Zhang *et al. Biochar* (2025) 7:61
https://doi.org/10.1007/s42773-025-00446-2

Biochar

ORIGINAL RESEARCH Open Access

Robust biochar yield and composition prediction via uncertainty-aware ResNet-based autoencoder

Yali Zhang[1], Bowen Lei[1], Amirhossein Mahdaviarab[1], Xiao Wang[1] and Zong Liu[1*]

Abstract

With growing demand for energy and fossil fuels, biomass and biochar are gaining attention due to their abundance and sustainability. However, there is a crucial need for optimizing production conditions and minimizing environmental risks associated with biochar production. Machine learning is an emerging strategy for predicting biochar yield and composition, optimizing production conditions and minimizing environmental risks. This paper presents a ResNet-based autoencoder model that utilizes biomass properties and pyrolysis conditions to more accurately and robustly predict biochar yield and composition. The developed model has the advantage of addressing the common data uncertainty concerns in training data. Our model outperforms commonly used baseline methods, including MLP-NN (mean $R^2 = 0.907$), Random Forest (RF, mean $R^2 = 0.798$), XGBoost (XGB, mean $R^2 = 0.826$), and Gaussian Process (GP, mean $R^2 = 0.786$), by achieving a mean R^2 of 0.974. The performance of the model was further improved by incorporating previously discarded data with high missing rates, achieving an average R^2 of 0.983. The addition of the three newly collected covariates resulted in an average R^2 of 0.985. Additionally, robust sensitivity analyses of the input covariates revealed the impact of data uncertainty on the performance of the model, emphasizing the robustness of the model. In advancing the application of machine learning in biochar research, this study provides a reliable method to determine optimal production conditions.

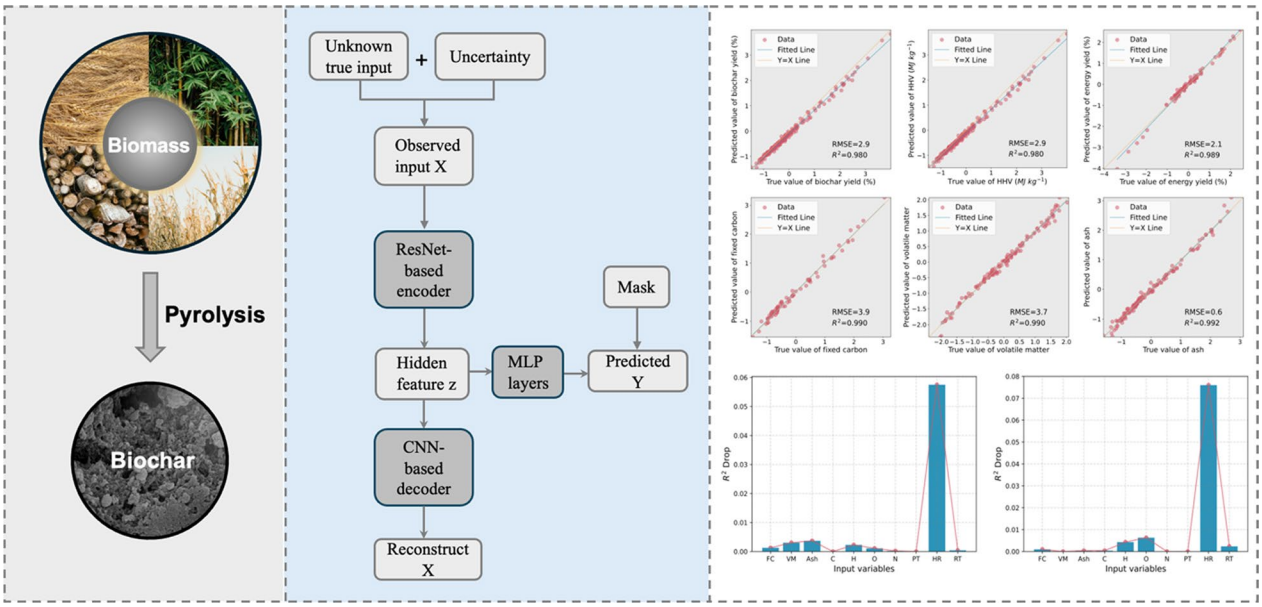
Highlights

- A ResNet-based model enhanced accuracy in biochar yield prediction, addressing data uncertainty robustly
- This model provided better predictive performance under various levels of data uncertainty, including missing values and deviations
- A masking strategy and additional covariates optimized data utilization, leading to more accurate biochar production condition predictions

Keywords Biochar yield, Data uncertainty, Robustness, ResNet, Autoencoder

*Correspondence:
Zong Liu
zong.liu@ag.tamu.edu
[1]Texas A&M University, 375 Olsen Blvd, College Station, TX 77845, USA.

Graphical Abstract



1 Introduction

Biochar is a carbon-rich material generated from the pyrolysis of organic feedstocks, which has attracted much attention due to its remarkable environmental friendliness and sustainability (Leng et al. 2021; Seow et al. 2022). Biochar is featured with high structural stability, large specific surface area, rich porous structure, and abundant functional groups (Liu et al. 2015; Wen et al. 2023). With these superior physicochemical properties, biochar has been widely utilized in many applications, such as water pollution treatment, soil improvement, catalyst utilization, waste management, etc. (Panwar and Pawar 2020; Qambrani et al. 2017; Yu et al. 2023). Despite the fact that biochar is often considered environmentally friendly, there are growing concerns about its potential environmental risks due to the production of harmful components such as heavy metals, polycyclic aromatic hydrocarbons, and dioxins (Qadeer et al. 2017). These toxic compounds can be formed if biomass feedstocks, preparation conditions, and pyrolysis methods are not carefully selected (Liu et al. 2017). Temperature and feedstock origin during pyrolysis have significant impacts on the formation of these harmful components (Qadeer et al. 2017; Yargicoglu et al. 2015). Therefore, selecting appropriate experimental conditions is crucial in maximizing biochar production and reducing their toxic effects.

However, it is necessary to find new methods for predicting the yield and composition of biochar in order to produce it more efficiently and to prevent the production of harmful components. Biochar production conditions exhibit considerable variability in different studies, including a wide range of techniques such as pyrolysis, hydrothermal carbonization, gasification, torrefaction and flash carbonization. These processes also vary by operating parameters, including temperature (200-1200 °C), residence time (0.01-240 min) and heating rate (0.1-300 °C min^{-1}) (Güleç et al. 2022; Safarian 2023; Yaashikaa et al. 2020). Since there are multiple factors to be considered in determining an effective method, it is inefficient to determine the optimal conditions for biochar preparation by conducting laboratory experiments on all possible conditions. Therefore, we need to find a new method to predict the yield and composition of biochar while saving time cost and laboratory experiment costs. Currently, machine learning algorithms are receiving increasing attention and can be used to predict biochar yield and composition, which will aid in identifying optimal production conditions and reducing the potential environmental risks associated with biochar production (Chen et al. 2023a, b). Utilizing these technologies can help to ensure that biochar remains an environmentally friendly solution.

While the data-driven machine learning algorithms have become more and more popular and studies have

been done to investigate the relationship between biochar yield and composition and pyrolysis conditions, the advances in machine learning have not been fully leveraged. Random forest (RF) is a popular ensemble learning method that constructs multiple weak learners and aggregates them into a stronger one (Breiman 2001). This model investigated influencing factors in the pyrolysis process, which showed $R^2=0.855$ for the biochar yield based on 245 datasets (Zhu et al. 2019). Extreme Gradient Boosting (XGBoost) is a widely used extended version of gradient boosting decision tree (Chen et al. 2015a, b). A prior work explored the XGB model and successfully predicted biochar yield from co-pyrolysis of biomass and plastics and it had a $R^2=0.96$ using 94 datasets (Alabdrabalnabi et al. 2022). Additionally, Gaussian Process (GP) is also a broadly used non-parametric Bayesian method to fit unknown functions (Williams & Rasmussen 1995). A recent work using GP model incorporated with the exponential kernel function had a preformation with $R^2=0.767$ and RMSE = 9.021 (Kanthasamy et al. 2023). However, they are shallow machine learning methods that are limited by the expressive power of their own models and may not fully capture complex data patterns. Deep learning methods are also explored. For example, multilayer Perceptron Neural Network (MLP-NN) is an MLP-layer based deep neural network, which provided state-of-art predictive performance in biochar prediction (Li et al. 2022). Li et al. combined data from different sources and for the first time comprehensively predicted biochar yield and composition (Li et al. 2022). However, MLP-based deep learning methods may suffer from the problem of gradient vanishing, especially when increasing the model depth. To avoid gradient vanishing, model depth as well as expressiveness would be limited.

In addition to model architecture, another important factor in the predictive performance of data-driven methods, dataset uncertainty, has not been fully explored in existing research. This may prevent us from using data-driven methods to accurately explore intrinsic relationships and may lead to performance degradation, especially when data uncertainty is high. In general, uncertainty arises from two sources: (1) missing values in the inputs and outputs, and (2) values that deviate from the true values, which is usually inevitable in real data, especially for multi-source datasets with mismatched variables or experimental environments. For example, Khan et al. did not pay enough attention to the biochar data uncertainty that makes the observed values deviate from the true values during data analysis (Khan et al. 2022), which might sacrifice some prediction accuracy. Alabdrabalnabi et al., in the same way, directlty utilized data-driven method on the data without properly handling the uncertainty within the data. Then, although Li et al. took data uncertainty into account when modeling multi-source data where missing data become more prevalent, they naively dropped variables with high missing rates, which can lead to the loss of important covariates and waste information valuable for biochar prediction (Li et al. 2022). Also, Zhu et al. (2019) treated the multi-source data uncertainty similarly and simply excluded the unsolved missing data, not fully leveraging the valuable information in the data and leading to accuracy degradation. Therefore, we need a robust method to unveil the underlying mechanism masked by the uncertainty in experimental findings.

