation of lignocellulosic biomass pyrolysis product yields using artificial neural networks. J Anal Appl Pyrolysis 157, 105180. https://doi.org/10.1016/j.jaap.2021.105180

Wehrle, R., Welp, G., Pätzold, S., 2021. Total and Hot-Water Extractable Organic Carbon and Nitrogen in Organic Soil Amendments: Their Prediction Using Portable Mid-Infrared Spectroscopy with Support Vector Machines. Agronomy 11, 659. https://doi.org/10.3390/agronomy11040659

Zhang, K., Zhong, S., Zhang, H., 2020. Predicting Aqueous Adsorption of Organic Compounds onto Biochars, Carbon Nanotubes, Granular Activated Carbons, and Resins with Machine Learning. Environ Sci Technol 54, 7008–7018. https://doi.org/10.1021/acs.est.0c02526

Zhang, Z., Schott, J.A., Liu, M., Chen, H., Lu, X., Sumpter, B.G., Fu, J., Dai, S., 2019. Prediction of Carbon Dioxide Adsorption via Deep Learning. Angewandte Chemie International Edition 58, 259–263. https://doi.org/10.1002/anie.201812363

Zhao, Y., Li, Y., Fan, D., Song, J., Yang, F., 2021. Application of kernel extreme learning machine and Kriging model in prediction of heavy metals removal by biochar. Bioresour Technol 329, 124876. https://doi.org/10.1016/j.biortech.2021.124876

Zhou, M., Gallegos, A., Liu, K., Dai, S., Wu, J., 2020. Insights from machine learning of carbon electrodes for electric double layer capacitors. Carbon N Y 157, 147–152. https://doi.org/10.1016/j.carbon.2019.08.090

Zhu, X., Li, Y., Wang, X., 2019a. Machine learning prediction of biochar yield and carbon contents in biochar based on biomass characteristics and pyrolysis conditions. Bioresour Technol 288, 121527. https://doi.org/10.1016/J.BIORTECH.2019.121527

Zhu, X., Tsang, D.C.W., Wang, L., Su, Z., Hou, D., Li, L., Shang, J., 2020. Machine learning exploration of the critical factors for CO2 adsorption capacity on porous carbon materials at different pressures. J Clean Prod 273, 122915. https://doi.org/10.1016/j.jclepro.2020.122915

Zhu, X., Wan, Z., Tsang, D.C.W., He, M., Hou, D., Su, Z., Shang, J., 2021. Machine learning for the selection of carbon-based materials for tetracycline and sulfamethoxazole adsorption. Chemical Engineering Journal 406, 126782. https://doi.org/10.1016/j.cej.2020.126782

Zhu, X., Wang, X., Ok, Y.S., Sik, Y., 2019b. 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