scientific reports

Check for updates

OPEN

Precision biochar yield forecasting employing random forest and XGBoost with Taylor diagram visualization

Sudhakar Uppalapati[1], Prabhu Paramasivam[2]✉, Naveen Kilari[3], Jasgurpreet Singh Chohan[4,8], Praveen Kumar Kanti[5], Harinadh Vemanaboina[3], Leliso Hobicho Dabelo[6]✉ & Rupesh Gupta[7]

Waste-to-energy conversion via pyrolysis has attracted increasing attention recently owing to its multiple uses. Among the products of this process, biochar stands out for its versatility, with its yield influenced by various factors. Extensive and labor-intensive experimental testing is sometimes necessary to properly grasp the output distribution from various feedstocks. Nonetheless, data-driven predictive models using large-scale historical experiment records can provide insightful analysis of projected yields from a variety of biomass materials, hence overcoming the challenges of empirical modeling. As such, five modern approaches available in modern machine learning are employed in this study to develop the biochar yield prediction models. The Lasso regression, Tweedie regression, random forest, XGBoost, and Gradient boosting regression were employed. Out of these five XGBoost was superior with a training mean squared error (MSE) of 1.17 and a test MSE of 2.94. The XGBoost-based biochar yield model shows excellent performance with a strong predictive accuracy of the R^2 values as 0.9739 (training) and 0.8875 (test). The mean absolute percentage error value was only 2.14% in the training phase and 3.8% in the testing phase. Precision prognostic technologies have broad effects on sectors including biomass logistics, conversion technologies, and effective biomass utilization as renewable energy. Leveraging SHAP based on cooperative game theory, the study shows that while ash and moisture lower biochar yield, FPT, nitrogen, and carbon content significantly boost it. Small variables like heating rate and volatile matter have a secondary impact on production efficiency.

Keywords Explainable machine learning, Biochar, Boosting, Random forest, Prediction, Sustainability, SHAP analysis

An expanding worldwide issue is the energy problem. About 80% of the global energy is sourced from fossil fuels, which are thus rather important. Still, this dependence is unsustainable. As fossil fuels are running out, their usage and extraction do major harm to the surroundings. Leading greenhouse gas emissions from coal, oil, and natural gas combustion propel climate change[1,2]. The demand for alternative energy sources arises as the world's population and energy consumption rise, therefore stressing these limited resources. Many nations have responded by adopting the "Net Zero" goal[3,4]. With carbon capture or another similar technology offset, this target seeks to almost eliminate greenhouse gas emissions by 2025. Reducing the consequences of climate change depends on net zero achievement. Still, the change to greener energy is nuanced[5]. It calls for significant financial commitment to grid upgrading, energy storage technology, and renewable energy sources. Nations have to

[1]Department of Mechanical Engineering, Marri Laxman Reddy Institute of Technology and Management, Hyderabad 500043, India. [2]Department of Research and Innovation, Saveetha School of Engineering, SIMATS, Chennai, Tamil Nadu 602105, India. [3]VEMU Institute of Technology, Chittoor, Andra Pradesh 517112, India. [4]School of Mechanical Engineering, Rayat Bahra University, Mohali 140104, India. [5]University Center for Research and Development (UCRD), Chandigarh University, Mohali 140413, Punjab, India. [6]Department of Mechanical Engineering, Mattu University, P.O. Box 318, Mettu, Ethiopia. [7]Chitkara University Institute of Engineering and Technology, Chitkara University, Rajpura 140401, Punjab, India. [8]Faculty of Engineering, Sohar University, 7119 Sohar, Oman. ✉email: lptprabhu@gmail.com; leliso.hobicho@meu.edu.et

strike a balance between environmental responsibility and economic progress; hence this issue has spurred fresh attention on sustainable development[6].

The waste-to-energy (WtE) method is one emerging option in the energy change. A rising problem especially in cities is waste generation. Unsustainability characterizes the present conventional waste disposal techniques including incineration and landfills. They poll the land in addition to consuming priceless resources[7]. It offers another energy source and lowers garbage volume. WtE infrastructure can have a significant capital cost, hence the type of waste determines the efficiency of waste-to-energy conversion. In an energy context, this entails creating more resource-efficient systems, thereby lowering energy use, and so extending the lifetime of products[8,9]. Resource recovery is fundamental in a circular economy. This is reintegrating byproducts into the manufacturing process and gathering energy from waste sources. Industrial waste heat, for example, can be gathered and applied to run factories or heat homes. Likewise, agricultural wastes including crop waste can be turned into biofuels, therefore lowering reliance on fossil fuels[10]. A waste product from biomass processing, biochar has great circular economy value. Made from organic resources like agricultural waste by pyrolysis, biochar is a carbon-rich compound with several uses. Biochar is useful in energy generation since it can effectively store carbon in the ground for a long time, thereby serving as a kind of carbon sequestration[11,12].

Biochar also offers agricultural advantages. Added to soils, it enhances water retention, nutrient availability, and soil structure. In damaged or parched soils, these qualities make it very valuable. Additionally improving crop yields, biochar forms a feedback loop supporting food security through carbon sequestration[13,14]. Moreover, biochar finds use in waste management. For gasification, for instance, biochar can produce energy from organic waste. Made using this technique, syngas—a versatile and sustainable energy source—can be employed as a source of electricity generation or as feedstock for biofuel synthesis. Biochar gasification is one of the promising technologies since it helps energy generation and offers a sustainable approach to dispose of waste[15]. More generally, biochar makes sense in keeping with the concepts of regenerative agriculture. Rather, than draining ecosystems, this approach stresses farming techniques meant to aid in repairing and enhancing them. These techniques can help the agriculture industry move over time toward more strong and sustainable systems[16–18].

Variations in the feedstock and the complex pyrolysis process make modeling of biochar production challenging. The production depends on numerous parameters. These are type of biomass, its moisture content, heating rate, temperature, and residence duration during pyrolysis. Each one of these elements interacts nonlinearly, which makes the exact projection of the final biochar yield difficult[19–21]. Biomass is fairly diverse by itself. Different kinds of biomass including algae, agricultural waste, and wood have different chemical compositions. The thermal breakdown process is influenced by lignin, cellulose, hemicellulose, and ash concentrations of various feedstocks. Variations in particle size, moisture content, and provenance even inside the same type of biomass could influence the pyrolyze reaction[22,23]. Conventional mechanistic models so find it challenging to depict this complexity. Sometimes they demand simplified assumptions, but it reduces their expected accuracy. Moreover, pyrolysis is quite sensitive to variations in operating conditions. Conventional modeling techniques including kinetic models rely on complete thermodynamic data and reaction rate equations. Time-consuming and costly, these models sometimes call for vast volumes of experimental data to calibrate and validate[24,25].

Machine learning (ML) is one very effective replacement for modeling biochar generation. Unlike conventional models, ML techniques can control complex, non-linear relationships between inputs and outputs. Employing large dataset analysis from prior pyrolyzed experiments, ML techniques enable the identification of trends not easily visible by mechanistic models. This data-driven method reduces the need for certain presumptions in the pyrolysis process. Depending on input parameters like biomass content and pyrolyzed conditions, support vector regression, decision trees, and linear regression among other supervised learning models can predict biochar output. These models increase their accuracy as more data becomes available by learning from experimental data. Moreover, they are flexible and accommodate numerous feedstocks and running circumstances[23,26]. ML can also facilitate optimization. Investigating the vast parameter space of biomass quality and pyrolyzed settings allows ML techniques to identify ideal conditions for optimizing biochar output. Since efficiency depends on exacting operational conditions, this is particularly useful for improving biochar output. Combining real-time data from pyrolyze reactors allows ML to additionally dynamically anticipate, thereby enabling more perfect control over the process[27,28].

The intricate, non-linear relationships between biomass properties, pyrolysis process parameters, and biochar yield pose significant challenges for accurate modeling. Conventional models often limit their forecast performance by depending on large databases and simplifying presumptions. This work uses advanced machine learning methods including random forest (RF), Gradient Boosting, Tweedie regression, and Extreme Gradient Boosting (XGBoost) to improve biochar yield forecasts, so addressing these constraints. Combining these models with Taylor diagrams offers benefits since it allows one to visually assess model performance in terms of accuracy and variability. This method improves the prediction system and provides a more strong and dependable instrument for biochar production optimization.

Materials and methods

Biochar data

In the present study data from a well-published journals was adopted for modeling[29–51]. Several important benefits arise from the researchers employing biochar data from published work to leverage already obtained experimental results. First, obtaining published data helps in a larger spectrum of biomass kinds, pyrolysis settings, and biochar yields all of which would be time-consuming and expensive to recreate through individual studies. This strategy lets the researchers concentrate on model development instead of the time-consuming data-collecting choreography. Moreover, published data has usually gone through peer review to guarantee a minimum of quality and dependability. Including different feedstocks and pyrolyzed conditions recorded over

several trials helps in improving the generalizability of the machine learning models. This diversity helps the models including Extreme Gradient Boosting and random forest to learn patterns from different datasets, hence enhancing their prediction ability. Using available data allows the researchers to validate their models against known outcomes, therefore enhancing the credibility and resilience of their conclusions and reducing the need for extensive experimental resources.

The dataset comprises three main elements: biomass types in the first column, eleven predictor variables (control factors), and the biochar yield as the response variable in the 13th column. The biomass names classify the feedstock, which differs in characteristics influencing pyrolysis. The eleven predictors comprise feedstock properties (e.g., carbon, hydrogen, oxygen, nitrogen content, and ash %) and operational parameters (e.g., temperature, heating rate, residence time, and particle size). These parameters were selected since they are important determinants of the pyrolysis process and biochar synthesis. While residence time defines the degree of conversion, temperature and heating rate affect thermal breakdown. Heat transport depends on particle size; elemental composition determines the thermal stability and production of the biochar. Analyzing these predictors against biochar yield helps scientists to better understand how various control elements interact, so enabling solutions for carbon sequestration, soil improvement, and process optimization for uses including sustainable energy.