To address the issue of data uncertainty and enhance biochar prediction accuracy, a multiobjective ResNet-based Autoencoder model is proposed. The model demonstrates enhanced robustness to data uncertainty and predicts biochar yield, proximate composition, and final composition with higher accuracy. Additionally, it distills more informative features through the convolutional neural network, enabling more precise modeling for multiple targets via task-specific multilayer perceptrons (MLPs), and further strengthens robustness against uncertainty disturbances through Autoencoder structure and masking strategy. In this way, valuable data can be preserved, enabling the utilization of all the data collected from various sources. Experimental results yielded a lower root mean squared error (RMSE) and higher R^2 across varying noise levels compared to widely used baseline methods, underscoring the effectiveness of the proposed method. This study also provides guidance for researchers in optimizing biochar treatment conditions, thereby expanding the application range of biochar and enhencing its potential value (Xiang et al. 2021). This study was carried out (1) to design a robust deep learning method ResNet-based Autoencoder to predict biochar yield and quality; (2) to investigate the effect of data uncertainty on the prediction performance of machine learning algorithms.

2 Materials and methods

2.1 Dataset and preprocessing

The modelling is based on the 226 samples collected from multiple sources by Li et al., along with three additional input covariates that is eliminated by Li et al. due to high data uncertainty. The 226 samples contained 30 different biomass feedstocks, such as agricultural residues, forestry residues and poultry manure. More specifically, the original input covariates were from biomass characteristics and pyrolysis conditions, which included (1) proximate analysis, i.e., ash, fixed carbon (FC), volatile matter (VM), and moisture (2) feedstock ultimate composition, i.e., H-O-N-S-C (3) feedstock structural composition, i.e., cellulose (Cel), hemicellulose (Hem),

and lignin (Lig) (4) feedstock properties, i.e., feedstock type, size, and pH (5) pyrolysis condition, i.e., pyrolysis temperature (PT), heating rate (HR), and residence time (RT). Our output variables came from (1) biochar proximate composition: ash, FC, and VM (2) biochar ultimate composition, H-O-N-S-C (3) process efficiency: biochar yield, energy yield, and higher heating value (HHV). In addition, we searched among the multiple sources used in Li et al. (2022) and collected three new input covariates about feedstock properties, including feedstock type, size (mm), and pH.

Data uncertainty is prevalent in these 226 samples. Previous (Li et al. 2022) directly eliminated variables with high data uncertainty, which could results in a waste of valuable data and missing important factors. In contrast to the direct elimination, we used a method that keeps all the variables with low and high uncertainty. Take the missing values as an example of data uncertainty, we summarized the missing rates for each input and output variable. As shown in Table 1, for eight of the 18 input variables, some data were missing from the original dataset, i.e., moisture missing rate was 57.96%. Cellulose, hemicellulose, and lignin data sets missing rates were 50.44%, 52.65%, 50.44%, respectively. For pH, the missing rate was high to 94.25%. The same is true for 10 of the 11 outputs. Sulfur (S), energy yield, HHV, the missing rate were 49.12%, 61.50%, 53.54%, respectively. More detailed information is listed in the Table 1. As we can see, the problem of missing input and output variables was significant and needed to be dealt with appropriately in the data modeling methods.

We also conducted the Pearson correlation coefficient analysis to quantify the correlation between input and output variables. As shown in Fig. 1, heating rate (HR) has large correlation with multiple output variables, like biochar yield (BY), fixed carbon (FC), volatile matter (VM), C, H, and O. Also, input ash is also highly related to output ash and C. In addition, input N is relavant to output N. This analysis demonstrates the existence of correlations between input and output variables.

The data preprocessing before modeling mainly included missing value processing and normalization. We first addressed the issue of missing data and recorded the location of missing data by setting masks of 0 and 1 for both input covariates and output data, respectively, where 0 denotes missing values at the corresponding location and 1 denotes observed values. In this way, the mask can be used as a gate during the training process to filter out data uncertainty caused by missing data without discarding valuable data. This made it possible to both fully utilize the data and use only the observations to guide the updating of the model parameters. Second, following the data preprocessing process widely used in machine learning methods, we normalized all inputs and outputs so that their mean and variance were equal to 0 and 1, respectively.

Table 1 Input and output variable for the raw dataset

Type	Input	Missing rate (%)	Type	Output	Missing rate (%)
Feedstock proximate analysis	ash (db.)	0.00	Biochar proximate composition	ash (db.)	29.65
	FC (% bd.)	0.00		FC (% bd.)	29.65
	VM (% bd.)	0.00		VM (% bd.)	29.65
	moisture (% bd.)	57.96			
Feedstock ultimate composition	H (% bd.)	0.00	Biochar ultimate composition	H (% bd.)	33.63
	O (% bd.)	7.08		O (% bd.)	33.63
	N (% bd.)	0.00		N (% bd.)	28.32
	S (% bd.)	32.74		S (% bd.)	49.12
	C (% bd.)	0.00		C (% bd.)	28.32
Feedstock structural composition	Cel (% bd.)	50.44	Process efficiency	Biochar yield (%)	0.00
	Hem (% bd.)	52.65		Energy yield (%)	61.50
	Lig (% bd.)	50.44		HHV (MJ/kg)	53.54
Feedstock properties	feedstock type	0.00			
	size (mm)	8.85			
	pH	94.25			
Pyrolysis conditions	PT (°C)	0.00			
	HR (°C/min)	0.00			
	RT (min)	0.00			

Missing rate: ratio of missing data to total data

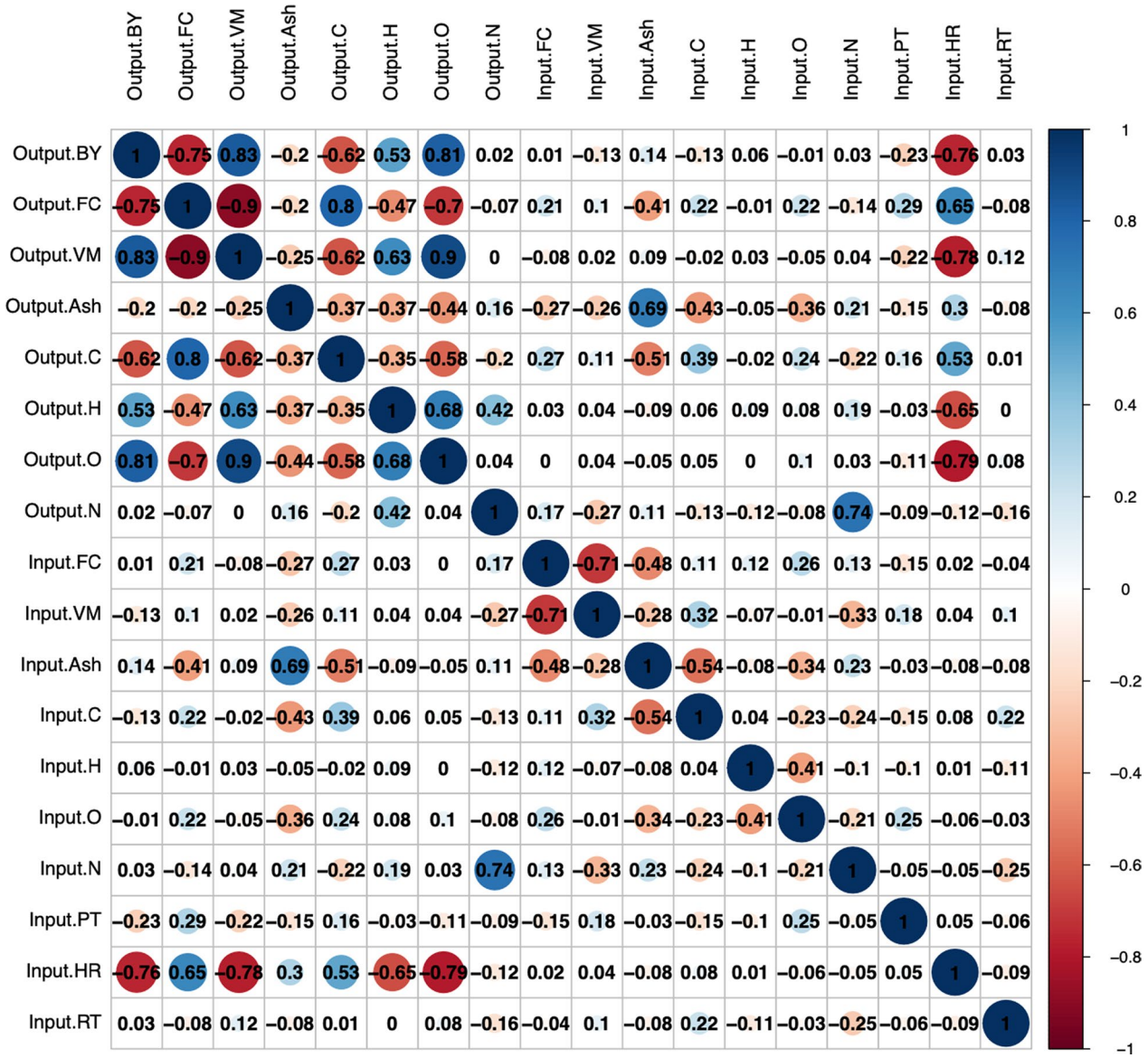


Fig. 1 Pearson correlation matrix between input variable and output variables, including Fixed carbon (FC), Volatile matter (VM), Ash, Carbon (C), Hydrogen (H), Oxygen (O), Nitrogen (N), Pyrolysis temperature (PT), Heating rate (HT), Residence time (RT), and Biochar yield (BY)

