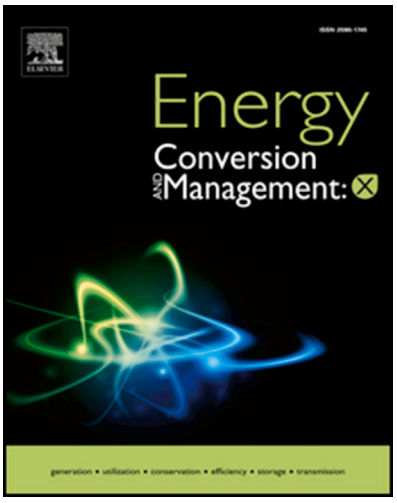
Journal Pre-proofs



On the construction of hybrid algorithms to predict biochar yield as a function of pyrolysis parameters

Kusum Yadav, Lulwah M. Alkwai, Shahad Almansour, Mehrdad Mottaghi

PII: S2590-1745(26)00530-1
DOI: https://doi.org/10.1016/j.ecmx.2026.102047
Reference: ECMX 102047

To appear in: *Energy Conversion and Management: X*

Received Date: 9 December 2025
Revised Date: 6 May 2026
Accepted Date: 5 June 2026

Please cite this article as: K. Yadav, L.M. Alkwai, S. Almansour, M. Mottaghi, On the construction of hybrid algorithms to predict biochar yield as a function of pyrolysis parameters, *Energy Conversion and Management: X* (2026), doi: https://doi.org/10.1016/j.ecmx.2026.102047

On the construction of hybrid algorithms to predict biochar yield as a function of pyrolysis parameters

Kusum Yadav [1], Lulwah M. Alkwai [1], Shahad Almansour [2], and Mehrdad Mottaghi [3]*

1 College of Computer Science and Engineering. University of Hail, Hail. Kingdom of Saudi Arabia

2 Applied College, University of Hail, Hail, Kingdom of Saudi Arabia

3 Faculty of chemistry, Kabul University, Kabul, Afghanistan

***Corresponding Author: mmottaghi41@gmail.com**

Abstract

Biochar yield prediction is essential for adjusting pyrolysis progressions and proceeding justifiable biomass utilization. This research develops a unified, interpretable framework that join in Gradient Boosting Decision Trees (GBDT) with metaheuristic optimizing algorithms, Ant Colony Optimization (ACO), Whale Optimization Algorithm (WOA), Coupled Simulated Annealing (CSA), and Particle Swarm Optimization (PSO), to enhance predictive accuracy. A comprehensive dataset of 211 experimentally validated pyrolysis observations, comprising 14 physicochemical and operational parameters, was curated from peer-reviewed literature. After preprocessing through outlier detection and normalization, all models were trained using 5-fold cross-validation and evaluated via R^2, MSE, and AARE%. SHAP analysis was employed to quantify feature contributions and elucidate mechanistic relationships between pyrolysis conditions and biochar yield. Among the optimized models, the GBDT-ACO model achieved the highest predictive accuracy, yielding test R^2 values of 0.709 and test MSE of 16.284. The results highlight the critical influence of ash content, residence time, and peak temperature on yield formation, while also revealing the trade-off between computational efficiency and predictive robustness across optimization strategies.

Keywords: SHAP interpretability; Machine learning; Hyperparameter tuning; Sustainable biomass utilization

1. Introduction

Biomass is a major renewable carbon resource, while continued fossil fuel use causes serious environmental problems, especially greenhouse gas emissions and global warming [1-5]. Consequently, the transition toward renewable energy alternatives has gained critical momentum. Among these alternatives, biomass is distinguished by its sustainability, widespread availability, and reduced net carbon footprint. Within this paradigm, biochar has garnered substantial academic and industrial interest owing to its multifaceted utility in agricultural enhancement, bioenergy generation, and sustainable waste management [5-9].

Biochar properties depend mainly on the biomass type and the pyrolysis conditions [10-13]. Critical operational variables include peak temperature, heating rate, ambient pressure, and residence time [14]. Elevated pyrolysis temperatures typically facilitate the volatilization of organic matter, which consequently diminishes the overall biochar yield while augmenting its aromaticity and structural recalcitrance [15-17]. Similarly, carbonization efficiency is profoundly governed by residence time; extended durations ensure the thorough thermal degradation of the biomass, systematically elevating the fixed carbon content within the final product [18-20]. Moreover, system pressure and the choice of pyrolytic atmosphere exert a significant influence on conversion dynamics [21-24].

The optimization of pyrolysis processes and the advancement of sustainable biomass utilization directly support several critical UN Sustainable Development Goals (SDGs). By offering a viable pathway to transition away from fossil fuels and reduce greenhouse gas emissions, biochar production contributes significantly to Affordable and Clean Energy (SDG 7) and Climate Action (SDG 13). Furthermore, the multifaceted utility of biochar in sustainable waste management and agricultural enhancement aligns closely with Responsible Consumption and Production (SDG 12) and supports efforts toward sustainable land use and food security (SDG 15 and SDG 2).

Recent studies have demonstrated that interpretable machine-learning frameworks can simultaneously achieve high predictive accuracy and provide mechanistic insights into pyrolysis behavior. For example, Paudel et al. (2025) [25] employed an ANN-PSO hybrid model to jointly predict biochar yield, higher heating value (HHV), and carbon content with an R^2 of 0.909. Their feature-importance, partial-dependence, and SHAP analyses consistently identified pyrolysis temperature and volatile matter as dominant drivers of biochar formation, underscoring the value of explainable ML tools for process optimization.

Contemporary research has increasingly utilized machine learning [26] to forecast biochar yield and evaluate biomass characteristics, with a pronounced shift toward explainable artificial intelligence (XAI) techniques [27], particularly SHAP [12, 28-31]. These studies emphasize the necessity of embedding interpretability within predictive frameworks, ensuring that robust statistical performance is augmented by a mechanistic comprehension of underlying pyrolysis dynamics [32]. Integrating these analytical tools positions the current investigation within the broader academic discourse, highlighting both the methodological novelty and the applied significance of the proposed GBDT-ACO framework [33].

However, a salient gap persists in the current literature: the scarcity of research that concurrently evaluates multiple metaheuristic and probabilistic optimization algorithms against a singular predictive baseline. This deficiency restricts both algorithmic progression and practical deployment, necessitating the development of a holistic, optimization-centric, and highly transparent paradigm for biochar yield forecasting.

Addressing this void, the present study establishes a sophisticated ML-based approach to optimize the production of biochar, thereby maximizing its environmental and agricultural utility. To further augment predictive precision, this study integrates a GBDT model with four distinct metaheuristic optimization architectures: Coupled Simulated Annealing (CSA), Ant Colony Optimization (ACO), Whale Optimization Algorithm (WOA), and Particle Swarm Optimization (PSO). This comparative methodology leverages the unique search capabilities of diverse optimization paradigms, ultimately identifying the GBDT-ACO configuration as the most efficacious approach. Through the integration of rigorous data curation, multi-algorithmic optimization, and SHAP-based interpretability, this study establishes a unified and transparent framework that advances beyond prior machine-learning approaches. The resulting model is physically interpretable, offering both methodological innovation and deeper practical insights into biochar synthesis. The complete workflow is depicted in Figure 1.

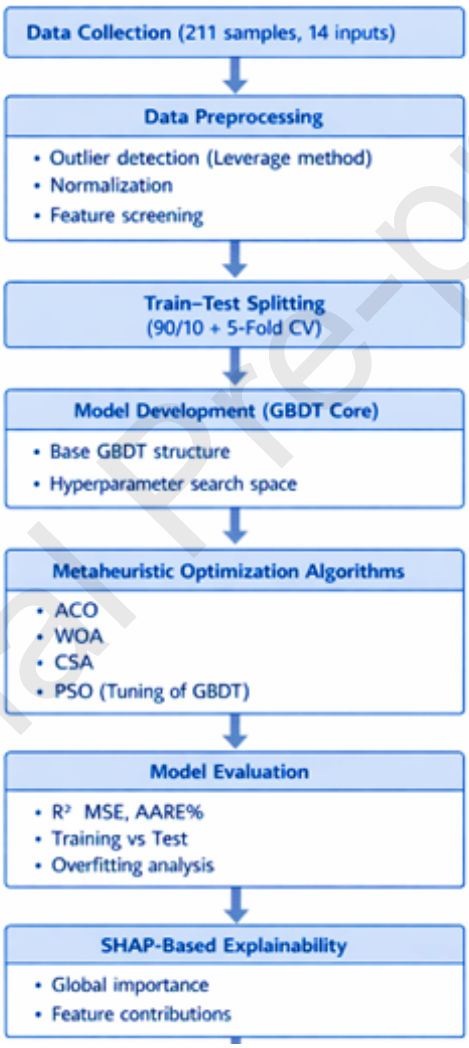


Figure 1. Schematic representation of the study workflow

2. Estimator and Optimizers Methodology

2.1. Gradient Boosting Decision Tree

This algorithm is a highly robust ensemble learning algorithm that constructs predictive models by sequentially combining multiple weak learners, typically decision trees. Unlike independent

ensemble methods, GBDT operates iteratively, where each successive tree is specifically trained to minimize the residual errors of the preceding ensemble. Mathematically, the model updates its predictions by optimizing a differentiable loss function, $L(y,F(x))$, using steepest gradient descent. At each iteration m, a new tree $h_m(x)$is fitted to the negative gradient of the loss function, and the model is updated according to $F_m(x) = F_{m-1}(x) + \nu \cdot h_m(x)$, where ν represents the learning rate [34]. This sequential error-correction mechanism allows GBDT to handle complex, non-linear relationships in high-dimensional datasets, making it exceptionally well-suited for predicting biochar yield based on varied pyrolysis and feedstock parameters.

2.2. Coupled Simulated Annealing

These Coupled Simulated Annealing (CSA) is an advanced probabilistic optimization algorithm designed to overcome the limitations of traditional Simulated Annealing (SA) by mitigating the risk of entrapment in local optima. While standard SA relies on a single search trajectory guided by a cooling schedule, CSA employs a multitude of parallel optimizers (probes) that are coupled together. The acceptance probability of a new solution is determined not only by the individual probe's energy state but also by the variance of the cost functions across all coupled probes. By evaluating the global state of the ensemble, denoted by the acceptance probability function $P(x_i \rightarrow y_i)$, CSA dynamically adjusts its search strategy, ensuring a more exhaustive exploration of the hyperparameter space and accelerating convergence toward the global optimum [35].

2.3. Whale Optimization Algorithm

The Whale Optimization Algorithm (WOA) is a nature-inspired metaheuristic based on the unique bubble-net hunting strategy of humpback whales. The algorithm mathematically models this behavior through three primary phases: shrinking encircling prey, bubble-net attacking (exploitation), and searching for prey (exploration). During the exploitation phase, whales update their positions based on a spiral equation mimicking the helix-shaped movement toward the best-known position, X^*. The position update is governed by the equation $X(t+1) = D' \cdot e^{bl} \cdot \cos(2\pi l) + X^*(t)$, where $D'$is the distance to the prey, b is a logarithmic spiral shape constant, and l is a random number in $[-1,1]$. By dynamically transitioning between these mathematically defined exploratory and exploitative phases, WOA efficiently navigates the hyperparameter space to optimize the predictive framework [36].

2.4. Particle Swarm Optimization

The Particle Swarm Optimization (PSO) is a population-based stochastic optimization technique inspired by the social dynamics of bird flocking and fish schooling. In the PSO framework, potential solutions are represented as "particles" navigating through a multi-dimensional search space. Each particle $i$updates its velocity $v_i(t)$and position $x_i(t)$at each iteration $t$based on its own best-known historical position, $P_{best,i}$, and the global best-known position achieved by the entire swarm, G_{best}. The velocity update formula is typically expressed as $v_i(t+1) = w \cdot v_i(t) + c_1 \cdot r_1 \cdot (P_{best,i} - x_i(t)) + c_2 \cdot r_2 \cdot (G_{best} - x_i(t))$, where w is the inertia weight, c_1 and c_2 are cognitive and social acceleration coefficients, and

$r_1$and r_2 are random variables [37]. This continuous exchange of individual and collective intelligence enables rapid convergence when tuning the GBDT model parameters.

2.5. Ant Colony Optimization

Inspired by the foraging behavior of natural ant colonies, Ant Colony Optimization (ACO) is a swarm intelligence metaheuristic utilized for solving complex combinatorial optimization problems. The algorithm relies on the concept of stigmergy, where artificial ants communicate indirectly by depositing virtual pheromones, denoted as τ_{ij}, on paths that lead to high-quality solutions. During the search process, the probability $p_{ij}^{k}$of an ant $k$transitioning from state $i$to state $j$is governed by a combination of the accumulated pheromone trail and a heuristic desirability value, η_{ij}. Over successive iterations, pheromone evaporation prevents premature convergence, while paths yielding superior predictive accuracy in the GBDT model receive stronger pheromone updates, effectively guiding the swarm toward optimal hyperparameter configurations [38].

3. Data explanation and processing

3.1. Data Elucidation

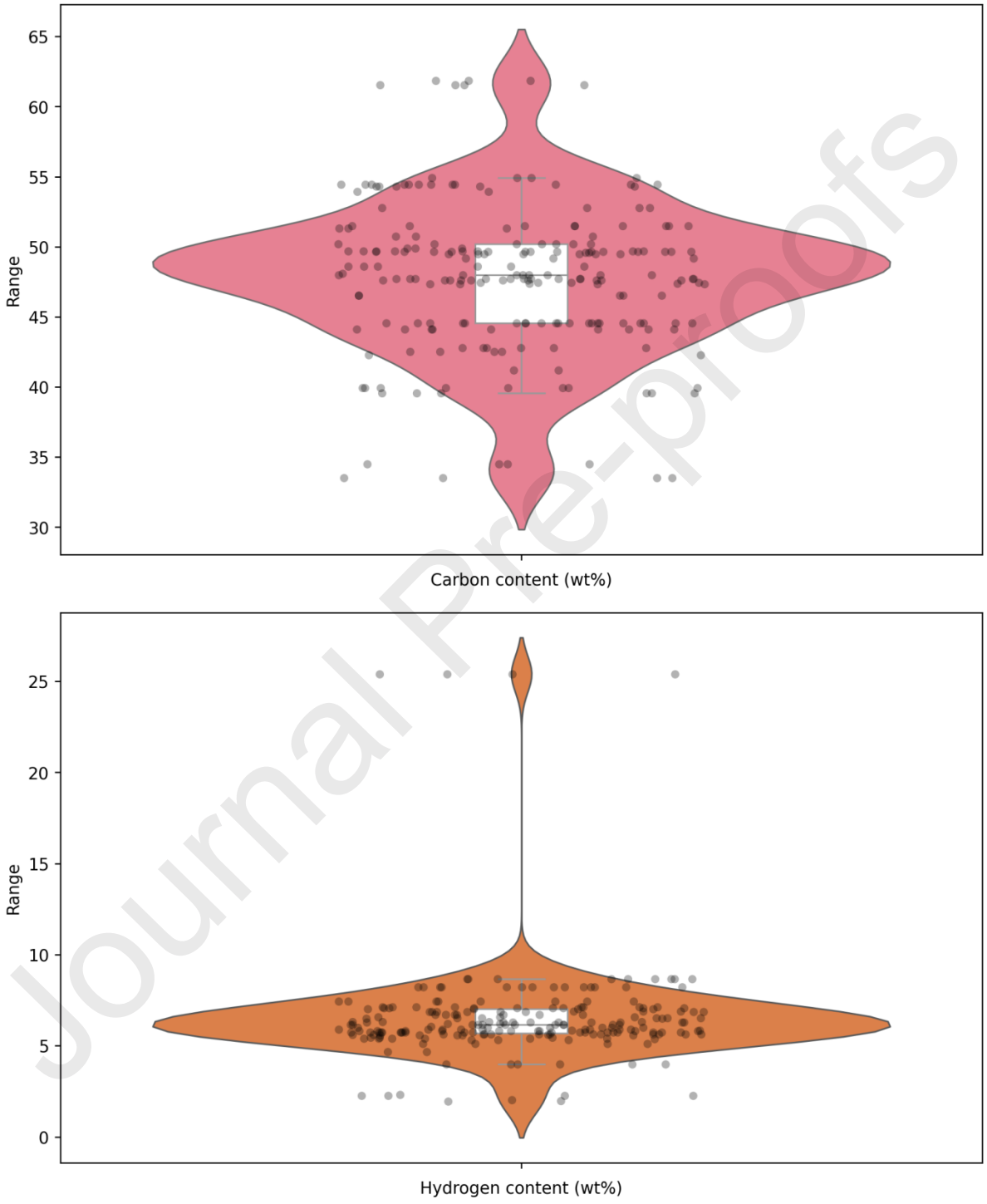
The empirical foundation for the predictive modeling framework was established through a comprehensive systematic review of existing literature concerning biochar synthesis via biomass pyrolysis [39-45]. The study utilizes a consolidated dataset comprising 211distinct data points, curated entirely from rigorously peer-reviewed experimental investigations. By excluding proprietary or unverified in-house data, the dataset maintains a high standard of empirical validity and encompasses a diverse spectrum of biomass feedstocks and pyrolytic operating conditions. Ultimately, the compiled matrix comprises 14 independent input features utilized to predict a singular target variable across various architectures.