Machine learning

Several advanced regression techniques are utilized for improving prediction accuracy and handling complex datasets. The popular Python libraries were employed for model development. Lasso regression, available in the 'scikit-learn' library, uses L1 regularization to shrink some coefficients to zero, aiding in variable selection and reducing overfitting[52,53]. It performs especially well for high-dimensional and sparse datasets; k-fold cross-valuation and grid search help to maximize its performance by adjusting the regularization parameter, {alpha}. Also available from 'scikit-learn', gradient boosting regression (GBR) reduces the loss function iteratively by including decision trees to handle residual errors. Key to preventing overfitting are regularization methods including shrinkage and subsampling; hyperparameters such learning rate and number of estimators can be tweaked with grid search with cross-valuation to produce reliable predictions in noisy data. Part of 'scikit-learn', Tweedie regression models datasets with mixed continuous and discrete outputs by modifying the Tweedie distribution index parameter, hence fitting for skewed or zero-inflated data; parameter adjustment similarly boosted by cross-valuation and grid search[54,55]. Using second-order Taylor expansion for effective gradient computing and including regularization to control tree complexity, XGBoost, housed in the 'xgboost' package, enhances gradient boosting. random forest (RF), accessible in 'scikit-learn', creates multiple decision trees using random subsets of data, aggregating predictions via voting or averaging to enhance generalization and prevent overfitting[56]. XGBoost's hyperparameters—including max depth, learning rate, and tree complexity—are optimized by grid search and k-fold validation to improve accuracy in high-dimensional tasks. Hyperparameters like maximum tree depth and the number of trees ('n_estimators') are fine-tune using cross-valuation and grid search[57,58]. These methods provide exact and strong forecasts across several data kinds and complexity when paired with solid optimization tactics.

Evaluation

Using statistical techniques, the developed models for biochar yield will be thoroughly tested to evaluate their performance and correctness. We will make use of important measures such as Coefficient of Determination (R^2), Mean Squared Error (MSE), and Mean Absolute Percentage Error (MAPE). Emphasizing the total prediction error, the MSE will offer insight into the average squared difference between measured and predicted values. Acting as a gauge of model fit, the R^2 score will show how effectively the model explains the variation in biochar output data. MAPE will also evaluate the percentage variance between actual and expected values, therefore providing a normalized gauge of predictive accuracy. These statistical measures taken together will give a complete assessment of the capacity of the model to forecast biochar yield, therefore guaranteeing dependability and resilience in practical uses. This method guarantees that the models are generalizable over many situations and accurate as well. The following expressions were used in this study[59–61]:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 \tag{1}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{2}$$

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \widehat{y}_i}{y_i}\right| \tag{3}$$

Herein, y_i denotes the measured value of biochar yield, $\widehat{y}_i$ Represents the predicted value of the forecasted value of biochar yield, and n is the total number of observations. Also $\overline{y}$ Is the mean of measured values.

Taylor diagram

A useful graphical tool for assessing machine learning (ML) models created to forecast biochar yield is a Taylor diagram. It graphically combines three significant statistical measures: the standard deviation (SD) between actual and predicted values, the correlation coefficient (R), and the root mean squared error (RMSE)[62]. The

Taylor diagram enables biochar yield models to evaluate how closely their forecasts match real-world data over several models. The x and y axes of this graph show the standard deviation; the distance from the origin shows RMSE. The angle from the x-axis represents the correlation coefficient; stronger correlations provide angles more near zero. Plotting several ML models on the same Taylor diagram allows one to readily evaluate their accuracy and variability. This method facilitates the selection and improvement of models by helping to identify which one best fits the trends in biochar yield.

Violin plots

An efficient visual aid for evaluating the distribution and density of expected biochar production estimates using machine learning (ML) models is violin plots. Combining the features of density plots and box graphs, they provide a more thorough understanding of the performance of the created models. Violin plots can show the distribution of both actual and expected yields, therefore enabling a comparison of their central tendency and spread for biochar yield projections[63,64]. The violin plot offers a graphic depiction of the variance in predictions over several datasets in the framework of ML models created for biochar production prediction. The breadth of the graphic highlights the areas with more frequent predictions by reflecting the density of the data points at every value. Furthermore, by overlaying several ML models on the same plot, scientists can evaluate model performances and pinpoint those that most closely match the actual biochar yield. This makes model dependability and robustness easily analyzed.

Explainable machine learning

SHAP (SHapley Additive exPlanations) analysis, implemented through the 'shap' Python library, is a powerful framework for interpreting machine learning models. Based on cooperative game theory, it assigns each feature an importance value commensurate with its influence on the prediction of a model[65]. By separating individual predictions into additive contributions of input features, this approach offers explainability and helps researchers to know how each feature affects the outcome. Within the framework of XML (eXtreme Multi-Label Learning), SHAP analysis clarifies feature importance in models intended for high-dimensional, multi-label data[66]. SHAP helps researchers to find important trends and interactions inside the data by showing feature contributions for every label. Especially in complicated models like Gradient Boosting or XGBoost, it guarantees interpretability while preserving great predictive accuracy, hence supporting openness. Furthermore guiding feature engineering and optimization, SHAP values help to improve model efficiency and trustworthiness in XML applications[67,68].

Results and discussion

Data analysis with correlation values

A comprehensive picture of the interactions among several factors, including volatile matter, ash content, fixed carbon, moisture content, elemental composition, and operating variables including temperature and residence time, is offered by the correlation matrix for biochar yield. Understanding how these variables affect biochar yield and each other by looking at their correlation coefficients helps one to maximize biochar synthesis techniques. The correlation heatmap is illustrated in Fig. 1. The correlation matrix is listed in Table 1. Volatile matter and biochar yield have a weak negative connection (−0.2), meaning that the biochar yield somewhat drops as the content of volatile matter rises. This probably results from more volatile matter content releasing more gaseous byproducts during pyrolysis, so reducing the solid material available to contribute to the ultimate biochar output. By contrast, the positive (0.2) correlation between fixed carbon and yield suggests that a larger biochar production results from increasing fixed carbon content. This makes sense since directly contributing to the biochar, fixed carbon is the solid residue left over after pyrolysis.

Ash concentration exhibits a meager positive connection with yield (0.23). Higher ash concentration naturally causes a minor improvement in yield since ash is the incombustible residue left after pyrolysis. With a correlation of 0.14, moisture content does not affect biochar output either significantly or otherwise. This could be because, during the pyrolysis process, moisture mostly evaporates and has little effect on the creation of solid biochar. With a negative (−0.13) connection between carbon content and biochar yield, it is clear that increasing carbon content does not always raise yield. This could be explained by pyrolyzes turning carbon into gaseous forms instead of staying in the solid phase. Likewise, hydrogen concentration exhibits a weak negative association (−0.13) with yield; this can be justified by its function in generating volatile chemicals during the pyrolysis process instead of helping to define the solid biochar fraction. With biochar yield, oxygen content reveals a somewhat weak positive association (0.05), suggesting that oxygen has minimal effect on the ultimate output. Although nitrogen content exhibits a rather greater positive connection (0.12), its impact on yield is still minor.

With a negative correlation of −0.49, the final pyrolyzed temperature and yield show one of the more important relationships in the matrix. This implies that biochar yield lowers as the pyrolyze temperature rises. Higher temperatures usually encourage the development of gaseous products; hence this tendency is expected: less solid biochar will result. With a weak negative correlation (−0.15), the heating rate also shows that quicker heating may somewhat lower yield by encouraging volatile development at the expense of the solid fraction. At last, residence time has a modest negative connection (−0.06) with biochar output. Longer residence lengths somewhat lower yield, presumably because biomass breaks down more broadly into gases and liquids rather than solid biochar. These interactions show generally important links between biomass content, process parameters, and biochar yield, therefore offering important information for maximum biochar output.

In this work, the data preprocessing activities comprised management of missing data, outlier identification and removal, and normalizing. Statistical approach interquartile ranges with predefined criteria was employed to detect and remove the outliers in the data. In this study no outlier was detected. Mean imputation was used to handle missing data points, selected to maintain the integrity of the dataset and reduce any bias. In the data used for model deveopmnet there was missing data. Min–Max scaling was also used to normalize all variables

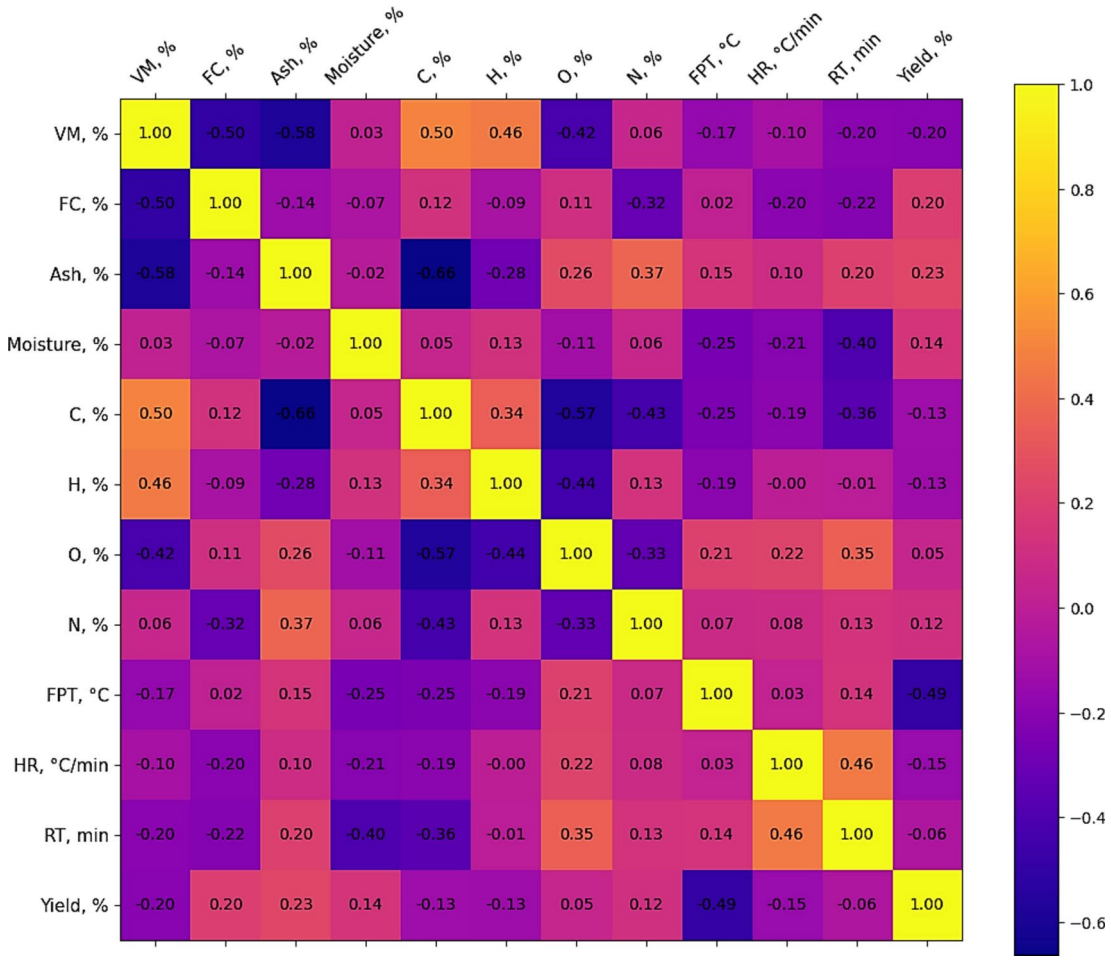


Fig. 1. Correlation heatmap.

