Artificial neural networks for the prediction of biochar yield: A comparative study of metaheuristic algorithms

Journal Pre-proofs

Artificial neural networks for the prediction of biochar yield: A comparative study of metaheuristic algorithms

Muzammil Khan, Zahid Ullah, Ondřej Mašek, Salman Raza Naqvi, Muhammad Nouman Aslam Khan

PII: S0960-8524(22)00544-2
DOI: https://doi.org/10.1016/j.biortech.2022.127215
Reference: BITE 127215

To appear in: *Bioresource Technology*

Received Date: 25 February 2022
Revised Date: 18 April 2022
Accepted Date: 19 April 2022

Please cite this article as: 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, *Bioresource Technology* (2022), doi: https://doi.org/10.1016/j.biortech.2022.127215

Artificial neural networks for the prediction of biochar yield: A comparative study of metaheuristic algorithms

Muzammil Khan[1] (engrmuzammilkhan@gmail.com)

Zahid Ullah[1] (engineerzahidullah@gmail.com)

Ondřej Mašek[2*] (ondrej.masek@ed.ac.uk)

Salman Raza Naqvi[1] (salman.raza@scme.nust.edu.pk)

Muhammad Nouman Aslam Khan[1] (mnouman@scme.nust.edu.pk)

[1]School of Chemical and Materials Engineering, National University of Sciences and Technology, Islamabad 44000, Pakistan

[2*]UK Biochar Research Centre, School of GeoSciences, University of Edinburgh, King's Buildings, Edinburgh EH9 3JN, UK

To whom correspondence may be addressed,

Ondřej Mašek, PhD

Email: (ondrej.masek@ed.ac.uk)

Abstract

In this study, an integrated framework of artificial neural networks (ANNs) and metaheuristic algorithms have been developed for the prediction of biochar yield using biomass characteristics and pyrolysis process conditions. Comparative analysis of six different metaheuristic algorithms was performed to optimize the ANN architecture and select important features. The results suggested that the ANN model coupled with the Rao-2 algorithm outperformed (R^2~0.93, RMSE~1.74%) all other models. Furthermore, the detailed information behind the models was acquired, identifying the most influencing factors as follows: pyrolysis temperature (56%), residence time (23%), and heating rate (8%). The partial dependence plot analysis revealed how each influencing factor affected the target variable. Finally, an easy-to-use software tool for predicting biochar yield was built using the ANN-Rao-2 model. This study demonstrates huge potential that machine learning presents in predictive modelling of complex pyrolysis processes, and reduces the time-consuming and expensive experimental work for estimating the biochar yield.

Keywords

Pyrolysis; Biochar; Optimization; Machine Learning; Neural Networks

1. Introduction

Diminishing fossil fuel reserves, environmental challenges, and an elevation in the market price of fossil fuels have accelerated the thrust of scientists and researchers to investigate commercial and environmental-friendly energy resources (Kang et al., 2020; Su et al., 2022). In this regard, biomass has received considerable interest as one of the most abundant and promising renewable raw materials that could serve as a potential alternative to fossil fuels. (Iaccarino et al., 2021). Among various conversion pathways of biomass, the thermochemical conversion process has been considered a promising technology for producing high value-added products from biomass due to the versatility in feedstock processing, lower reaction time, low capital cost, range of products, high process yield, CO2 neutrality, and product up-gradation (Won et al., 2022). Pyrolysis, a type of thermochemical conversion process, is a widely used process for converting solid biomass into solid (biochar), liquid (bio-oil), and gaseous (pyrolytic gas) products in an oxygen-free environment (Abnisa et al., 2013; Khuenkaeo & Tippayawong, 2020). Among these pyrolysis products, biochar has received much attention due to its unique properties such as high surface area and surface functional groups (Wang & Wang, 2019; Weber & Quicker, 2018). It has a carbon-rich microporous structure that has been used in various applications such as carbon sequestration in soil (Windeatt et al., 2014), pollutant sorption (Sizmur et al., 2017), pollution remediation (Wang et al., 2019), and carbon-based products (Zhou et al., 2021). However, the properties and productivity of biochar vary depending on the type of biomass and pyrolysis conditions. As a result, researchers have investigated the biochar production process utilizing diverse biomass feedstock and developed ways for maximizing biochar yield, which could eventually play a substantial role in biomass exploitation (Mani et al., 2011; Panwar et al., 2019).

The biochar yield mainly depends on characteristics of biomass feedstock and pyrolysis process parameters (Hasan et al., 2017; Şensöz & Angin, 2008). As the most generally used approaches, proximate analysis (fixed carbon (FC), volatile matter (VM), moisture contents (MC), ash) and ultimate analysis (carbon (C), hydrogen (H), nitrogen (N), oxygen (O) contents) could reflect the diversity and internal attributes of biomass (Ding et al., 2017). Meanwhile, the structural information (cellulose, hemicellulose, lignin, and some extractive) of biomass feedstock and pyrolysis process conditions such as highest treatment temperature (HTT), heating rate (HR), residence time (RT), particle size (PS), and sweep gas flow rate were assumed to be the most critical parameters in the pyrolysis process (Chen et al., 2015). Since then, significant research has focused on the impact of the aforementioned parameters on biochar yield. For instance, it was found by Demirbas (Demirbas, 2004) that pyrolysis of different biomass feedstock (olive husk, corncob, and tea waste) under similar experimental conditions resulted in different biochar yields. The olive husk with the highest lignin content produced the most biochar, whereas the corncob with the lowest lignin content produced the least. Moreover, the presence of water or moisture in the biomass feedstock negatively impacts the biochar yield. The pyrolysis of wood with 5% moisture content resulted in a higher biochar yield than the wood biomass with 20% moisture content (Fang et al., 2014). It was discovered that the greater the concentration of oxygen in the biomass, the greater the reactivity (Gani & Naruse, 2007). The higher biochar production requires a low temperature and a lengthy vapor residence time (Encinar et al., 1996). Furthermore, the increase in pyrolysis temperature from 400 to 700 °C has reduced biochar yield by 10% for hazelnut shells (Pu & Pu, 1999) and 17% for sesame stalk (Ateş et al., 2004). The effect of particle size of various biomass wastes (olive husk, corncob) on the biochar yield has been reported by Demirbas (Demirbas, 2004). It was reported that changing the particle size of

olive husk and corncob from 0.5 mm to 2.2 mm enhanced the biochar yield by 16.2% and 10.9%, respectively. This is because as particle size increases, the distance between the surface of the input biomass and its core increases, slowing the fast heat flow from the hot to cold end. The char yield is favored by this temperature gradient (Encinar et al., 2000).

Although numerous research studies have been reported for investigating the impact of various parameters on biochar yield, identifying the suitable and optimum parameters for biochar production is a difficult task. Traditional approaches of exploring biochar yield and its interaction with inducing factors need complicated experimentation, which is time-consuming, costly, and challenging to carry out. To productively examine the total impact of feedstock attributes and pyrolysis process conditions on biochar yield, it is important to analyze the process of biomass pyrolysis with consolidated contemplations of feedstock characteristics and pyrolysis conditions implementing emerging techniques such as big data, artificial intelligence, and machine learning approaches.

Fortunately, machine learning algorithms can predict and evaluate relationships between input and output data (Ahmad et al., 2021; Khan et al., 2021). In this regard, various machine learning-based studies have been conducted in the last decade for forecasting biochar yield (Aziz et al., 2017; Cao et al., 2016; Ewees & Elaziz, 2020; Pathy et al., 2020; Zhu et al., 2019). For instance, Cao et al., (Cao et al., 2016) have successfully developed a least square support vector machine model (coefficient of determination (R^2)~0.96) to predict biochar yield using cattle manure as a feedstock. Additionally, an adaptive neuro-fuzzy inference system (ANFIS) coupled with the optimization method was proposed by Aziz et al., (Aziz et al., 2017) for biochar yield prediction using cattle manure and has achieved good performance (R^2=0.98). Furthermore, Zhu et al., (Zhu et al., 2019) have successfully employed a random forest model for the biochar yield (R^2=0.85)

and carbon content (R^2=0.85) prediction using the composition of lignocellulosic biomass and pyrolysis conditions. Pathy et al., (Pathy et al., 2020) have used the extreme gradient boosting model for the prediction of biochar based on algal feedstock with (R^2=0.84). Ewees and Elaziz, (Ewees & Elaziz, 2020) have developed an improved ANFIS model combined with the grey wolf optimization method for estimating the biochar produced from manure pyrolysis with (R^2=0.98).

Although considerable research has been performed on the application of machine learning in biochar yield prediction, most of these studies considered only one type of biomass feedstock (mostly cattle manure), a very few influencing parameters, and a limited dataset. Additionally, various features or parameters are involved in biomass pyrolysis for the production of biochar and the identification of the most influential features is critical. The selection of important features not only saves time and resources but could sometimes enhance the product yield as well. Various methods are used for the selection of important features including filter, wrapper, and embedded methods (Chandrashekar & Sahin, 2014). These methods help identify the best features that could not only help the ML models to train fast but also boost the accuracy of the model and reduce the overfitting. Among these, the wrapper method trains the real ML algorithm on each set of features to select the most significant features. It saves time and enhances results by utilizing cross-validation performance (Wah et al., 2018). Motivated by this, a genetic algorithm-based approach has been successfully developed in our previous study for the selection of features for bio-oil yield prediction (Ullah et al., 2021). One of the motivations of this study was to extend and develop a detailed framework for biochar yield prediction. However, the study by Ullah et al. (2021) was limited to the analysis of single optimization method (genetic algorithm (GA)) which makes it hard to evaluate and generalize the

performance of optimization methods. Besides GA, the comparative analysis of various other metaheuristic optimization techniques including grey wolf optimization (GWO), Rao algorithms (RA), particle swarm optimization (PSO), and sine cosine algorithm (SCA) should be implemented not only the feature selection but also for the ANN architecture selection to identify the best algorithm for biochar yield prediction. Moreover, the compilation of robust dataset containing detailed information on pyrolysis process conditions, proximate analysis, ultimate analysis, and structural information (cellulose, hemicellulose, and lignin contents) was also lacking in most of the studies.