2.2 Feature extraction: ResNet-based autoencoder

To extract valuable features from biochar data, we chose to use the residual network (ResNet) (He et al. 2016), which is a popular deep convolutional network that has been widely used for various tasks (Lu et al. 2020; Ma et al. 2020; Liu et al. 2020). Previously, deep neural networks (DNNs) could not fully demonstrate their strengths due to gradient vanishing (i.e., gradients become zero at deeper layers) and accuracy saturation challenges (i.e., deeper networks have higher training errors and testing errors). To accomplish the limitations, researchers designed shortcuts to jump over blocks of layers. A regular block of DNN mapped input feature x to the response via a combination of weight layer and activation function. In contrast to the regular block, ResNet utilized a residual block which added x to the final output by the shortcut and the weight layers actually are fitting the f (x) – x. The main feature of residual connectivity is that it provides short paths from early to late layers. In this way, every additional layer was able to more easily contain the identity function as one of its elements, enabling the deeper model to perform at least as well as the shallower model. Therefore, the counter-intuitive phenomena in accuracy saturation will not appear any more.

More specifically, the architecture of ResNet (He et al. 2016) we utilized includes 50 layers. The first layer has a convoultion with a 7×7×64 kernel and a stride of size 2 and a max pooling operation. And 9 layers follow and are grouped into 3 similar blocks. Each block has a convolution with a 1×1×64 kernel, a convolution with a 3×3×64 kernel, and a convolution with a 1×1×256 kernel. Then, 12 layers follow and are grouped into 4 similar blocks. Each block has a convolution with a 1×1×128 kernel, a convolution with a 3×3×128 kernel, and a convolution with a 1×1×512 kernel. After this, 18 layers follow and are grouped into 6 similar blocks. Each block has a convolution with a 1×1×256 kernel, a convolution with a 3×3×256 kernel, and a convolution with a 1×1×1024 kernel. Next, 9 layers follow and are grouped into 3 similar blocks. Each block has a convolution with a 1×1×512 kernel, a convolution with a 3×3×512 kernel, and a convolution with a 1×1×2048 kernel. Lastly, a dense layer at the end is included to map the feature to the output size.

Additionally, we also proposed to incorporate the ResNet in an Autoencoder (Blaschke et al. 2018; Yang et al. 2022; Minarno et al. 2021) structure to better deal with the data uncertainty. Since the data were collected from several independent efforts, there were some inconsistencies in the experimental conditions, which usually leaded to data uncertainty. An autoencoder is an artificial neural network that distills key information from data X in an unsupervised manner and discards unimportant knowledge. The features represent the key information and was verified and refined by trying to reconstruct the input X from the feature, where the unimportant information (noise) was usually ignored and key feature was distilled. In this way, the data uncertainty like the value deviation would be ignored by the model and data uncertainty issue is better handled.

We further incorporated both unsupervised and supervised losses in the ResNet-based Autoencoder to accommodate our supervised task. Specifically, the unsupervised loss was defined as the l_2 difference between the original input X and the reconstructed X′, i.e., $||X - X\prime||_2^2$. For the supervised loss, it was defined as the l_2 difference between the labels Y and predicted Y′ which was obtained from the hidden features via additional layers. In this way, with a designed masking strategy (in Sect. 2.4), we can simultaneously deal with the label prediction and the data uncertainty.

Our proposed ResNet-based Autoencoder focuses on model architecture and is orthogonal to the study of different loss function, like Huber Loss (Huber 1992). Huber loss can make model more robust to one kind of data uncertainty, which is shown as values deviating from the true values. Future research could be done to incorporate Huber loss in our ResNet-based Autoencoder model and see how robust the model would be. However, Huber loss usually is not good at handling another type of data uncertainty, i.e., missing data. The robustness of the Huber loss mainly comes from the replacement of l_2 loss with l_1 loss when faced with outliers, which does not have advantages in dealing with missing data. On the contrary, our method is able to tackle with both of the two types of data uncertainty.

2.3 Masking strategy: handling missing values

With the masks obtained during the data preprocessing step, we designed a masking strategy to deal with data missing in our collected data during training. Specifically, taking output Y as an example, we defined a mask M that has the same shape as Y. If certain element of Y is missing, we set the corresponding element of M to zero. Otherwise, the element of M is one. Then, when we calculated the l_2 difference between the Y and predicted Y′, we changed it to $||\mathrm{Y} \odot \mathrm{M} - \mathrm{Y}' \odot \mathrm{M}||^2$, where ⊙ means multiplying the corresponding elements. In this way, we do not need to throw away inputs and outputs with high missing ratios so as to fully extract the information in the data. The data uncertainty is properly handled given that the missing elements play no role in the loss meaning that they could not lead the model to any wrong direction. The workflow of the ResNet-based autoencoder framework and how data uncertainty affects observed input variables and responses is illustrated in Fig. 2.

2.4 Model performance evaluation metrics

To compare the predicted performance for biochar, we followed Li et al. and chose two metrics, i.e., RMSE and R^2, which were widely used to measure the prediction capacity of regression models (Li et al. 2022). Specifically, RMSE represents the difference between predicted labels and the true labels as shown in Eq. 1, where smaller RMSE indicates better regression model. As for R^2, it shows what percentage of information with the true labels can be explained by the model as shown in Eq. 2, where larger RMSE R^2 implies stronger regression model.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \hat{Y}_i\right)^2} \tag{1}$$

$$R^2 = \left[\frac{\frac{1}{N}\left(Y_i - \hat{Y}_i\right)^2}{\frac{1}{N}\left(Y_i - \overline{Y}\right)^2}\right] \tag{2}$$

where $\{Y_1, \cdots, Y_N\}$ is the true labels, $\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$ representing the true labels' mean, $\{\widehat{Y}_1, \cdots, \widehat{Y}_N\}$ denotes the

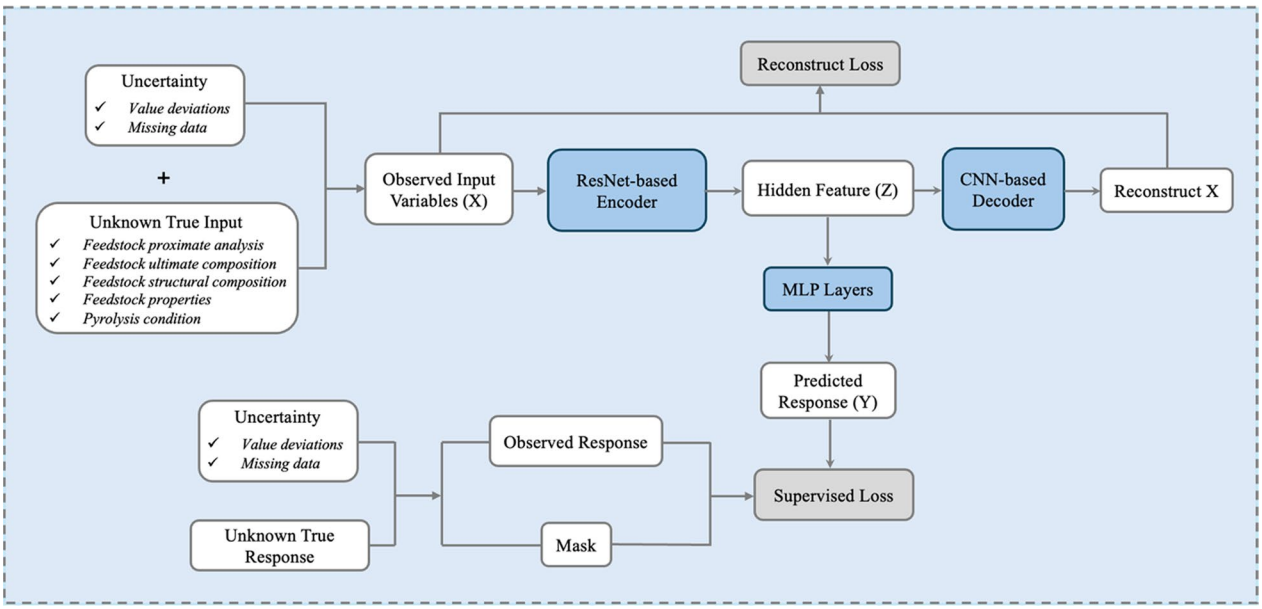


Fig. 2 Research logic diagram of our robust biochar yield and composition prediction. It illustrates how data uncertainty affects observed input variables and responses, and describes the workflow of the ResNet-based autoencoder framework

predicted labels from the regression model, and N is the number of testing samples.