	VM, %	FC, %	Ash, %	Moisture, %	C, %	H, %	O, %	N, %	FPT, °C	HR, °C/min	RT, min	Yield, %
VM, %	1	−0.5	−0.58	0.03	0.5	0.46	−0.42	0.06	−0.17	−0.1	−0.2	−0.2
FC, %	−0.5	1	−0.14	−0.07	0.12	−0.09	0.11	−0.32	0.02	−0.2	−0.22	0.2
Ash, %	−0.58	−0.14	1	−0.02	−0.66	−0.28	0.26	0.37	0.15	0.1	0.2	0.23
Moisture, %	0.03	−0.07	−0.02	1	0.05	0.13	−0.11	0.06	−0.25	−0.21	−0.4	0.14
C, %	0.5	0.12	−0.66	0.05	1	0.34	−0.57	−0.43	−0.25	−0.19	−0.36	−0.13
H, %	0.46	−0.09	−0.28	0.13	0.34	1	−0.44	0.13	−0.19	0	−0.01	−0.13
O, %	−0.42	0.11	0.26	−0.11	−0.57	−0.44	1	−0.33	0.21	0.22	0.35	0.05
N, %	0.06	−0.32	0.37	0.06	−0.43	0.13	−0.33	1	0.07	0.08	0.13	0.12
FPT, Â°C	−0.17	0.02	0.15	−0.25	−0.25	−0.19	0.21	0.07	1	0.03	0.14	−0.49
HR, Â°C/min	−0.1	−0.2	0.1	−0.21	−0.19	0	0.22	0.08	0.03	1	0.46	−0.15
RT, min	−0.2	−0.22	0.2	−0.4	−0.36	−0.01	0.35	0.13	0.14	0.46	1	−0.06
Yield, %	−0.2	0.2	0.23	0.14	−0.13	−0.13	0.05	0.12	−0.49	−0.15	−0.06	1

Table 1. Correlation analysis.

such that their compatibility for model training was maximized. These preprocessing processes were carried out to improve the analytical repeatability and dependability.

The descriptive statistics (Table 2) of biochar offer a complete picture of the distribution, central tendency, and variability over the parameters studied. Key variables comprising volatile matter (VM), ash content, fixed carbon (FC), moisture, elemental composition (C, H, O, N), final pyrolyzed temperature (FPT), heating rate (HR), residence time (RT), and biochar production constitute 134 samples. With a standard deviation of 6.45%, the mean biochar yield is 33.88%, suggesting somewhat modest variation. The skewness (0.66) indicates a minor positive skew, so most yield values are concentrated toward the lower end but with some higher outliers. The

	VM, %	FC, %	Ash, %	Moisture, %	C, %	H, %	O, %	N, %	FPT, °C	HR, °C/min	RT, min	Yield, %
Count	134.00	134	134	134	134	134	134	134	134	134	134	134.00
Mean	77.94	12.66	5.73	7.55	46.25	6.21	44.66	2.06	478.36	8.76	51.33	33.88
Std	6.55	4.32	4.17	3.32	5.83	1.13	6.00	4.16	106.28	3.19	15.77	6.45
Min	60.00	2.05	0.23	1	33.4	2.8	27.04	0.09	300	2	4	21.40
25%	73.26	9.08	2.85	4.9	42.3	5.69	41.5	0.6	400	10	30	29.38
50%	78.24	12.04	4.6	8.11	48.3	6.24	42.84	1.04	500	10	60	33.46
75%	83.69	15.4	7.96	10.3	50.9	6.9	49.12	1.7	550	10	60	38.11
Max	87.30	26.6	17.96	17.53	56.56	9.31	56.39	22.1	700	15	90	57.58
Skewness	−0.58	0.39	0.96	−0.17	−0.48	−0.14	0.26	4.23	0.26	−0.85	−0.67	0.66
Kurtosis	−0.19	−0.20	0.08	−0.77	−0.55	1.93	−0.31	17.55	−0.69	0.28	0.07	0.92

Table 2. Descriptive statistical analysis.

smallest yield value is 21.40%, while the maximum is 57.58%. Reflecting a concentration of values close to the top range, volatile matter (VM) has a mean of 77.94% and is negatively skewed (−0.58). On the other hand, FC and ash content show positive skewness (0.39 and 0.96, respectively), therefore indicating the presence of some samples with really high values. Reflecting the occurrence of extreme outliers, nitrogen (N) shows noteworthy skewness (4.23) and strong kurtosis (17.55). Generally speaking, the other parameters carbon, hydrogen, oxygen, and operational factors like FPT, HR, and RT show somewhat mild fluctuation. Except for a few outliers, especially concerning nitrogen concentration and residency period, the skewness and kurtosis values imply that most distributions are near normal. Depending on biomass type and pyrolysis settings, this data distribution emphasizes the complexity and variety of biochar output.

Model development and evaluation

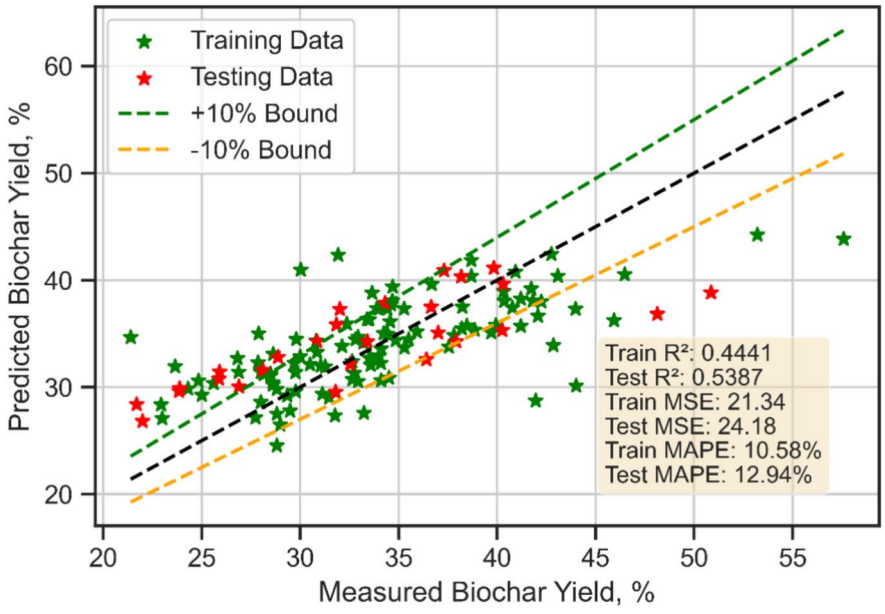
Figure 2a presents a performance evaluation of a LASSO regression model for predicting biochar yield, comparing predicted values with measured values for both training and testing datasets. Green stars show training data; red stars show testing data. The dashed lines, which provide a ± 10% bound around the black dashed line the ideal prediction line demonstrate how well the model forecasts fit inside this range. The statistical evaluation of the model is listed in Table 3. The training hyperparameters used are listed in Table 4. With an R^2 value of 0.4441 for the training data and 0.5387 for the testing data, the model demonstrates modest performance showing that, respectively, roughly 44.41% and 53.87% of the variance in biochar yield are explained by the model. While the mean squared error (MSE) for the testing data is 24.18, for the training data it is 21.34, therefore indicating a minor rise in prediction error for unknown data. With the test data revealing somewhat higher inaccuracy, the Mean Absolute Percentage inaccuracy (MAPE) for training is 10.58% while for testing it is 12.94%. This indicates reasonable prediction accuracy. The model performs typically within reasonable bounds, with some prediction error.

Comparing projected values with measured values for both training and testing datasets, Fig. 2b shows the performance of a gradient-boosting regression model for estimating biochar output. The R^2 values of 0.9770 for the training data and 0.8072 for the testing data point to a high degree of accuracy for the model. This implies that the model clarifies 80.72% of the variance in the testing set and 97.70% of the variance in the training set. Though it stays rather low, the MSE for training is 0.88 and for testing is 10.10, indicating a minor rise in error for the unseen data. Reflecting the model's superb prediction accuracy, the MAPE for training is 2.05%, and for testing is 6.48%. Especially in the training data, most expected values that show good performance and low overfitting fall within the ± 10% threshold.