Motivated by the above-mentioned gaps, ANN-based models coupled with various metaheuristic optimization algorithms have been developed for predicting biochar yield utilizing biomass feedstock characteristics and pyrolysis conditions. we have tried our best to include most of the influential parameters in our study to investigate the impact of various biomass feedstock and input parameters on the biochar yield. The input parameters considered in this study include ultimate analysis, proximate analysis, biomass structural information, and pyrolysis conditions. In addition, various metaheuristic optimization methods (GWO, PSO, RA-1, RA-2, GA, and SCA) have been employed for features selection as well as for the selection of ANN architecture (number of hidden layers and number of associated neurons), which was the key novelty of current study. The reason for employing comparative analysis of multiple optimization methods was to identify the best method that selects the best input features and ANN architecture which eventually boosts the prediction performance of the ANN model. Furthermore, to examine the impact of each input parameter on the biochar yield, a partial dependence plot analysis was carried out. Lastly, a graphical user interface (GUI) based software package was developed for the prediction of biochar yield.

2. Materials and methods

2.1. Data collection and pre-processing

The development of a robust dataset specific to the application is mandatory before developing the ML model. In this study, various biomass feedstock for the biochar pyrolysis process was considered including lignocellulosic biomass, herbaceous plants, sewage sludge, and animal manure. The biochar yield is primarily determined by the properties of the biomass feedstock and the pyrolysis process parameters. Proximate and ultimate analyses, as the most commonly used approaches, could reflect the diversity and intrinsic attributes of biomass. Meanwhile, biomass feedstock structural information and pyrolysis process conditions were considered to be the most essential parameters in the pyrolysis process (Hasan et al., 2017; Şensöz & Angin, 2008). Therefore, the proximate analysis (MC, FC, VM, and ash content), ultimate analysis (C, H, N, and O content), and structural analysis (cellulose, hemicellulose, and lignin) of the biomass feedstock, as well as the pyrolysis conditions (HTT, HR, RT, and PS), were selected as input parameters. While the corresponding biochar yield was considered as output or target variable.

For the data collection, a systematic literature review was conducted in the citation databases of Google Scholar, Web of Science, and Science Direct using multiple search terms such as machine learning, biochar, pyrolysis products, lignocellulosic biomass, sewage sludge, and animal wastes, among others. Following the retrieval of relevant articles, data extraction and compilation was performed to develop and evaluate machine learning-based predictive models. A total of 402 data points were extracted from 89 articles in the timeframe ranging from 2004-2021 using 15 different types of lignocellulosic biomass, 10 different types of herbaceous plants, and 7 other types of feedstock (see supplementary materials). The feedstock was normally dried

and crushed into smaller-size particles before being analyzed for feedstock structural information and proximate and ultimate analyses. In the studies, the biochar yield was reported on a dry ash-free basis. The pyrolysis experimentations were performed at varying HR, HTT, and RT values. Finally, the biochar was gathered so that the yield could be calculated. Following data collection, the data was pre-processed to normalize all input and output units. Except for variables that were deleted owing to unresolved missing data, each set of data contained valid values for all variables.

2.2. Employed meta-heuristic algorithms

i. Grey-wolf optimization (GWO)

GWO was inspired by the social hierarchy and the hunting technique of grey wolves. It mimics the leadership and hunting mechanism of grey wolves. In the mathematical model, the best solution is designated as alpha (α), the second-best as beta (β), and the third-best as delta (δ). The remaining potential solutions are considered to be omegas (ω). To replicate hunting behavior, three steps are used: (i) encircling, (ii) tracking, and (iii) assaulting the prey. The equation for encircling the prey is given in the equation (Mirjalili et al., 2014).

$$D(t+1) = |A.Xp(t+1) - X(t+1)|$$

$$X(t+1) = |Xp(t+1) - F.D(t+1)|$$

where D is the distance between the wolf X and the prey Xp, while F and A are the coefficient vectors.

$$F = 2f.r1 - f$$

$$A = 2r$$

where r1 and r2 are random vectors in the interval [0, 1], and the value of parameter a is reduced from 2–0 in linear form with each iteration.

The location of any wolf X in the present population can be updated based on the positions of α, β, and δ.

$$X = \frac{X1 + X2 + X3}{3}$$

$$X1 = |X\alpha - A.D\alpha|,\ X2 = |X\beta - A.D\beta|,\ X3 = |X\delta - A.D\delta|$$

$$D\alpha = |C2.X\alpha - X|,\ D\beta = |C2.X\beta - X|,\ D\delta = |C3.X\delta - X|$$

ii. Rao algorithms (RA)

Ravipudi Venkata Rao introduced metaphor-free optimization methods named Rao algorithms in 2020. It employs both the best and the worst alternatives in each generation, as well as random encounters among the possible solutions, to quickly locate an optimum solution. They require two common parameters: population size and a maximum number of evaluations that can be readily adjusted. Rao algorithms (Rao-1 and Rao-2) employ the three equations listed below (Rao, 2020).

$$X'_{j,k,i} = X_{j,k,i} + r_{1,j,i}(X_{j,best,i} - X_{j,worst,i})$$

$$X'_{j,k,i} = X_{j,k,i} + r_{1,j,i}(X_{j,best,i} - X_{j,worst,i}) + r_{2,j,i}(|X_{j,k,i}\ or\ X_{j,l,i}| - |X_{j,l,i}\ or\ X_{j,k,i}|)$$

Where $X_{j,best,i}$ and $X_{j,worst,i}$ signify the best and worst choices as the value of parameter j during the ith iteration, respectively. $X'_{j,k,i}$ is the updated value following the equation, and $r_{1,j,i}$ and

$r_{2,j,i}$ are produced at random in the range [0,1] for the jth variable during the ith iteration. The phrase $|X_{j,k,i}\ or\ X_{j,l,i}|$ refers to the comparison of a candidate solution k to another candidate l chosen at random the available candidates in the population.

iii. Sine Cosine algorithm (SCA)

SCA was inspired by the principles of trigonometric sine and cosine functions. It updates individual locations toward the ideal solution using the principle of trigonometric sine and cosine functions. The solutions in SCA are updated using the following equations (Mirjalili, 2016):

$$X_i^{t+1} = X_i^t + r_1 * Sin(r_2) * |r_3 P_i^t - X_i^t|)$$

$$X_i^{t+1} = X_{ij}^t + r_1 * Cos(r_2) * |r_3 P_i^t - X_{ii}^t|)$$

$$r1 = a - a * \frac{iter}{itermax}$$

where X_i^t is the position of the existing solution, r1/r2/r3 are random numbers, Pi is the position of the destination point, a is a constant parameter and its value is taken 2 in the present study. While iter represents the present iteration number and itermax shows the total number of iterations (taken 100 in the present study).

iv. Genetic algorithm (GA)

GA replicates the process of biological evolution and is based on natural selection and genetic inheritance. GA employs a fitness (objective) function to measure the fitness of each member of the population. Following that, GA operators (selection, crossover, mutation) are utilized to generate subsequent populations (Katoch et al., 2021). To fix poor solutions, the best alternatives are chosen at random using a selection technique. The crossover produces a pair of offspring

having features from both parents (members), to improve fitness values. The mutation is the appearance of a random change in the value of a string position with the minimum possible probability. Its objective is to keep critical information contained inside strings from being lost prematurely. The above technique is continued until an optimal solution(s) is (are) identified or the relative difference between solutions is less than a given limit is reached. The parametric setting in the present study is as follows; elite count = 2, population type = bilstring, crossover-type = crossover arithmetic, mutation type = mutation uniform, selection type = selection roulette, mutation probability = 0.1, and crossover probability = 0.8 (Ewees & Elaziz, 2020).

v. Particle swarm optimization (PSO)

PSO was inspired by the cooperative behavior of some animals, such as bird flocks or fish schools. Each solution in PSO is referred to as a "bird" in the search space, which is termed as a "particle." A swarm of these particles traverses the search space in search of the optimal location. In an N-dimensional space problem, the particle carries the position vector as well as the velocity vector. The location vector is denoted by Xi = (xi1, xi2, xi3, ... xiN), whereas the velocity vector is denoted by Vi=(vi1,vi2,...,viN). where i represents the ith particle and N represents the problem dimension or the number of unknown variables. The position and velocity of the particles are updated using the two equations below (Clerc, 2010).

$$V_i^k = w.V_i^k + c_1 r_1 \left(pbest_i^k - X_i^k\right) + c_2 r_2 (gbest^k - X_i^k)$$

$$X_i^{k+1} = X_i^k + V_i^{k+1}$$

where 'pbest' represents the ith particle's personal best position, and 'gbest' represents the swarm's best-so-far position search. The inertia weight w represents the influence of the prior velocity

vector on the new vector. The velocity in all dimensions is given an upper restriction Vmax. In this study, the parametric setting was as follows; minimum adopted neighborhood = 0.25, self-adjustment weight = 1.49, social adjustment weight = 1.49, and inertia weight = 1 (Ewees & Elaziz, 2020).

2.3. Artificial neural networks (ANN)

ANN possess the self-learning ability; they may enhance their performance as additional information becomes available. They may acquire and store knowledge (information-based) and can be described as a group of processing units depicted by artificial neurons, unified by huge interconnections (artificial synapses), and executed by synaptic weights vectors and matrices. ANN builds a framework by connecting mathematical ‘‘neurons," or ‘‘nodes," to simulate complicated processes. Generally, ANN contains input, hidden, and output layers. Each neuron in the hidden layer and output layer has a linear or nonlinear activation function. The most used activation functions are sigmoid, log sigmoid, and purlin (Abiodun et al., 2018).

$$Y_i = \mathrm{f}\left(\sum_{j=1}^{n} X_j W_{ij}\right)$$

Where X represents the input features, W represents the weight associated with the respective input feature. Each input parameter is multiplied with its weight and is passed through the summing junction to give the output Y. In this study, all the ANN models were developed using MATLAB R2021b software

2.4. Performance evaluation measures

The root means squared error (RMSE) and coefficient of determination (R^2) criteria were used to assess the performance of all six artificial neural network (ANN) based models (Zhang et al.,

2019).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i^{exp} - Y_i)^2}{\sum_{i}^{n}(Y_i^{exp} - Y_{avg}^{exp})^2}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i}^{n}(Y_i^{exp} - Y_i)^2}$$

Where n stands for the number of test samples, Y_i^{exp}stands for actual value, and Y_i denotes the predicted value. Y_{avg}^{exp} represents the mean value of actual and predicted values.