2.5 Comparison with data enhancement

Data enhancement is needed to avoid model overfitting and improve generalization when faced with small data (Chen, et al 2023a, b). Our study is orthogonal to data enhancement and can be combined with data enhancement methods. More specifically, our study focuses on dealing with data uncertainty including (1) missing values in the inputs and outputs, and (2) values that deviate from the true values. However, data enhancement is not designed to handle these two types of data uncertainty. In the data modeling, if the variables with high data uncertainty are eliminated, using the data enhancement method cannot generate new information and make up the valuable information lost. Thus, it is desirable to have our model to better deal with data uncertainty during data analysis.

2.6 Computational complexity

For the computational complexity, since the biochar data size is usually not big data and our model architecture is not a big network, the training and inference are fast and have good feasibility in industrial applications. Take one NVIDIA RTX A6000 as an example in our study. It only takes 5 min to train our model with 250 epochs, and only takes 0.15 s to do inference on the testing set, which demonstrate the feasibility of our method in industrial applications.

3 Results

3.1 Predictive performance results and discussion

3.1.1 Statistical analysis of dataset

The main characteristics of the dataset is summarized in Table 1. The FC, VM, and ash contents varied in the range of 4.33% to 27.8%, 68.2-91.6%, and 0.16-15.14%. The C and O in the biomass feedstock ranged from 35.7% to 64.23%, and 27.61% to 53.1%, respectively. The H, N, S ranged from 4.10% to 10.18%, 0 to 9.61%, and 0.09% to 0.92%, respectively. The distribution of cellulose, hemicellulose and lignin ranged from 17.89% to 47.67%, 11.48–56.29%, and 4.99–32.26%, respectively. The pyrolysis process parameters including RT, PT and HR ranged from 1.00 to 90.00 min, 200-800 °C, and 5.00-25.00 °C min^{-1}, respectively. These data points allowed optimization of the operating parameters of the feedstock and pyrolysis processes to achieve the potential for high biochar yields. The wide range of values for each variable gave the model enough data to learn from a variety of data and thus helped in generalization.

3.1.2 Predictive performance of ResNet-based autoencoder

The predictive performance of our ResNet-based Autoencoder is demonstated via parity plots (Fig. 3). In contrast to previous work that could only predict eight outputs with low deletion rates, our method can predict all eleven outputs with low RMSE and large R^2. In Fig. 3, we showed the parity plots for process efficiency (i.e., biochar yield, HHV, and energy yield), biochar proximate composition (i.e., fixed carbon, volatile matter, and ash), and ultimate

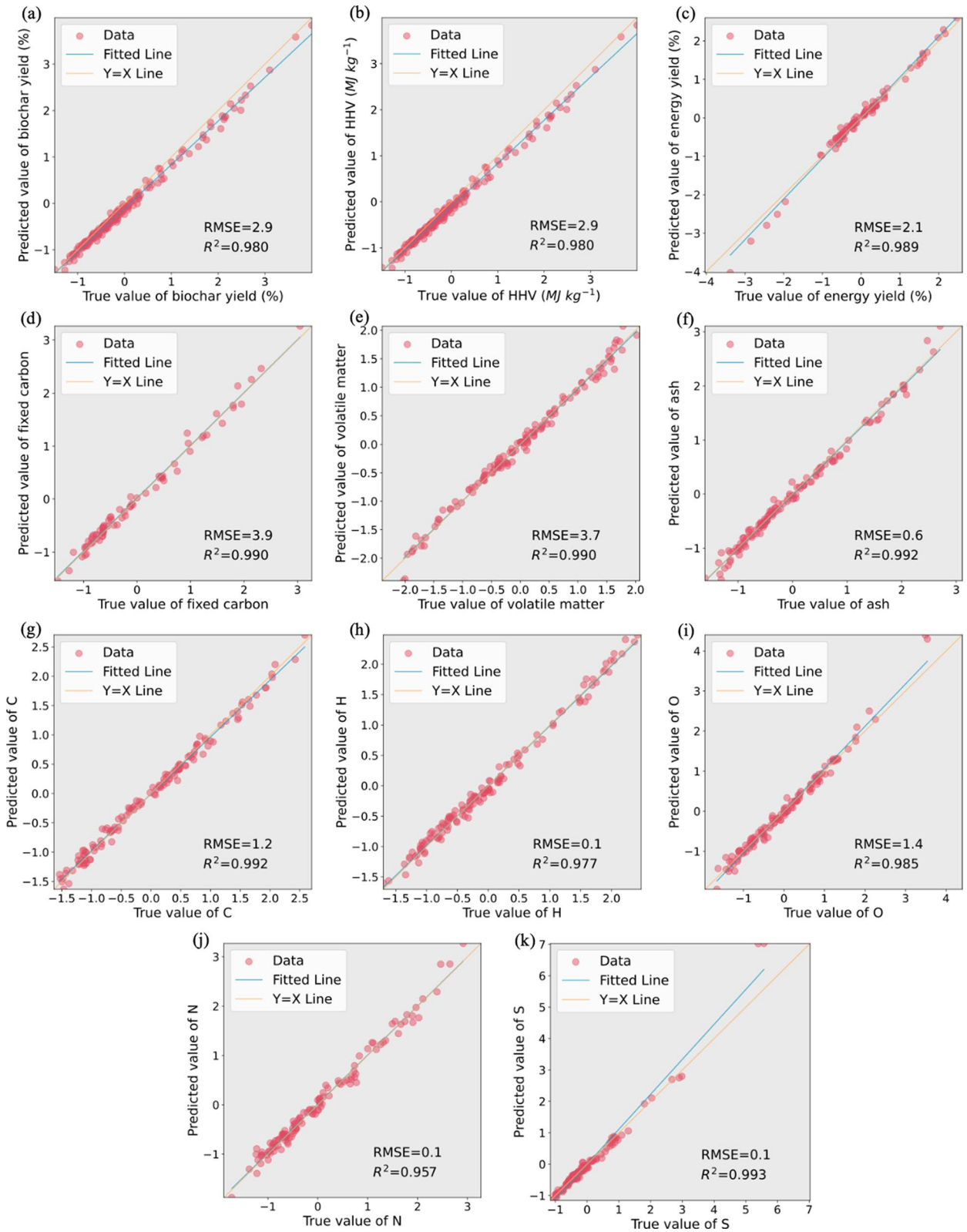


Fig. 3 Parity plots for ResNet-based Autoencoder comparing the actual and predicted values of process efficiency, biochar proximate composition, and ultimate analysis including (**a**) Biochar yield (%), (**b**) HHV (MK/kg), (**c**) Energy yield (%), (**d**) Fixed Carbon (vol. % db.), (**e**) Volatile Matter (vol. % db.), and (**f**) Ash (vol. % db.), (**g**) C (vol. % db.), (**h**) H (vol. % db.), (**i**) O (vol. % db.), (**j**) N (vol. % db.), and (**k**) S (vol. % db.)

analysis. Our method maintained good predictive performance across different outputs. Specifically, as depicted in Fig. 3, we achieved (a) RMSE = 2.9 & R^2 = 0.980 for biochar yield (%), (b) RMSE = 0.6 & R^2 = 0.983 for HHV (MJ/kg), (c) RMSE = 2.1 & R^2 = 0.989 for energy yield (%), (d) RMSE = 3.9 & R^2 = 0.990 for fixed carbon, (e) RMSE = 3.7 & R^2 = 0.990 for volatile matter (vol. % db.), and (f) RMSE = 0.6 & R^2 = 0.992 for ash (vol. % db.). As summarized in Fig. 3, we obtained (a) RMSE = 1.2 & R^2 = 0.992 for C (vol. % db.), (b) RMSE = 0.1 & R^2 = 0.977 for H (vol. % db.), (c) RMSE = 1.4 & R^2 = 0.985 for O (vol. % db.), (d) RMSE = 0.1 & R^2 = 0.957 for N (vol. % db.), and (e) RMSE = 0.1 & R^2 = 0.993 for S (vol. % db.).

3.1.3 Comparison of baseline methods

The predictive performance of the ResNet-based Autoencoder was compared with popular baseline methods, i.e., MLP-NN (Li et al. 2022), Random Forest (Ho 1995), XGBoost (Chen et al. 2015a, b), and Gaussian Process (Williams & Rasmussen 1995). The results of MLP-NN (MLP), Random Forest (RF), XGBoost (XGB), Gaussian Process (GP), and our ResNet-based Autoencoder (Ours) are summarized in Table 2.

To show better predictive performance in our method, we first compared models trained on the same covariates that have lower missing rates. As we can see, the popular shallow machine learning methods (i.e., RF, XGB, and GP) tend to give smaller R^2 and larger RMSE compared to deep neural network methods (i.e., MLP and ours), implying stronger feature extraction ability in deep models. In addition, despite improved performance in MLP, our method (Ours) further significantly increased R^2 and reduced RMSE of MLP, suggesting more accurate biochar prediction in our method.

To further show the ability to handle missing data in our method, we included models trained on full data containing high missing-rate covariates. The previous method (i.e., MLP) droped covariates with missing rates greater than 30% (Li et al. 2022), and we believed that much useful information was thrown away and wasted. Since the data were collected from a wide range of literature, the inconsistencies between datasets were inevitably, leading to data missing. As a result, a portion of the valuable covariates had to be discarded, thus losing important information. In contrast to previous method, our ResNet-based Autoencoder can better handle missing and noisy data. We used all the data without throwing covariates with high missing rate. The results based on full data are denoted by "Ours (f)". As shown in Table 2, Ours (f) can further greatly increase R^2 and reduce RMSE from Ours. This shows the useful information in the dropped covariates and demonstrates the power of our method in handling missing and noisy data.