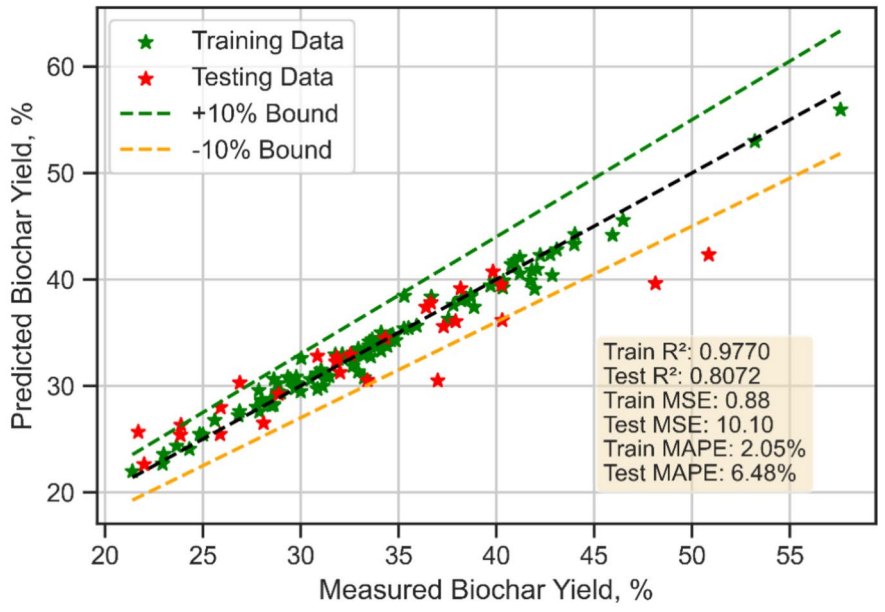
Figure 2c shows how well a Tweedie regression model forecasts biochar yield. With an R^2 value of 0.7755 for the training set and 0.6260 for the testing set, the Tweedie model exhibits somewhat modest performance. The model thus explains roughly 77.55% of the variance in the training data and 62.60% of the variation in the test data. Indicating a clear rise in error for the unseen data, the Mean Squared Error (MSE) for the training data is 8.62, and for the test data is 19.60. With a greater error margin for the test set than the training set, the Mean Absolute Percentage Error (MAPE) for training and testing shows reasonable predictive accuracy, however. Particularly in the middle range of measured biochar output, most of the expected values are between ± 10% limits. Several forecasts for the testing data, particularly for higher yield values, however, lie outside this bound and suggest some difficulties in precisely forecasting these outliers. The Tweedie model offers a reasonable fit overall; yet, especially for testing data, additional tuning could improve prediction accuracy.

Measured data indicates the performance of an XGBoost model built to anticipate biochar yield. The comparative performance of the XGBoost-based biochar yield model is depicted in Fig. 2d. The performance metrics of the darkening area present significant fresh perspectives on model accuracy. Although the training data gets an R^2 value of 0.9739, showing a high degree of accuracy, the testing data has an R^2 value of 0.8875, so it promises excellent performance on unknown data. The mean squared error for the training set is 1.17, whereas for the test set it is 2.94, indicating far greater errors in the test data. MAPE for the training and testing sets shows low prediction errors: 2.14% for the former and 3.80% for testing, respectively. The model exhibits good overall predictive capacity; most of the projections fall below the tolerable error bounds.

Figure 2e shows a comparison between the anticipated and observed biochar yields, therefore illustrating the performance of a random forest model applied to forecast biochar output. The right half of the graph



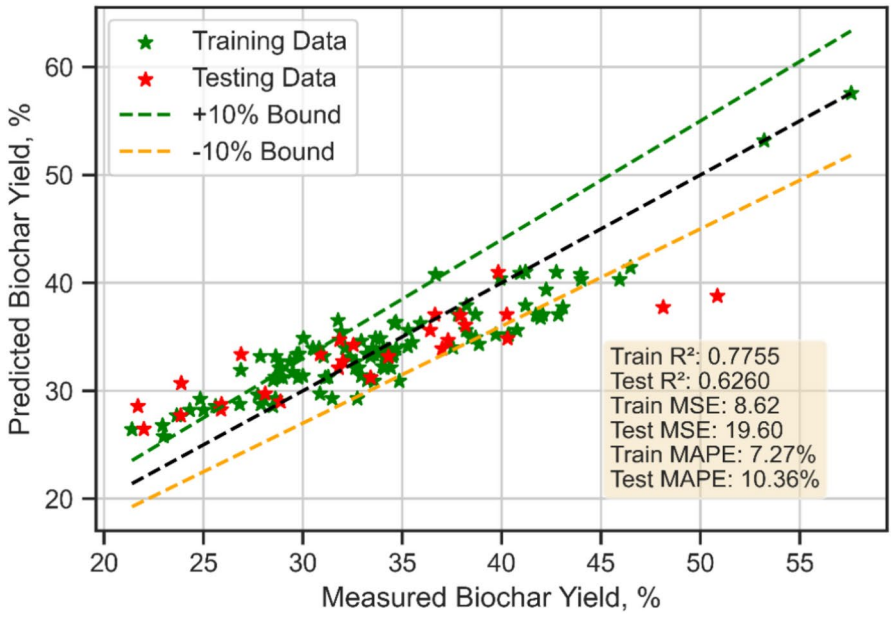
(a)



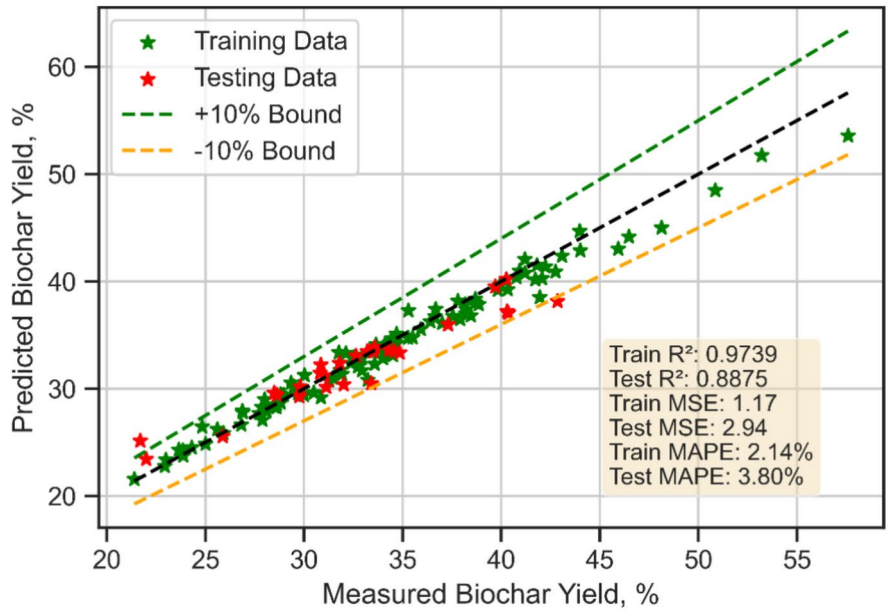
(b)

Fig. 2. Measured vs model predicted yield for (**a**) Lasso Regression, (**b**) Gradient Boosting Regression, (**c**) Tweedie regression, (**d**) XGboost and (**e**) random forest.

summarizes the model's performance measures. When projecting biochar yields based on the training data, the model achieves a coefficient of determination (R^2) of 0.9430, therefore suggesting a great degree of accuracy and goodness of fit. The test R^2 falls to 0.7396, however, showing less robust performance of the model on the test data and possible overfitting or difficulties generalizing to more data. The random forest model has more prediction errors on the test data; the MSE for the training set is 2.19 and for the test set is 13.65. Further revealing that prediction errors are significantly greater on the test set than the training set, the MAPE for the training data is 3.27% and for the testing data is 8.54%. The larger distribution of the test data points (red stars) and the higher test MAPE and MSE show that generally the random forest model shows a solid fit on the training data but suffers with generalizing. The difficulty of obtaining accurate predictions for unknown data with this model is highlighted by many of the test data points falling beyond the ± 10% limits.



(c)



(d)

Figure 2. (continued)

Among the models tested for biochar yield, XGBoost and gradient boosting regression show the greatest general performance. Concerning 97.70% of the variance in the training data and 80.72% of the variance in the test data, the gradient boosting model achieves good prediction accuracy. With just a small increase in prediction error, it preserves a reasonable degree of error for unknown data. The error margins of the model show a balance between test performance and training; most of the predictions lie within the allowed range of error. Analogous great accuracy is displayed by the XGBoost model, which explains 88.75% of the variance in the test data and 97.39% of the variation in training data. Its error measures reveal that the model reduces errors than the gradient boosting model and indicates good generalization to unprocessed data. Most of the expected values also stay within the limits of reasonable errors. Though it performs well on the training data, the random forest model suffers from generalize with more points falling outside acceptable error bounds and more errors

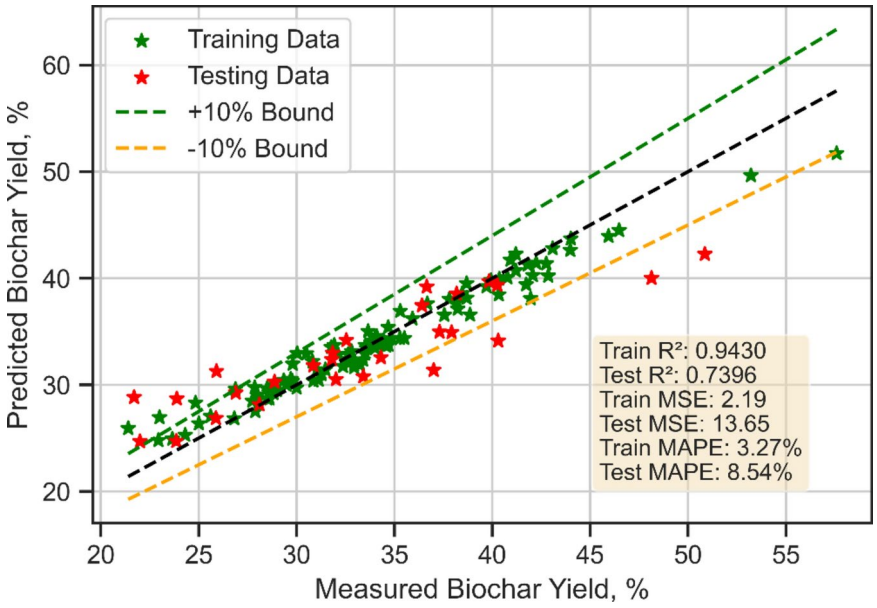


Figure 2. (continued)

Model	Train MSE	Test MSE	Train R^2	Test R^2	Train MAPE, %	Test MAPE, %
Lasso Regression	21.3392	24.1760	0.4441	0.5387	10.59	12.94
Gradient Boosting	0.8841	10.1038	0.9770	0.8072	2.05	6.48
Tweedie Regression	8.6189	19.6041	0.7755	0.6260	7.27	10.36
XGBoost	1.17	2.94	0.9739	0.8875	2.14	3.8
Random forest	2.19	13.65	0.9430	0.7396	3.27	8.54

Table 3. Statistical evaluation of developed models.