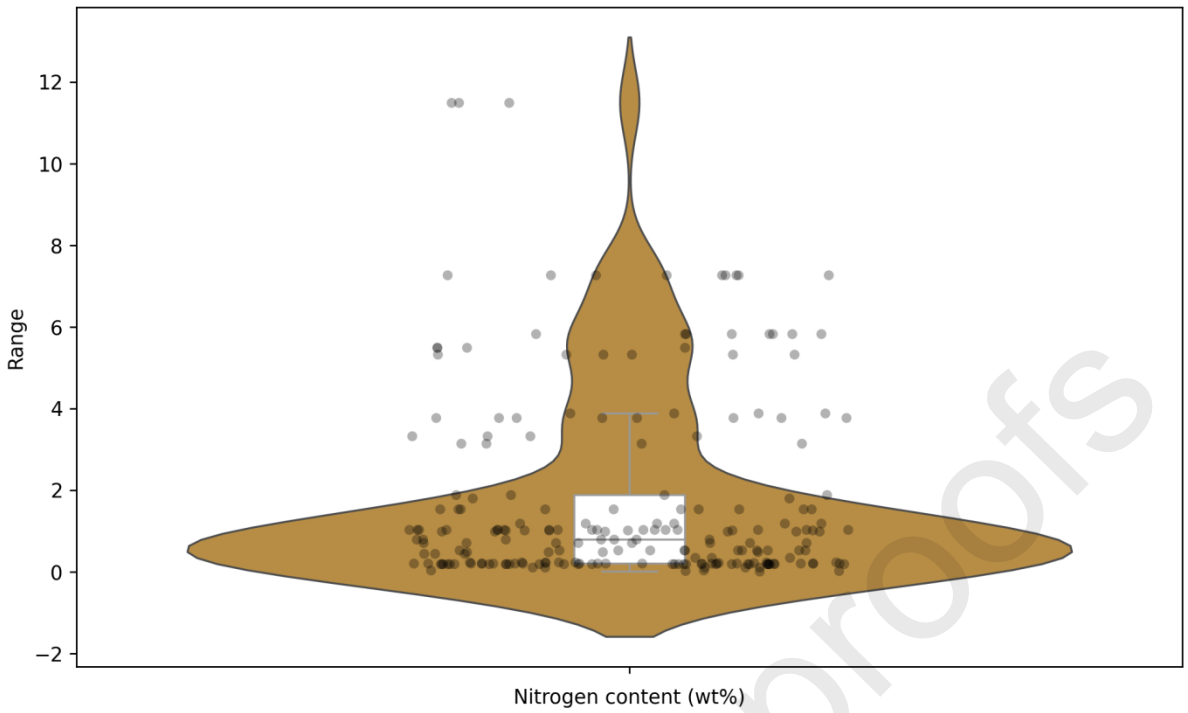
The 14 independent parameters were systematically classified into three distinct categories based on their physicochemical and operational roles. The first category encompasses the thermochemical and elemental composition of the raw biomass, specifically including the mass fractions of carbon (C), hydrogen (H), nitrogen (N), oxygen (O), ash, fixed carbon, and volatile matter, in addition to the acidic site density of the employed catalyst (quantified in $mmol/g$). The second classification pertains to the physical and structural properties of the feedstock, defined by the crystallinity index and the Brunauer-Emmett-Teller (BET) surface area. The final category encompasses the operational features of pyrolysis, namely the catalyst-to-biomass ratio, peak temperature, and residence time (minutes). The dependent variable is biochar yield, reported as the weight percentage (wt%) of the solid carbonaceous residue obtained after pyrolysis. While the pyrolytic process concurrently generates syngas and bio-oil fractions, this study exclusively targets biochar yield due to its critical, well-established applications in agronomy, particularly concerning soil amelioration and sustainable food production.

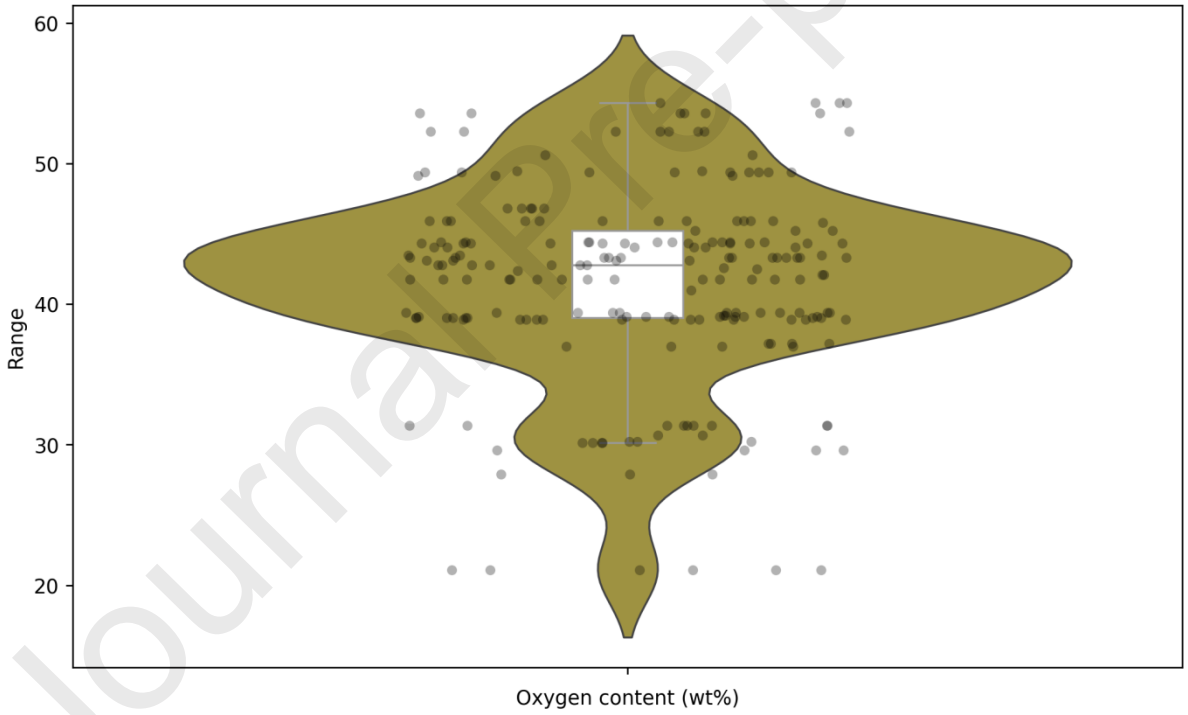
To construct and evaluate the machine learning models, the complete dataset of 211 observations was partitioned utilizing a 90:10split ratio. Specifically, 90% of the data was allocated to the training subset for algorithm learning and optimization, while the remaining 10

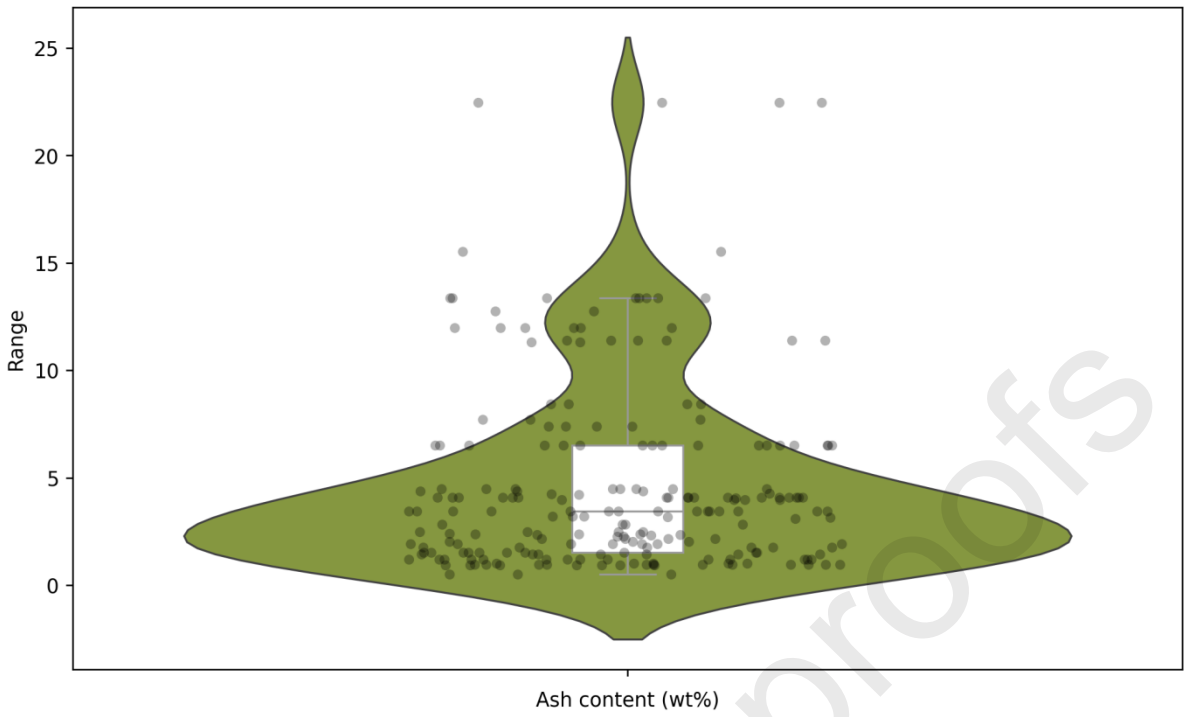
%was sequestered as an independent testing subset to rigorously evaluate model generalization and predictive validity. A statistical summary of the dataset's overall parametric distribution is graphically depicted via the violon plot presented in Figure 2.

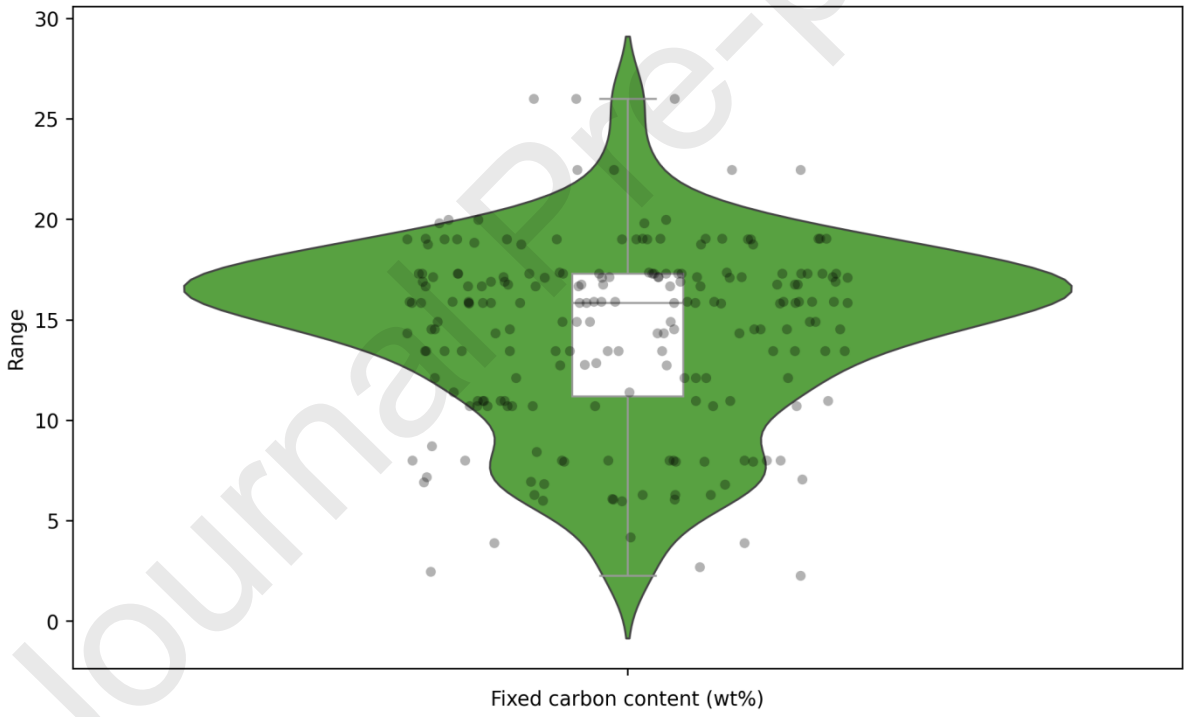
.