3.1.4 Performance with extra input variables

We added three new input variables, as described in Sect. 2.1, from multiple data sources that were not selected in the existing work (Li et al. 2022). We compared the model performance trained on data with and without the extra input variables. We denoted the model with the augmented data as "Ours (aug)". As shown in Table 3, Ours (aug) can further slightly increase the R^2 and decrease the RMSE for most of the output variables compared to Ours (f). This shows the useful information in the additional covariates and demonstrates that our method can effectively distill useful information from data with high uncertainty.

Table 2 Predictive performance comparion

		$Yield_{BC}$	FC_{BC}	VM_{BC}	ash_{BC}	C_{BC}	H_{BC}	O_{BC}	N_{BC}
R^2	MLP	0.964	0.903	0.899	0.940	0.918	0.855	0.887	0.888
	RF	0.807	0.836	0.840	0.918	0.905	0.329	0.885	0.869
	XGB	0.810	0.878	0.841	0.941	0.941	0.583	0.776	0.842
	GP	0.866	0.828	0.651	0.606	0.877	0.843	0.849	0.771
	Ours	**0.965**	**0.988**	**0.986**	**0.982**	**0.979**	**0.976**	**0.983**	**0.930**
	Ours (f)	**0.980**	**0.990**	**0.990**	**0.992**	**0.992**	**0.977**	**0.985**	**0.957**
RMSE	MLP	3.4	5.1	6.1	2.0	3.1	0.5	3.3	0.4
	RF	39.1	62.4	58.5	6.8	14.2	1.3	10.9	0.2
	XGB	38.6	46.4	58.1	4.9	8.9	0.8	21.3	0.3
	GP	27.2	65.5	79.0	32.8	18.4	0.3	14.3	0.4
	Ours	**3.4**	**4.4**	**5.0**	**1.5**	**2.9**	**0.1**	**1.6**	**0.1**
	Ours (f)	**2.9**	**3.9**	**3.7**	**0.6**	**1.2**	**0.1**	**1.4**	**0.1**

Bold values represent predicted effects that are greater than or equal to the effects of other methods

Table 3 Predictive performance comparion

		$Yield_{BC}$	FC_{BC}	VM_{BC}	ash_{BC}	C_{BC}	H_{BC}	O_{BC}	N_{BC}
R^2	Ours	0.965	0.988	0.986	0.982	0.979	0.976	0.983	0.930
	Ours (f)	0.980	0.990	0.990	0.992	0.992	0.977	0.985	0.957
	Ours (aug)	**0.989**	**0.990**	**0.990**	**0.992**	**0.986**	**0.983**	**0.986**	**0.960**
RMSE	Ours	3.4	4.4	5.0	1.5	2.9	0.1	1.6	0.1
	Ours (f)	2.9	3.9	3.7	0.6	1.2	0.1	1.6	0.1
	Ours (aug)	**2.3**	**3.9**	**3.7**	**0.5**	**1.0**	**0.1**	**1.3**	**0.1**

Bold values show the best performance in our methods

3.2 Stress test and discussion about robustness

A stress test was designed to further demonstrate the robustness of the method and to evalute its ability to handle data uncertainty. In the stress test, we added two types of data uncertainty to the input covariates and output variables, respectively, simulating the deviation of the observed values from the true values and missing values in the data. With the additional perturbation, we simulated a more stressful scenario for the models where the noise level was increased and the signal noise ratio was reduced. The underlying mapping from input covariates to outputs becomes more difficult to discover, often leading to degradation of model performance. The ideal model with greater robustness should have less performance degradation compared to other models.

3.2.1 Robustness to deviation from true value in input covariates

First, to simulate the increase in deviation from the true value in input covariates, we added Gaussian noise to the input covariates. More specifically, we set the standard deviation of the random Gaussian noise to $\sigma = 0.1$. As described in the data collection and preprocessing section, the inputs had been normalized to have a variance equal to one. Thus, the added uncertainty represented a 10% variance perturbation.

To compare the model performance, we used R^2 as the metric. As shown in Fig. 4a-c, the red curves (raw data) were significantly lower than the blue curves (raw data + additional Gaussian noise) for Random Forest (RF), XGBoost (XGB), and Gaussian process (GP), showing the degraded performance of these three methods in the face of additional noise. This indicated that the baseline methods were not robust and unable to maintain the performance when faced with stressful scenarios about deviation from the true value. On the contrary, compared with the baseline methods, as shown in Fig. 4 (d), the red and blue curves of our method were very close to each other, presenting a much smaller performance degradation in our method. This demonstrated the robustness of our method under stress scenarios, which is crucial for modeling biochar data where data uncertainty from deviation cannot be avoided.

3.2.2 Robustness to missing value in input covariates

Second, to mimic a more severe missing data problem in input covariates, we added additional missing values to the input covariates thereby increasing the missingness rate. More specifically, we randomly selected 10% values in all of the covariates and set the selected values as 0, which represented an additional 10% increase in missing rate.

For the model performance comparison, we used R^2 as an example. As shown in Fig. 5 a-c, the red curves (raw data) were significantly lower than the blue curves (raw data + additional missing values) for Random Forest (RF), XGBoost (XGB), and Gaussian process (GP), depicting the degraded performance of these three methods in the face of additional missing values. This implied that the baseline methods were not robust and unable to maintain the performance when faced with stressful scenarios about missing values. Compared with the baseline methods, as shown in Fig. 5 (d), the red and blue curves of our method were closer to each other, showing a much smaller performance degradation in our method. This demonstrated the robustness of our method to missing values, which was essential for modeling biocarbon data, where data uncertainty due to missing data was prevalent, especially when we combined data from multiple sources.

3.2.3 Robustness to deviation from true value in output variables

When turning to the output variables, to simulate the increase in deviation from the true value, we added Gaussian noise to the output variables. Similar to the setup in Sect. 3.2.1, we set the standard deviation of the random Gaussian noise to $\sigma = 0.1$, thus increasing the uncertainty level by 10%. To compare the model performance, we continued using R^2 as the metric. As presented in Fig. 6a-c, the red curves (raw data) were significantly lower than the blue curves (raw data + additional Gaussian noise) for Random Forest (RF), XGBoost (XGB), and

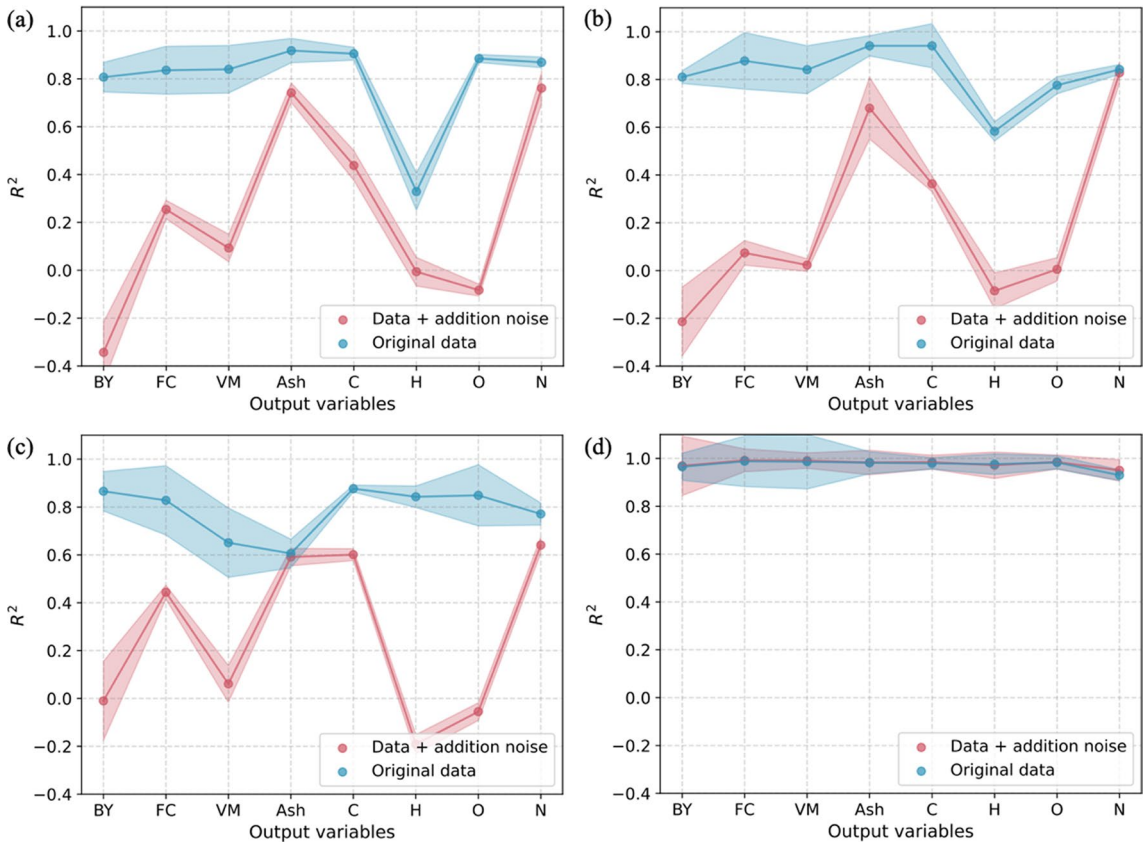


Fig. 4 R^2 comparison between raw input data and raw input data plus additional Gaussian noise (sigma = 0.1) when predicting output variables including Biochar yield (BY), Fixed Carbon (FC), Volatile Matter (VM), Ash, C, H, O, N, S. Methods include (**a**) Random Forest, (**b**) XGBoost, (**c**) Gaussian Process, and (**d**) our ResNet-based Autoencoder (Ours)

Gaussian process (GP), implying the degraded performance of these three baselines in the face of additional uncertainty. In contrast, as depicted in Fig. 4d, the red and blue curves of our method were much closer to each other compared with the other baseline methods, suggesting that the performance degradation of our method was much smaller. This suggested that our method was more robust under stressful scenarios, which was crucial for modeling biochar data, as data uncertainty due to deviations was unavoidable. We also explored the threshold where the performance of our model significantly drops due to the values deviating from the true values. Specifically, we begin to observe model performance degradation when the sigma of additional noise is larger than 4.0, and we see model divergence when the sigma is larger than 10.