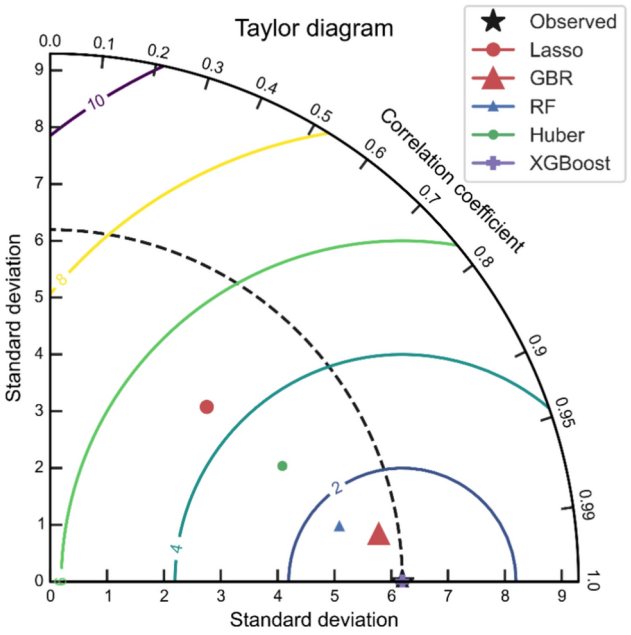
Model	Parameter range	Best parameters
Lasso Regression	{'alpha': [0.01, 0.1, 1, 10, 100]}	{'alpha': 1}
Gradient Boosting	{'n_estimators': [50, 100, 200], 'learning_rate': [0.01, 0.1, 0.2], 'max_depth': [3, 5, 7]}	{'n_estimators': 200, 'learning_rate': 0.2, 'max_depth': 3 }
XGBoost	{'n_estimators': [50, 100, 200], 'learning_rate': [0.01, 0.1, 0.2], 'max_depth': [3, 5, 7], 'subsample': [0.7, 0.8, 1.0]}	{'n_estimators': 200, 'learning_rate': 0.2, 'max_depth': 3, 'subsample': 1.0}
Random forest	{'n_estimators': [50, 100, 200], 'max_depth': [None, 10, 20], 'min_samples_split': [2, 5, 10]}	{'n_estimators': 200 'max_depth': None, 'min_samples_split': 2, }
Tweedie Regressor	{'power': [0, 1, 1.5, 2], 'alpha': [0.01, 0.1, 1, 10]}	{'power': 1.5, 'alpha': 10, }

Table 4. Training hyperparameters used.

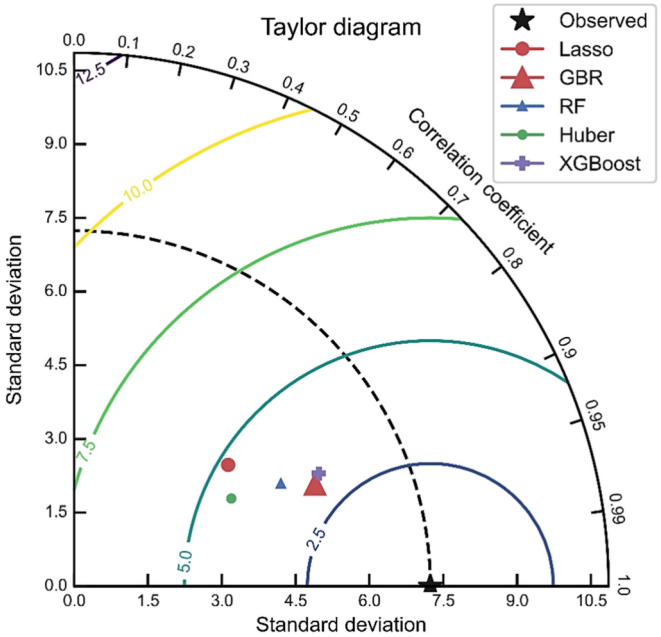
on the test data. With more modest performance, the Tweedie and LASSO regression models exhibit higher prediction errors and help to explain less data variance. With few errors and great generalization, XGBoost offers the best prediction performance overall; gradient boosting follows closely.

Model comparison with Taylor diagrams and Violin plots

This Taylor diagram Fig. 3a visualizes the performance of different biochar yield prediction models during the training phase by comparing them with observed data. Three important statistical measures correlation coefficient, standard deviation, and root mean square error shown by dashed contours can be simultaneously evaluated using the diagram. Where a greater number indicates a better linear connection between the observed and anticipated values, the correlation coefficient is displayed along the curved axis at the top of the picture between 0 (no correlation) and 1 (perfect correlation). Plotting the standard deviation along both the radial and horizontal axes reflects the variability in the expected data relative to the observed data; values nearer the



(a)



(b)

Fig. 3. Taylor diagram for (**a**) Training, (**b**) Testing phase.

observed standard deviation are more accurate. The dashed circles extending out from the origin show the root mean square error (RMSE); lower RMSE values indicate better prediction accuracy. With a high correlation coefficient approaching 0.95 and a standard deviation almost matching the observed value, XGBoost (purple line and marker) gets the highest overall performance demonstrating that XGBoost tightly captures both the variability and linear relationship of the data. With a correlation coefficient of around 0.9, random forest (RF) (blue triangle) exhibits good performance; yet, its standard deviation is somewhat higher than that of the actual data, thereby showing it overestimates the variability. With a correlation coefficient of around 0.9 but with more

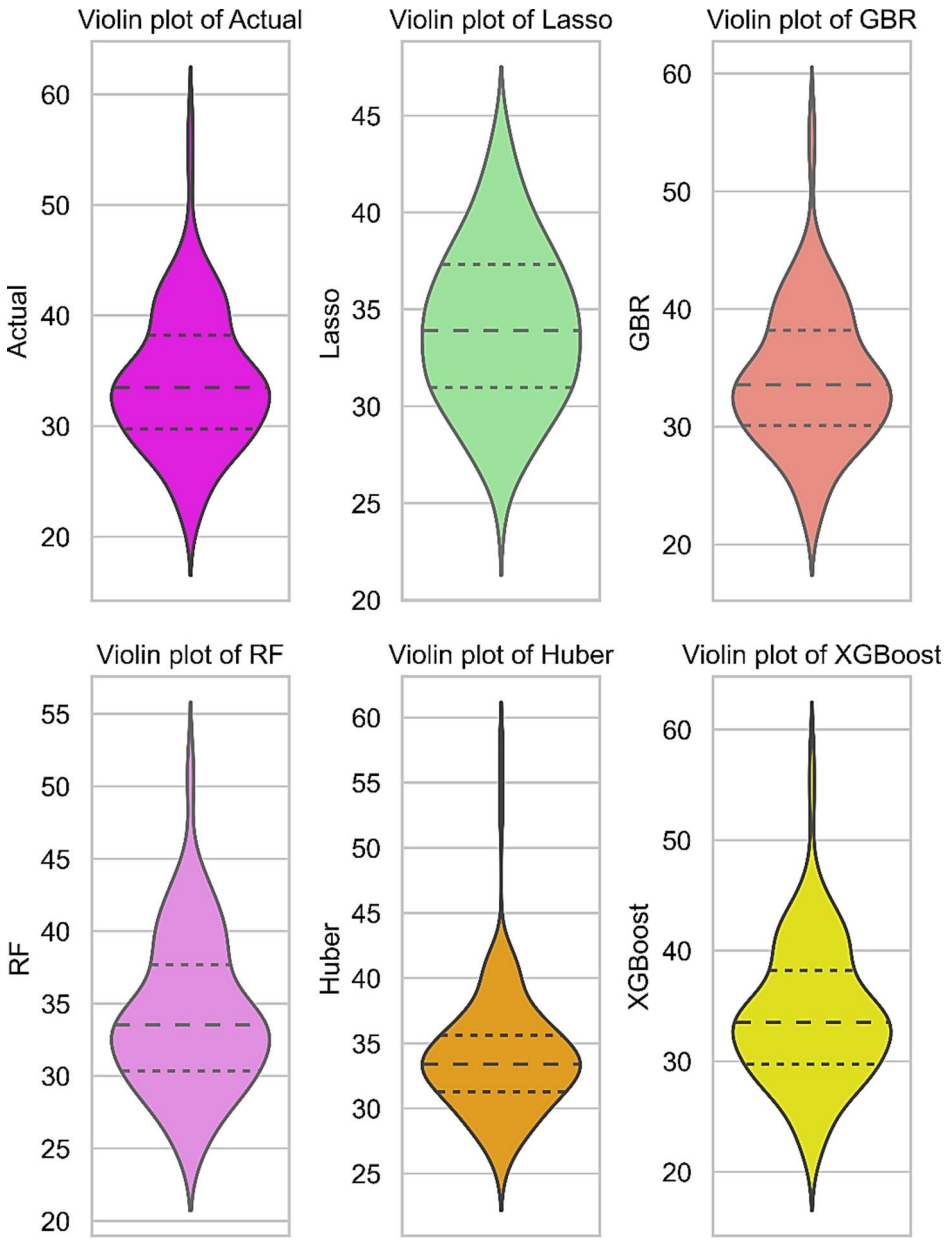


Fig. 4. Violin plots for (**a**) Training and (**b**) Testing phase.

fluctuation, akin to random forest, Gradient Boosting (GBR) (red triangle) also performs well. With lower correlation coefficients (below 0.7) and bigger standard deviations than the observed data, Huber (green dot) and Lasso (red dot) show poorer performance. XGBoost's excellent correlation and accurate standard deviation make it the best model overall throughout the training phase; random forest and Gradient Boosting follow closely.

Utilizing Lasso, GBR, RF, Huber, and XGBoost models against observed values, the Taylor diagram in Fig. 3b shows the performance of biochar yield models throughout the test phase. With XGBoost and RF showing the highest correlations, near 0.98, the correlation coefficients for the models range from 0.85 to 0.98, demonstrating varied degrees of predictive accuracy. Standard deviation values vary among models; the observed data has a standard deviation of roughly 7.5. Most models show that they understate the variability in biochar output relative to the observed data by clustering around standard deviations of 2.5 to 3.0. With standard deviations at 2.5 and 4.0 respectively, lasso and GBR models exhibit different alignments with the observed values. With a standard deviation near 3.0 and XGBoost positioned closer to the observed point, it shows better balance in capturing both the correlation and variability in biochar yield estimates. This shows XGBoost as the most successful model tested.