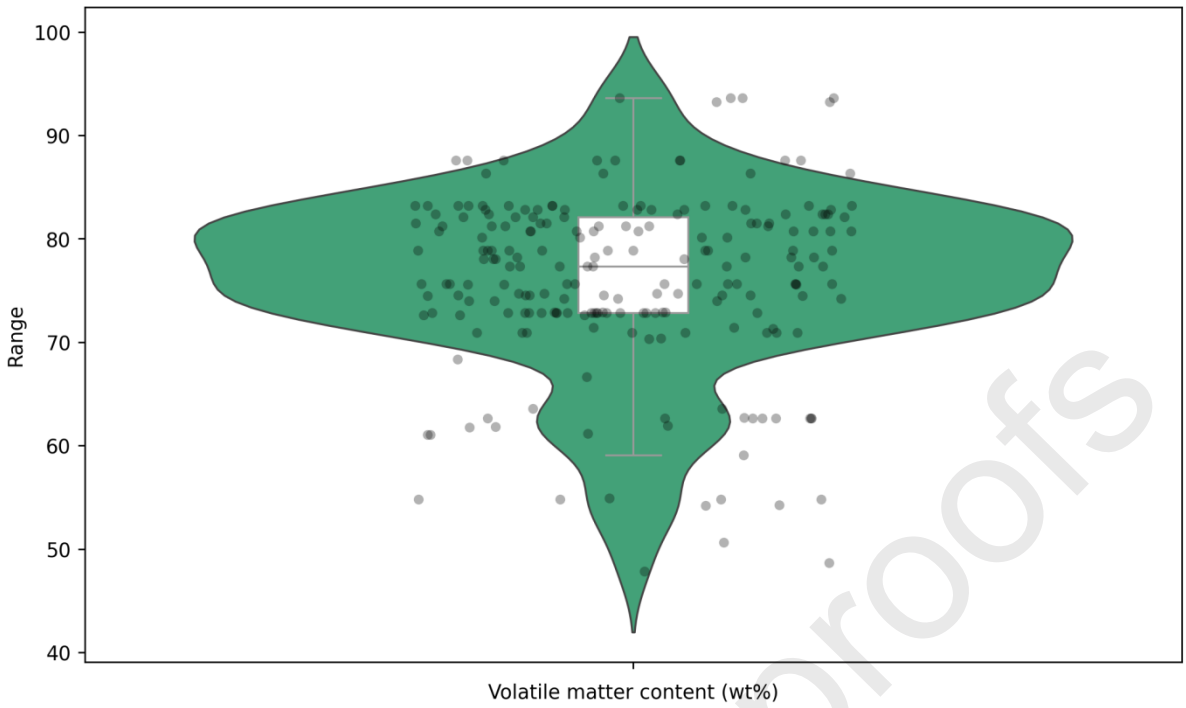


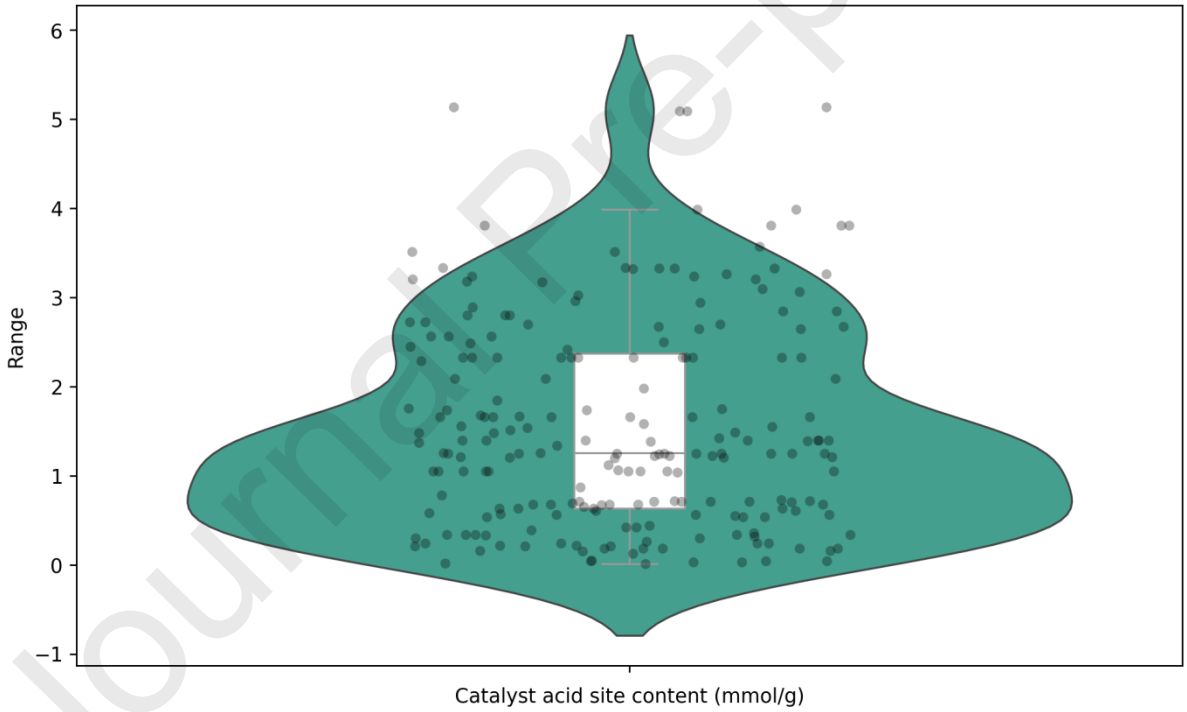


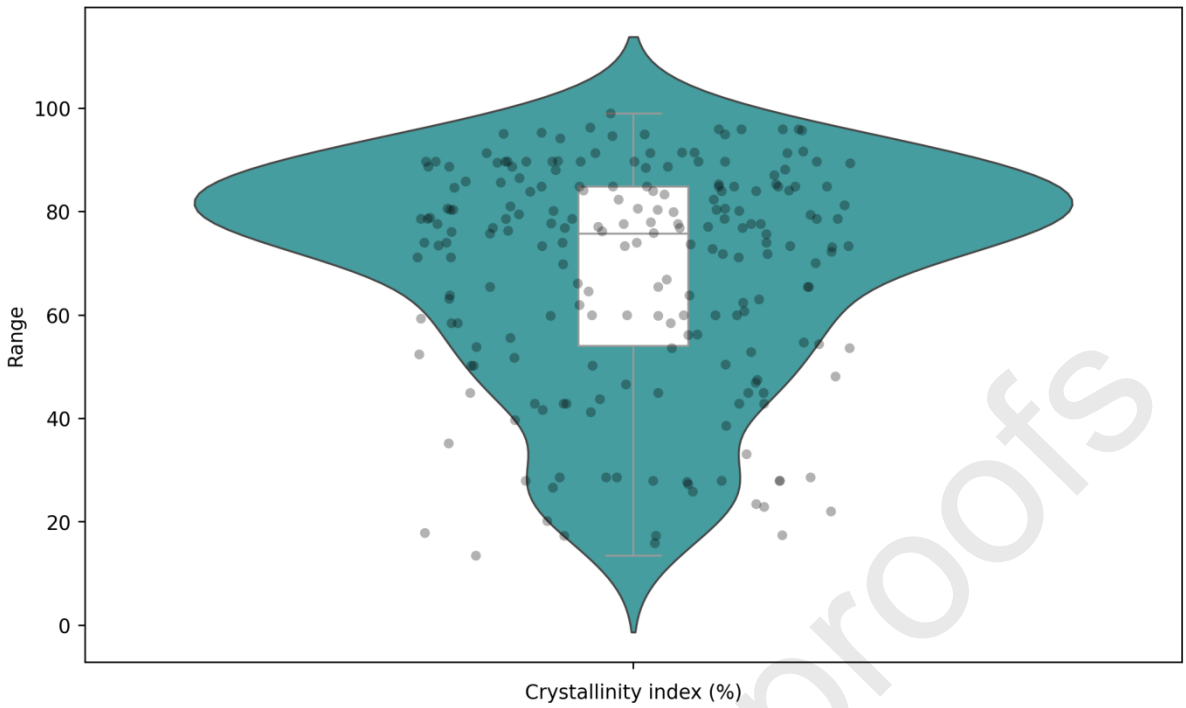


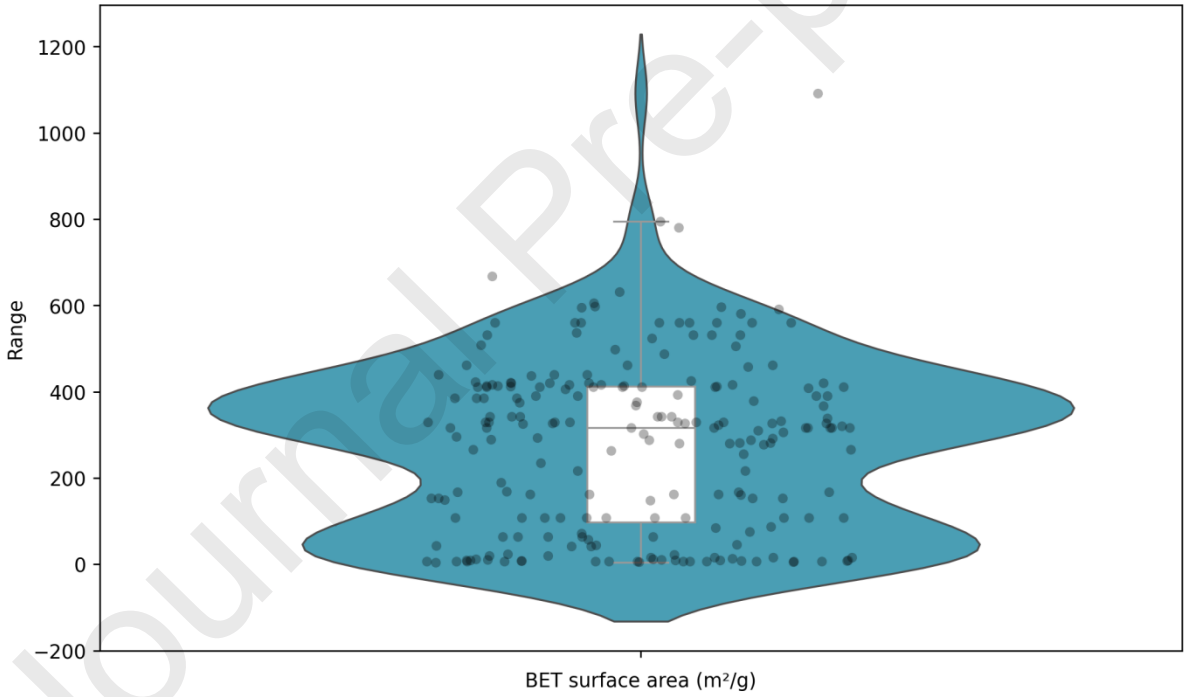


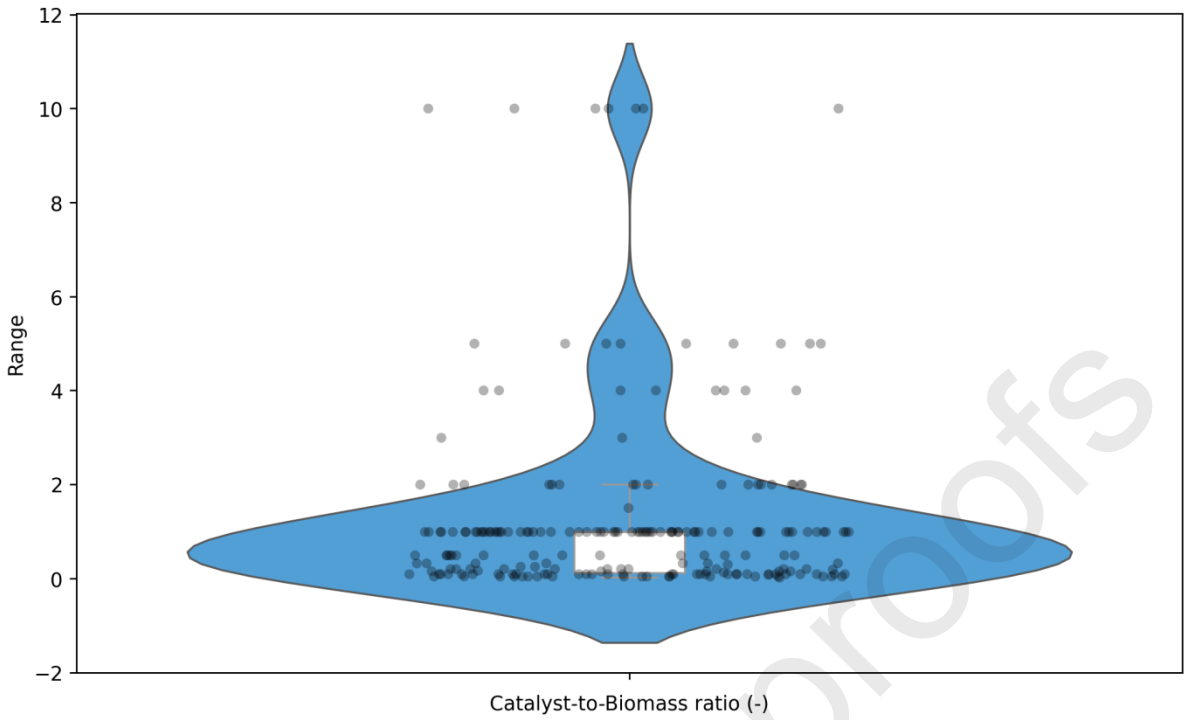


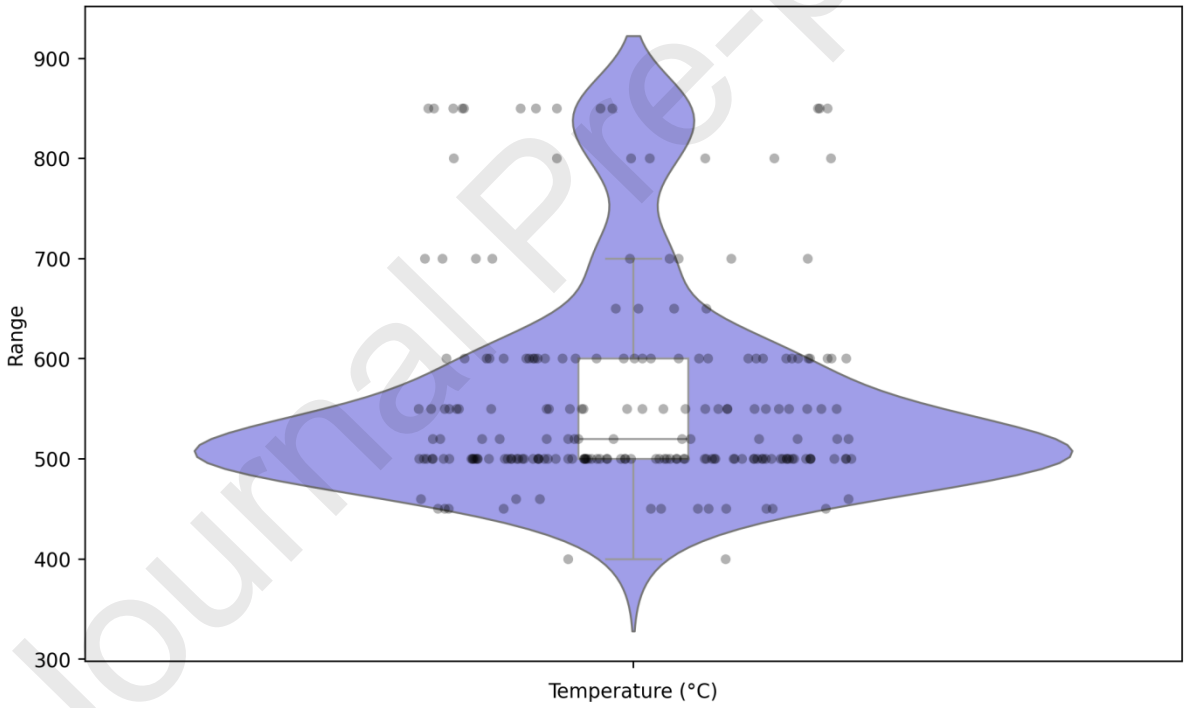


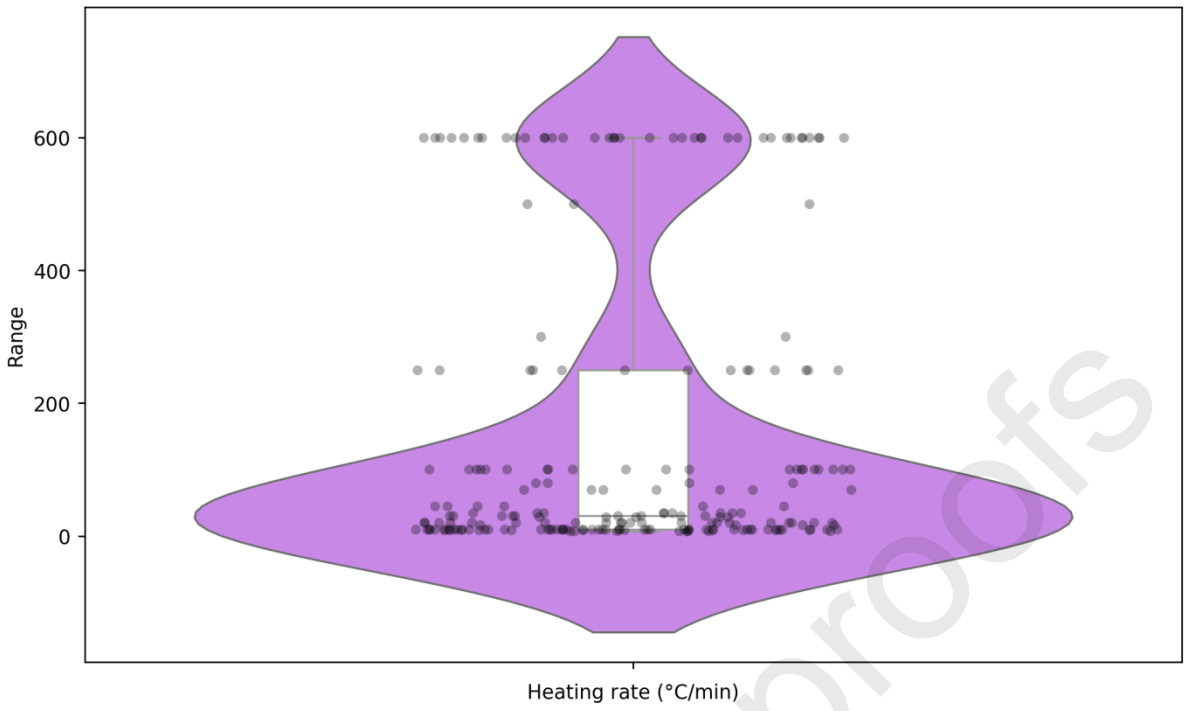












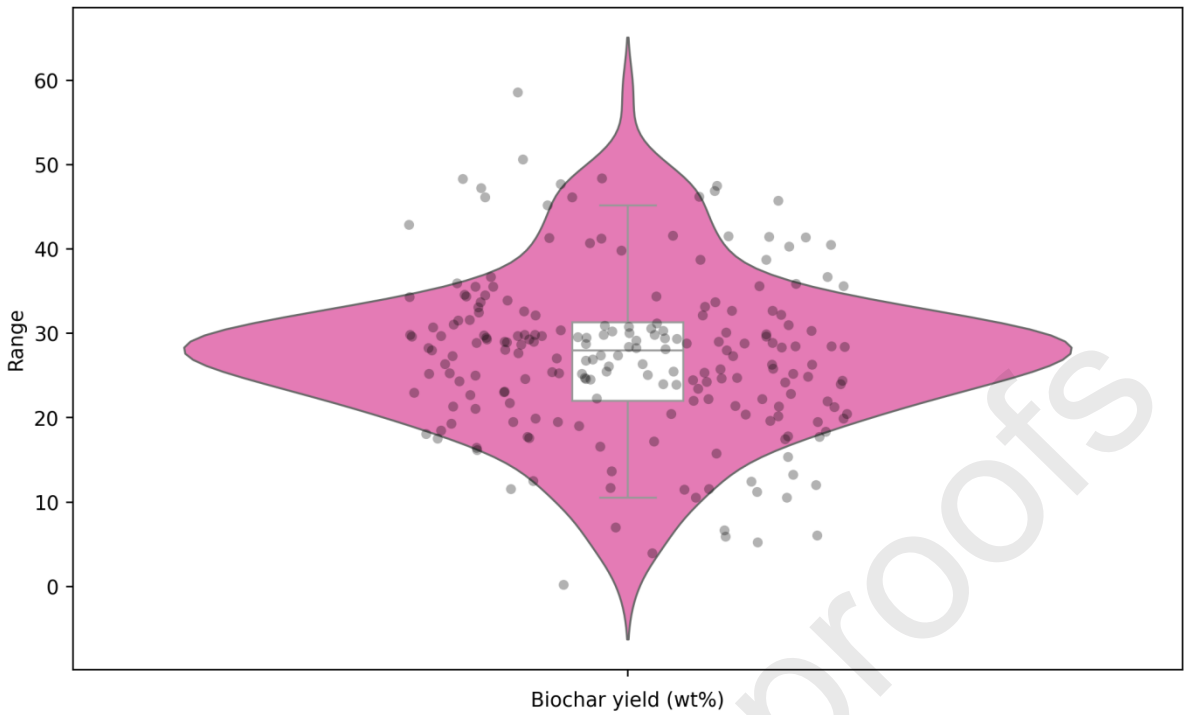


Figure 2. Data points' violon plot

All variables were harmonized prior to model development. Biochar yield values were converted to dry-basis weight percent, residence time was standardized to minutes, and catalyst loading was normalized to a mass ratio (g catalyst per g biomass). When multiple operating conditions were reported, only steady-state pyrolysis data were retained. Feedstock composition (C, H, N, O, ash, volatile matter, fixed carbon) was extracted directly from proximate or ultimate analyses as reported in each study.