3. Results and discussion

3.1. Statistical analysis of the dataset

Boxplot showed the data distribution and the presence of outliers in the collected dataset. The data was pre-processed to handle the outlier and improve the quality of the dataset. The in-built functions of MATLAB R2021b software were used for handling missing values and outliers in the data. The boxplot analysis in Figure 1 shows the visualization of the data distribution for various features including biomass feedstock composition and pyrolysis conditions. The C and O in the biomass feedstock varied in the range of 29.42% to 64.23 and 18.47% to 57.2%, respectively. The distribution of H and N fell within the accepted norms of 4.28–11.6 % and 0–8.7 %, respectively. The weight fraction of lignin contents ranged from 1.96% to 50.6%, whereas the weight fractions of cellulose and hemicellulose ranged from 2.4% to 60.57% and 3.3% to 52.75%, respectively. The FC, VM, MC, and ash contents ranged from 0-72.9%, 19.4-86.6%, 0-70.16%, and 0.2-44.61%, respectively. The VM fluctuated the most due to the presence of diverse biomass feedstock. Moreover, the sewage sludge contains numerous inorganics, its ash content was significant (Manara & Zabaniotou, 2012).

The pyrolysis process parameters including HR, PS, HTT, and RT ranged from 3.5-232.5 ^{o}C/min, 0.15-15 mm, 350-927 ^{o}C, 0.5-120 min, respectively; given the fact that we considered many cases of pyrolysis from various articles. The biochar yield ranged from 11.7% to 56%. These data points revealed the possibility of optimizing the feedstock and operating parameters of the pyrolysis process to attain a high biochar yield. It is worth noting that the wide-ranging range of values for each variable showed that the compiled dataset will allow the model to learn from various data, hence contributing to generalization.

3.2. Proposed frameworks for features and ANN architecture selection

Various metaheuristic optimization algorithms including GWO, RA-1, RA-2, SCA, PSO, and GA were used in this study to select the most influential features as well as the ANN architecture. Figure 2 depicts the schematic framework used for the execution of these algorithms coupled with the ANN model. The number of iterations and population size was kept constant for all methods to evaluate the relative performance and to conduct a fair comparison of various optimization methods (population size = 25, maximum iterations = 100). The working flow of the proposed framework is given as follows;

- The dataset comprising of 402 data points was first distributed into 80% training and 20% testing data.
- The population size of 25 was initialized
- The ANN model was developed using generated set of solutions
- The fitness function was evaluated (RMSE in both cases)
- If the number of iterations exceeded the limit, the loop was terminated. Otherwise, it would go for the next iteration to generate the new population

- Once the loop was terminated, the best input features and ANN architecture were selected by the metaheuristic optimization methods based on the lowest RMSE value.

To begin, the GWO method was used to select features for both the proximate and ultimate analysis cases, followed by the selection of an ANN architecture using the same optimization method. All other optimization methods (PSO, GA, RA-1, RA-2, SCA) followed the same strategy. Table 1 shows the features selected by various optimization methods for both the proximate and ultimate cases. For proximate analysis, the GWO method selected 8 input features out of 11 features, while for ultimate analysis, a total of 9 features were chosen. Similarly, the RA-1, RA-2, SCA, GA, and PSO methods selected 8, 9, 8, 7, and 5 features for proximate analysis, respectively, whereas 4, 9, 7, 8, and 9 features were chosen for ultimate analysis. Likewise, these metaheuristic algorithms were also successfully employed for ANN architecture (hidden layers and associated neurons) selection as shown in Table 2. The set of hidden layers and associated neurons were selected based on the lowest RMSE values of the ANN model.

3.3. Prediction performance of ANN models

After the selection of optimal input features and ANN architectures using metaheuristic optimization algorithms (discussed in previous sections 3.1 and 3.2), the ANN-based models were successfully trained and tested using the set of these selected features and architectures to predict the biochar yield. These ANN models coupled with various optimization methods are termed ANN-GWO, ANN-RA 1, ANN-RA 2, ANN-SCA, ANN-GA, and ANN-PSO. The ANN models were constructed considering proximate analysis data with structural information and pyrolysis conditions, and the ultimate analysis data with structural information and pyrolysis conditions. A total of 402 data points were used for developing the ANN models with 80%

training and 20% being used as testing data. A feed-forward neural network (ANN) with backpropagation using the Levenberg-Marquardt algorithm was used for training the model. In the hidden and output layers, the tangent sigmoid and purelin transfer functions were employed, respectively. The coefficient of determination (R^2) and root mean squared error (RMSE) values were used to assess the performance of ANN models as shown in Table 3. The prediction performance of ANN-based models was validated using a test dataset as shown in Figure 3.

In the case of proximate analysis (case 1), the ANN-RA 1, ANN-GA, and ANN-PSO showed almost comparable performance (R^2~0.92, RMSE<2) for biochar yield prediction. However, the analysis based on selected features and ANN architecture as shown in Table 1 and 2 showed that the PSO model selected 5 features out of 12 features in the case of proximate analysis while GA and RA 1 models have selected 7 and 8 features, respectively. Similarly, the PSO, GA, and RA 1 models have selected 2 hidden layers while RA 2 has selected 3 hidden layers, hence required comparatively more computational time. In addition, the ANN-RA 2 model exhibited high performance (R^2~0.93, RMSE~1.74) followed by the ANN-PSO, ANN-GA, and RA 1 models (R^2~0.92, RMSE<2 each). This is because, the best model is the model with least number of input parameters, hidden layers, and number of neurons along with highest R^2 and lowest RMSE values, which are the performance evaluation criterion. As a result, the model with the highest R^2 and lowest RMSE values is not deemed the best of all models as long as it has the least input parameters, hidden layers, and associated neurons. Therefore, in terms of R^2 and RMSE values, ANN-RA 2 model outperformed ANN-PSO, ANN-GA, and RA 1 models. However, in terms of a number of features and ANN architecture, GA and PSO have selected least features (7 features out of 11 input features) followed by RA 1 and GWO (8, 8 features), and RA 2 and PSO (9, 9 features). Similarly, GA, PSO, and RA 1 models have selected 2 hidden layers while RA 2 has

selected 3 hidden layers. Hence, in terms of number of in terms of a number of features and ANN architecture, GA, PSO, and RA 1 performed better than RA 2.

In the case of ultimate analysis (case 2), the ANN-RA 1 model showed poor results (R^2~0.70, RMSE~3.55) followed by ANN-SCA (R^2~0.78, RMSE~3.43). However, the ANN-RA 2, ANN-GA and ANN-PSO-based models showed identical performance (R^2~0.88, RMSE<3). In terms of a number of features and ANN architecture selection, the RA 1 has selected 4 features out of 11 input features followed by SCA and GA with 7 and 8 features, respectively. While GWO, RA 2, and PSO have selected 9 features each. All of the models performed better than the previously reported ML-based studies for biochar yield prediction (Pathy et al., 2020; Zhu et al., 2019).

In summary, various ANN models coupled with metaheuristic optimization algorithms were successfully trained and tested for the prediction of biochar yield. Results indicated that ANN-SCA model showed poor performance compared to ANN-GWO, ANN-RA 1, ANN-RA 2, ANN-GA, and ANN-PSO. The reason behind ANN-SCA's poor performance in comparison to other models is that for the same number of iterations and population, the model was unable to find a comparatively satisfactory solution and may require additional iterations. Furthermore, the ANN-GA and ANN-PSO models showed almost identical results in both proximate and ultimate cases. It was observed that the ANN-RA 2 and ANN-PSO outperformed all other models in terms of R^2 and RMSE values as well as based on the least number of features and ANN architecture selection.

In general, the trend of prediction performances of various ANN models (most to least accurate) based on R^2 and RMSE values for case 1 (proximate analysis) was as follows; (ANN-RA 2 > ANN-PSO > ANN-GA > ANN-RA-1 > ANN-GWO > ANN-SCA > ANN-RA-2). Similarly, in case 2 (ultimate analysis) the overall trend was as follows; (ANN-PSO > ANN-RA 2 > ANN-GA

> ANN-GWO > ANN-SCA> ANN-RA 1). Overall, the proximate analysis-based ANN models outperformed ultimate analysis-based models. It is clear from the findings that the novel approach for features and ANN architecture selection using optimization algorithms helped select the optimal features and ANN architecture which would help reduce the number of features, and computational time with better ANN prediction performance. Thus, it is suggested based on this study that ANN models may be employed well for the prediction of biochar yield without the need for lab-based experiments, hence reducing the time and expense of testing with the minimum features.