3.2.4 Robustness to missing value in output variables

In addition, we also considered the case of more stressful missingness in the output variables, i.e., we simulated an additional 10% increase in the missingness rate by randomly selecting 10% of the values in all the output variables and setting the selected value to zero. Aligned with previous sections, we used R^2 as a comparison metric. As shown in Fig. 7a-c, the red curves (raw data) were significantly lower than the blue curves (raw data + additional missing values) for Random Forest (RF), XGBoost (XGB), and Gaussian process (GP), indicating that the performance of these three methods degraded when faced with additional missing values. In contrast to the baseline methods, as shown in Fig. 7d, the red and blue curves of our method were closer to each other, depicting a much smaller performance degradation in our method. This suggested that our method was more robust to missing values, which was critical for modeling biochar data because data uncertainty due to missing data was very prevalent, especially when we combined data from multiple sources. We also explored the threshold where the performance of our model significantly drops due to the missing data. Specifically, we begin to see slightly larger performance degradation when the additional missing rate is larger than 25%, and we see model divergence when the additional missing rate is larger than 40%.

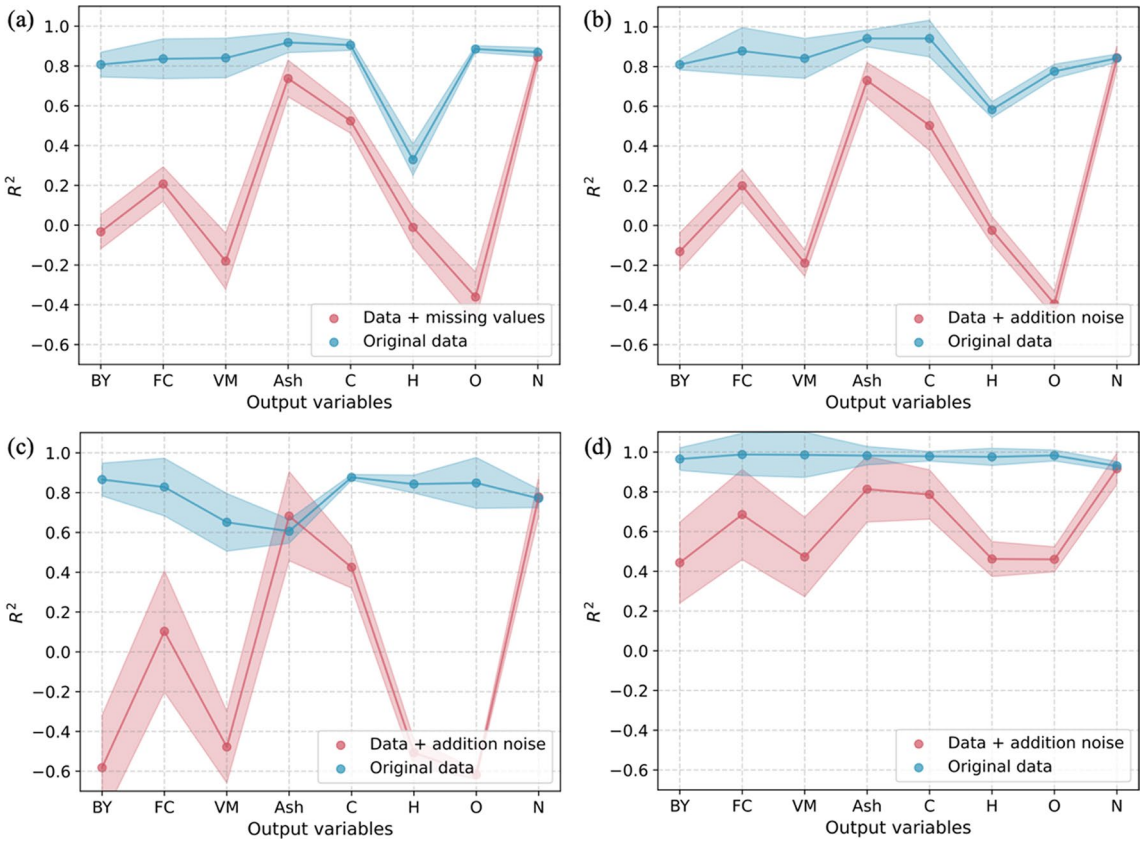


Fig. 5 R^2 comparison between raw input data and raw input data with additional 10% missing values when predicting output variables including Biochar yield (BY), Fixed Carbon (FC), Volatile Matter (VM), Ash, C, H, O, N, S. Methods include (**a**) Random Forest, (**b**)XGBoost, (**c**) Gaussian Process, and (**d**) our ResNet-based Autoencoder (Ours)

3.2.5 Robust sensitivity analysis of input covariates

In order to investigate the effect of increased data uncertainty in each input covariate on the model performance, we also performed robust sensitivity analyses on the input covariates. We took into account both the deviation from true values and the missing values. More specifically, on the one hand, we first added Gaussian noise with $\sigma = 0.2$ to each of the input covariates individually to increase the uncertainty level by 20%. As shown in Fig. 8a, as this type of data uncertainty increased further in the heating rate (HR), the average R^2 of the model decreased the most, by more than 5%. In addition, the R^2 of the model decreased slightly when the uncertainty in the data for volatile matter (VM) and ash was large. Other input covariates were relatively more robust to this data uncertainty, i.e., deviation from the true values. On the other hand, we also added an additional 20% missing rate in each input covariate by randomly selecting 20% of the values in all output variables and setting the selected values to zero. To analyze the sensitivity, we calculated the average R^2 drop for different input covariate uncertainty increments. As shown in Fig. 8b, an increase in the missing rate of the heating rate (HR) also leaded to a relatively large drop in R^2 of more than 7%. Additionally, the average R^2 decreased slightly when the missing rates of H and O increased. Other input covariates were relatively more robust to this data uncertainty, i.e., missing values.

Sensitivity analysis shows how the model performance changes if we increase the data uncertainty of an input variable. If the model performance decreases a lot when increasing the data uncertainty of a variable, it indicates that the model is more sensitive to the data uncertainty of that variable. Therefore, in practice, if the data uncertainty of a variable is large in the data we collect, the model performance is more likely to decrease. In this case, we can spend more time and resources to ensure that the variables are accurately collected and not missed. For other variables that are not sensitive, with limited resources, we can reduce the accuracy a bit or allow partial missingness.