To ensure robust model evaluation and mitigate the risk of overfitting, K-fold cross-validation was implemented as a foundational validation strategy. This statistical resampling technique systematically partitions the overall dataset into Kmutually exclusive subsets. Across K iterations, the algorithms are trained on $K-1$folds, while the single withheld fold functions as an independent validation set. The model's overarching predictive efficacy is subsequently determined by averaging the performance metrics across all Kiterations. By neutralizing the bias inherent in arbitrary, static train-test data splits, this method yields a highly reliable assessment of a model's capacity to generalize to unseen data. As graphically depicted in Figure 3, this study applied a 5-fold cross-validation protocol to all predictive algorithms during the training phase, thereby guaranteeing a rigorous statistical evaluation and fostering the development of highly stable and reliable models.

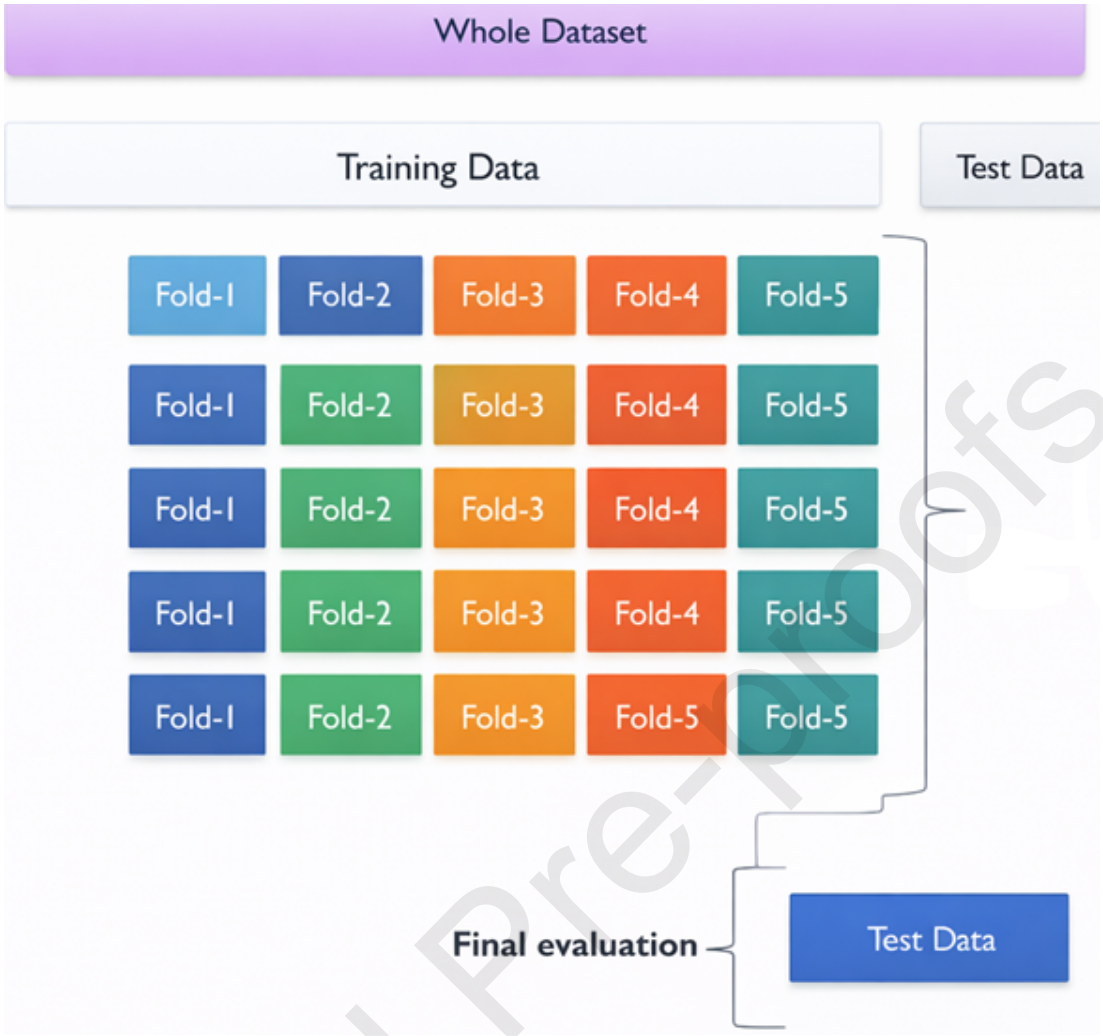


Figure 3. Schematic representation of the K-fold cross-validation procedure

3.2. Models' assessment

The predictive efficacy of the proposed models was quantitatively evaluated utilizing three primary statistical metrics. The Mean Squared Error (MSE) calculates the average of the squared residuals between the predicted and observed values. By squaring the deviations, this metric inherently penalizes larger forecast errors, thereby providing a clear indication of error magnitude. The coefficient of determination (R^2) evaluates the model's goodness-of-fit, denoting the proportion of variance in the dependent variable that can be systematically explained by the independent input features. Furthermore, the Average Absolute Relative Error ($AARE\%$) was incorporated alongside MSE and R^2 to measure the mean relative deviation between the forecasted and empirical yields. $AARE\%$ supplies a critical, scale-independent perspective on predictive accuracy. This attribute is exceptionally pertinent in biochar research, where actual yield values fluctuate significantly depending on the diverse biomass feedstocks and pyrolytic conditions utilized. Consequently, relative error serves as a highly interpretable indicator of a model's practical reliability. The integration of these three criteria ensures a comprehensive appraisal of model robustness, capturing both the absolute magnitude of predictive errors and the overarching explanatory power.

To mitigate the adverse effects of disparate scaling and inherent data variability, all input features and target variables were subjected to a rigorous normalization process prior to model

training. This transformation was executed according to the following standard scaling equation [46]:

$$n^{norm} = \frac{n - n^{\min}}{n^{\max} - n^{\min}} \quad (1)$$

where n_{norm} designates the resulting normalized value, n represents the original empirical data point, and $n_{min}$and n_{max}correspond to the minimum and maximum recorded values within the specific parameter subset, respectively.

Notice that we implemented a 90/10 % for train and test stages, respectively, with 5 folds of k-fold cross validation algorithm. In addition, we used fixed seed number (n=42) and no stopping criteria was applied. Also, we used Python V3 with the required libraries (NumPy V 2.4, Scikit-learn V1.7, etc.). The optimization process was done using grid search, using the most significant hyperparameter for each machine learning algorithm.

4. Results and discussion

4.1. Linear sensitivity examination

The linear interdependencies among the dataset variables are quantitatively evaluated via a Pearson correlation matrix, as depicted in Figure 4. The Pearson correlation coefficient, constrained within the mathematical interval of $[-1, +1]$, delineates both the magnitude and direction of the linear associations. Specifically, coefficients of $+1$and -1 denote perfect positive and negative correlations, respectively, whereas a value of 0 indicates the absolute absence of a linear relationship.

This analytical matrix yields critical preliminary insights into the dynamic relationships between the independent input parameters and the targeted dependent variable, biochar yield. As illustrated, specific parameters, most notably feedstock ash content, pyrolytic residence time, and catalyst acidic site concentration, exhibit robust positive correlations with the final solid yield. Conversely, pronounced negative correlations are evident for both the heating rate and the peak pyrolysis temperature. The remaining input features demonstrate a comparatively marginal linear influence on the target variable.

Furthermore, the derived Pearson correlation matrix (Figure 4) reflects phenomenological realities highly consistent with established thermochemical principles. The strong inverse relationship between pyrolysis temperature and biochar yield corroborates the well-documented trend wherein elevated thermal conditions enhance the volatilization of organic matter, thereby reducing the solid residue. Similarly, the moderate positive correlation associated with residence time underscores its fundamental role in facilitating extensive carbonization and secondary repolymerization reactions. Regarding initial biomass composition, intrinsic ash content demonstrates a distinct positive correlation with the final yield, operating in direct contrast to the negative influence exerted by volatile matter fractions. Collectively, these bivariate statistical trends establish an initial empirical baseline identifying

the parameters most likely to dictate model performance, a foundational understanding that is subsequently corroborated and mechanistically expanded through SHAP analysis.

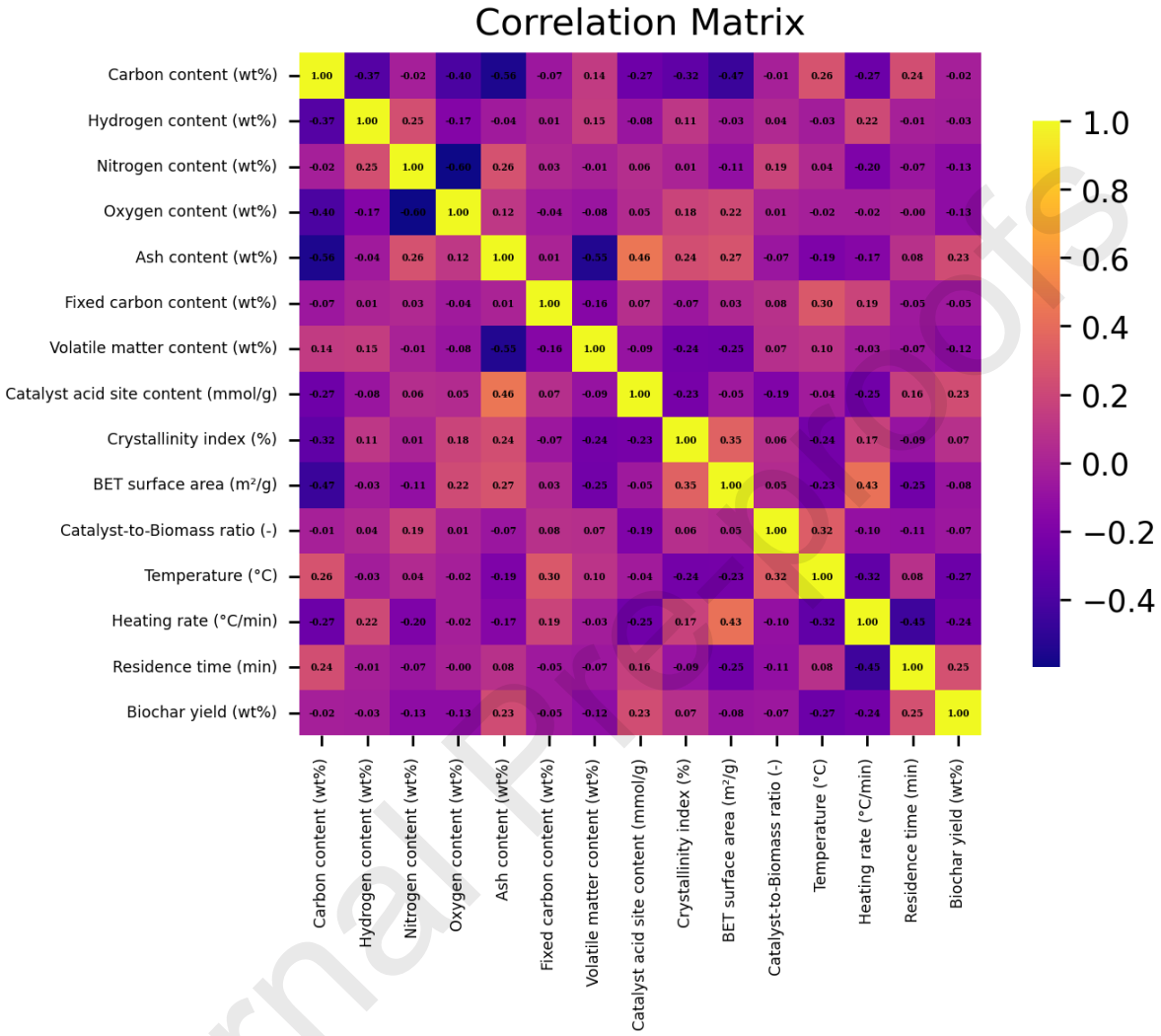


Figure 4. Pearson correlation matrix illustrating the pairwise linear associations among all analyzed variables.

4.2. Data quality justification

Leverage analysis serves as a robust statistical diagnostic procedure for identifying anomalous or disproportionately influential data points within regression frameworks. This methodological approach is fundamentally predicated on the Hat matrix, a mathematical construct that quantifies the degree to which each individual observation dictates the fitted model. The Hat matrix, denoted as H, is formally defined as:

$$H = X(X^T X)^{-1} X^T \quad (2)$$

where X represents the design matrix of the input features and X^T designates its transpose [47]. The principal diagonal elements of H yield the leverage values for the respective data points; abnormally elevated values indicate observations that may exert an undue, disproportionate influence on the model calibration process.

To systematically isolate these high-leverage instances, a critical threshold limit is conventionally established according to the following criterion:

$$H^* = 3(n + 1)/m \quad (3)$$

where n represents the total number of input variables and m denotes the aggregate sample size [48]. Observations demonstrating leverage values that surpass this defined threshold ($H > H^*$) are flagged for rigorous secondary evaluation, as they represent potential outliers capable of unduly skewing the regression trajectory. As Figure 5 shows, application of the leverage criterion identified three observations with $H > H^*$. These points were removed prior to model training to prevent disproportionate influence on the regression structure.

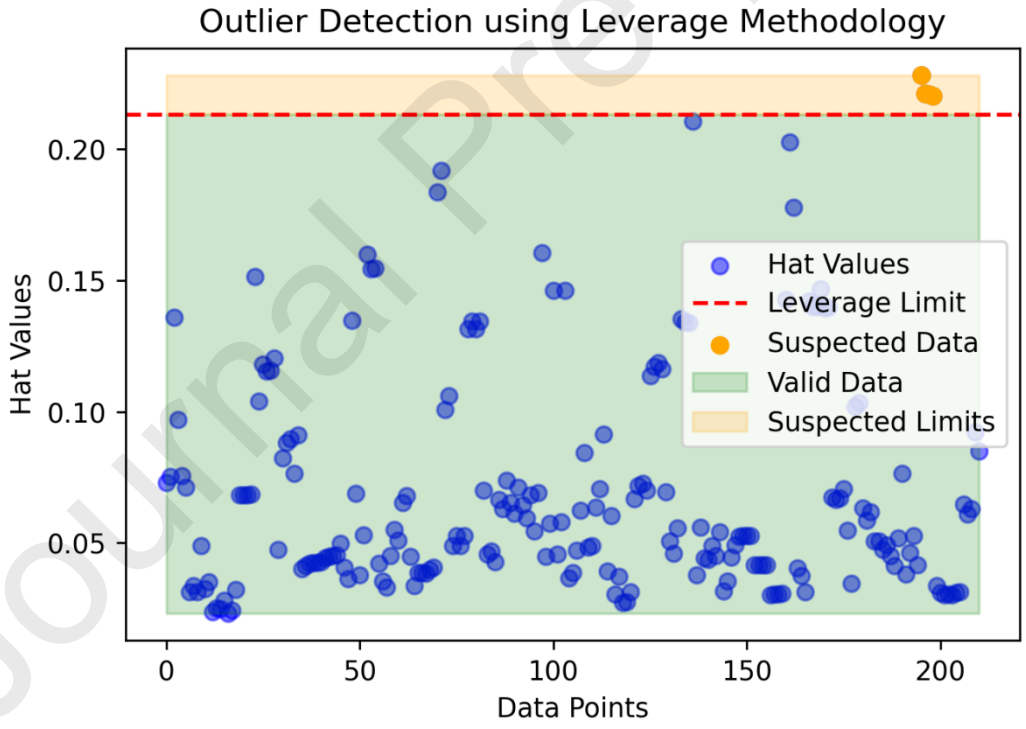


Figure 5. Leverage-based outlier detection using the Hat matrix. The plot displays the hat values for all data points, with the red dashed line indicating the critical leverage threshold.

Outlier detection was performed using a dual-criterion approach combining leverage analysis, standardized residuals, and Monte Carlo resampling (Figure 6). Monte Carlo simulations of the target variable distribution were conducted, with anomalous points defined as those falling outside ±2σ bounds in more than 95% of random draws. Three observations simultaneously exceeded the leverage threshold ($H > H^*$) and were repeatedly flagged by Monte Carlo analysis.