The prediction performance of the employed proximate analysis-based models (ANN-GWO, ANN-RA 1, ANN-RA 2, ANN-SCA, ANN-GA, and ANN-PSO) over various data points are shown in Figure 4. The black dotted line shows the actual data, blue circles show ANN-GWO, the green line represents ANN-RA 1, red diamond-shaped points represent ANN-RA 2, purple square-shaped points denote ANN-SCA, pink line denotes ANN-GA, and the light blue line represents the variations in the ANN-PSO model, respectively. It is observed that ANN-RA 2 closely followed the same trend over various data points as followed by the experimental data. After ANN-RA 2, the ANN-PSO has also shown a smaller deviation from the actual data points. Whereas, ANN-SCA showed clear predicting deviation from the actual data and it is also proved from its lower R^2 and higher RMSE values as shown in Table 1, hence it is not recommended this case due to its poor performance. Finally, the results of all ANN models (ANN-GWO, ANN-Rao 1, ANN-Rao 2, ANN-SCA, ANN-PSO, and ANN-GA) were compared and validated with 10 different experimental studies taken from the literature (Abnisa et al., 2013; Angin, 2013; Chen et al., 2015; Demirbas, 2004; Iaccarino et al., 2021; Khuenkaeo & Tippayawong, 2020; Mani et al., 2011; Oginni et al., 2017; Şensöz & Angin, 2008; Windeatt et al., 2014). The

validation of developed models with the experimental studies showed an average deviation of <2% (see supplementary materials). Consequently, based on these comparisons, we may infer that the ANN-RA 2 method predicts biochar yield more accurately than the other methods. The results show that the suggested ANN-based models may be utilized successfully for the prediction of biochar yield without the need for lab-based experiments, hence saving time and resources. Furthermore, the significance level (p-value, a statistical test used to check the validity of a hypothesis against observable data) of all models was evaluated using the T-test. The threshold level of significance (alpha) was selected to be 0.01. The null hypothesis was that the sample mean value is equal to the coefficient of determination (R^2) values of each model, but the alternative hypothesis was that the sample mean R^2 value is not equal to the R^2 values of the models shown in Table 3. The T-test was performed on R^2 value calculated form 20 random samples from the dataset each having 20 data points. The model with p-value greater than 0.01 means that the null hypothesis is accepted while the model with p-value lower than 0.01 means that the null hypothesis is rejected. In the case of proximate analysis, ANN-GWO, ANN-RA 2, and ANN-SCA models have p-value greater than 0.01 hence the null hypothesis was accepted while ANN-RA 1, ANN-GA, and ANN-PSO models have p-value lower than 0.01 hence the null hypothesis was rejected. Similarly, in the case of ultimate analysis, ANN-RA 2, ANN-SCA, ANN-GA, and ANN-PSO models have p-value greater than 0.01 hence the null hypothesis was accepted while ANN-RA 1 and ANN-GWO models have p-value lower than 0.01 hence the null hypothesis was rejected. In summary, the significance analysis indicated how significant the average R^2 values from 20 runs of each ANN model.

3.4. ANN model-based feature analysis

The influence of various input parameters on the biochar yield can be evaluated using feature importance based on the ANN-RA 2 model (best model). The impact of selected features for ANN-RA 2 is shown in Figure 5. The features were ranked by their corresponding importance (from least to most important). It can be seen from Figure 5, that in the case of proximate analysis based ANN- RA 2 model, the most influential parameter on the biochar yield was HTT (56% importance), followed by RT (23%) and HR (8%) on the second and third rank based on their relative importance. The ranking of the relative importance of HR, HTT, and RT agreed well with the literature (Angin, 2013; Tripathi et al., 2016). Furthermore, the remaining features based on the ANN-RA 2 model were ranked as follows; FC (5%), Hem (3%), VM (2%), cellulose (1%), MC (0.3%), and Ash (0.3%).

Although the feature importance plot illustrates the relative importance of input parameters for the biochar yield, it could not interpret how the input parameters impact the biochar yield. As a result, partial dependence analysis (PDA) was carried out to elucidate the impact of the most important parameters on biochar yield. Figure 6 shows the PDA plot for the evaluation of the individual impact of input features on the biochar yield. The tick marks on the x-axis indicate the feature data density, and without them, the trendline may be meaningless. The influence of increasing HTT showed a negative impact on biochar yield, as shown in Figure 6a. The rapid decline in the biochar yield with an increase in the value HTT from 350-410 ^{o}C was due to the breakdown and devolatilization of volatile matter and moisture content confined in biomass feedstock (Zhao et al., 2018). In the range of 410-600 ^{o}C, the biochar yield was decreased but at a gradual pace which may be due to the breakdown of biomass structural contents (cellulose, hemicellulose, and lignin) (Stefanidis et al., 2014). Furthermore, the biochar yield decreases with an increase in the HR (Figure 6b), as the high HR results in thermal cracking of biomass and

supports the production of gaseous and liquid products. Whereas at low HR, no thermal cracking or depolymerization of biomass takes place leading to more biochar yield (Tripathi et al., 2016). Cellulose and hemicellulose contents in biomass feedstock are likely to produce volatile products while lignin is liable to the biochar yield. Therefore, the biochar yield decreases with an increase in cellulose and hemicellulose contents (Figure 6d and 6f) while increases with an increase in lignin content (Figure 6e) (Tomczyk et al., 2020). Nevertheless, some studies have also reported that biochar yield increases with an increase in the cellulose and lignin contents but the presence of lignin content in biomass results in more biochar compared to the cellulose content (Demirbas, 2004). In addition, Figure 6c shows that with an increase in ash content the biochar yield increases due to the presence of inorganic components in biomass feedstock (K. Wang et al., 2015). Overall, the PDP examination was useful for examining the influencing parameters of biomass pyrolysis and easing the process optimization in engineering applications. Future research is needed to better understand the relationship between various other parameters (such as the detailed structures of cellulose-hemicellulose-lignin, the categories of various woody and grassy biomass, the mass of biomass samples, the ash composition, and the pyrolyzer design parameters, among others) and the output parameter. To get a better understanding of the interaction between various input and output characteristics, deep learning-based prediction models for biochar production should be implemented. This research provides a practical technique for future comprehensive biomass conversion research.

4. Conclusions

The biochar yield was successfully predicted with an integrated framework of ANN and metaheuristic algorithms using biomass characteristics and pyrolysis conditions. The ANN-RA 2 model showed good performance (R^2~0.93) as compared to other models ANN-PSO (R^2~0.92), ANN-GA (R^2~0.92), ANN-RA 1 (R^2~0.92), SCA (R^2~0.84), and ANN-GWO (R^2~0.85). Furthermore, the validation of developed models with the experimental studies showed an average deviation of <2%. PDP analysis indicated that the optimal parameters selected using various metaheuristic algorithms have a significant impact on biochar yield. This study paves the way for future implementation of ML methods in research and industrial applications, beyond predicting yields.

"E-supplementary data of this work can be found in online version of the paper"

CRediT authorship contribution statement

Muzammil Khan: Investigation, Writing - original draft. **Zahid Ullah:** Formal analysis, Writing - original draft. **Ondřej Mašek:** Methodology, Supervision, Writing - review & editing. **Salman Raza Naqvi:** Validation, Writing - review & editing. **Muhammad Nouman Aslam Khan:** Formal analysis, Validation.

References

[1] Abiodun, O. I., Jantan, A., Omolara, A. E., Dada, K. V., Mohamed, N. A. E., & Arshad, H. (2018). State-of-the-art in artificial neural network applications: A survey. *Heliyon*, *4*(11), e00938. https://doi.org/10.1016/j.heliyon.2018.e00938

[2] Abnisa, F., Arami-Niya, A., Daud, W. M. A. W., & Sahu, J. N. (2013). Characterization of Bio-oil and Bio-char from Pyrolysis of Palm Oil Wastes. *Bioenergy Research*, *6*(2), 830–840. https://doi.org/10.1007/s12155-013-9313-8

[3] Ahmad, I., Sana, A., Kano, M., Cheema, I. I., Menezes, B. C., Shahzad, J., Ullah, Z., Khan, M., & Habib, A. (2021). *Machine Learning Applications in Biofuels' Life Cycle: Soil, Feedstock, Production, Consumption, and Emissions*. 1–27.

[4] Angin, D. (2013). Effect of pyrolysis temperature and heating rate on biochar obtained from pyrolysis of safflower seed press cake. *Bioresource Technology*, *128*, 593–597. https://doi.org/10.1016/j.biortech.2012.10.150

[5] Ateş, F., Pütün, E., & Pütün, A. E. (2004). Fast pyrolysis of sesame stalk: Yields and structural analysis of bio-oil. *Journal of Analytical and Applied Pyrolysis*, *71*(2), 779–790. https://doi.org/10.1016/j.jaap.2003.11.001

[6] Aziz, M. A. El, Hemdan, A. M., Ewees, A. A., Elhoseny, M., Shehab, A., Hassanien, A. E., & Xiong, S. (2017). Prediction of biochar yield using adaptive neuro-fuzzy inference system with particle swarm optimization. *Proceedings - 2017 IEEE PES-IAS PowerAfrica Conference: Harnessing Energy, Information and Communications Technology (ICT) for Affordable Electrification of Africa, PowerAfrica 2017*, 115–120. https://doi.org/10.1109/PowerAfrica.2017.7991209

[7] Cao, H., Xin, Y., & Yuan, Q. (2016). Prediction of biochar yield from cattle manure pyrolysis via least squares support vector machine intelligent approach. *Bioresource*

Technology, *202*, 158–164. https://doi.org/10.1016/j.biortech.2015.12.024

[8] Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers and Electrical Engineering*, *40*(1), 16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024

[9] Chen, D., Liu, D., Zhang, H., Chen, Y., & Li, Q. (2015). Bamboo pyrolysis using TG-FTIR and a lab-scale reactor: Analysis of pyrolysis behavior, product properties, and carbon and energy yields. *Fuel*, *148*, 79–86. https://doi.org/10.1016/j.fuel.2015.01.092

[10] Clerc, M. (2010). Particle Swarm Optimization. *Particle Swarm Optimization*, 1942–1948. https://doi.org/10.1002/9780470612163

[11] Demirbas, A. (2004). Effects of temperature and particle size on bio-char yield from pyrolysis of agricultural residues. *Journal of Analytical and Applied Pyrolysis*, *72*(2), 243–248. https://doi.org/10.1016/j.jaap.2004.07.003

[12] Ding, Y., Ezekoye, O. A., Lu, S., Wang, C., & Zhou, R. (2017). Comparative pyrolysis behaviors and reaction mechanisms of hardwood and softwood. *Energy Conversion and Management*, *132*, 102–109. https://doi.org/10.1016/j.enconman.2016.11.016

[13] Encinar, J. M., Beltrán, F. J., Bernalte, A., Ramiro, A., & González, J. F. (1996). Pyrolysis of two agricultural residues: Olive and grape bagasse. Influence of particle size and temperature. *Biomass and Bioenergy*, *11*(5), 397–409. https://doi.org/10.1016/S0961-9534(96)00029-3

[14] Encinar, J. M., González, J. F., & González, J. (2000). Fixed-bed pyrolysis of Cynara cardunculus L. Product yields and compositions. *Fuel Processing Technology*, *68*(3), 209–222. https://doi.org/10.1016/S0378-3820(00)00125-9

[15] Ewees, A. A., & Elaziz, M. A. (2020). Improved Adaptive Neuro-Fuzzy Inference System Using Gray Wolf Optimization: A Case Study in Predicting Biochar Yield. *Journal of Intelligent Systems*, *29*(1), 924–940. https://doi.org/10.1515/jisys-2017-0641