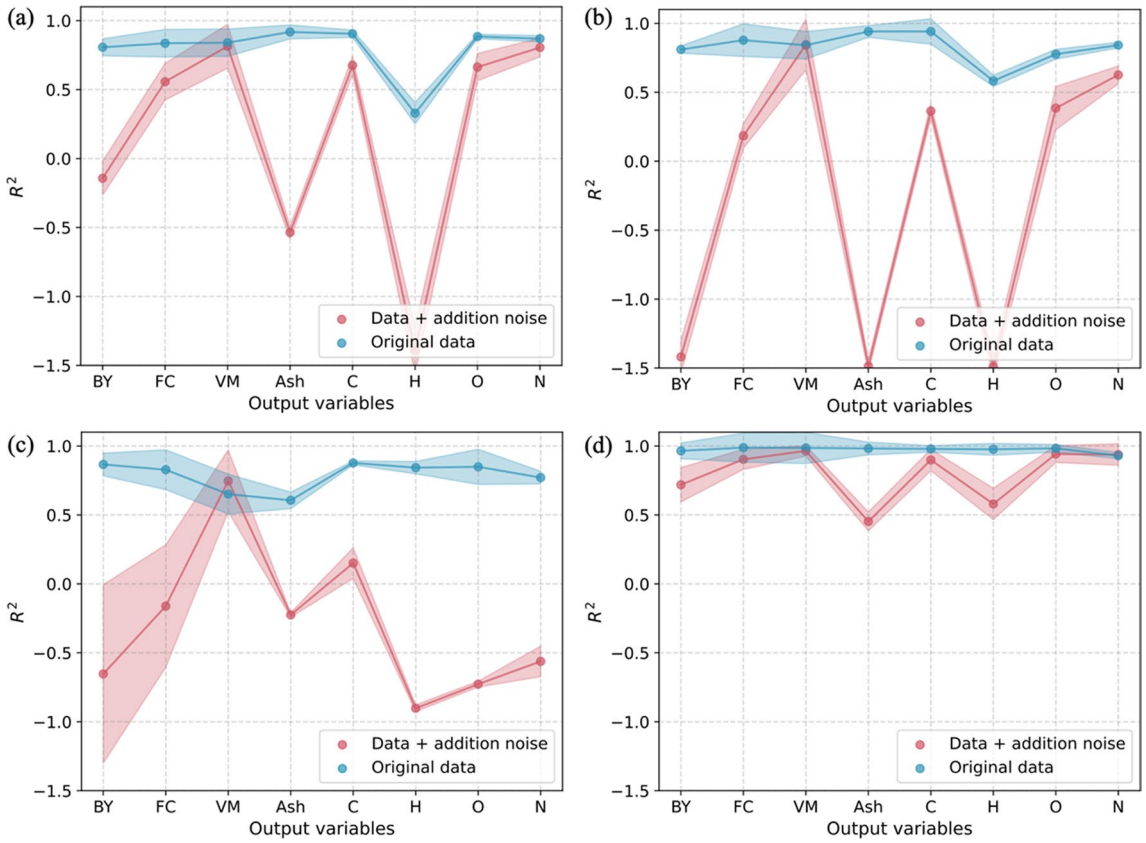


Fig. 6 R^2 comparison between raw output data and raw output data plus additional Gaussian noise (sigma = 0.1) when predicting output variables including Biochar yield (BY), Fixed Carbon (FC), Volatile Matter (VM), Ash, C, H, O, N, S. Methods include (**a**) Random Forest, (**b**) XGBoost, (**c**) Gaussian Process, and (**d**) our ResNet-based Autoencoder (Ours)

4 Discussions

This study highlights the potential of a ResNet-based autoencoder for predicting biochar yield and composition, addressing significant challenges such as data uncertainty and missing values. The proposed method outperforms baseline models, such as MLP-NN, Random Forest, XGBoost, and Gaussian Process, by delivering superior predictive accuracy (average $R^2 = 0.974$) and robustness. Incorporating previously discarded high-missing-rate data further enhanced the performance of model, achieving an average R^2 of 0.983. Additional covariates from diverse data sources increased the average R^2 to 0.985, emphasizing the value of preserving and utilizing data rather than eliminating variables with high uncertainty. The robustness of the model was validated through stress tests, where it demonstrated minimal performance degradation under simulated scenarios of added Gaussian noise and increased missing values. This capability underscores the reliability of model in handling real-world biochar datasets, which are often characterized by inconsistencies and uncertainty due to the diverse experimental environments. The sensitivity analysis revealed that certain covariates, such as heating rate and volatile matter, significantly impact model performance under increased uncertainty. This insight suggests that prioritizing the accurate measurement and inclusion of these variables can further improve predictive outcomes.

The findings are particularly valuable because they demonstrate the feasibility of leveraging machine learning to address inherent challenges in biochar production research. By effectively handling missing and noisy data, the model facilitates more reliable and efficient optimization of production conditions, which is crucial for reducing environmental risks and improving biochar quality. Moreover, the robustness demonstrated under various stress tests makes this approach practical for real-world scenarios, where data inconsistencies are inevitable.

Despite its robust performance, the study has certain limitations. The impact of varying dataset sizes on model performance was not explored, leaving room for further investigation into the scalability of the approach. Additionally, the interaction between the proposed method

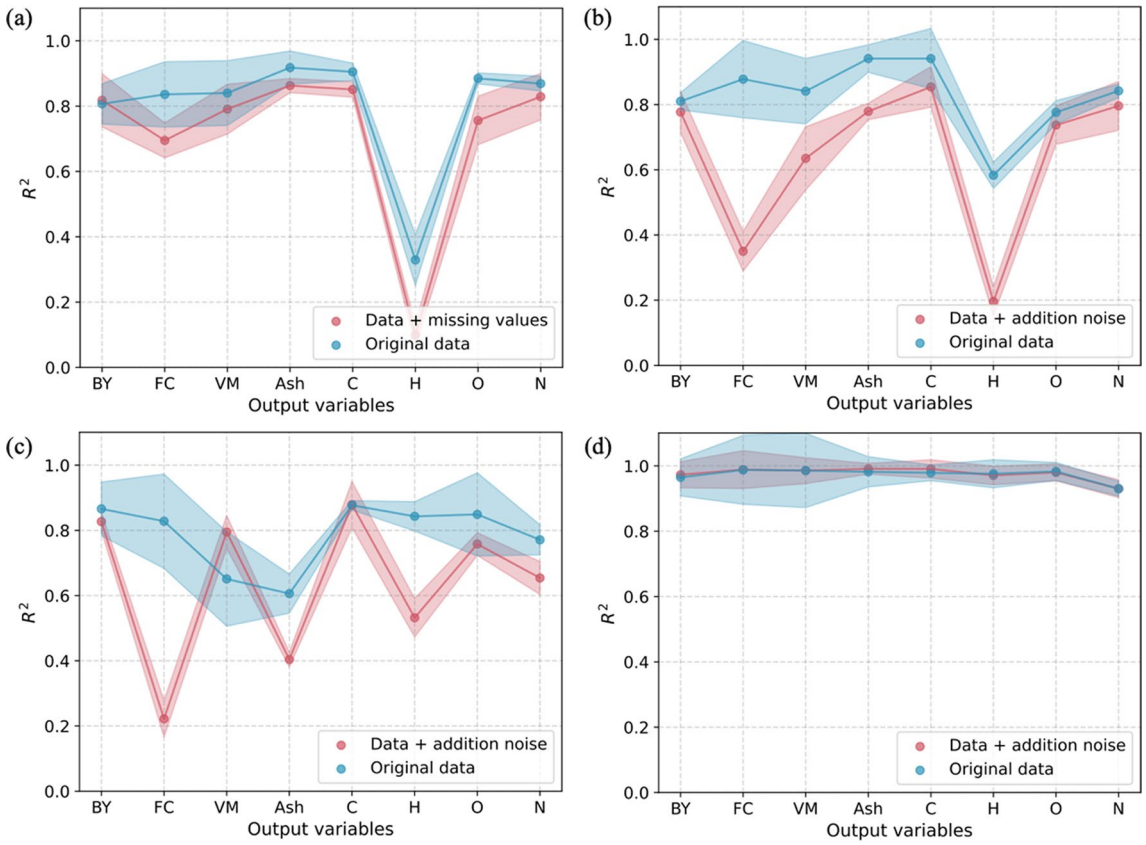


Fig. 7 R^2 comparison between raw output data and raw output data with additional 10\% missing values when predicting output variables including Biochar yield (BY), Fixed Carbon (FC), Volatile Matter (VM), Ash, C, H, O, N, S. Methods include (**a**) Random Forest, (**b**) XGBoost, (**c**) Gaussian Process, and (**d**) our ResNet-based Autoencoder (Ours)

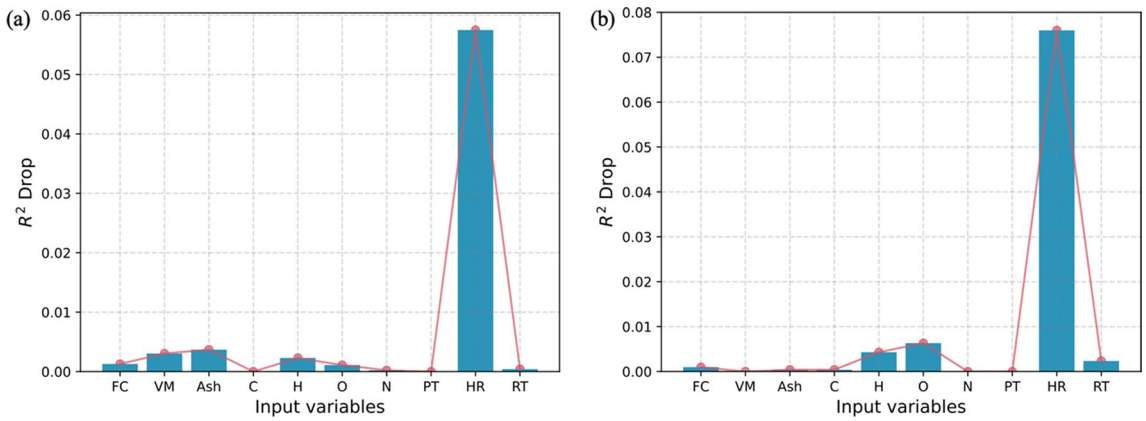


Fig. 8 Robust Sensitivity Analysis of Input Covariates: average R^2 drop for different uncertainty increments in each input covariate including Fixed Carbon (FC), Volatile Matter (VM), Ash, C, H, O, N, Pyrolysis Temperature (PT), Heating Rate (HR), Residence Time (RT). The data analysis is based on our multiobjective ResNet-based Autoencoder model. (**a**) Average R^2 drop with 20% additional deviation from the true values in each input covariate, (**b**) Average R^2 drop with 20% additional missing values in each input covariate

and data augmentation techniques was not examined. Finally, while the model effectively handles numerical and categorical data, integrating unstructured language data remains a future research opportunity.

5 Conclusions

The ResNet-based autoencoder presented in this study provides a reliable and efficient approach for predicting biochar yield and composition under varying levels of data uncertainty. Its ability to maintain high predictive accuracy with low RMSE across diverse datasets marks a significant advancement over traditional and baseline machine learning methods. By effectively addressing missing values and deviations from true values, this model ensures the utilization of all available data, preserving valuable information and enhancing robustness. The findings demonstrate the potential of this method to optimize biochar production conditions, reducing experimental costs and environmental risks. The proposed approach is computationally efficient, making it highly applicable to real-world industrial settings where rapid and accurate predictions are critical.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1007/s42773-025-00446-2.