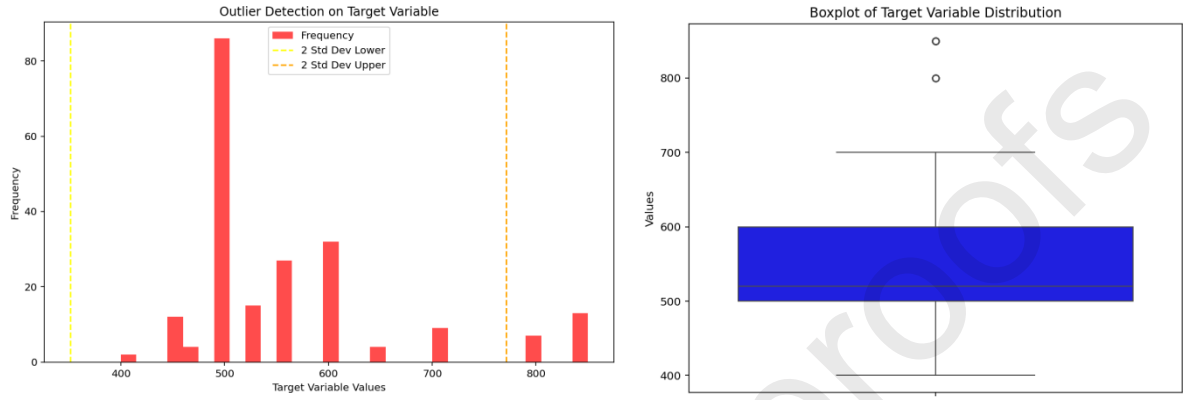


Figure 6. Monte Carlo–based outlier detection of the target variable distribution

4.3. Models' optimizing tuning and assessment

To ensure a fair and comprehensive comparison among the metaheuristic algorithms, a consistent hyperparameter search space was defined for all optimization procedures. The selected ranges reflect commonly adopted bounds in gradient boosting applications and were chosen to balance model flexibility with computational efficiency. For optimization algorithms operating in continuous search spaces (CSA, PSO), the raw optimal values for integer-constrained hyperparameters (e.g., n_estimators, max_depth) were continuous. Prior to model training, these values were rounded to the nearest integer to comply with the GBDT implementation. Table 1 reports the rounded values actually used in the final models.

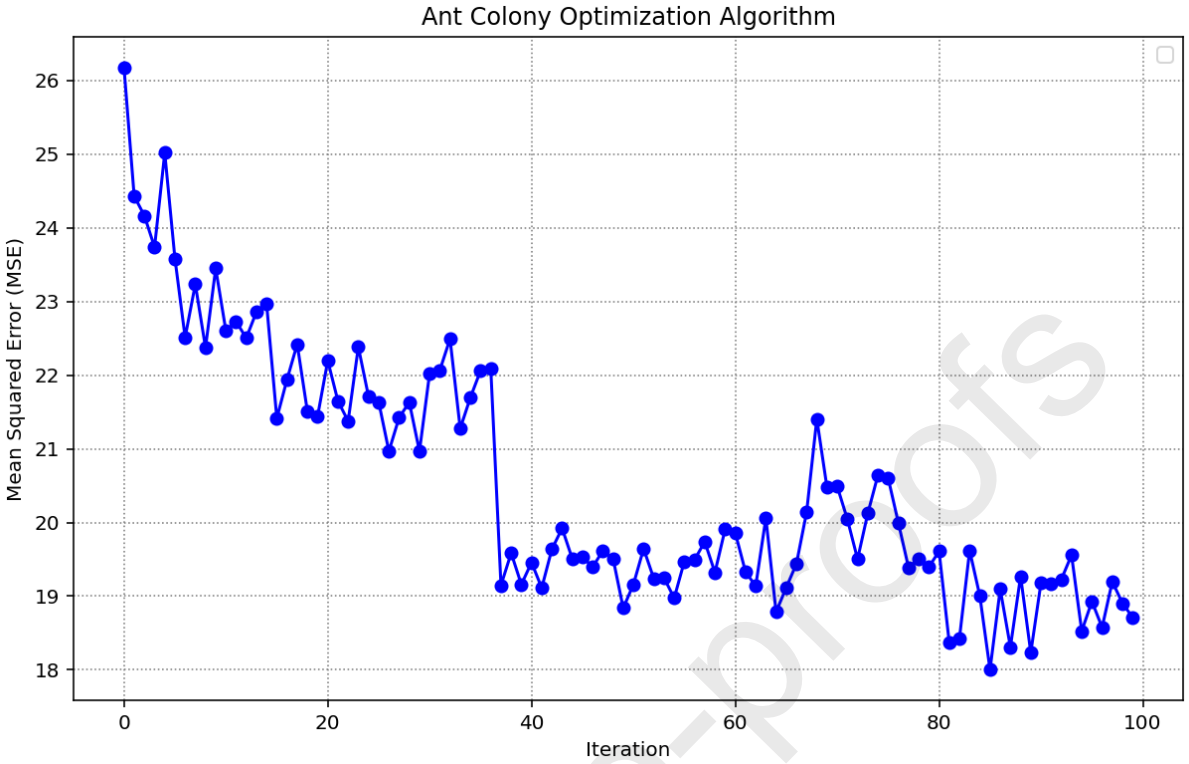
Table 1. Summary of the hyperparameter search space and the best-performing configurations

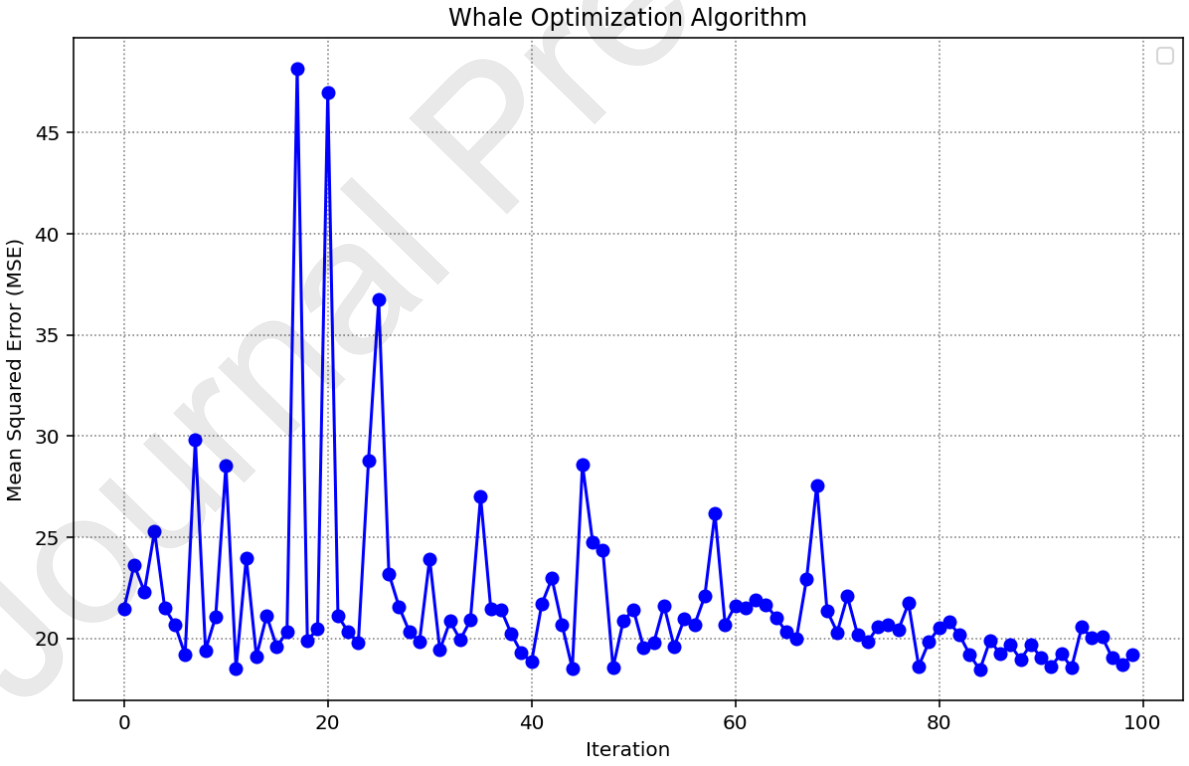
Hyperparameter	ACO	WOA	CSA	PSO
n_estimators	237	162	147	69
max_depth	5	15	14	17
max_features	0.3195	0.3946	0.4916	0.5415
min_samples_split	0.4346	0.0346	0.2938	0.2137
learning_rate	0.2333	0.0955	0.1964	0.2693

subsample	1.0	0.6222	0.5038	0.9572
min_samples_leaf	0.0188	1	0.01377	0.1455

Figure 7 illustrates the comparative optimization trajectories of the algorithms by plotting MSE over 100 iterations, enabling evaluation of their relative efficiency in converging toward optimal hyperparameters. The convergence curves reveal distinct behavioral patterns among the algorithms: ACO exhibits a steady and monotonic reduction in MSE, indicating stable exploration and consistent improvement throughout the search process. WOA shows a more irregular trajectory with pronounced oscillations in the early and mid-stage iterations, reflecting its characteristic balance between exploration and exploitation before gradually settling toward lower error values. PSO demonstrates a rapid decline in MSE during the initial iterations, followed by a smooth and stable convergence phase, suggesting fast adaptation of particle positions toward promising regions of the search space. In contrast, CSA achieves a sharp drop in MSE within the first 20 iterations and then maintains a nearly constant error level, highlighting its ability to quickly locate a high-quality solution and preserve it through controlled perturbations. Together, these convergence profiles provide a clear visual comparison of the dynamic search behaviors and convergence efficiencies of the four metaheuristic optimizers.

The computational runtime analysis in Figure 8 highlights substantial differences in the efficiency of the four metaheuristic algorithms. As shown in Figure 8, WOA exhibits the highest computational cost, requiring nearly 900 s to complete the optimization process, followed by ACO with a runtime of approximately 670 s. In contrast, CSA and PSO demonstrate markedly superior efficiency, each completing the optimization in roughly 50 s. This sharp disparity indicates that while ACO and WOA may offer competitive search capabilities, they do so at a significantly higher computational expense. Conversely, CSA and PSO achieve rapid convergence with minimal runtime overhead, making them more suitable for applications where computational efficiency is a critical constraint.





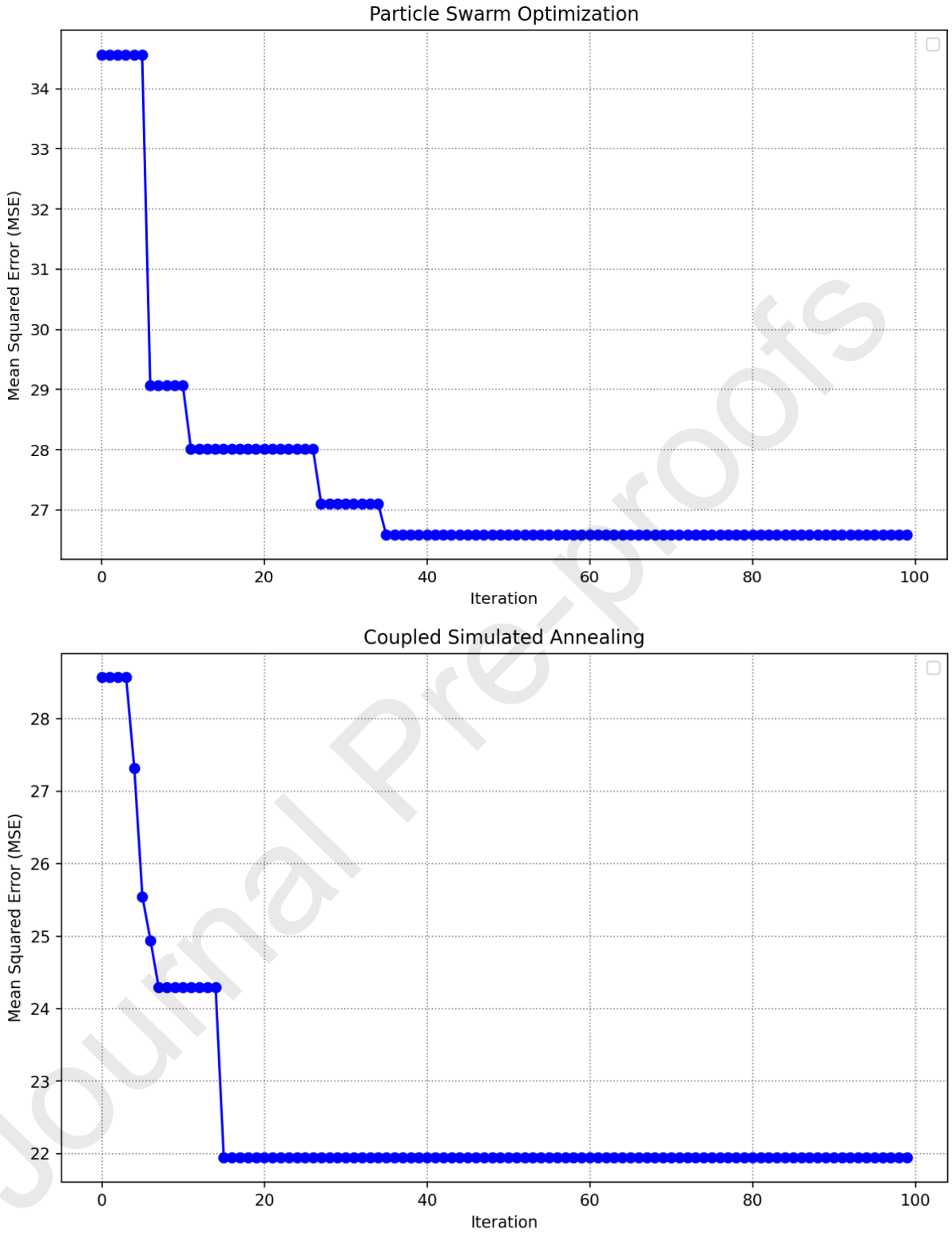


Figure 7. Iterative optimization paths showing the reduction of MSE over 100 iterations

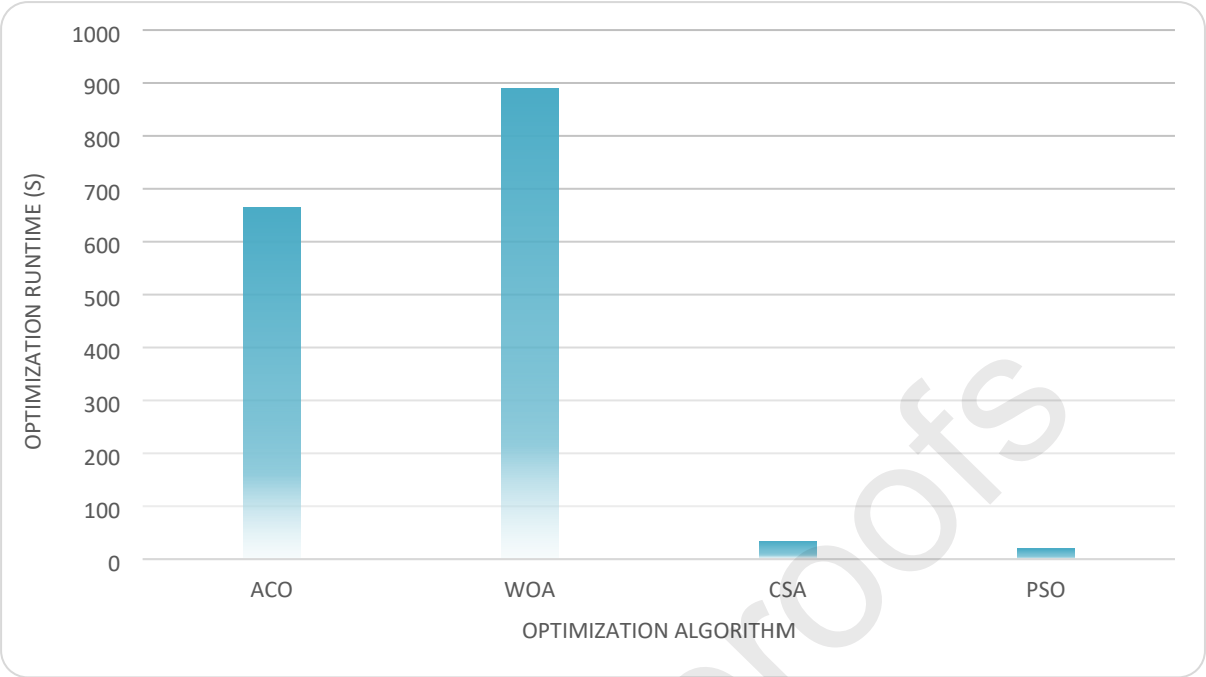


Figure 8. Runtime comparison illustrating the computational cost associated with each optimization method.

The predictive performance of the four optimized GBDT models in Table 2 shows clear differences in accuracy across the training, test, and total datasets. A comparison of the four optimized GBDT models based solely on test-set performance reveals clear differences in their generalization capabilities. Among all configurations, GBDT-ACO demonstrates the strongest overall test-stage behavior, achieving the highest test-set R^2 (0.709) and the lowest test-set MSE (16.284) among the four models. Its test-set AARE% (9.635%) is also competitive, indicating that ACO provides a balanced trade-off between predictive accuracy and stability when exposed to unseen data. These results suggest that the ACO optimizer is particularly effective at identifying hyperparameters that generalize well beyond the training subset.