[16] Fang, J., Leavey, A., & Biswas, P. (2014). Controlled studies on aerosol formation during biomass pyrolysis in a flat flame reactor. *Fuel*, *116*, 350–357. https://doi.org/10.1016/j.fuel.2013.08.002

[17] Gani, A., & Naruse, I. (2007). Effect of cellulose and lignin content on pyrolysis and combustion characteristics for several types of biomass. *Renewable Energy*, *32*(4), 649–661. https://doi.org/10.1016/j.renene.2006.02.017

[18] Hasan, M. M., Wang, X. S., Mourant, D., Gunawan, R., Yu, C., Hu, X., Kadarwati, S., Gholizadeh, M., Wu, H., Li, B., Zhang, L., & Li, C. Z. (2017). Grinding pyrolysis of Mallee wood: Effects of pyrolysis conditions on the yields of bio-oil and biochar. *Fuel Processing Technology*, *167*, 215–220. https://doi.org/10.1016/j.fuproc.2017.07.004

[19] Iaccarino, A., Gautam, R., & Sarathy, S. M. (2021). Bio-oil and biochar production from halophyte biomass: effects of pre-treatment and temperature onSalicornia bigeloviipyrolysis. *Sustainable Energy and Fuels*, *5*(8), 2234–2248. https://doi.org/10.1039/d0se01664k

[20] Kang, Y., Yang, Q., Bartocci, P., Wei, H., Liu, S. S., Wu, Z., Zhou, H., Yang, H., Fantozzi, F., & Chen, H. (2020). Bioenergy in China: Evaluation of domestic biomass resources and the associated greenhouse gas mitigation potentials. *Renewable and Sustainable Energy Reviews*, *127*(April), 109842. https://doi.org/10.1016/j.rser.2020.109842

[21] Katoch, S., Chauhan, S. S., & Kumar, V. (2021). A review on genetic algorithm: past, present, and future. In *Multimedia Tools and Applications* (Vol. 80, Issue 5). Multimedia Tools and Applications. https://doi.org/10.1007/s11042-020-10139-6

[22] Khan, M., Mehran, M. T., Haq, Z. U., Ullah, Z., Naqvi, S. R., Ihsan, M., & Abbass, H. (2021). Applications of artificial intelligence in COVID-19 pandemic: A comprehensive review. *Expert Systems with Applications*, *185*(July), 115695. https://doi.org/10.1016/j.eswa.2021.115695

[23] Khuenkaeo, N., & Tippayawong, N. (2020). Production and characterization of bio-oil and biochar from ablative pyrolysis of lignocellulosic biomass residues. *Chemical Engineering Communications*, *207*(2), 153–160. https://doi.org/10.1080/00986445.2019.1574769

[24] Manara, P., & Zabaniotou, A. (2012). Towards sewage sludge based biofuels via thermochemical conversion - A review. *Renewable and Sustainable Energy Reviews*, *16*(5), 2566–2582. https://doi.org/10.1016/j.rser.2012.01.074

[25] Mani, T., Murugan, P., & Mahinpey, N. (2011). Pyrolysis of oat straw and the comparison of the product yield to wheat and flax straw pyrolysis. *Energy and Fuels*, *25*(7), 2803–2807. https://doi.org/10.1021/ef200546v

[26] Mirjalili, S. (2016). SCA: A Sine Cosine Algorithm for solving optimization problems. *Knowledge-Based Systems*, *96*, 120–133. https://doi.org/10.1016/j.knosys.2015.12.022

[27] Mirjalili, S., Mirjalili, S. M., & Lewis, A. (2014). Grey Wolf Optimizer. *Advances in Engineering Software*, *69*, 46–61. https://doi.org/10.1016/j.advengsoft.2013.12.007

[28] Oginni, O., Singh, K., & Zondlo, J. W. (2017). Pyrolysis of dedicated bioenergy

crops grown on reclaimed mine land in West Virginia. *Journal of Analytical and Applied Pyrolysis*, *123*, 319–329. https://doi.org/10.1016/j.jaap.2016.11.013

[29] Panwar, N. L., Pawar, A., & Salvi, B. L. (2019). Comprehensive review on production and utilization of biochar. *SN Applied Sciences*, *1*(2), 1–19. https://doi.org/10.1007/s42452-019-0172-6

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

[31] Pu, E., & Pu, A. E. (1999). *Pyrolysis of hazelnut shells in a fixed-bed.pdf*. *52*, 33–49.

[32] Rao, R. V. (2020). Rao algorithms: Three metaphor-less simple algorithms for solving optimization problems. *International Journal of Industrial Engineering Computations*, *11*(1), 107–130. https://doi.org/10.5267/j.ijiec.2019.6.002

[33] Şensöz, S., & Angin, D. (2008). Pyrolysis of safflower (Charthamus tinctorius L.) seed press cake: Part 1. The effects of pyrolysis parameters on the product yields. *Bioresource Technology*, *99*(13), 5492–5497. https://doi.org/10.1016/j.biortech.2007.10.046

[34] Sizmur, T., Fresno, T., Akgül, G., Frost, H., & Moreno-Jiménez, E. (2017). Biochar modification to enhance sorption of inorganics from water. *Bioresource Technology*, *246*, 34–47. https://doi.org/10.1016/j.biortech.2017.07.082

[35] Stefanidis, S. D., Kalogiannis, K. G., Iliopoulou, E. F., Michailof, C. M., Pilavachi, P. A., & Lappas, A. A. (2014). A study of lignocellulosic biomass pyrolysis via the pyrolysis of cellulose, hemicellulose and lignin. *Journal of Analytical and Applied*

Pyrolysis, *105*, 143–150. https://doi.org/10.1016/j.jaap.2013.10.013

[36] Su, G., Ong, H. C., Mofijur, M., Mahlia, T. M. I., & Ok, Y. S. (2022). Pyrolysis of waste oils for the production of biofuels: A critical review. *Journal of Hazardous Materials*, *424*(PB), 127396. https://doi.org/10.1016/j.jhazmat.2021.127396

[37] Tomczyk, A., Sokołowska, Z., & Boguta, P. (2020). Biochar physicochemical properties: pyrolysis temperature and feedstock kind effects. *Reviews in Environmental Science and Biotechnology*, *19*(1), 191–215. https://doi.org/10.1007/s11157-020-09523-3

[38] Tripathi, M., Sahu, J. N., & Ganesan, P. (2016). Effect of process parameters on production of biochar from biomass waste through pyrolysis: A review. *Renewable and Sustainable Energy Reviews*, *55*, 467–481. https://doi.org/10.1016/j.rser.2015.10.122

[39] Ullah, Z., khan, M., Raza Naqvi, S., Farooq, W., Yang, H., Wang, S., & Vo, D. V. N. (2021). A comparative study of machine learning methods for bio-oil yield prediction – A genetic algorithm-based features selection. *Bioresource Technology*, *335*(April), 125292. https://doi.org/10.1016/j.biortech.2021.125292

[40] Wah, Y. B., Ibrahim, N., Hamid, H. A., Abdul-Rahman, S., & Fong, S. (2018). Feature selection methods: Case of filter and wrapper approaches for maximising classification accuracy. *Pertanika Journal of Science and Technology*, *26*(1), 329–340.

[41] Wang, J., & Wang, S. (2019). Preparation, modification and environmental application of biochar: A review. *Journal of Cleaner Production*, *227*, 1002–1022. https://doi.org/10.1016/j.jclepro.2019.04.282

[42] Wang, K., Zhang, J., Shanks, B. H., & Brown, R. C. (2015). The deleterious effect of inorganic salts on hydrocarbon yields from catalytic pyrolysis of lignocellulosic biomass and its mitigation. *Applied Energy*, *148*, 115–120.

https://doi.org/10.1016/j.apenergy.2015.03.034

[43] Wang, R. Z., Huang, D. L., Liu, Y. G., Zhang, C., Lai, C., Wang, X., Zeng, G. M., Gong, X. M., Duan, A., Zhang, Q., & Xu, P. (2019). Recent advances in biochar-based catalysts: Properties, applications and mechanisms for pollution remediation. *Chemical Engineering Journal*, *371*(December 2018), 380–403. https://doi.org/10.1016/j.cej.2019.04.071

[44] Weber, K., & Quicker, P. (2018). Properties of biochar. *Fuel*, *217*(January), 240–261. https://doi.org/10.1016/j.fuel.2017.12.054

[45] Windeatt, J. H., Ross, A. B., Williams, P. T., Forster, P. M., Nahil, M. A., & Singh, S. (2014). Characteristics of biochars from crop residues: Potential for carbon sequestration and soil amendment. *Journal of Environmental Management*, *146*, 189–197. https://doi.org/10.1016/j.jenvman.2014.08.003

[46] Won, M., Hoon, S., Nam, H., Lee, D., Tokmurzin, D., Wang, S., & Park, Y. (2022). Bioresource Technology Recent advances of thermochemical conversion processes for biorefinery. *Bioresource Technology*, *343*(August 2021), 126109. https://doi.org/10.1016/j.biortech.2021.126109

[47] Zhang, X., Kano, M., & Matsuzaki, S. (2019). A comparative study of deep and shallow predictive techniques for hot metal temperature prediction in blast furnace ironmaking. *Computers and Chemical Engineering*. https://doi.org/10.1016/j.compchemeng.2019.106575

[48] Zhao, B., O'Connor, D., Zhang, J., Peng, T., Shen, Z., Tsang, D. C. W., & Hou, D. (2018). Effect of pyrolysis temperature, heating rate, and residence time on rapeseed stem derived biochar. *Journal of Cleaner Production*, *174*, 977–987.

https://doi.org/10.1016/j.jclepro.2017.11.013

[49] Zhou, Y., Qin, S., Verma, S., Sar, T., Sarsaiya, S., Ravindran, B., Liu, T., Sindhu, R., Patel, A. K., Binod, P., Varjani, S., Rani Singhnia, R., Zhang, Z., & Awasthi, M. K. (2021). Production and beneficial impact of biochar for environmental application: A comprehensive review. *Bioresource Technology*, *337*(June), 125451. https://doi.org/10.1016/j.biortech.2021.125451