Supplementary material 1.

Acknowledgements
We thank the USDA-NIFA for their support through the grant "HBCUH-SIRIU Consortium: A Synergistic Paradigm for Training the Next Generation Agriculture Workforce for a Sustainable Future" (Prime Award Number 2023-70440-40147).

Author contributions
Zong Liu: Conceptualization, Funding acquisition; Yali Zhang: Formal analysis; Methodology;Validation;Writing-original draft; Bowen Lei: Writing-review & editing; Amirhossein Mahdaviarab: Writing-review & editing; Xiao Wang: Methodology; Writing-review & editing.

Funding
The authors declare the following financial interests/personal relationships which may be considered as potential competing interests. This work was supported by the United States Department of Agriculture, National Institute of Food and Agriculture (USDA-NIFA) under Prime Award Number 2023-70440-40147.

Availability of data and materials
The datasets used or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Competing interests
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Received: 10 September 2024 Revised: 17 January 2025 Accepted: 6 February 2025
Published online: 18 March 2025

References

Alabdrabalnabi A, Gautam R, Sarathy S (2022) Machine learning to predict biochar and bio-oil yields from co-pyrolysis of biomass and plastics. Fuel 328:125303. https://doi.org/10.1016/j.fuel.2022.125303
Blaschke T, Olivecrona M, Engkvist O, Bajorath J, Chen H (2018) Application of generative autoencoder in de novo molecular design. Mol Inf 37(1–2):1700123. https://doi.org/10.1002/minf.201700123
Breiman L (2001) Random forests. Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324
Chen D, Liu D, Zhang H, Chen Y, Li Q (2015a) Bamboo pyrolysis using TG–FTIR and a lab-scale reactor: analysis of pyrolysis behavior, product properties, and carbon and energy yields. Fuel 148:79–86. https://doi.org/10.1016/j.fuel.2015.01.092
Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H, Chen K, Mitchell R, Cano I, Zhou T (2015) Xgboost: extreme gradient boosting. R Packag Vers 1(4):1–4
Chen C, Wang Z, Ge Y, Liang R, Hou D, Tao J, Yan B, Zheng W, Velichkova R, Chen G (2023a) Characteristics prediction of hydrothermal biochar using data enhanced interpretable machine learning. Biores Technol 377:128893. https://doi.org/10.1016/j.biortech.2023.128893
Chen MW, Chang MS, Mao Y, Hu S, Kung CC (2023b) Machine learning in the evaluation and prediction models of biochar application: a review. Sci Prog 106(1):00368504221148842. https://doi.org/10.1177/00368504221148842
Güleç F, Williams O, Kostas ET, Samson A, Lester E (2022) A comprehensive comparative study on the energy application of chars produced from different biomass feedstocks via hydrothermal conversion, pyrolysis, and torrefaction. Energy Convers Manage 270:116260. https://doi.org/10.1016/j.enconman.2022.116260
He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition. 770–778.
Ho TK (1995). Random decision forests. In Proceedings of 3rd international conference on document analysis and recognition. IEEE, 278–282.. https://doi.org/10.1109/ICDAR.1995.598994
Huber PJ (1992) Robust estimation of a location parameter. Breakthroughs in statistics: methodology and distribution. Springer, New York
Kanthasamy R, Almatrafi E, Ali I, Sait HH, Zwawi M, Abnisa F, Peng L, Ayodele BV (2023) Biochar production from valorization of agricultural wastes: data-driven modelling using machine learning algorithms. Fuel 351:128948. https://doi.org/10.1016/j.fuel.2023.128948
Khan M, Ullah Z, Mašek O, Naqvi SR, Khan MNA (2022) Artificial neural networks for the prediction of biochar yield: a comparative study of metaheuristic algorithms. Biores Technol 355:127215. https://doi.org/10.1016/j.biortech.2022.127215
Leng L, Xiong Q, Yang L, Li H, Zhou Y, Zhang W, Jiang S, Li H, Huang H (2021) An overview on engineering the surface area and porosity of biochar. Sci Total Environ 763:144204. https://doi.org/10.1016/j.scitotenv.2020.144204
Li Y, Gupta R, You S (2022) Machine learning assisted prediction of biochar yield and composition via pyrolysis of biomass. Biores Technol 359:127511. https://doi.org/10.1016/j.biortech.2022.127511
Liu WJ, Jiang H, Yu HQ (2015) Development of biochar-based functional materials: toward a sustainable platform carbon material. Chem Rev 115(22):12251–12285. https://doi.org/10.1021/acs.chemrev.5b00195
Liu WJ, Li WW, Jiang H, Yu HQ (2017) Fates of chemical elements in biomass during its pyrolysis. Chem Rev 117(9):6367–6398. https://doi.org/10.1021/acs.chemrev.6b00647
Liu X, Zhou Y, Zhao J, Yao R, Liu B, Ma D, Zheng Y (2020) Multiobjective ResNet pruning by means of EMOAs for remote sensing scene classification. Neurocomputing 381:298–305. https://doi.org/10.1016/j.neucom.2019.11.097
Lu Z, Bai Y, Chen Y, Su C, Lu S, Zhan T, Hong X, Wang S (2020) The classification of gliomas based on a pyramid dilated convolution resnet model. Pattern Recogn Lett 133:173–179. https://doi.org/10.1016/j.patrec.2020.03.007
Ma L, Shuai R, Ran X, Liu W, Ye C (2020) Combining DC-GAN with ResNet for blood cell image classification. Med Biol Eng Compu 58:1251–1264. https://doi.org/10.1007/s11517-020-02163-3
Minarno AE, Ghufron KM, Sabrila TS, Husniah L, Sumadi FDS (2021) Cnn based autoencoder application in breast cancer image retrieval. In 2021 International Seminar on Intelligent Technology and Its Applications (ISITIA), IEEE, 29–34. https://doi.org/10.1109/ISITIA52817.2021.9502205

Panwar NL, Pawar A (2020) Influence of activation conditions on the physicochemical properties of activated biochar a review. Biomass Convers Biorefinery. https://doi.org/10.1007/s13399-020-00870-3
Qadeer S, Anjum M, Khalid A, Waqas M, Batool A, Mahmood T (2017) A dialogue on perspectives of biochar applications and its environmental risks. Water Air Soil Pollut 228:1–26. https://doi.org/10.1007/s11270-017-3428-z
Qambrani NA, Rahman MM, Won S, Shim S, Ra C (2017) Biochar properties and eco-friendly applications for climate change mitigation, waste management, and wastewater treatment: a review. Renew Sustain Energy Rev 79:255–273. https://doi.org/10.1016/j.rser.2017.05.057
Safarian S (2023) Performance analysis of sustainable technologies for biochar production: a comprehensive review. Energy Rep 9:4574–4593. https://doi.org/10.1016/j.egyr.2023.03.111
Seow YX, Tan YH, Mubarak NM, Kansedo J, Khalid M, Ibrahim ML, Ghasemi M (2022) A review on biochar production from different biomass wastes by recent carbonization technologies and its sustainable applications. J Environ Chem Eng 10(1):107017. https://doi.org/10.1016/j.jece.2021.107017
Wen C, Liu T, Wang D, Wang Y, Chen H, Luo G, Zhou Z, Li C, Xu M (2023) Biochar as the effective adsorbent to combustion gaseous pollutants: preparation, activation, functionalization and the adsorption mechanisms. Prog Energy Combust Sci 99:101098. https://doi.org/10.1016/j.pecs.2023.101098
Williams C, Rasmussen C (1995) Gaussian processes for regression. Adv Neural Inf Process Syst 8:2
Xiang L, Liu S, Ye S, Yang H, Song B, Qin F, Shen M, Tan C, Zeng G, Tan X (2021) Potential hazards of biochar: the negative environmental impacts of biochar applications. J Hazard Mater 420:126611. https://doi.org/10.1016/j.jhazmat.2021.126611
Yaashikaa PR, Kumar PS, Varjani S, Saravanan A (2020) A critical review on the biochar production techniques, characterization, stability and applications for circular bioeconomy. Biotechnol Rep 28:e00570. https://doi.org/10.1016/j.btre.2020.e00570
Yang Z, Xu B, Luo W, Chen F (2022) Autoencoder-based representation learning and its application in intelligent fault diagnosis: a review. Measurement 189:110460. https://doi.org/10.1016/j.measurement.2021.110460
Yargicoglu EN, Sadasivam BY, Reddy KR, Spokas K (2015) Physical and chemical characterization of waste wood derived biochars. Waste Manage 36:256–268. https://doi.org/10.1016/j.wasman.2014.10.029
Yu S, Zhang W, Dong X, Wang F, Yang W, Liu C, Chen D (2023) A review on recent advances of biochar from agricultural and forestry wastes: Preparation, modification and applications in wastewater treatment. J Environ Chem Eng. https://doi.org/10.1016/j.jece.2023.111638
Zhu X, Li Y, Wang X (2019) Machine learning prediction of biochar yield and carbon contents in biochar based on biomass characteristics and pyrolysis conditions. Biores Technol 288:121527. https://doi.org/10.1016/j.biortech.2019.121527