In contrast, GBDT-WOA exhibits the weakest test-set R^2 (0.449) and the highest test-set MSE (27.588), despite achieving the best training performance (R^2 = 0.998 and MSE = 0.094). This large discrepancy between training and test metrics indicates substantial overfitting, where the model captures training-set patterns too precisely and fails to generalize effectively. Although GBDT-WOA achieves the lowest total AARE% (1.734%), this metric is dominated by the training subset and does not reflect true generalization. Therefore, WOA's optimization trajectory appears to favor aggressive fitting rather than robust predictive behavior.

The GBDT-CSA model shows moderate generalization, with a test-set R^2 of 0.547 and a test-set MSE of 22.708. Its test-set AARE% (12.329%) is higher than those of ACO and WOA, indicating weaker relative-error performance. While CSA avoids the extreme overfitting observed in WOA, its predictive accuracy remains limited compared to ACO. These results suggest that CSA converges to acceptable but suboptimal hyperparameter regions, offering neither the strongest accuracy nor the fastest runtime.

Finally, GBDT-PSO performs similarly to CSA, with a test-set R^2 of 0.468 and a test-set MSE of 26.662. Its test-set AARE% (11.358%) is slightly better than CSA but still inferior to ACO. Like WOA, PSO shows a noticeable gap between training and test performance, though less severe. This indicates that PSO also struggles to maintain generalization, likely due to premature convergence or insufficient exploration of the hyperparameter space.

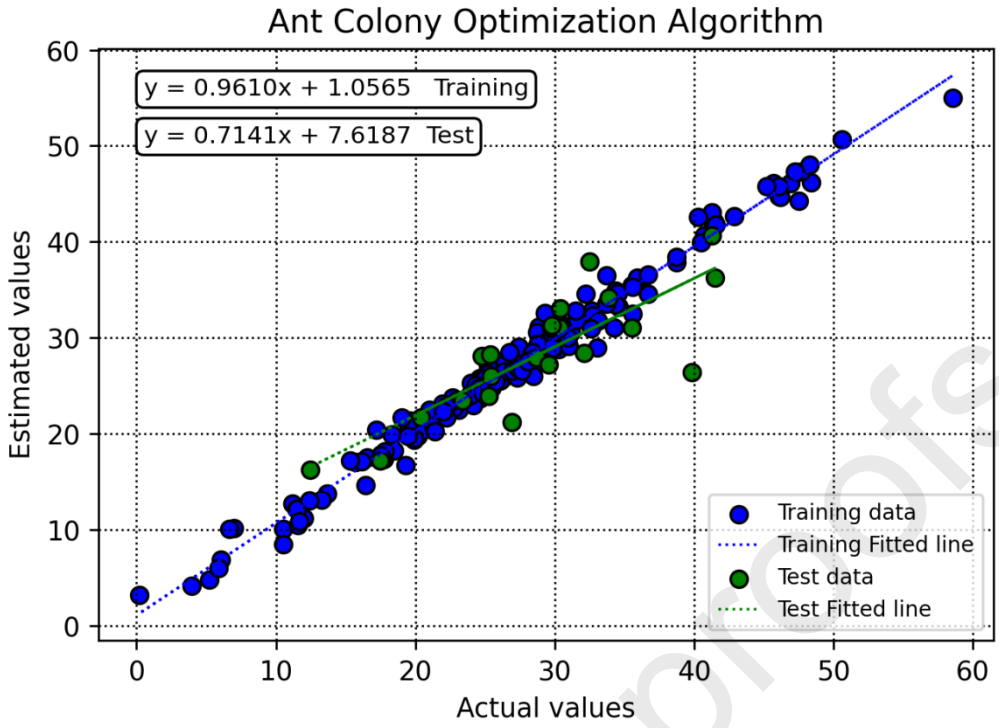
Table 2. Comparative assessment of model accuracy following algorithmic optimization

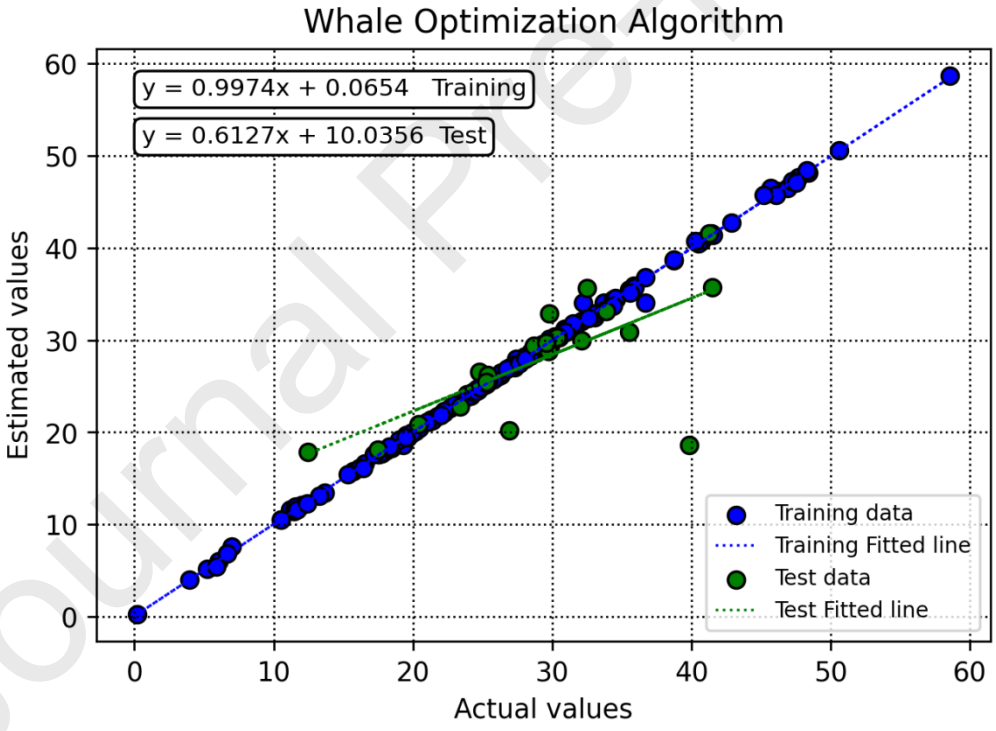
Model	R2			MSE			AARE%		
	Training	Test	Total	Training	Test	Total	Training	Test	Total
GBDT-ACO	0.984	0.709	0.966	1.463	16.284	3.008	11.145	9.635	10.987
GBDT-WOA	0.998	0.449	0.967	0.094	27.588	2.961	0.850	9.328	1.734
GBDT-CSA	0.931	0.547	0.909	6.503	22.708	8.193	18.627	12.329	17.970
GBDT-PSO	0.863	0.468	0.841	12.826	26.662	14.268	20.372	11.358	19.432

Figure 9 illustrates the cross-plots comparing observed and predicted outputs for all models, offering a clear visual benchmark of predictive accuracy and model agreement. The proximity of the data points to the 1:1 reference line indicates the degree of fidelity in each model's predictions. Among the evaluated approaches, the ACO-optimized model demonstrates the strongest alignment, with points tightly concentrated around both the regression fit and the ideal reference line, indicating minimal dispersion and high predictive reliability.

Figure 10 presents the RE% associated with each individual data point across the different modeling strategies, providing insight into the distribution of prediction errors and the stability of each method. The zero-error baseline serves as a reference for evaluating consistency. Once again, the ACO-based model shows superior behavior, with RE% values clustered closely around zero, highlighting its robustness and enhanced generalization capability compared with the other optimization techniques.

.





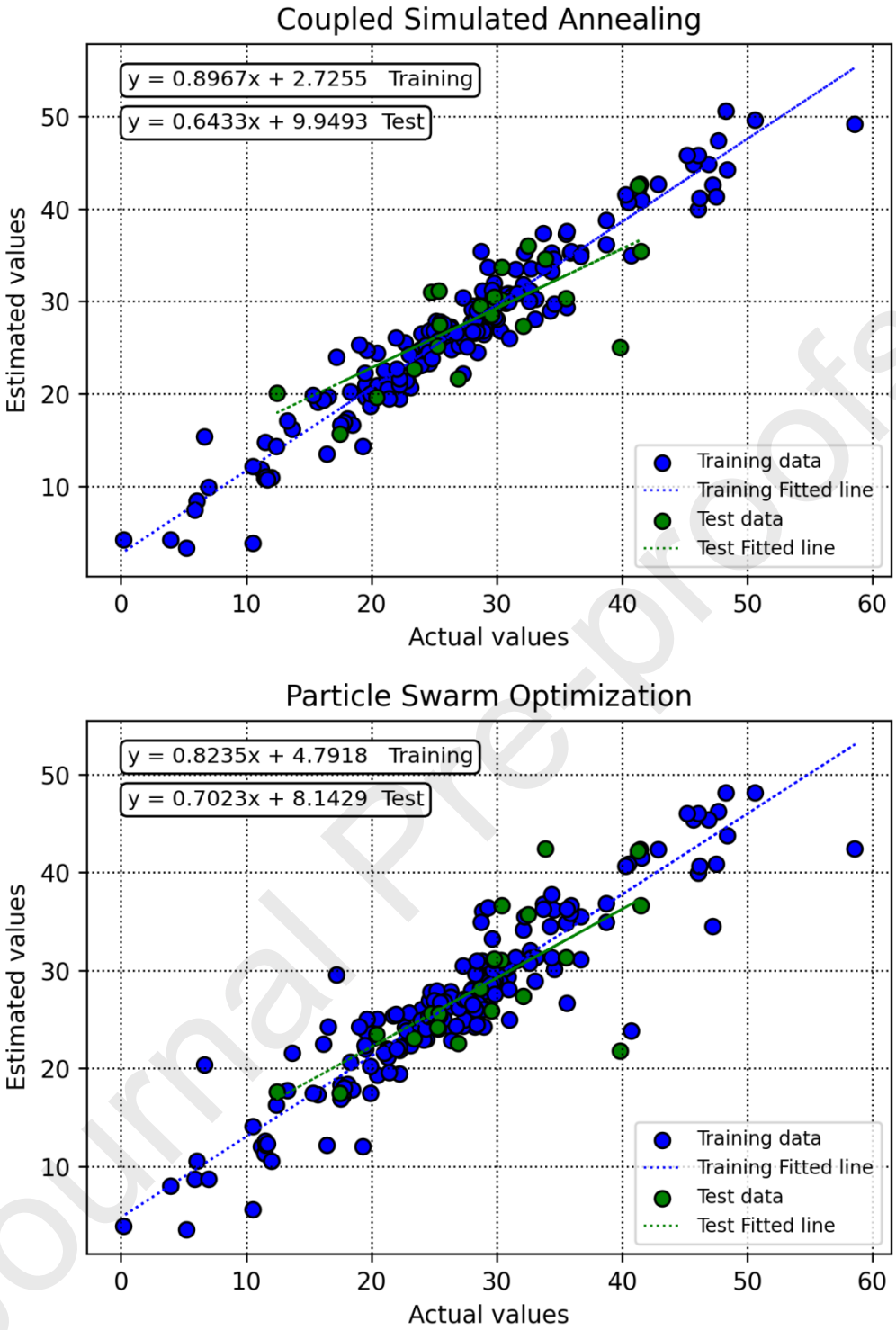
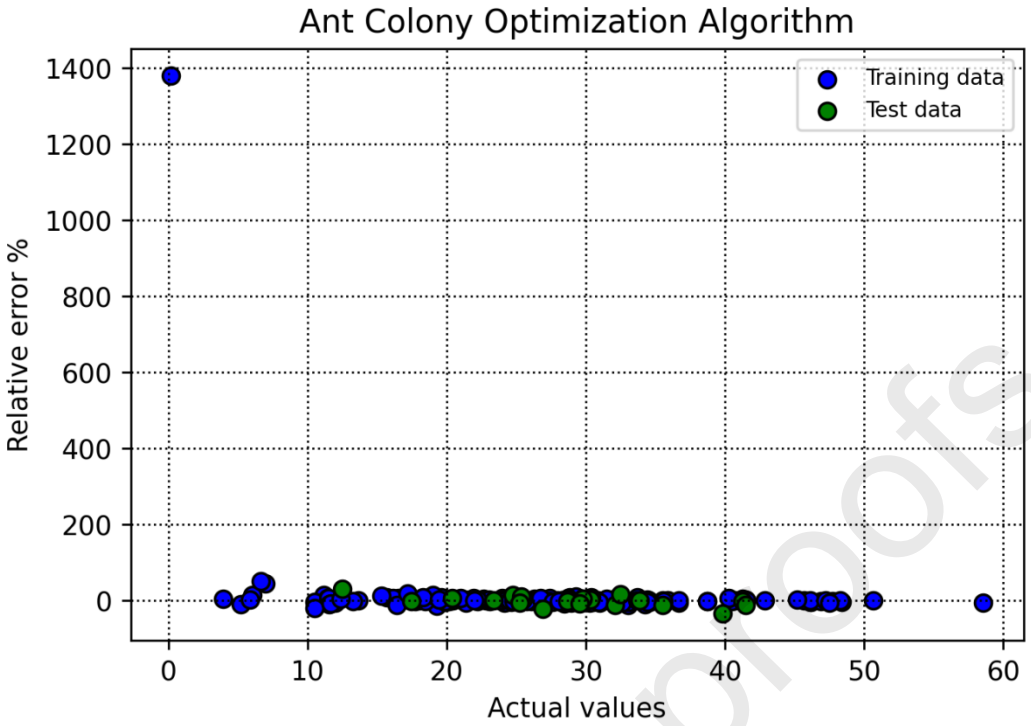
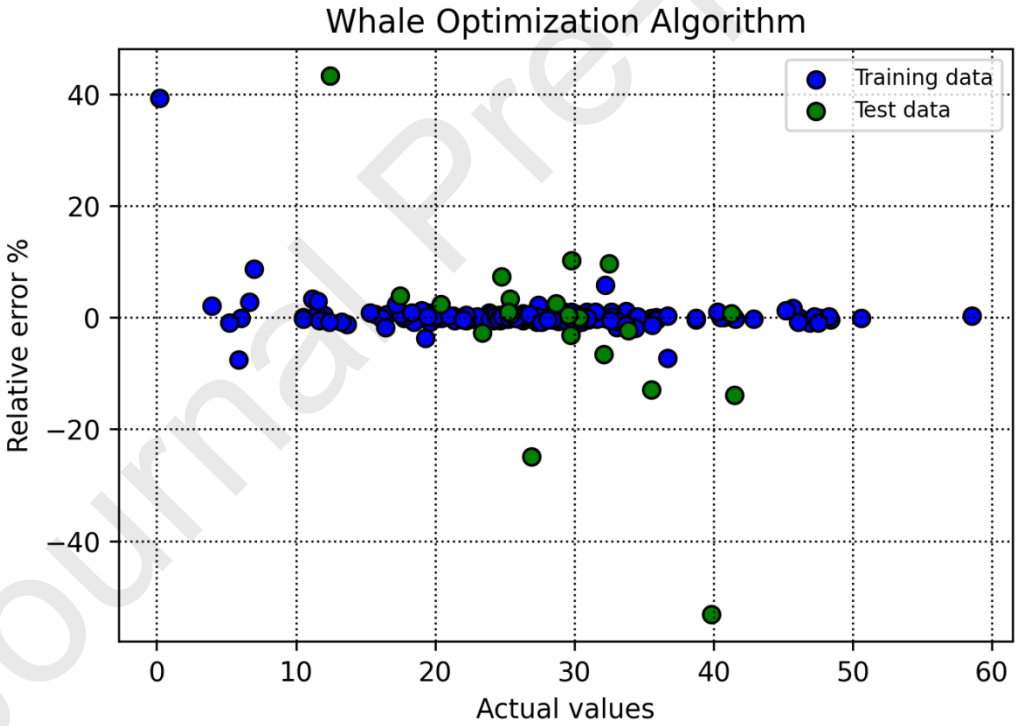


Figure 9. Cross-plots illustrating the relationship between observed and predicted values.