[50] Zhu, X., Li, Y., & Wang, X. (2019). Machine learning prediction of biochar yield and carbon contents in biochar based on biomass characteristics and pyrolysis conditions. *Bioresource Technology*, *288*(May), 121527. https://doi.org/10.1016/j.biortech.2019.121527

1 Table 1: Features selected by metaheuristic optimization algorithms

Optimization Algorithms	Proximate Analysis	Ultimate Analysis
GWO	MC, ash, cellulose, hemicellulose, HTT, HR, RT, PS	C, H, O, lignin, cellulose, hemicellulose, HTT, HR, PS
RA-1	FC, VM, ash, lignin, cellulose, HTT, RT, PS	C, O, lignin, PS
RA-2	FC, VM, MC, ash, cellulose, lignin, HTT, HR, RT	C, N, O, lignin, cellulose, hemicellulose, HTT, HR, RT
SCA	FC, VM, ash, cellulose, HTT, HR, RT, PS	C, N, cellulose, HTT, HR, RT, PS
GA	VM, MC, ash, hemicellulose, HTT, HR, PS	N, H, lignin, cellulose, HTT, HR, RT, PS
PSO	FC, VM, ash, lignin, PS	C, H, O, lignin, cellulose, HTT, HR, RT, PS

2

3

4

5

6

7

Table 2: ANN architecture selected by metaheuristic optimization algorithms

Model	Proximate Analysis		Ultimate Analysis	
	Hidden Layers	Number of Neurons	Hidden Layers	Number of Neurons
GWO	3	30, 30, 23	3	23, 27, 30
RA 1	2	27, 30	3	23, 20, 18
RA 2	3	30, 30, 12	3	30, 26, 5
SCA	2	20, 25	2	9, 5
GA	2	22, 18	3	28, 23, 18
PSO	1	18	1	15

Table 3: Prediction performance of ANN models

Model	Proximate Analysis			Ultimate Analysis		
	R^2	RMSE	P-value[a]	R^2	RMSE	P-value[a]
ANN-GWO	0.85	2.82	0.414	0.87	2.52	0.001
ANN-RA 1	0.92	1.89	0.004	0.70	3.55	0.034
ANN-RA 2	0.93	1.74	0.030	0.88	2.23	0.212
ANN-SCA	0.84	2.86	0.276	0.78	3.43	0.742
ANN-GA	0.92	1.87	0.009	0.88	2.25	0.259
ANN-PSO	0.92	1.82	0.007	0.88	2.18	0.011

a: P-value based on the R^2 values of corresponding models.