Against the actual yield values, the violin plots Fig. 4a for training and Fig. 4b for test phase contrast the distribution of expected biochar yield values among several models Lasso, GBR, RF, Huber, and XGBoost. Though most data points fall between 30 and 40, the actual biochar yield distribution has a median of around

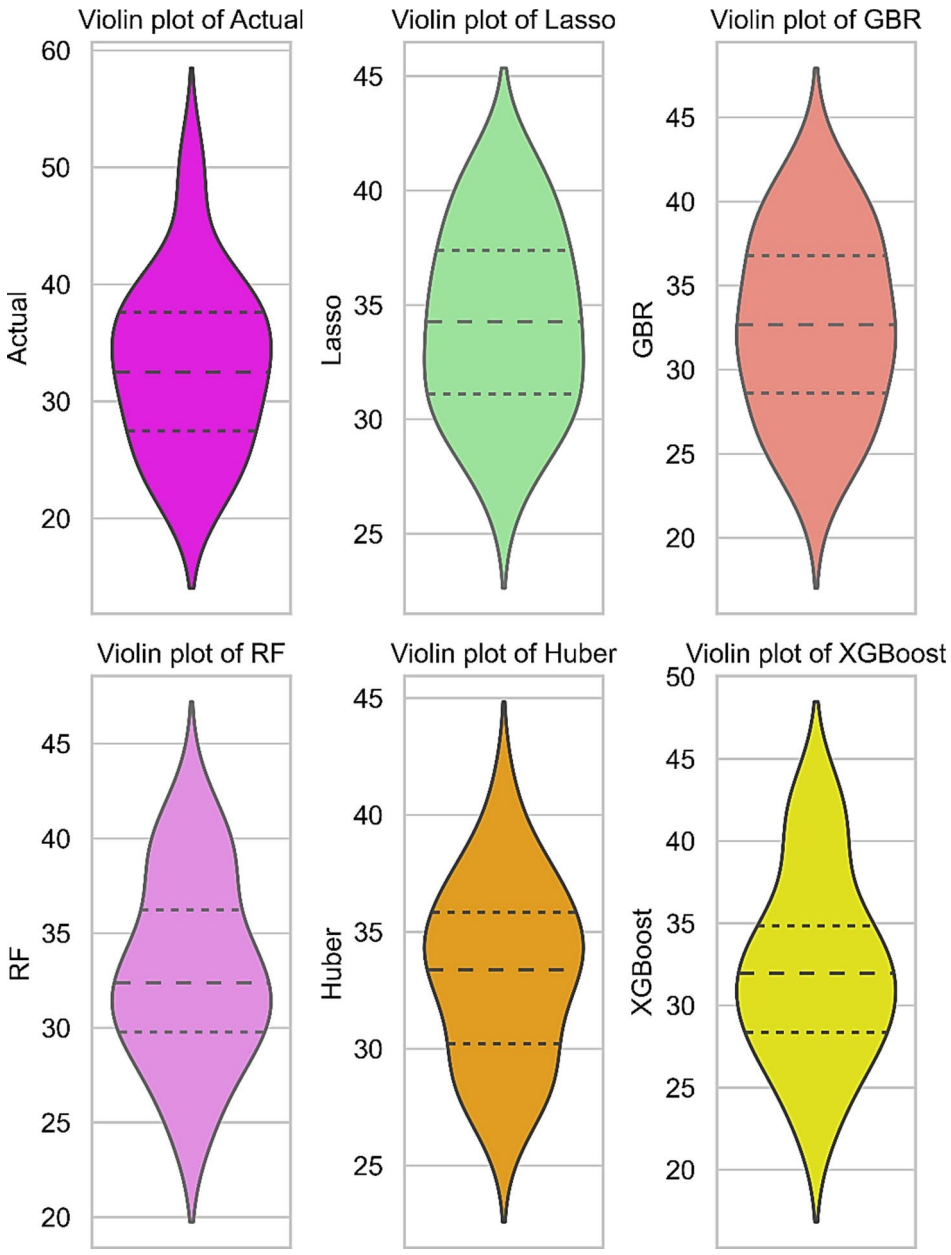


Figure 4. (continued)

35. Around 33–34, lasso and Huber models reveal smaller distributions with medians somewhat below the actual values. Though with GBR exhibiting a somewhat higher median yield around 35–36, both GBR and XGBoost reflect more broadly similar actual yield distribution more closely. Although RF displays considerable compression at the extremes, its forecasts cluster around 35, therefore it presents a distribution like to the real data. With a well-centered median around 35 and a range that reflects the observed values, XGBoost shows to be most suited to capturing the spread of the actual yield, hence providing a more accurate general prediction of biochar yield variability. In the case of the training phase, as depicted in Fig. 4b, the actual yield distribution is generally between 30 and 40 and exhibits a median near 35. Both Lasso and Huber models have somewhat narrow spreads, suggesting more concentrated predictions around their medians; their distributions resemble those with slightly lower medians around 33–34. With a median closer to 35, GBR has a larger distribution than Lasso and Huber; yet, the model reveals a rather more scattered range of values. RF shows a median between 34 and 35 and catches the yield range with a rather larger spread. With a median close to 35 and a more extensive, more even distribution, XGBoost most fairly resembles the real yield distribution. This implies that, among other models, XGBoost excels in capturing the central tendency as well as the variability of biochar yield projections.

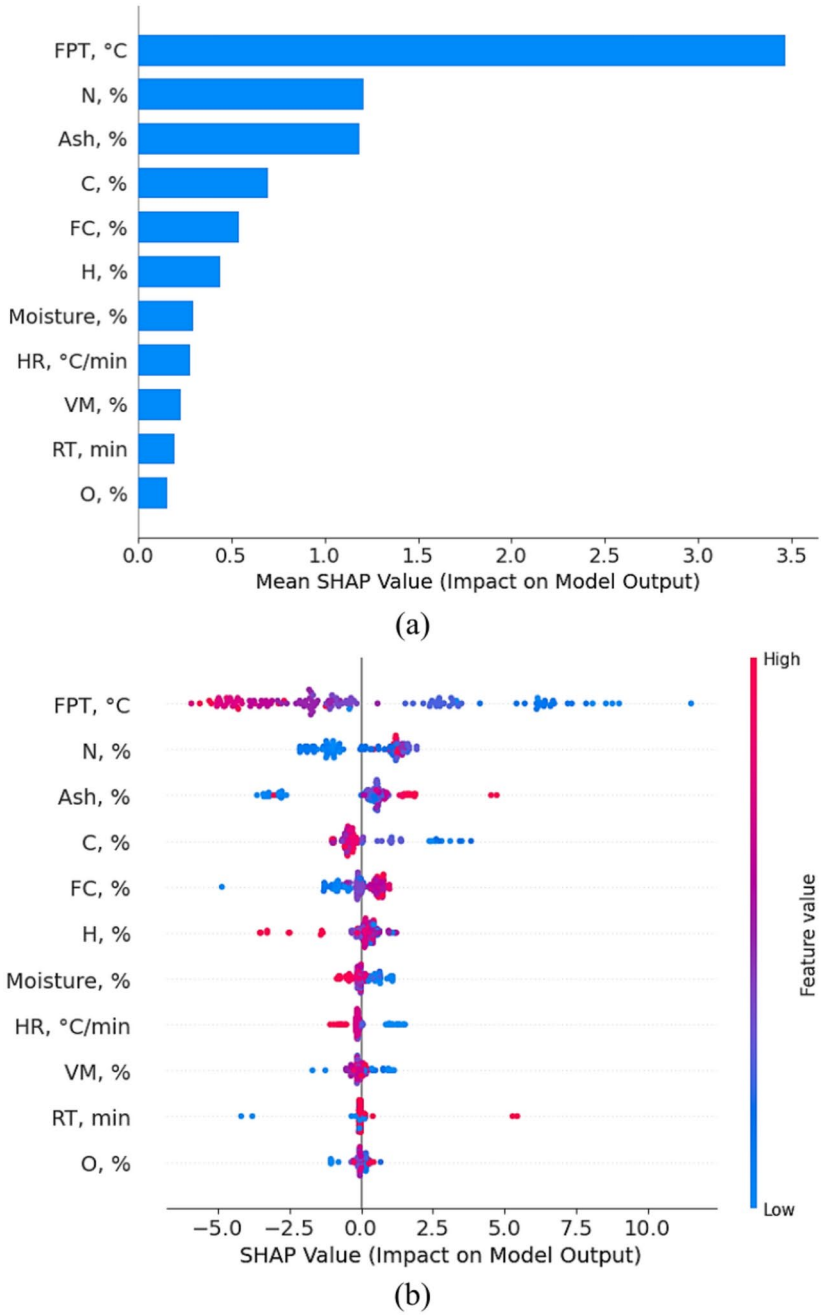


Fig. 5. SHAP based feature analysis showing (**a**) mean SHAP value, (**b**) impact on model output.

SHAP based feature analysis

The SHAP analysis of the biochar yield model highlights the critical role of process and material parameters in determining biochar production. With a mean SHAP value of almost 3.5 9 (Fig. 5a), the Pyrolysis Temperature turns up as the most important characteristic. Higher PT values as shown by red points in the summary plot (Fig. 5b) positively correlate with biochar yield, meaning that higher temperatures help to convert biomass into biochar by so encouraging the carbonization process. With a typical SHAP value of about 2.5, nitrogen content (N,%) comes second most significantly. Owing to its effect on biomass breakdown and stability, a larger nitrogen proportion encourages biochar generation. Third with a mean SHAP value almost 2 is Ash content (Ash, %), which differs from PT and N in that increased ash percentage reduces biochar output. This is attributed to the result of organic matter's decreased availability for thermal conversion in biomass high in ashes.

With typical SHAP values almost 1.5, carbon content (C,%) and fixed carbon (FC,%) exhibit little effects. These characteristics emphasize the need of organic content in increasing biochar yield since carbon-rich biomass is more suitable for carbonization. Concurrently, moisture content and hydrogen (H, %) show less yet significant impacts (mean SHAP values around 1). While hydrogen's importance may relate to its effect on reaction kinetics, high moisture content can prevent biochar production by raising energy demands for water

evaporation. Lesser important elements with mean SHAP values below 1 indicate they play secondary roles: heating rate (HR, °C/min), volatile matter (VM, %), and residence time (RT, min). Confirming its limited effect in this process, oxygen concentration (O, %) has little influence. The biochar yield model is dominated overall by compositional characteristics (N, Ash, C, and FC).