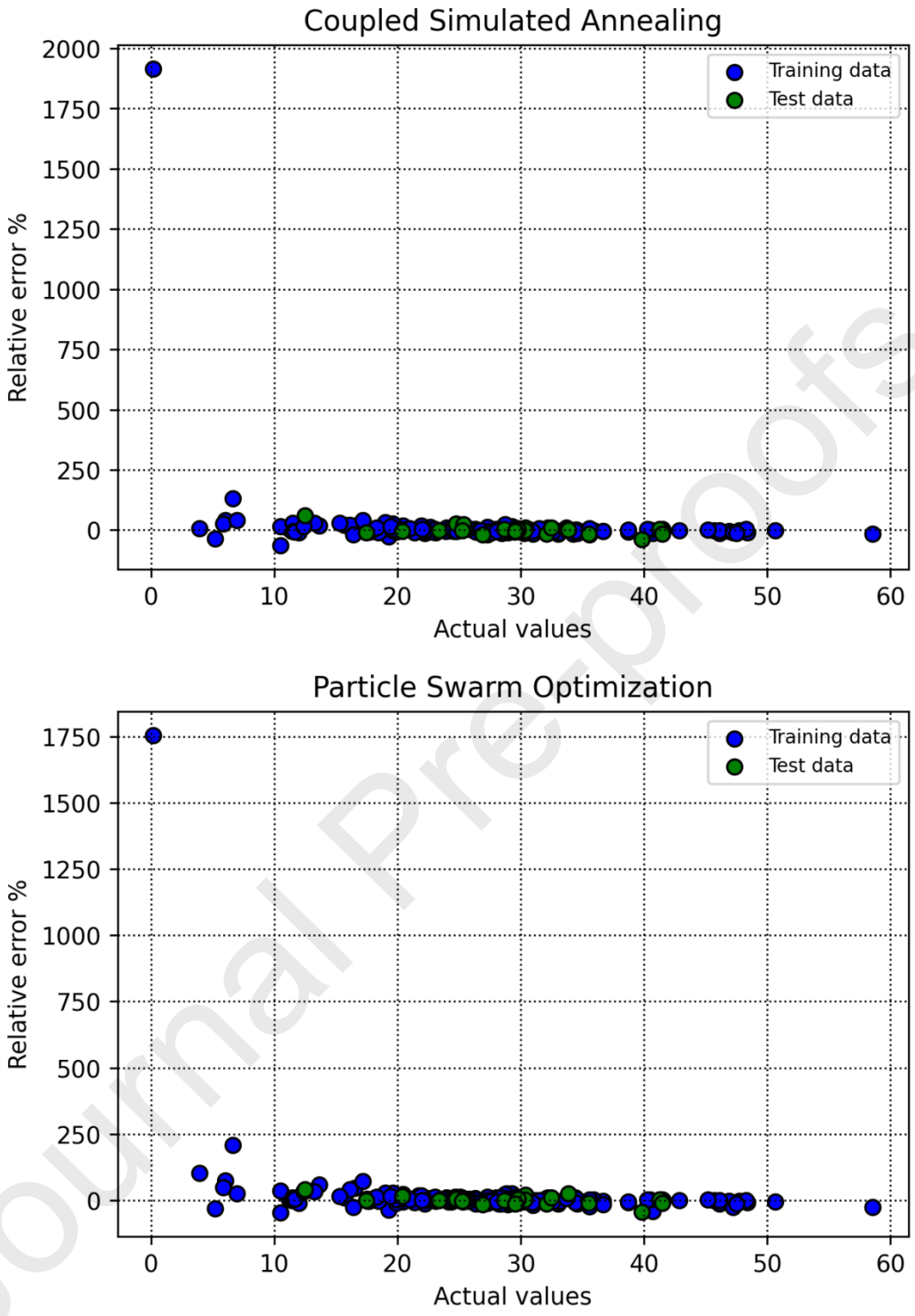


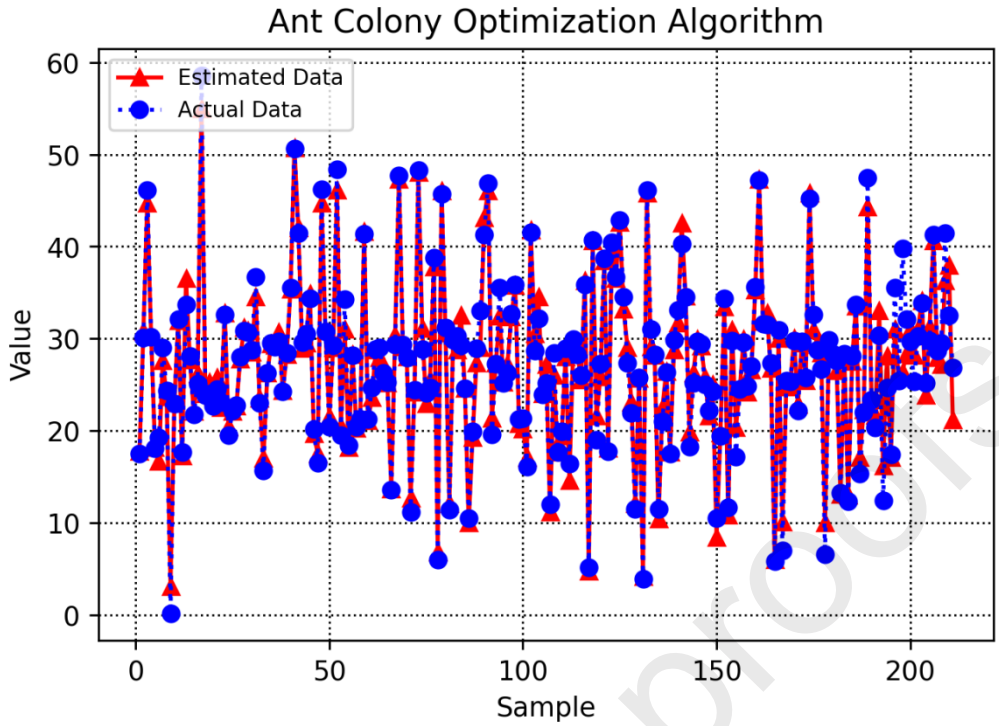
Figure 10. Relative error values for each data point as predicted by all evaluated models.

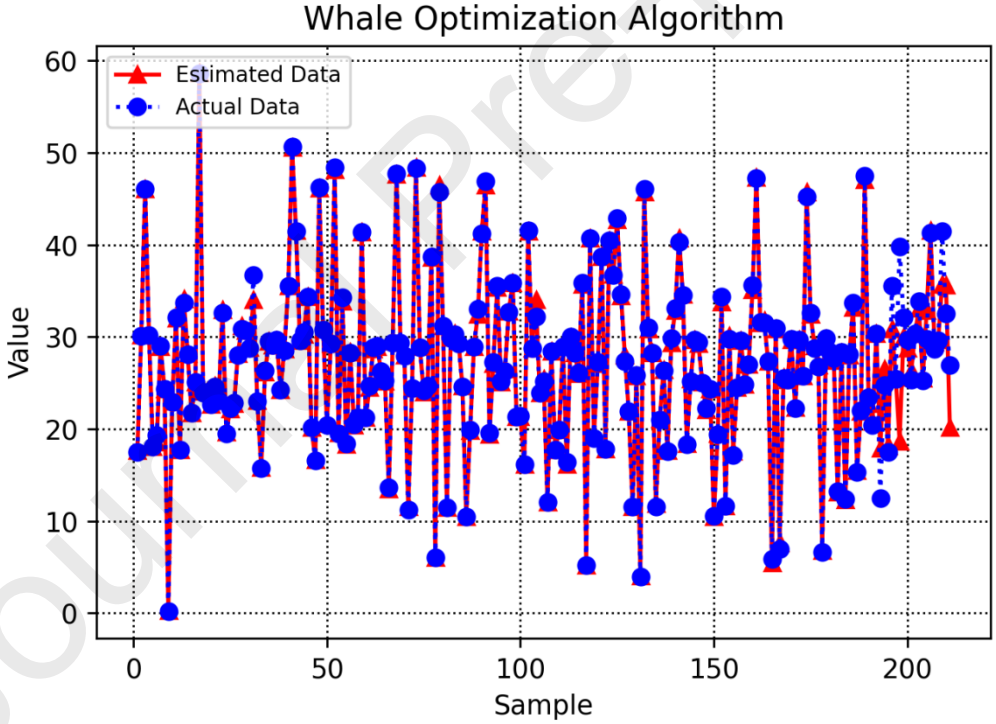
Figure 11 presents a comprehensive visualization that juxtaposes predicted target values against their experimentally observed counterparts across the entire dataset. The plot integrates the outputs of all models, each refined through distinct optimization strategies, into a single unified diagram. This consolidation facilitates a direct and systematic evaluation of predictive fidelity, allowing readers to discern how closely each modeling approach reproduces the true measurements. By displaying the collective results in one frame, the figure highlights differences in accuracy, scatter, and alignment with the identity line, thereby offering an

intuitive basis for comparing the relative effectiveness of the optimization methods in enhancing model generalization and reliability. Among the tested approaches, the ACO-optimized model shows the closest clustering along the reference line, reflecting minimal scatter and superior alignment with experimental data.

This integrated visualization provides a clear perspective on the relative effectiveness of each optimization method in improving model generalization and predictive fidelity. The contrast in dispersion patterns across the models highlights the extent to which different tuning strategies influence accuracy, offering an intuitive and objective basis for comparing their overall performance.

.





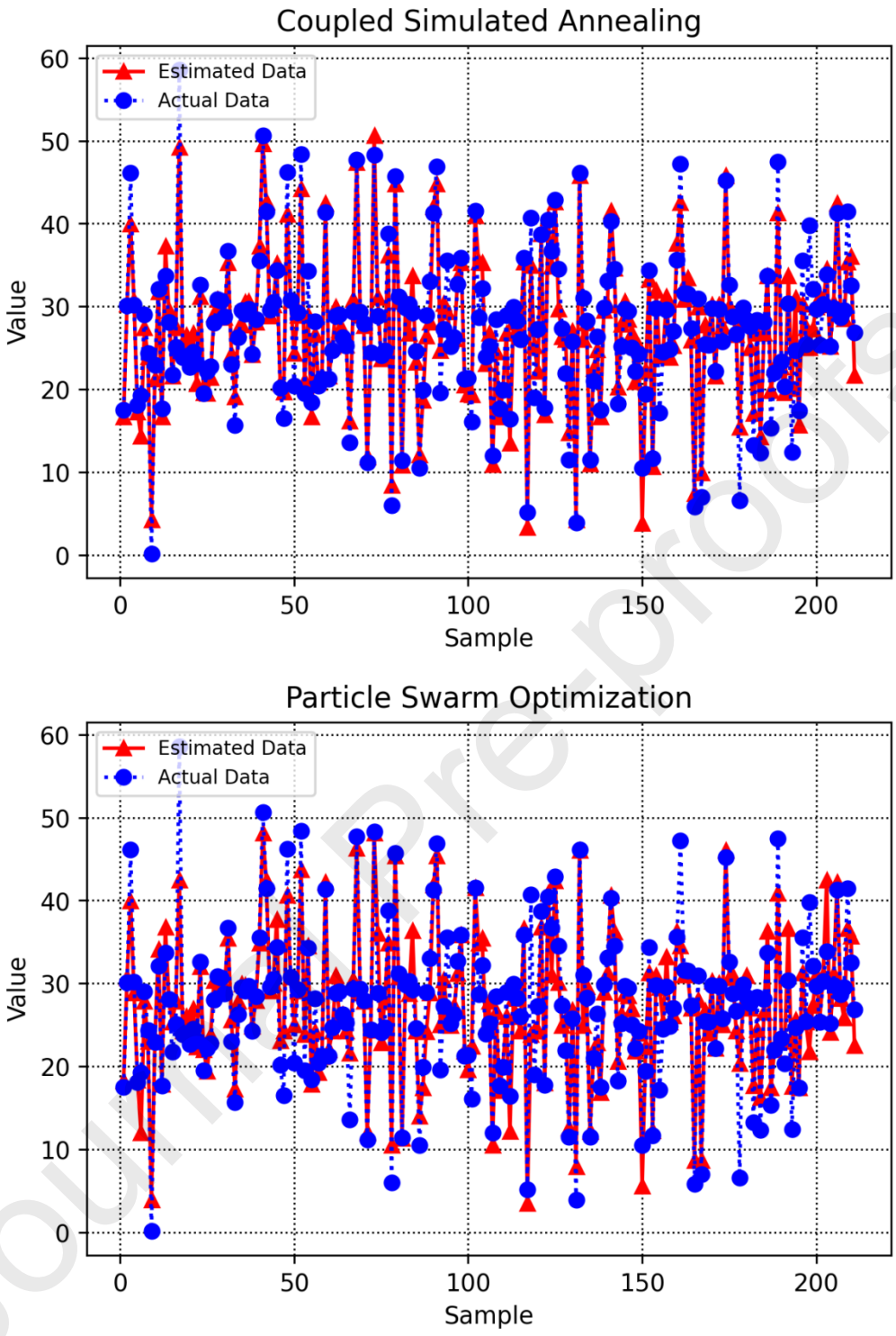


Figure 11. Combined plot illustrating the correspondence between observed and predicted target values

Figure 12 presents the global SHAP summary plot, which ranks all input variables according to their mean absolute SHAP values. This analysis shows that fixed carbon content, residence time, and temperature are the most influential predictors of biochar yield, followed by ash content and oxygen content. These variables consistently exhibit the largest SHAP magnitudes, indicating that they contribute the most to the model's predictive structure across all GBDT configurations. Features such as volatile matter, catalyst-to-biomass ratio, BET surface area, and elemental composition (C, H, N) show progressively smaller SHAP values, suggesting a

weaker but still measurable influence on the model output. The global ranking in Figure 12 therefore provides a clear hierarchy of feature importance, highlighting which parameters the model relies on most heavily when forming yield predictions.

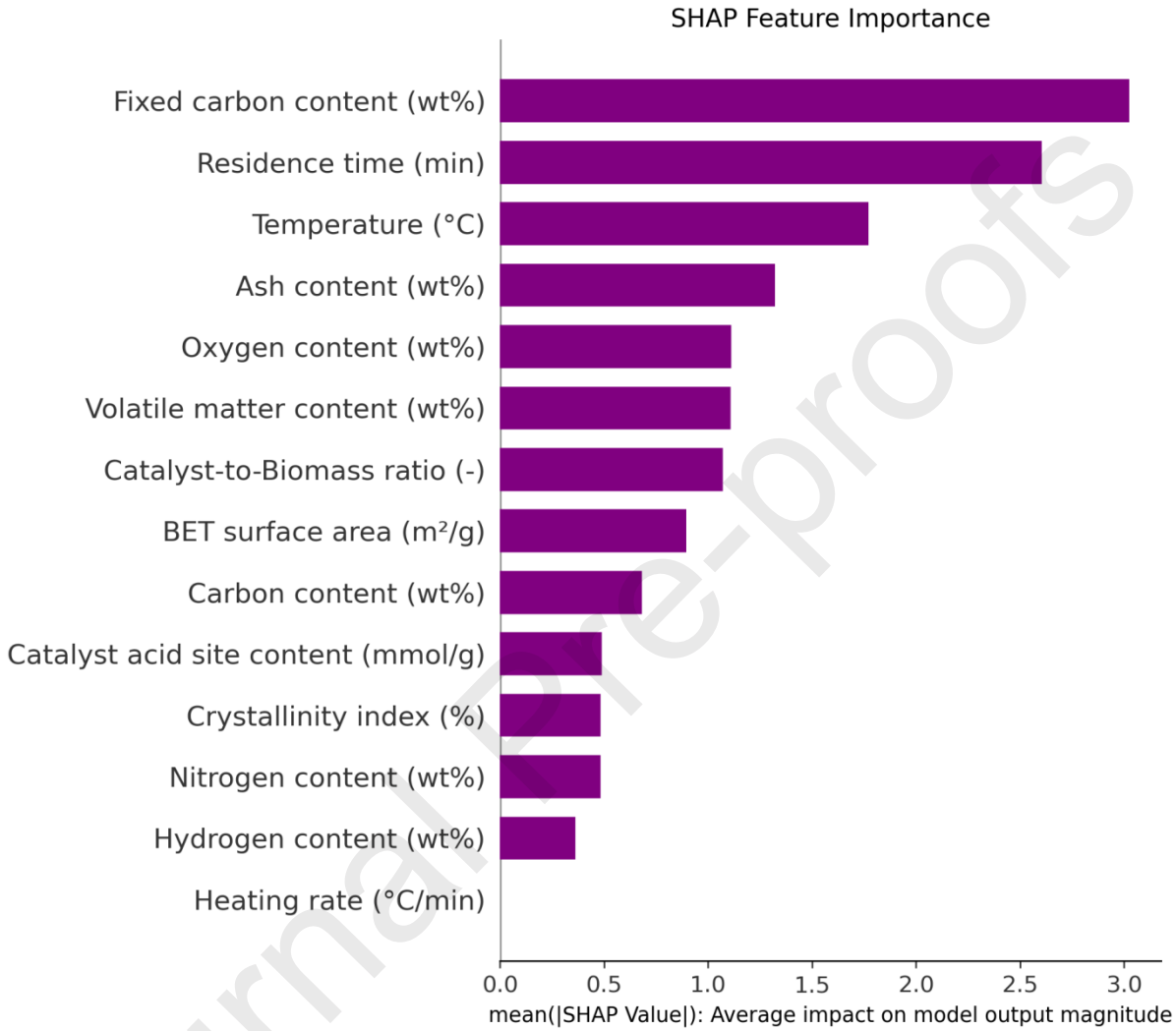


Figure 12. Feature importance profile based on SHAP analysis, illustrating the influence of each predictor on model output.

Figure 13 displays the SHAP dependence plots, which illustrate how individual features affect the predicted yield across their full value ranges. These plots reveal both the directionality and the nonlinearity of each feature's contribution. For example, increasing temperature produces predominantly negative SHAP values, confirming its inverse relationship with yield at higher pyrolysis intensities. Conversely, higher fixed carbon content and longer residence times generate positive SHAP contributions, consistent with their experimentally observed roles in enhancing carbon retention. The color gradient in Figure 13 further clarifies how low versus high feature values shift the model output: high fixed-carbon values (pink points) cluster on the positive SHAP side, while high temperature values cluster on the negative side. Intermediate-importance features such as ash content, oxygen content, and volatile matter exhibit mixed SHAP patterns, indicating nonlinear or interaction-driven effects. By combining global importance (Figure 12) with feature-level contribution behavior (Figure 13), the SHAP

analysis provides a transparent and mechanistic interpretation of how the GBDT models internalize the underlying pyrolysis relationships.