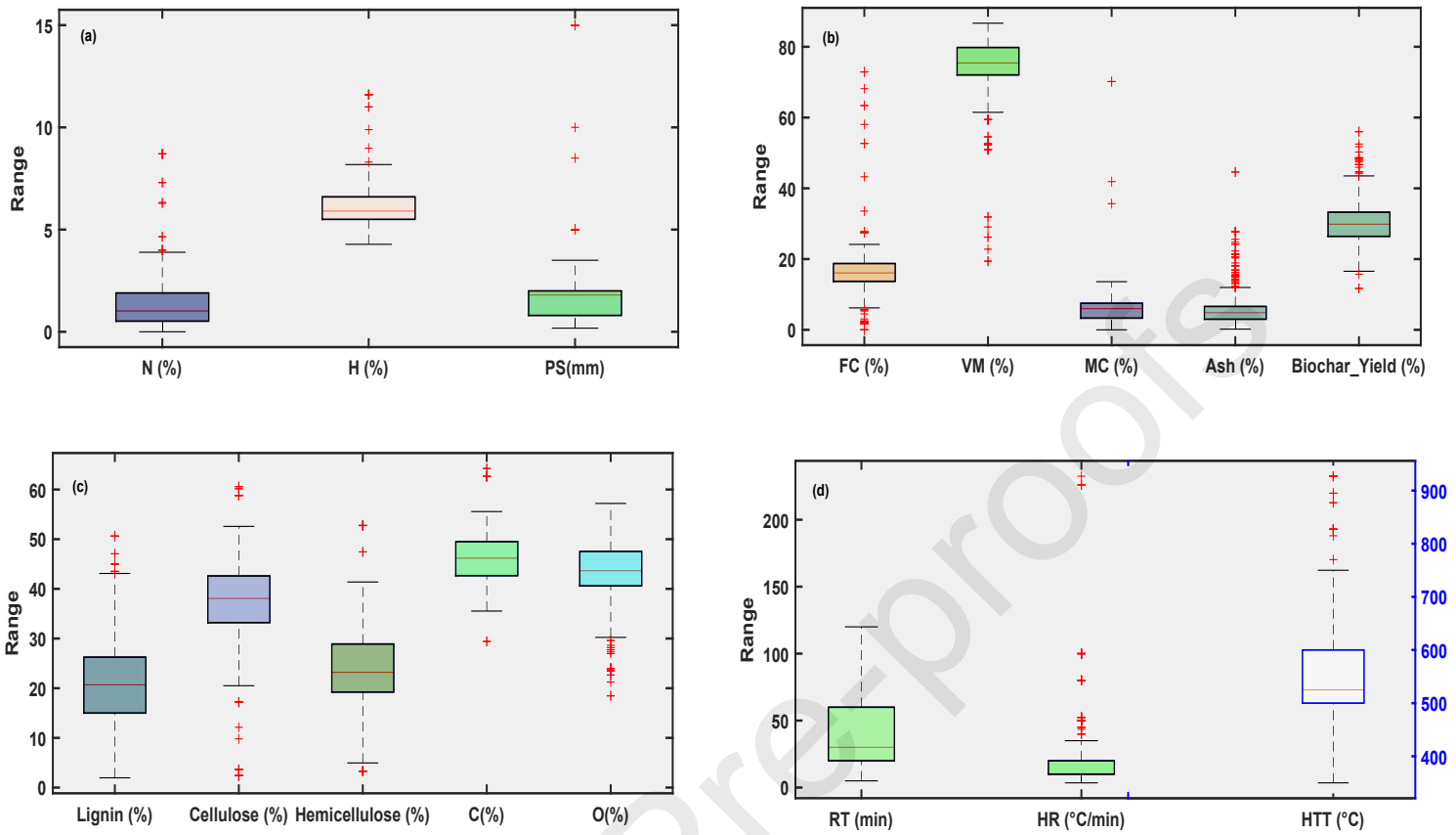


Figure 1: Box plot of the raw dataset

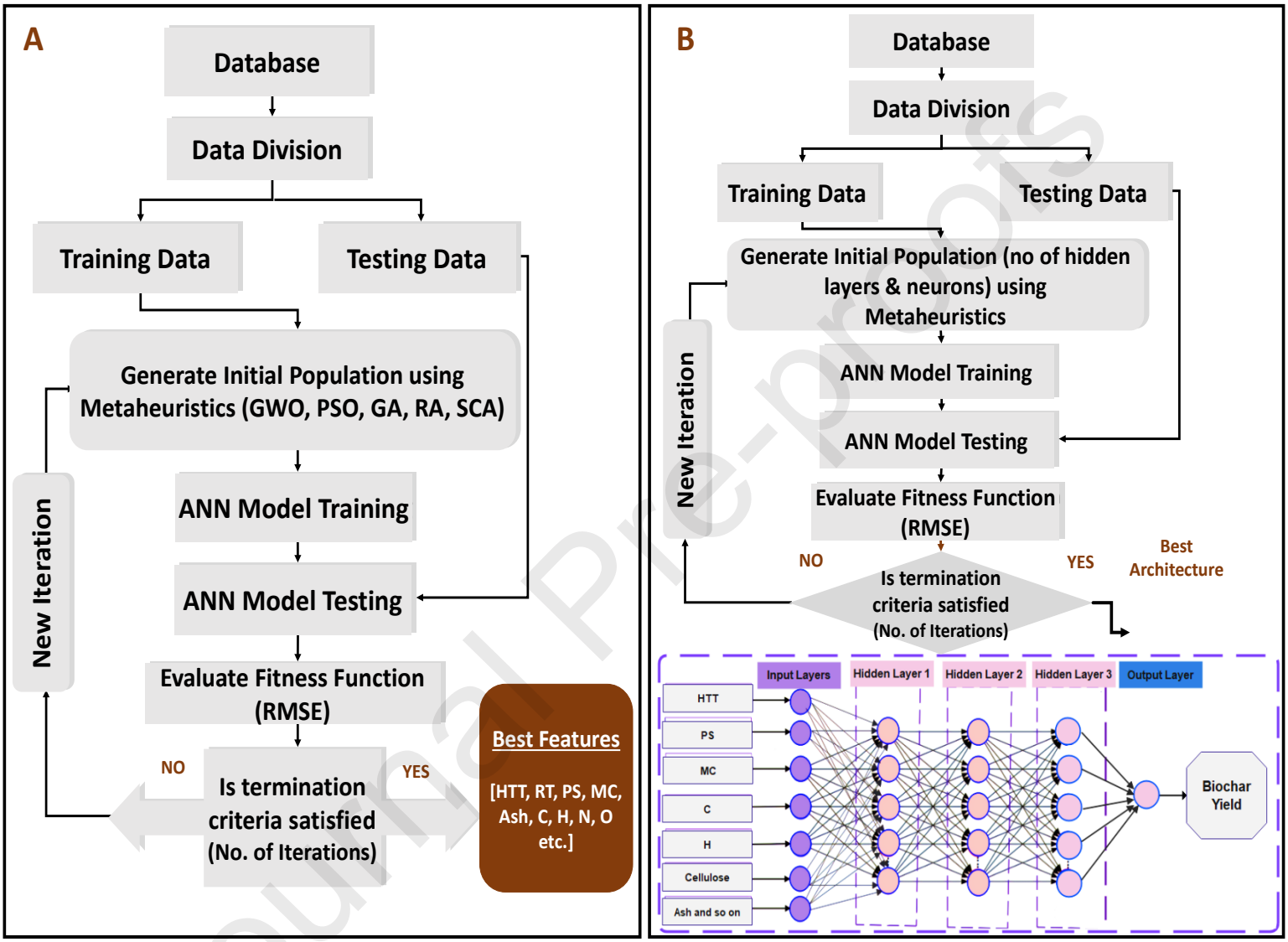
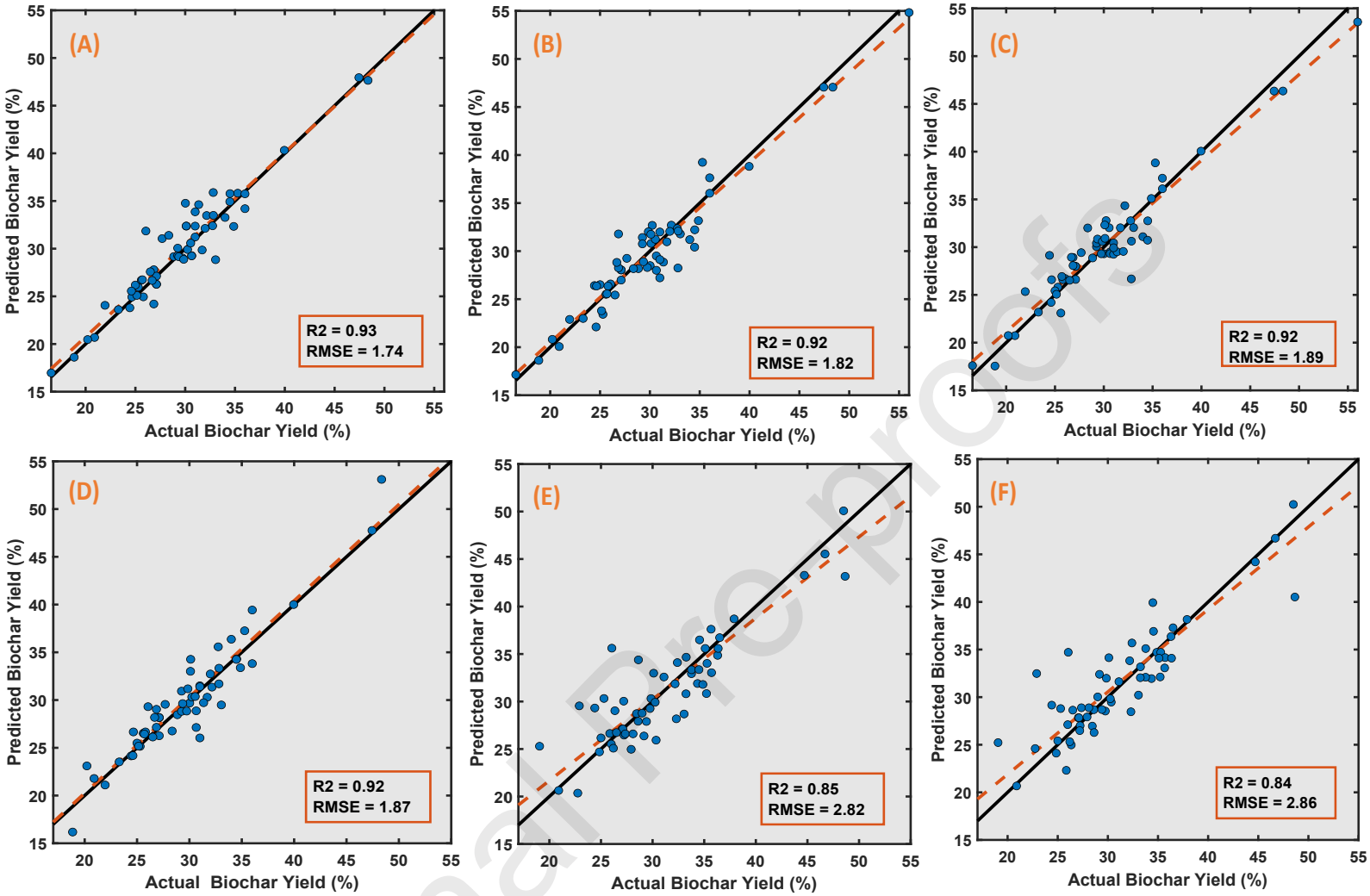


Figure 2: Proposed framework for features (Fig. 2A) and ANN architecture selection (Fig. 2B).

1



2

Figure 3: Comparison of experimental results and predicted model outputs (dashed line) trained using test data (solid blue circles): (A) ANN-GWO (B) ANN-RA 1 (C) ANN-RA 2 (D) ANN-SCA (E) ANN-GA (F) ANN-PSO. The black lines show the line $y = x$ (i.e. predicted values = true values).

3

4

5

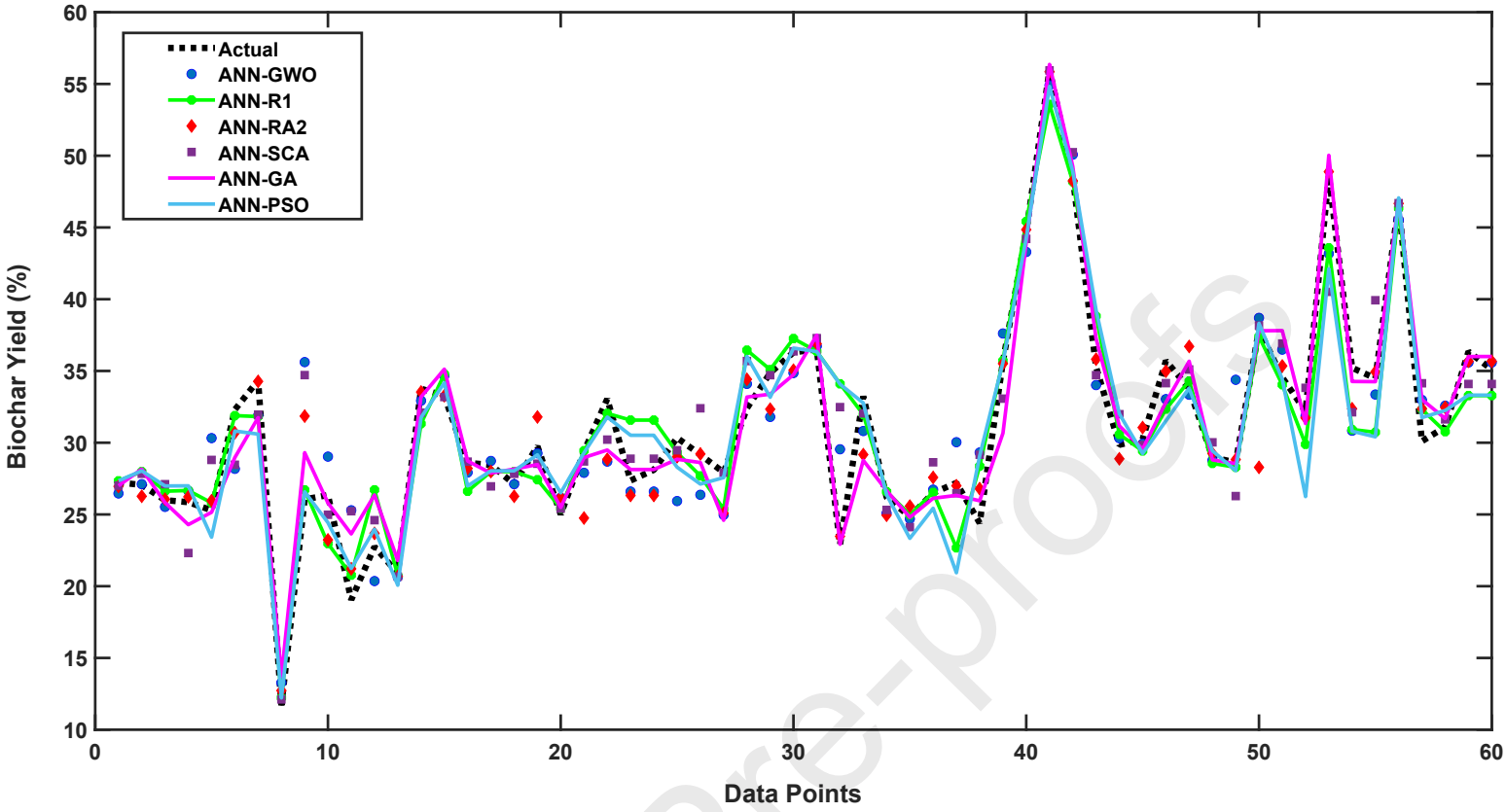


Figure 4: Variations of biochar yield predicted by ANN models over various data points