Conclusion

In conclusion, this study effectively demonstrates the application of machine learning models for biochar yield prediction, with a particular focus on random forest and XGBoost algorithms. Among the five models tested Lasso, Tweedie regression, random forest, Gradient Boosting Regression (GBR), and XGBoost. XGBoost regularly outperformed the others, showing extraordinary accuracy in both the training and test phases. With low mean squared error (MSE) values of 1.17 and 2.9 respectively, XGBoost caught 97.39% of the variance in training data and 88.75% in test data. Reflecting its accuracy and resilience, its mean absolute percentage error (MAPE) was likewise modest, registering only 2.14% in training and 3.8% in testing. Explaining 97.70% of the variance in the training data and 80.72% in the test data, gradient boosting regression also did rather well. It kept great predictive power even if its error margin marginally changed in testing. With more errors on test data and a reduced capacity to manage unknown data, random forest suffered with generalization as well. Less appropriate for reliable biochar yield prediction the Lasso and Tweedie models showed rather larger prediction errors and less capacity to explain data variance. Striking a mix of low prediction error and great generalization ability, XGBoost turned up as the most dependable model overall. Its exceptional performance has significant consequences for waste-to-energy conversion since it provides a prediction tool that can maximize biomass use and support sustainability initiatives in biochar manufacture. The XML based SHAP analysis revealed that pyrolysis temperature, nitrogen, and carbon content are key drivers of biochar yield, while ash content and moisture negatively influence the production. In conclusion, while the current study provides valuable insights into biochar yield prediction using machine learning models, the relatively small sample size of 134 data points could limit the generalizability of the findings. Future research with more experimental data could help to better grasp the fundamental trends and enable the creation of more dependable models fit for many real-world situations.

Data availability

The datasets during and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Received: 14 October 2024; Accepted: 20 February 2025
Published online: 27 February 2025

References

1. Neogi, S. et al. Sustainable biochar: A facile strategy for soil and environmental restoration, energygeneration, mitigation of global climate change and circular bioeconomy. *Chemosphere* **293**, 133474. https://doi.org/10.1016/j.chemosphere.2021.133474 (2022).
2. Trinh, V. L. & Chung, C. K. Renewable energy for SDG-7 and sustainable electrical production, integration, industrial application, and globalization: Review. *Clean Eng. Technol.* **15**, 100657. https://doi.org/10.1016/j.clet.2023.100657 (2023).
3. Gangadhari, R. K., Karadayi-Usta, S. & Lim, W. M. Breaking barriers toward a net-zero economy. *Nat. Resour. Forum* https://doi.org/10.1111/1477-8947.12378 (2023).
4. Tan, X. & Tan, T. Biofuels from biomass toward a net-zero carbon and sustainable world. *Joule* **6**, 1396–1399. https://doi.org/10.1016/j.joule.2022.06.005 (2022).
5. Olufemi, A. S., Osundare, O. S., Odeyemi, İ. O. & Kakar, M. Desulphurization of syngas produced from biomass using dolomite as adsorbent. *Turk. J. Eng.* **4**(3), 142–153. https://doi.org/10.31127/tuje.644597 (2020).
6. He, Y., Jaiswal, D., Long, S. P., Liang, X. Z. & Matthews, M. L. Biomass yield potential on U.S. marginal land and its contribution to reach net-zero emission. *GCB Bioenergy* https://doi.org/10.1111/gcbb.13128 (2024).
7. Anonas, S. D. S. et al. From waste to renewable energy: A policy review on waste-to-energy in the Philippines. *Sustainability (Switzerland)* **15**, 12963. https://doi.org/10.3390/su151712963 (2023).
8. Hu, Q. et al. Biochar industry to circular economy. *Sci. Total Environ.* **757**, 143820. https://doi.org/10.1016/j.scitotenv.2020.143820 (2021).
9. Natrayan, L. et al. Eco-friendly zinc oxide nanoparticles from *Moringa oleifera* leaf extract for photocatalytic and antibacterial applications. *Clean Technol. Environ. Policy* https://doi.org/10.1007/s10098-024-02814-1 (2024).
10. Kathi, S., Singh, S., Yadav, R., Singh, A. N. & Mahmoud, A. E. D. Wastewater and sludge valorisation: a novel approach for treatment and resource recovery to achieve circular economy concept. *Front. Chem. Eng.* **5**, 1129783. https://doi.org/10.3389/fceng.2023.1129783 (2023).
11. Preisner, M. et al. Indicators for resource recovery monitoring within the circular economy model implementation in the wastewater sector. *J. Environ. Manag.* **304**, 114261. https://doi.org/10.1016/j.jenvman.2021.114261 (2022).
12. Gregson, N., Crang, M., Fuller, S. & Holmes, H. Interrogating the circular economy: the moral economy of resource recovery in the EU. *Econ. Soc.* **44**, 1013353. https://doi.org/10.1080/03085147.2015.1013353 (2015).
13. Wani, I., Ramola, S., Garg, A. & Kushvaha, V. Critical review of biochar applications in geoengineering infrastructure: Moving beyond agricultural and environmental perspectives. *Biomass Convers. Biorefin.* **14**, 5943–5971. https://doi.org/10.1007/s13399-021-01346-8 (2024).
14. Wang, Y. et al. Research status, trends, and mechanisms of biochar adsorption for wastewater treatment: a scientometric review. *Environ. Sci. Eur.* **36**, 25. https://doi.org/10.1186/s12302-024-00859-z (2024).
15. Tian, H. et al. Optimizing the gasification reactivity of biochar: The composition, structure and kinetics of biochar derived from biomass lignocellulosic components and their interactions during gasification process. *Fuel* **324**, 124709. https://doi.org/10.1016/j.fuel.2022.124709 (2022).
16. Yuan, X. et al. Recent advancements and challenges in emerging applications of biochar-based catalysts. *Biotechnol. Adv.* **67**, 108181. https://doi.org/10.1016/j.biotechadv.2023.108181 (2023).
17. Yang, G. et al. Hydrogen-rich syngas production from biomass gasification using biochar-based nanocatalysts. *Bioresour. Technol.* **379**, 129005. https://doi.org/10.1016/j.biortech.2023.129005 (2023).