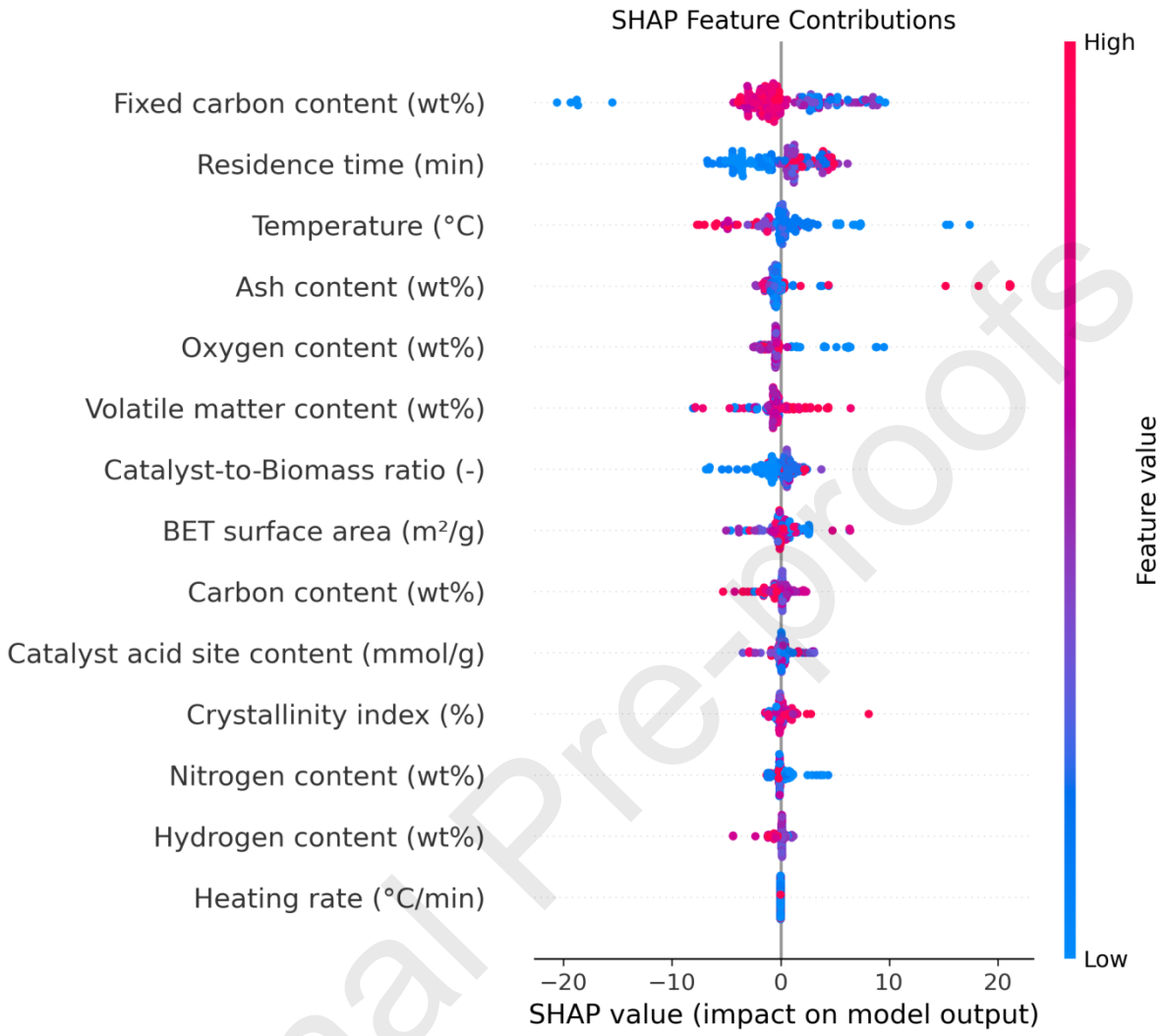


Figure 13. summary of SHAP-based feature contributions, highlighting the directional impact of each predictor on model estimates.

5. Conclusions

This study develops an interpretable machine learning framework for predicting biochar yield using Gradient Boosting Decision Trees (GBDT) optimized through four metaheuristic algorithms. By systematically comparing ACO, WOA, CSA, and PSO, the work provides a detailed assessment of how optimization strategy influences predictive accuracy, convergence behavior, and computational efficiency. The curated dataset, derived exclusively from peer-reviewed experimental studies, ensures strong empirical grounding and broad representation of biomass feedstocks and pyrolysis conditions.

The comparative analysis demonstrates that only the GBDT-ACO model achieves acceptable generalization, with a test-set R^2 of 0.709 and the lowest corresponding test-set MSE and AARE%. In contrast, WOA, PSO, and CSA exhibit limited predictive reliability on unseen

data, despite in some cases achieving high training or total R^2 values. These discrepancies highlight the importance of evaluating model performance based on independent test-set metrics rather than aggregated or training-dominated statistics. The runtime analysis further shows that CSA and PSO offer substantial computational efficiency, though at the cost of reduced predictive fidelity.

SHAP interpretability analysis consistently identifies ash content, residence time, and peak temperature as the dominant predictors of biochar yield, in agreement with established thermochemical principles. These insights reinforce the value of integrating explainable AI tools to elucidate mechanistic relationships within pyrolysis systems.

Overall, the findings indicate that GBDT-ACO represents the most balanced configuration, offering the best trade-off between accuracy, stability, and generalization among the tested models. Rather than claiming uniformly robust predictive performance, this study contributes a transparent comparative framework that clarifies the strengths and limitations of different optimization strategies. The workflow combining rigorous data curation, multi-algorithmic optimization, and SHAP-based interpretation provides a reproducible foundation for future research aimed at improving predictive modeling and mechanistic understanding of biochar production processes.

Data availability statement

Data is available in the supplementary material Excel file.

Funding

None

Conflicts of interests

None

Ethics statement for the use of human and animal subjects

Not applicable

Consent for publication

Not applicable

Author's Contribution

Formal study: Kusum Yadav

Software: Kusum Yadav, Shahad Almansour

Data curation: Lulwah M. Alkwai

Visualization: Shahad Almansour

Writing: Lulwah M. Alkwai, Mehrdad Mottaghi

Supervision: Mehrdad Mottaghi

References

1. Chun, Y., et al., *Recent advancements in biochar production according to feedstock classification, pyrolysis conditions, and applications: A review.* BioResources, 2021. **16**(3): p. 6512.

2. Zhu, J., X. Fei, and K. Yin, *Assessment of waste-to-energy conversion technologies for biomass waste under different shared socioeconomic pathways.* Energy & Environmental Sustainability, 2025. **1**(2): p. 100021.

3. da Silva Medeiros, D.C.C., et al., *Biochar-enhanced removal of naphthenic acids from oil sands process water: Influence of feedstock and chemical activation.* Energy & Environmental Sustainability, 2025. **1**(2): p. 100028.

4. Hu, K., et al., *A modified SLD-ESD model for characterizing and predicting supercritical CH4 adsorption on shale at wide temperatures and pressures.* Fuel, 2026. **416**: p. 138561.

5. Ibitoye, S.E., et al., *An overview of biochar production techniques and application in iron and steel industries.* Bioresources and bioprocessing, 2024. **11**(1): p. 65.

6. Puri, L., Y. Hu, and G. Naterer, *Critical review of the role of ash content and composition in biomass pyrolysis.* Frontiers in Fuels, 2024. **2**: p. 1378361.

7. Ge, Q., et al., *Removal of methylene blue by porous biochar obtained by KOH activation from bamboo biochar.* Bioresources and Bioprocessing, 2023. **10**(1): p. 51.

8. Guo, P., et al., *Enhanced Ultramicropore of Biomass-Derived Porous Carbon for Efficient and Low-Energy CO 2 Capture: Integration of Adsorption and Solar Desorption.* Energy & Environmental Materials, 2025: p. e70140.

9. Rahic, E., et al., *Biomass pyrolysis-derived aqueous phase as a laccase inducer in Pleurotus ostreatus: laccase production, properties, and applications.* Bioresources and Bioprocessing, 2026. **13**(1): p. 41.

10. Manyà, J.J., M. Azuara, and J.A. Manso, *Biochar production through slow pyrolysis of different biomass materials: Seeking the best operating conditions.* Biomass and Bioenergy, 2018. **117**: p. 115–123.

11. Wallace, C.A., M.T. Afzal, and G.C. Saha, *Effect of feedstock and microwave pyrolysis temperature on physio-chemical and nano-scale mechanical properties of biochar.* Bioresources and Bioprocessing, 2019. **6**(1): p. 1–11.

12. Birhanu, A., et al., *Optimization of pyrolysis conditions for Catha edulis waste-based biochar production using response surface methodology.* Bioresources and Bioprocessing, 2025. **12**(1): p. 62.

13. Zhou, Q., et al., *Co3S4-pyrolysis lotus fiber flexible textile as a hybrid electrocatalyst for overall water splitting.* Journal of Energy Chemistry, 2024. **89**: p. 336–344.

14. Park, S.-W. and C.-H. Jang, *Effects of pyrolysis temperature on changes in fuel characteristics of biomass char.* Energy, 2012. **39**(1): p. 187–195.

15. Paudel, P.P., et al., *Comprehensive study on microwave pyrolysis process variables and operating modes for optimized biochar production.* Bioresource Technology, 2025: p. 133120.

16. long Li, W., Y. Guo, and B. ming Chen, *Research progress on pyrolysis gasification and resource utilization of tobacco waste: Component analysis and recovery strategies.* Journal of Analytical and Applied Pyrolysis, 2025: p. 107425.

17. Kang, Z., et al., *Hydrogen transfer and reaction mechanism during in-situ pyrolysis of Fushun oil shale with steam injection.* Fuel, 2025. **381**: p. 133583.

18. Herath Bandara, S.J., *Exploring the Potential of Biochar in Enhancing US Agriculture.* Regional Science and Environmental Economics, 2025. **2**(3): p. 23.

19. Xue, Y., et al., *Bio-based benzoxazine containing phthalonitrile: nonsolvent synthesis, curing behavior and pyrolysis mechanism.* Polymer Degradation and Stability, 2025: p. 111542.

20. Xiao, C., et al., *Strong and tough multilayer heterogeneous pyrocarbon based composites.* Advanced Functional Materials, 2024. **34**(51): p. 2409881.

21. Franzon, G., *Exploring the biochar contribution in sustainable agriculture through techno-economic assessment and water impact simulations.* 2025.

22. He, Y., et al., *Stepwise Removal Mechanism of Impurity Using Multi-Flux During Blister Copper Pyrometallurgical Refining.* Metallurgical and Materials Transactions B, 2025: p. 1–15.

23. Wang, G., et al., *Data Analysis and Prediction Model for Copper Matte Smelting Process.* Metallurgical and Materials Transactions B, 2024. **55**(4): p. 2552–2567.

24. Li, X., et al., *Hydrogen-rich gas formation from catalytic pyrolysis of biomass tar by aluminum dross coupled HZSM-5 co-loaded Ni-Fe bimetallic catalysts: Influence of co-carrier characteristics.* Journal of Environmental Management, 2025. **389**: p. 126016.

25. Paudel, P.P., et al., *Artificial neural network modeling for prediction and optimization of biochar yield and properties.* Journal of Analytical and Applied Pyrolysis, 2025: p. 107421.

26. Deka, M.J., et al., *An approach towards building robust neural networks models using multilayer perceptron through experimentation on different photovoltaic thermal systems.* Energy Conversion and Management, 2023. **292**: p. 117395.

27. Nguyen, V.N., et al., *Potential of explainable artificial intelligence in advancing renewable energy: challenges and prospects.* Energy & Fuels, 2024. **38**(3): p. 1692–1712.

28. Le, A.T., et al., *Precise prediction of biochar yield and proximate analysis by modern machine learning and shapley additive explanations.* Energy & Fuels, 2023. **37**(22): p. 17310–17327.

29. Abdelfattah, W., et al., *Accurate modeling of biochar yield based on proximate analysis.* Energy Exploration & Exploitation, 2025. **43**(6): p. 2397–2423.

30. Abdelfattah, W., et al., *Predicting Biochar Yield from Biomass Pyrolysis: A Comprehensive Data-Driven Approach Using Machine Learning and SHAP Analysis.* Results in Engineering, 2025: p. 105389.

31. Hou, G., et al., *Development of robust machine learning models to estimate hydrochar higher heating value and yield based upon biomass proximate analysis.* Bioresources and Bioprocessing, 2025. **12**(1): p. 138.

32. Nguyen, V.G., et al., *Machine learning for the management of biochar yield and properties of biomass sources for sustainable energy.* Biofuels, Bioproducts and Biorefining, 2024. **18**(2): p. 567–593.

33. Nguyen, V.G., et al., *Improving the prediction of biochar production from various biomass sources through the implementation of eXplainable machine learning approaches.* International Journal of Green Energy, 2024. **21**(12): p. 2771–2798.

34. Lucchese, C., et al. *Selective gradient boosting for effective learning to rank*.

35. Spettel, P. and H.-G. Beyer, *Analysis of the (μ/μ I, λ)-CSA-ES with Repair by Projection Applied to a Conically Constrained Problem.* Evolutionary Computation, 2020. **28**(3): p. 463–488.

36. Du, Y., et al., *Whale optimization algorithm with applications to power allocation in interference networks.* Information Technology and Control, 2021. **50**(2): p. 390–405.

37. Wang, D., D. Tan, and L. Liu, *Particle swarm optimization algorithm: an overview.* Soft computing, 2018. **22**(2): p. 387–408.

38. Sagban, R., K.R. Ku-Mahamud, and M.S. Abu Bakar, *ACOustic: A Nature-Inspired Exploration Indicator for Ant Colony Optimization.* The Scientific World Journal, 2015. **2015**(1): p. 392345.

39. Shen, Y. and L. Chen, *Novel synthesis of activated biochar-supported catalysts for pyrolysis of cardboard waste derived from express package.* Fuel, 2023. **332**: p. 126136.

40. Li, Y., et al., *Deactivation mechanism and regeneration effect of bi-metallic Fe-Ni/ZSM-5 catalyst during biomass catalytic pyrolysis.* Fuel, 2022. **312**: p. 122924.

41. Xue, X., et al., *Dual-catalyst catalytic pyrolysis of poplar sawdust: A systematic study on first-layered catalysts.* Chemical Engineering Journal, 2022. **431**: p. 134251.

42. Wei, X., et al., *Hierarchical gallium-modified ZSM-5@SBA-15 for the catalytic pyrolysis of biomass into hydrocarbons.* Renewable Energy, 2022. **200**: p. 1037–1046.

43. Yang, C., et al., *Pyrolytic behaviors of Scenedesmus obliquus over potassium fluoride on alumina.* Fuel, 2020. **263**: p. 116724.

44. Zeng, K., et al., *Catalytic pyrolysis of Eupatorium adenophorum by sodium salt.* Journal of Material Cycles and Waste Management, 2021. **23**(4): p. 1626–1635.

45. Yadav, K., et al., *A robust hybrid data driven approach to model biochar yield in terms of biomass pyrolysis.* Bioresources and Bioprocessing, 2026. **13**(1): p. 62.

46. Aksu, G., C.O. Güzeller, and M.T. Eser, *The effect of the normalization method used in different sample sizes on the success of artificial neural network model.* International journal of assessment tools in education, 2019. **6**(2): p. 170–192.

47. Belsley, D.A., E. Kuh, and R.E. Welsch, *Regression diagnostics: Identifying influential data and sources of collinearity*. 2005: John Wiley & Sons.

48. Zhang, Z., L.A. Tell, and Z. Lin, *Development of Machine Learning and Chemical Language Model-Based QSAR Models for Predicting Drug Residue Depletion Half-Lives in Plasma and Tissues of Cattle Across Various Administration Routes.* Journal of Veterinary Pharmacology and Therapeutics, 2026. **49**(2): p. 150–171.

- ✓ Biochar yield prediction is essential for optimizing pyrolysis processes and advancing sustainable biomass utilization. This study develops a unified, interpretable machine learning framework that integrates Gradient Boosting Decision Trees (GBDT) with four metaheuristic optimization algorithms including Ant Colony Optimization (ACO), Whale Optimization Algorithm (WOA), Coupled Simulated Annealing (CSA), and Particle Swarm Optimization (PSO), to enhance predictive accuracy
- ✓ GBDT-ACO model achieved the highest predictive accuracy, yielding test R2 values of 0.709 and test MSE of 16.284
- ✓ The results highlight the critical influence of ash content, residence time, and peak temperature on yield formation, while also revealing the trade off between computational efficiency and predictive robustness across optimization strategies

Declaration of 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.