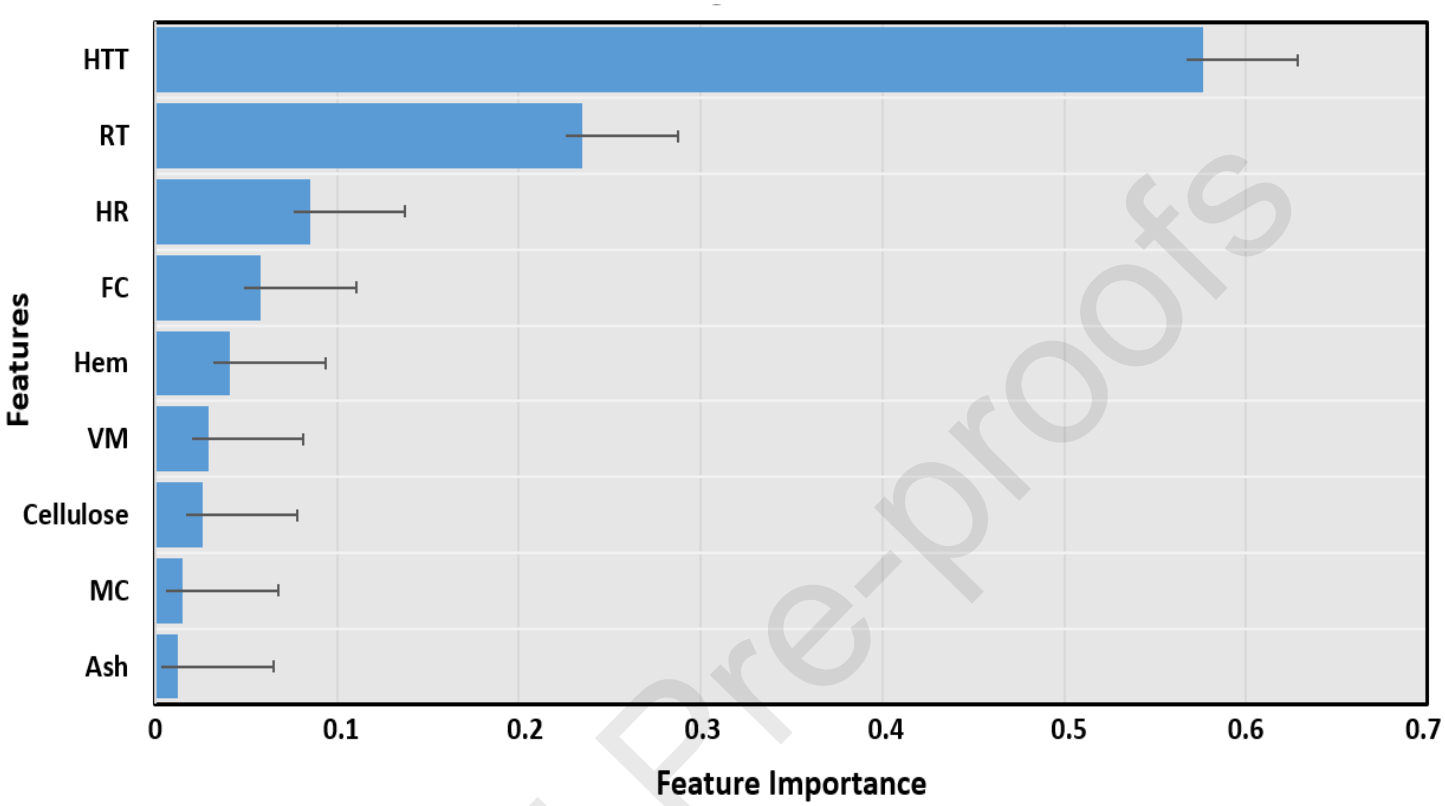


Figure 5:Schematic of feature importance on the predicted yield of biochar

(The black lines on each bar shows the positive error bars)

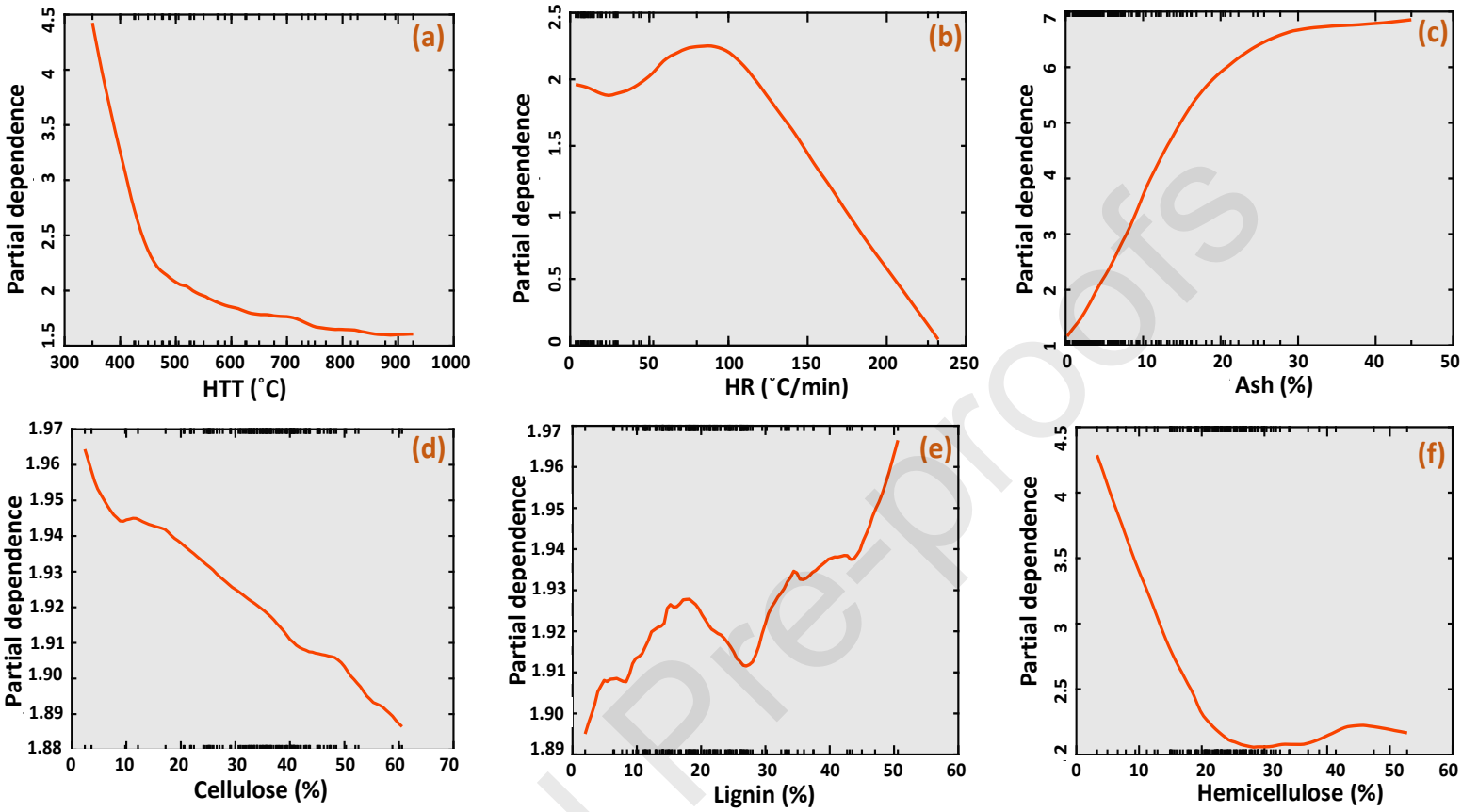


Figure 6: Partial differential plots for biochar yield: (a) HTT (b) HR (c) Ash (d) Cellulose (e) Lignin (f) Hemicellulose

Credit authorship contribution statement

Muzammil Khan: Investigation, Writing - original draft. **Zahid Ullah:** Formal analysis, Writing - original draft. **Ondřej Mašek:** Methodology, Supervision, Writing - review & editing. **Salman Raza Naqvi:** Validation, Writing - review & editing. **Muhammad Nouman Aslam Khan:** Formal analysis, Validation.

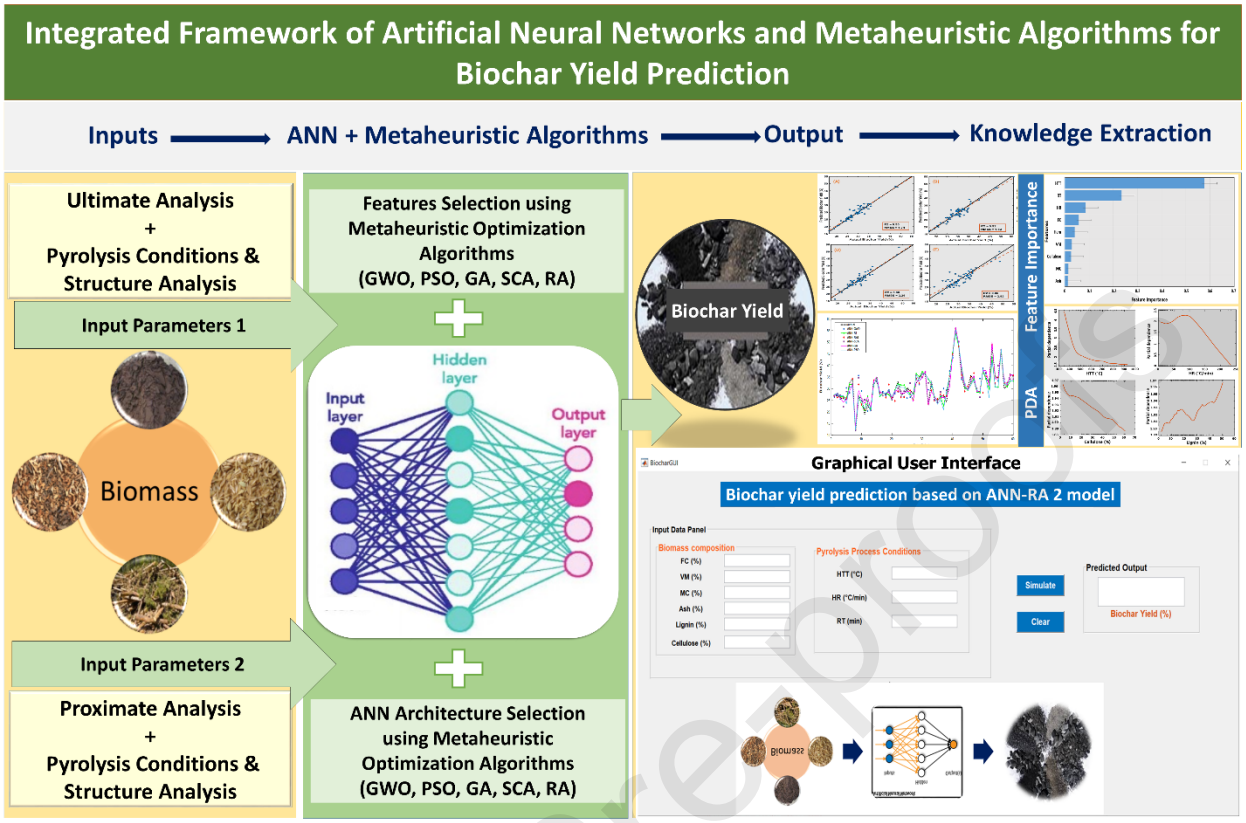
Declaration of interests

☒ 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.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Highlights

- ANN coupled with metaheuristic algorithms were used to predict biochar yield
- ANN integrated with Rao-2 algorithm outperformed all other models (R2~0.93)
- Pyrolysis temperature and residence time were the most important features
- Partial dependence analysis revealed inside details for the pyrolysis process
- An easy-to-use graphical user interface was developed for biochar yield prediction



1