18. Muralidaran, V. M. et al. Grape stalk cellulose toughened plain weaved bamboo fiber-reinforced epoxy composite: load bearing and time-dependent behavior. *Biomass Conv. Bioref.* **14**, 14317–14324. https://doi.org/10.1007/s13399-022-03702-8 (2024).
19. Sahoo, D. & Remya, N. Influence of operating parameters on the microwave pyrolysis of rice husk: Biochar yield, energy yield, and property of biochar. *Biomass Convers. Biorefin.* **12**, 3447–3456. https://doi.org/10.1007/s13399-020-00914-8 (2022).
20. Batista, R. R. & Gomes, M. M. Effects of chemical composition and pyrolysis process variables on biochar yields: Correlation and principal component analysis. *Floresta e Ambiente* https://doi.org/10.1590/2179-8087-FLORAM-2021-0007 (2021).
21. Khan, M., Ullah, Z., Mašek, O., Raza Naqvi, S. & Nouman Aslam Khan, M. Artificial neural networks for the prediction of biochar yield: A comparative study of metaheuristic algorithms. *Bioresour. Technol.* **355**, 127215. https://doi.org/10.1016/J.BIORTECH.2022.127215 (2022).
22. Grafmüller, J. et al. Wood ash as an additive in biomass pyrolysis: Effects on biochar yield, properties, and agricultural performance. *ACS Sustain. Chem. Eng.* **10**, 2720–2729. https://doi.org/10.1021/acssuschemeng.1c07694 (2022).
23. Leng, L. et al. Machine learning predicting and engineering the yield, N content, and specific surface area of biochar derived from pyrolysis of biomass. *Biochar* **4**, 63. https://doi.org/10.1007/s42773-022-00183-w (2022).
24. Nguyen, V. G. et al. Machine learning for the management of biochar yield and properties of biomass sources for sustainable energy. *Biofuels Bioprod. Biorefining* https://doi.org/10.1002/bbb.2596 (2024).
25. Le, A. T. et al. Precise prediction of biochar yield and proximate analysis by modern machine learning and shapley additive explanations. *Energy Fuels* **37**, 17310–17327. https://doi.org/10.1021/acs.energyfuels.3c02868 (2023).
26. Chen, J., Zhang, M., Xu, Z., Ma, R. & Shi, Q. Machine-learning analysis to predict the fluorescence quantum yield of carbon quantum dots in biochar. *Sci. Total Environ.* **896**, 165136. https://doi.org/10.1016/j.scitotenv.2023.165136 (2023).
27. Kaliappan, S. et al. Mechanical, fatigue, and hydrophobic properties of silane-treated green pea fiber and egg fruit seed powder epoxy composite. *Biomass Conv. Bioref.* **14**, 24061–24068. https://doi.org/10.1007/s13399-023-04534-w (2024).
28. Li, Y., Gupta, R. & You, S. Machine learning assisted prediction of biochar yield and composition via pyrolysis of biomass. *Bioresour. Technol.* **359**, 127511. https://doi.org/10.1016/J.BIORTECH.2022.127511 (2022).
29. Pathak, S., Pant, K. K. & Kaushal, P. Analysis of naphthalene adsorption from wastewater using activated and non-activated biochar produced from bagasse. *Biomass Convers. Biorefin.* https://doi.org/10.1007/s13399-023-04070-7 (2023).
30. Abdullah, N., Mohd Taib, R., Mohamad Aziz, N. S., Omar, M. R. & Md, D. N. Banana pseudo-stem biochar derived from slow and fast pyrolysis process. *Heliyon* **9**, e12940. https://doi.org/10.1016/j.heliyon.2023.e12940 (2023).
31. Azizan, M. T. et al. Catalytic reforming of oxygenated hydrocarbons for the hydrogen production: an outlook. *Biomass Convers. Biorefin.* https://doi.org/10.1007/s13399-020-01081-6 (2020).
32. Hosseinzaei, B. et al. Effect of heating rate and H3PO4 as catalyst on the pyrolysis of agricultural residues. *J. Anal. Appl. Pyrolysis* **168**, 105724. https://doi.org/10.1016/j.jaap.2022.105724 (2022).
33. Rathod, N., Jain, S. & Patel, M. R. Thermodynamic analysis of biochar produced from groundnut shell through slow pyrolysis. *Energy Nexus* **9**, 100177. https://doi.org/10.1016/j.nexus.2023.100177 (2023).
34. Premchand, P., Demichelis, F., Chiaramonti, D., Bensaid, S. & Fino, D. Study on the effects of carbon dioxide atmosphere on the production of biochar derived from slow pyrolysis of organic agro-urban waste. *Waste Manag.* **172**, 308–319. https://doi.org/10.1016/j.wasman.2023.10.035 (2023).
35. Dhar, S. A., Sakib, T. U. & Hilary, L. N. Effects of pyrolysis temperature on production and physicochemical characterization of biochar derived from coconut fiber biomass through slow pyrolysis process. *Biomass Convers. Biorefin.* **12**, 2631–2647. https://doi.org/10.1007/s13399-020-01116-y (2022).
36. Devi, P. & Dalai, A. K. Occurrence, distribution, and toxicity assessment of polycyclic aromatic hydrocarbons in biochar, biocrude, and biogas obtained from pyrolysis of agricultural residues. *Bioresour. Technol.* **384**, 129293. https://doi.org/10.1016/j.biortech.2023.129293 (2023).
37. Aqsha, A., Tijani, M. M., Moghtaderi, B. & Mahinpey, N. Catalytic pyrolysis of straw biomasses (wheat, flax, oat and barley) and the comparison of their product yields. *J. Anal. Appl. Pyrolysis* **125**, 201–208. https://doi.org/10.1016/j.jaap.2017.03.022 (2017).
38. Yousefian, F., Babatabar, M. A., Eshaghi, M., Poor, S. M. & Tavasoli, A. Pyrolysis of Rice husk, Coconut shell, and Cladophora glomerata algae and application of the produced biochars as support for cobalt catalyst in Fischer-Tropsch synthesis. *Fuel Process. Technol.* **247**, 107818. https://doi.org/10.1016/j.fuproc.2023.107818 (2023).
39. Sahoo, S. S., Vijay, V. K., Chandra, R. & Kumar, H. Production and characterization of biochar produced from slow pyrolysis of pigeon pea stalk and bamboo. *Clean Eng. Technol.* **3**, 100101. https://doi.org/10.1016/j.clet.2021.100101 (2021).
40. Sun, Y. et al. Pyrolysis of soybean residue: Understanding characteristics of the products. *Renew Energy* **174**, 487–500. https://doi.org/10.1016/j.renene.2021.04.063 (2021).
41. Ma, C. et al. Comprehensive investigation on the slow pyrolysis product characteristics of waste tobacco stem: Pyrolysis reaction mechanism and conversion mechanism of N. *Fuel* **350**, 128902. https://doi.org/10.1016/j.fuel.2023.128902 (2023).
42. Sakhiya, A. K., Anand, A., Aier, I., Vijay, V. K. & Kaushal, P. Suitability of rice straw for biochar production through slow pyrolysis: Product characterization and thermodynamic analysis. *Bioresour. Technol. Rep.* **15**, 100818. https://doi.org/10.1016/j.biteb.2021.100818 (2021).
43. Setter, C. et al. Slow pyrolysis of coffee husk briquettes: Characterization of the solid and liquid fractions. *Fuel* **261**, 116420. https://doi.org/10.1016/j.fuel.2019.116420 (2020).
44. Kaur, R., Tarun Kumar, V., Krishna, B. B. & Bhaskar, T. Characterization of slow pyrolysis products from three different cashew wastes. *Bioresour. Technol.* **376**, 128859. https://doi.org/10.1016/j.biortech.2023.128859 (2023).
45. Kaur, R., Kumar, A., Biswas, B., Krishna, B. B. & Bhaskar, T. Investigations into pyrolytic behaviour of spent citronella waste: Slow and flash pyrolysis study. *Bioresour. Technol.* **366**, 128202. https://doi.org/10.1016/j.biortech.2022.128202 (2022).
46. Vicente Rubi, R. et al. Slow pyrolysis of buri palm: Investigation of pyrolysis temperature and residence time effects. *Mater Today Proc.* https://doi.org/10.1016/j.matpr.2023.04.454 (2023).
47. Sarkar, J. K. & Wang, Q. Different pyrolysis process conditions of South Asian waste coconut shell and characterization of gas, biochar, and bio-oil. *Energies (Basel)* **13**, 1970. https://doi.org/10.3390/en13081970 (2020).
48. Vilas-Boas, A. C. M. et al. Valorisation of residual biomass by pyrolysis: influence of process conditions on products. *Sustain Energy Fuels* **8**, 379–396. https://doi.org/10.1039/d3se01216f (2023).
49. Chantanumat, Y. et al. Characterization of bio-oil and biochar from slow pyrolysis of oil palm plantation and palm oil mill wastes. *Biomass Convers. Biorefin.* **13**, 1–13. https://doi.org/10.1007/s13399-021-02291-2 (2023).
50. El kourdi, S., Chaabane, A., Abderafi, S. & Abbassi, M. A. Valorizing argan residues into biofuels and chemicals through slow pyrolysis. *Results Eng.* **21**, 101659. https://doi.org/10.1016/j.rineng.2023.101659 (2024).
51. Kaur, R., Kumar, A., Biswas, B., Krishna, B. B. & Bhaskar, T. Py-GC/MS and slow pyrolysis of tamarind seed husk. *J. Mater Cycles Waste Manag.* **26**, 1–16. https://doi.org/10.1007/s10163-024-01888-9 (2024).
52. Li, Y., Yang, R., Wang, X., Zhu, J. & Song, N. Carbon price combination forecasting model based on lasso regression and optimal integration. *Sustainability (Switzerland)* **15**, 9354. https://doi.org/10.3390/su15129354 (2023).
53. Ayyıldız, E. & Murat, M. A lasso regression-based forecasting model for daily gasoline consumption: Türkiye Case. *Turk. J. Eng.* **8**, 162–174. https://doi.org/10.31127/tuje.1354501 (2024).
54. Bonat, W. H. & Kokonendji, C. C. Flexible Tweedie regression models for continuous data. *J. Stat. Comput. Simul.* **87**, 2138–2152. https://doi.org/10.1080/00949655.2017.1318876 (2017).
55. Kokonendji, C. C., Bonat, W. H. & Abid, R. Tweedie regression models and its geometric sums for (semi-)continuous data. *Wiley Interdiscip. Rev. Comput. Stat.* https://doi.org/10.1002/wics.1496 (2021).

56. Said, Z., Sharma, P., Bora, B. J. & Pandey, A. K. Sonication impact on thermal conductivity of f-MWCNT nanofluids using XGBoost and Gaussian process regression. *J. Taiwan Inst. Chem. Eng.* **145**, 104818. https://doi.org/10.1016/J.JTICE.2023.104818 (2023).
57. Alhakeem, Z. M. et al. Prediction of ecofriendly concrete compressive strength using gradient boosting regression tree combined with GridSearchCV hyperparameter-optimization techniques. *Materials* **15**, 7432. https://doi.org/10.3390/ma15217432 (2022).
58. Xiong, X., Guo, X., Zeng, P., Zou, R. & Wang, X. A short-term wind power forecast method via XGBoost hyper-parameters optimization. *Front. Energy Res.* https://doi.org/10.3389/fenrg.2022.905155 (2022).
59. Kanti, P. K., Shrivastav, A. P., Sharma, P. & Maiya, M. P. Thermal performance enhancement of metal hydride reactor for hydrogen storage with graphene oxide nanofluid: Model prediction with machine learning. *Int. J. Hydrogen Energy* https://doi.org/10.1016/J.IJHYDENE.2023.03.361 (2023).
60. Gupta, H. V., Kling, H., Yilmaz, K. K. & Martinez, G. F. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *J. Hydrol. (Amst.)* **377**, 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003 (2009).
61. Gupta, H. V. & Kling, H. On typical range, sensitivity, and normalization of Mean Squared Error and Nash-Sutcliffe Efficiency type metrics. *Water Resour. Res.* **47**, 10601. https://doi.org/10.1029/2011WR010962 (2011).
62. Elvidge, S., Angling, M. J. & Nava, B. On the use of modified Taylor diagrams to compare ionospheric assimilation models. *Radio Sci.* **49**, 737–745. https://doi.org/10.1002/2014RS005435 (2014).
63. Nwafor, E. O. & Akintayo, F. O. Predicting trip purposes of households in Makurdi using machine learning: A comparative analysis of Decision Tree, CatBoost, and XGBoost algorithms. *Eng. Appl.* **3**(3), 260–274 (2024).
64. Hintze, J. L. & Nelson, R. D. Violin plots: A box plot-density trace synergism. *Am. Stat.* **52**, 181–184. https://doi.org/10.2307/2685478 (1998).
65. Jamei, M. et al. Application of an explainable glass-box machine learning approach for prognostic analysis of a biogas-powered small agriculture engine. *Energy* **288**, 129862. https://doi.org/10.1016/j.energy.2023.129862 (2024).
66. Kumar, K. P., Deepthi Jayan, K., Sharma, P. & Alruqi, M. Thermo-electro-rheological properties of graphene oxide and MXene hybrid nanofluid for vanadium redox flow battery: Application of explainable ensemble machine learning with hyperparameter optimization. *FlatChem* **43**, 100606. https://doi.org/10.1016/j.flatc.2023.100606 (2024).
67. Hajderi, A., Bozo, L. & Basholli, F. The impact of alternative fuel on diesel in reducing of pollution from vehicles. *Adv. Eng. Sci.* **4**, 15–24 (2024).
68. Futagami, K., Fukazawa, Y., Kapoor, N. & Kito, T. Pairwise acquisition prediction with SHAP value interpretation. *J. Finance Data Sci.* **7**, 22–44. https://doi.org/10.1016/J.JFDS.2021.02.001 (2021).

Acknowledgements

The authors would like to thank their respective institutions for their extended support throughout this research work.

Author contributions

S.U.: conceptualization, methodology, data curation, writing—original draft. PP: Software, resources, supervision N.K., J.S.C.: writing—review and editing, resources, methodology. P.K.K., H.V.: Former Analysis, Validation, writing—review, and editing. L.H.D., R.G.: writing—review and editing, Project Administration, supervision, Validation.

Funding

No Fundings received for this Research work.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to P.P. or L.H.D.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.